



## Article

# EEG Dataset Collection for Mental Workload Predictions in Flight-Deck Environment

Aura Hernández-Sabaté <sup>1,2,\*</sup> , José Yauri <sup>1</sup> , Pau Folch <sup>2</sup>, Daniel Álvarez <sup>3</sup> and Debora Gil <sup>1,2</sup> 

<sup>1</sup> Computer Vision Center (CVC), C/ Sitges, Edifici O, 08193 Bellaterra, Spain; jyauri@cvc.uab.cat (J.Y.); debora@cvc.uab.cat (D.G.)

<sup>2</sup> Engineering School, Universitat Autònoma de Barcelona, C/ Sitges, Edifici Q, 08193 Bellaterra, Spain; pau.folch@uab.cat

<sup>3</sup> Aslogic, Av. Electricitat, 1-21, 08191 Rubí, Spain; dalvarez@aslogic.es

\* Correspondence: aura@cvc.uab.cat

**Abstract:** High mental workload reduces human performance and the ability to correctly carry out complex tasks. In particular, aircraft pilots enduring high mental workloads are at high risk of failure, even with catastrophic outcomes. Despite progress, there is still a lack of knowledge about the interrelationship between mental workload and brain functionality, and there is still limited data on flight-deck scenarios. Although recent emerging deep-learning (DL) methods using physiological data have presented new ways to find new physiological markers to detect and assess cognitive states, they demand large amounts of properly annotated datasets to achieve good performance. We present a new dataset of electroencephalogram (EEG) recordings specifically collected for the recognition of different levels of mental workload. The data were recorded from three experiments, where participants were induced to different levels of workload through tasks of increasing cognition demand. The first involved playing the N-back test, which combines memory recall with arithmetical skills. The second was playing Heat-the-Chair, a serious game specifically designed to emphasize and monitor subjects under controlled concurrent tasks. The third was flying in an Airbus320 simulator and solving several critical situations. The design of the dataset has been validated on three different levels: (1) correlation of the theoretical difficulty of each scenario to the self-perceived difficulty and performance of subjects; (2) significant difference in EEG temporal patterns across the theoretical difficulties and (3) usefulness for the training and evaluation of AI models.

**Keywords:** mental workload; serious games; flight simulation; EEG physiological data; deep learning; transfer learning



**Citation:** Hernández-Sabaté, A.; Yauri, J.; Folch, P.; Álvarez, D.; Gil, D. EEG Dataset Collection for Mental Workload Predictions in Flight-Deck Environment. *Sensors* **2024**, *24*, 1174. <https://doi.org/10.3390/s24041174>

Academic Editor: Christian Baumgartner

Received: 24 December 2023

Revised: 26 January 2024

Accepted: 5 February 2024

Published: 10 February 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Abnormal cognitive states reduce human performance and diminish their ability to solve tasks. There is a wide variety of anomalous mental states that highlight mental workload, fatigue, distraction, and stress, as they decrease task performance, delay response capacity time, can block physical actions, and can lead to health and psychological disorders. Mental workload (MW or WL) has become of special interest in several areas because it affects overall human productivity and efficiency.

MW refers to the amount of mental resources required to perform a cognitive task [1]. Due to the inherent differences between subjects, MW strongly depends on each individual's ability, psychological motivation, and the surrounding environment [2,3]. In general, the more difficult a task is, the greater the mental workload and its impact on correlated mental states [4]. For example, when a high MW is in place for a long time, fatigue appears, and stress arises [5]. On the contrary, when the MW is low for an extended period of time, the mind may become distracted and bored, which can lead to drowsiness [4]. Both cases can be harmful, especially in some activities that require a mental effort to

succeed in a task [6], such as flying a plane [7] or driving a car [8], which can lead to catastrophic accidents.

The multifaceted nature of MW prevents direct evaluation, but it can be feasibly inferred from other quantifiable variables. On the one hand, a common method to evaluate MW is based on the measurement of the performance achieved during task loading, usually using questionnaires to capture the self-perceived workload of each participant. The NASA Task Load Index (NASA-TLX) [9] is among the most used questionnaires that capture the self-subject perception of a performance, complexity, time demand, and effort of a certain task. On the other hand, MW can also be evaluated using the physiological responses of the subject during the task [1]. Physiological data provides a more reliable measurement than the psychologically dependent self-subject reports. Among the great variety of physiological sensors, several studies have been carried out using a wearable electroencephalogram (EEG) due to their low cost and easy use [10,11].

Usually, MW studies focus on assessing and detecting MW in specific human communities, such as pilots during flights [1], drivers on the road [12], and other activities [13]. N-back tests and other serious specifically designed games have been proven to be useful in investigating MW since the use of working memory that they demand can provoke workload [14,15]. Although many datasets have been published using the N-back test [16,17], they are usually too short and mostly restricted to two highly differentiable mental tasks. Other researchers have collected specific datasets to study MW in specific areas, mainly in aeronautical and automotive scenarios, due to catastrophic results in aviation and car accidents. For example, in aeronautics, datasets are collected from computer-based flight simulators [18], immersive cockpit simulators [19], and real flights [20]. However, these datasets are generally private and have restricted access. Analogously, the datasets collected in the automotive industry are also too limited [21].

In this work, we present a publicly available dataset (<https://doi.org/10.5565/ddd.uab.cat/259591> (accessed on 26 December 2023)) to recognize different levels of MW. It has different levels of workload, including a baseline (BL) or normal cognitive state. Part of this dataset has previously been used in our related work [22], and we claim that it can be used for research purposes to test new methods for analyzing and evaluating MW. Furthermore, this dataset has been specifically designed to enable the validation of models able to transfer knowledge to flight scenarios, which are hard and expensive to collect.

The current repository contains physiological EEG recordings from subjects facing tasks of different complexity in three different scenarios. In the first scenario, data are collected from subjects performing three variants of the N-back test to induce low, medium, and high MW. In the second scenario, data are collected from the Heat-the-Chair game, a specifically designed serious game that combines simple and simultaneous task modes, emphasizing attention and multitasking abilities. In the third scenario, data are collected in an Airbus A-320 flight simulator cockpit, in which the pilot addresses several real flight situations. The total amount of data collected is 48 sessions from 16 participants in the N-back test for a total of 34 h of recordings, 34 sessions from 17 volunteers in the Heat-the-Chair game for a total of 7 h of recordings, and 5 flight sessions with 2 pilots for a total of 95 min. In addition, the ground truth of each task, the theoretical MW complexity, the self-perceived MW complexity, the scores achieved in the games, and the NASA-TLX answers are provided.

The dataset has been validated in three aspects. First, the validity of the theoretical MW complexity has been assessed by correlating the performance obtained by the subjects with their self-perceived difficulty and game scores. Second, the quality of the WL assessment of EEG recordings has been validated by correlating their temporal patterns to the theoretical MW. Finally, the usefulness of the whole dataset for the implementation of AI systems has been assessed using the presented dataset for training and validating a DL method for the recognition of MW.

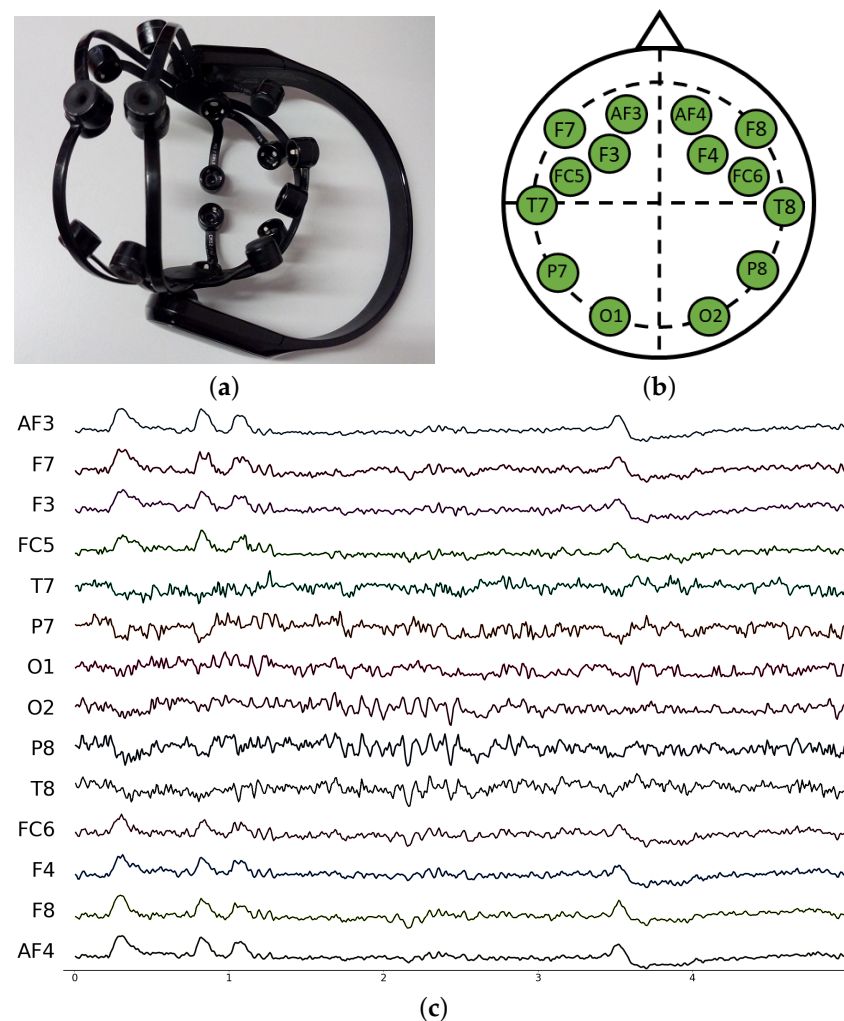
## 2. Data Acquisition Description

In this section, the main characteristics of the EEG device, participants, and experimental scenarios (design, implementation, and experiment structure) used to collect the database are described.

### 2.1. EEG Device Characteristics

The dataset described in this paper contains EEG signals recorded for a set of subjects and a set of experiments. Signals have been recorded with an Emotiv Epoc X EEG. As shown in Figure 1, it consists of a portable, wireless, high-resolution, 14-electrode EEG system, according to the International 10–20 System, that communicates via Bluetooth in real time. The electrodes are placed over the head scalp and record the electrical activity of the brain. This device provides the raw signals in  $\mu\text{V}$  at a sampling rate of 128 Hz. Furthermore, the sensor provides the power band for the major brain rhythms (beta: 4–8 Hz, alpha: 8–12 Hz, beta low: 12–18 Hz, beta high: 18–25 Hz and Gamma: >25 Hz). Emotiv gives 8 power samples per second computed over the previous 2 s.

Signals are recorded directly from the headset and undergo significant signal processing and filtering to remove mains noise and harmonic frequencies. In particular, signals are sampled at 2048 Hz, a dual-notch filter is applied at 50 Hz and 60 Hz, and a low-pass filter is computed at 64 Hz. Finally, data are filtered down to 128 Hz or 256 Hz for transmission.



**Figure 1.** EEG Emotiv Epoc X headset used for recording the data (a), spatial distribution of sensors for the EEG Emotiv Epoc X (b), signal pattern recorded during 5 s (c).

## 2.2. Participants

For all the experiments, written consent was obtained from each participant. The consent form explains the goal of the experiment and describes what kind of data are collected and the terms of privacy in the use of personal data. Additionally, it emphasizes that the data released to the general public does not contain information that can directly identify the subject and that any data and research results already shared with other investigators or the general public cannot be destroyed, withdrawn, or recalled. Each consent was hand-signed by each subject on the day of the first experiment.

The participants in all experiments were healthy people without any condition that might have caused an imbalance in the recorded data. The characteristics of all the participants are detailed as follows:

1. The N-back test experiment: 16 subjects (all male), with ages ranging from 20 to 60 years, participated in the experiments. The volunteers belonged to three different university research centers and shared a scientific background with different levels of expertise (students, junior researchers, senior researchers, or professors).
2. The Heat-the-Chair game experiment: 17 subjects (12 male and 5 female), with ages ranging from 20 to 60 years, participated in the experiment. The volunteers shared the same characteristics as participants in the previous experiment, and seven of them completed the preceding test.
3. The flight simulation experiment: two professional pilots, but with different experience levels, participated in all flight missions, but they exchanged roles depending on the mission. Table 1 details the information of the pilots.

**Table 1.** Pilots information.

Pilot	Gender	Age	Flight Hours
Pilot 1	Male	51	4000
Pilot 2	Male	32	1700

Figure 2 illustrates some of the participants in the different experiments. All of them signed a written consent to publish their images.



**Figure 2.** Volunteers, during the experiments, performing: (a) N-back test, (b) Heat-the-Chair game and (c) flight simulation in the cockpit of an A320.

## 2.3. N-Back Test Experiment

We used the N-back test game to induce different levels of mental workload in participants. This type of experiment requires the ability to manage one or two N-back tasks simultaneously, taking into account the insights shown in the n-trial before, so it demands a high usage of memory to complete the tasks. In particular, we designed three experiments with different levels of complexity (low, medium, and high), and each subject performed all the experiments, randomly assigned, using a computer. The three variants of the N-back test to induce mental workload were implemented as follows:

1. Low mental workload—position 1-back: As Figure 3 shows, a square appears every few seconds in one of nine different positions on a regular  $3 \times 3$  grid over the screen. The player must press a key on the keyboard when the position of the square on the current screen is the same as the position of the square that appeared on the previous screen.
2. Medium mental workload—arithmetic 1-back: As Figure 4 shows, an integer number between 0 and 9 appears every few seconds on the screen, while an arithmetic operation (plus, minus, times, and divide) is audibly presented. The player must solve this operation using the number that appeared on the previous screen and the current one. Results must be typed using the numerical keys.
3. High mental workload—dual position and arithmetic 2-back: This test combines the two previous ones. As Figure 5 shows, an integer number between 0 and 9 appears every few seconds in one of nine different positions on a regular grid. At the same time, an operator is presented visually. As before, players must type the solution of this operation using the number that appeared on the two screens before and the current one. In addition, players must press a key in case the position of the current number is the same as the position of the number shown two screens before.

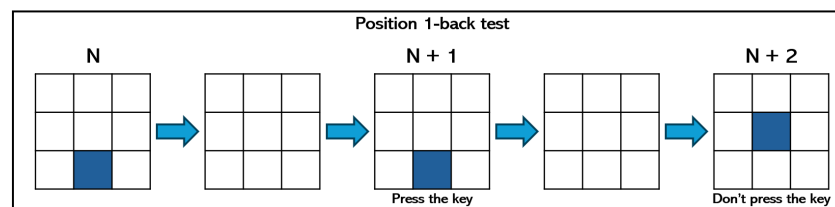


Figure 3. Example of position 1-back test.

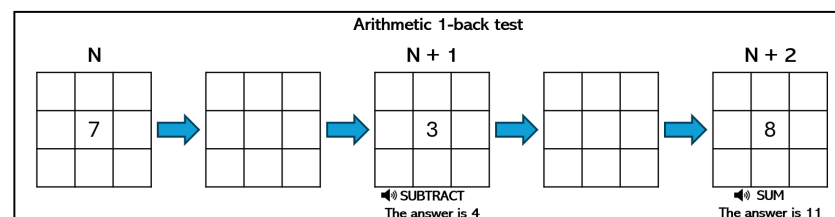


Figure 4. Example of arithmetic 1-back test.

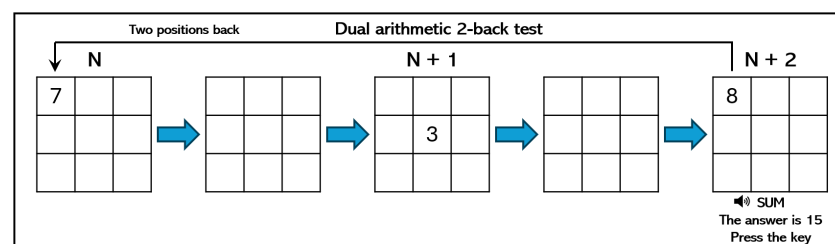
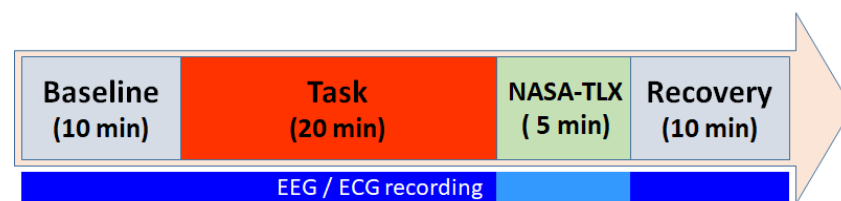


Figure 5. Example of dual position and arithmetic 2-back test.

**Experiment Structure.** Before playing and recording the data, the subject was informed about the rules and trained in the game for five minutes. For training, the dual position and audio 1-back mode were used, which simultaneously combines position and audio, taking into account the 1-back step, i.e., a number between 0 and 9 is audibly presented, and the player must press a key if it matches the one emitted in the previous screen, another one if its position matches, and another one if both matches occur.

We assume that, in the absence of any required mental effort, subjects will exhibit a baseline mental workload, and their physiological responses will accordingly be at a minimum scale. Additionally, we expect that the baseline levels will vary among individuals. To induce this baseline state, participants watched a relaxing video for 10 min before engaging in the N-back tests. Subsequently, they played the game for 20 min. After completing the game, they responded to the NASA-TLX questionnaire [9] to provide their subjective perceptions of the mental workload and effort demanded by the game. Finally, to come back to a calm state after the task, subjects underwent a 10-minute recovery step, mirroring the baseline stage. The experimental protocol is described in Figure 6. During each session, all neuro-physiological responses were continuously recorded. The dataset only contains signals from the baseline, the game task, and the recovery phase, removing the parts not strictly belonging to the experiment. The dataset also contains the results from the NASA-TLX questionnaire and the achieved scores of the player, which correspond to the number of hits.



**Figure 6.** Timeline of the N-back test experimental protocol.

#### 2.4. Heat-the-Chair Experiment

This game was specifically designed to create a scenario in which simultaneous tasks must be performed, replicating the demand for concentration and alertness of pilots while flying. The game consists of completing as many objectives as possible in 10 min. Completing an objective consists of obtaining and using the necessary pieces to form a 4-digit number, which appears at the top left of the screen for 10 s and then disappears, reappearing for 5 s every 1 min while the objective is not achieved. Once the correct pieces have been obtained and the target puzzle has been completed (the 4-digit number), the player increases the punctuation, and a new target number to be completed appears automatically. Figure 7 shows the game user interface. The target number appears in the upper left panel, while the pieces that the player obtains are in the lower right panel. Notice that the bottom row is designated for storing the rewarded pieces (in cyan), while the top row is dedicated to dragging and dropping the pieces to replicate the 4-digit number.

To obtain pieces, the player must perform two main tasks:

1. **Bars with sliders:** As we can observe in Figure 7, there are two colored bars in the bottom central-hand panel with sliders that move in the horizontal and vertical directions. The player must keep the sliders in the center of the bars using the directional keys of the keyboard.
2. **Dots:** In the same panel, there is a large square that will be filled with dots. To avoid this, the player must drag them to the dashed-line box shown in the center.

In the top central panel, there is a circular button with a depleting energy bar below it. The difficulty of tasks will increase proportionally to the depletion of the energy bar: the emptier the bar, the more challenging the game will become. Thus, the player must regularly recharge the power bar using the circular button.

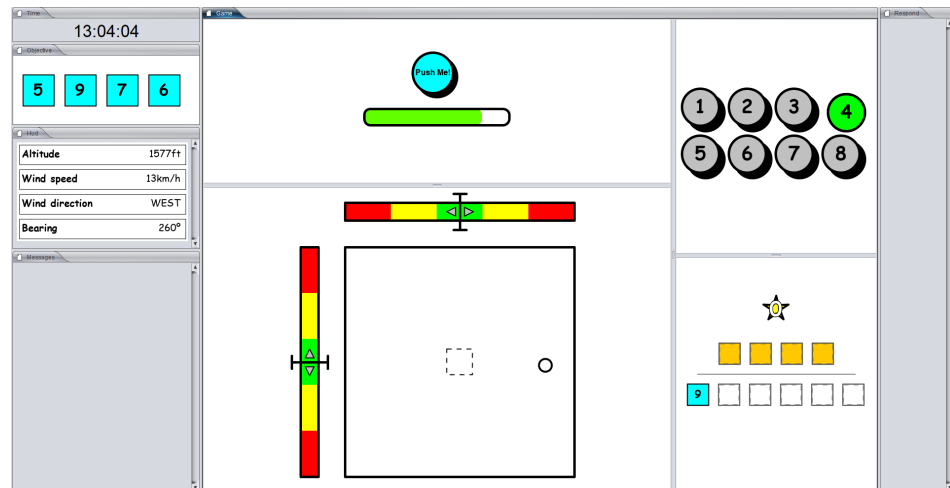


Figure 7. The Heat-the-Chair game user interface.

The key point is that the game supports two modes of operation: with or without interruptions. Interruptions are introduced in the game design to emulate the interruptions that pilots receive when interacting with Air Traffic Control (ATC). In this mode, incoming events randomly appear to be solved. In particular, five different tasks, in random order, are required to be completed by the player. Tasks can be either to report a current flight parameter (altitude, wind speed, wind direction, and bearing) or to change the number of the switch box (the switch box starts randomly at each game). Flight information is shown on the left-center portion of the screen, and the switch box is shown on the top-right portion of the screen. When an interruption arrives, an alert of messaging is shown on the bottom-left portion of the screen. The player must click and read the message. Each required task has a starting and an ending time to be completed, beyond which the player is penalized. Figure 8 depicts an interruption asking for a change to the current switch box. The start and end times to complete the task are highlighted in green and red, respectively. If the player does not complete the task or inserts incorrect information as an answer, one rewarded piece is lost.

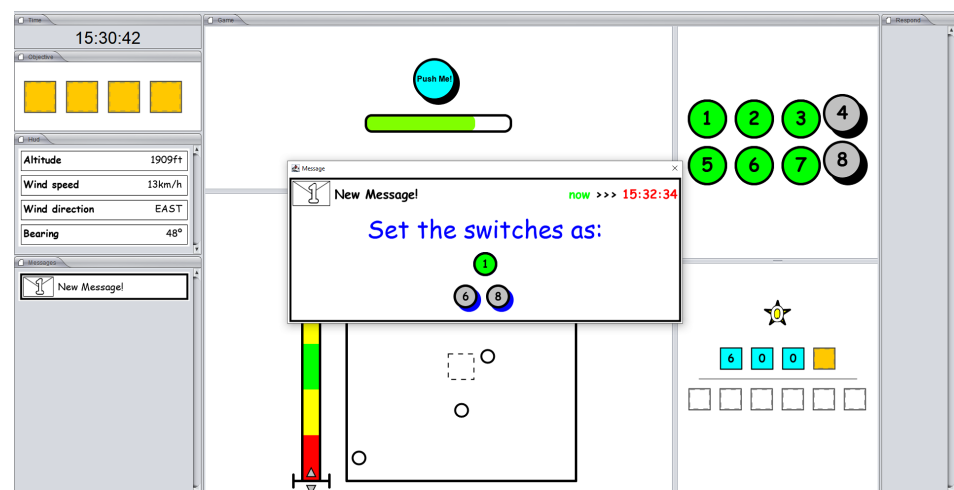


Figure 8. The Heat-the-Chair game with an interruption message.

**Experiment Structure.** Before playing and without recording data, the subject was informed about the rules and trained in the Heat-the-Chair game without interruptions for 5 min to familiarize themselves with controls. The game mode was chosen randomly before starting the experiment.

Given that each subject randomly faces the two modes of the game, each game is recorded in separate sessions. As Figure 9 shows, a session consists of three phases. The baseline lasts 3 min, in which the subject drags balls that randomly appear on the screen and drops them to the dashed square in the center. The game has the subject play the randomly selected game mode for 10 min, either with interruptions or without interruptions. In addition, finally, there is the NASA-TLX questionnaire, which the subject fills out, indicating his/her subjective perceived game complexity. Neurophysiological responses are continuously recorded for the entire session using the Emotiv. The dataset provides the signals from the baseline and the task phase, removing the time intervals that do not take part in the experiment. The dataset also contains the player's achieved scores and the results from the NASA-TLX questionnaire.

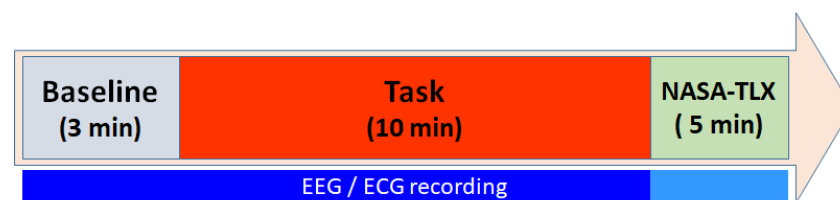


Figure 9. Timeline of the Heat-the-Chair experimental protocol.

### 2.5. Flight Simulation Experiment

The goal of this experiment was to collect experimental data useful for quantifying the impact of an increase in the mental workload of pilots during the performance of routine flight tasks (i.e., within reference parameters) and when they must manage additional unexpected phenomena, such as wind shears, machine failures, equipment warnings, and unusual traffic. In these situations, interaction between the crew and the ATC increases, and pilots are more likely to make mistakes due to the mental workload. For that, five flight-simulated scenarios were designed to evaluate the pilots' task load changes while they solve unexpected situations. Each flight scenario is an experiment, and each experiment is unique, with its own characteristics. The flight simulation was carried out in an immersive Airbus-320 cockpit simulator, and the chosen flight route was from Barcelona to Lleida in Spain, with an approximate duration of 14 minutes. Figure 10 illustrates the route followed by the pilot. Weather, weight, and speed conditions were fixed for all flights.



Figure 10. Flight simulation route.



An expert flying pilot defined five flying scenarios with different levels of complexity and events:

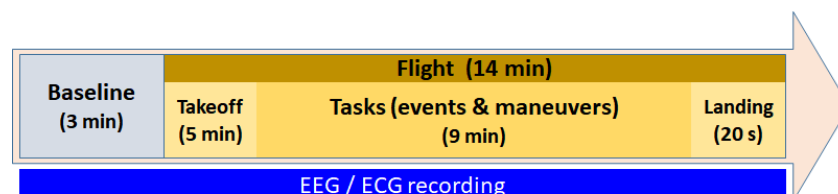
1. Flight 1 [easy difficulty]: The pilot performs a standard flight to be used as reference parameters.
2. Flight 2 [medium difficulty]: During the flight, the ATC reports much traffic, so the pilot is asked to change the position of the airplane above the glide slope at high speed.
3. Flight 3 [hard difficulty]: During the final stage of the flight, the airplane is hard destabilized by a strong wind shear, so the pilot must maneuver, recover the plane stability, and land it.
4. Flight 4 [medium-hard difficulty]: During the flight, a malfunction during the approach provokes an engine failure that increases the crew workload.
5. Flight 5 [medium difficulty]: This flight is similar to Flight 2 with a little variation.

Two flying pilots participated in the experiment. Before starting the experiment, pilots received a printed description of the assigned flight mission, and one was chosen as the pilot, whereas the other remained as the copilot/observer. A third pilot was monitoring the whole flight for the annotation of the pilot's perceived complexity in time stamps of 30 s. Table 2 reports the distribution of the roles of the two flying pilots for the five flying scenarios.

**Table 2.** Pilots roles.

Experiment	Pilot = Pilot 1 Observer = Pilot 2	Pilot = Pilot 2 Observer = Pilot 1
Easy	-	Flight 1
Medium	Flight 2	Flight 5
Medium-Hard	-	Flight 4
Hard	Flight 3	-

**Experiment Structure.** The experimental protocol of each flight was divided into two phases. First was the baseline phase, in which the pilot stays on the runway awaiting the order to take off. Second was the flight phase, which at the same time can be split into three stages: the takeoff, when the flight starts, and the plane climbs; the task phase itself; and a short time for landing the plane on the ground. The task phase encompasses cruise, descent, and approach tasks, along with standard communication with the ATC, and includes the specific tasks requested for each flight simulation. Figure 11 shows the timeline of the flight simulation experiment. The dataset contains the neurophysiological responses from all phases of the experiment. It also contains the difficulty perceived by pilots obtained from a modification of the NASA-TLX questionnaire to obtain a dynamic perceived difficulty across the flight phase collected by the third pilot named above. The level of difficulty was graded from 0 to 3, with 0 being the easiest one.



**Figure 11.** Timeline of the flight simulation experimental protocol.

### 3. Data Format and Structure

The dataset described in this paper has been made publicly available on the Digital Document Deposit of the Universitat Autònoma de Barcelona, accessible at <https://doi.org/10.5565/ddd.uab.cat/259591> (accessed on 26 December 2023). No registration is required.

The dataset is provided in a compressed file `workload\_dataset.zip`. After decompression, the dataset contains three main folders that store the collected data for the N-back test data (folder `data_n_back_test`), the Heat-the-Chair data (folder `data_heat_the_chair`), and the flight simulation data (folder `data_flight_simulator`), respectively.

The next subsections explain the data format and structure for each experiment.

### 3.1. N-Back Test

EEG signals, game performance, and TLX questionnaires were stored in 3 `.parquet` files containing the data obtained from all participants in the following structure.

The acquisition software saves the raw data in a CSV file that has 139 columns with datetime, timestamps, the 14-electrode data, the five frequency power band data, and quality metrics supplied by the sensor. Since a session itself is split into phases, the recording is a continuous signal. We used manual annotations of the starting and ending times of each phase to remove data outside the phases. Also, all the annotations were synchronized by means of a specifically designed application. The data recorded for the session phases and all participants was stored in a `.PARQUET` file that included metadata added as three additional columns: `SUBJECT`, `TEST`, `PHASE`. The `SUBJECT` column is the identifier of the volunteer. Participants are labeled as `subject_xx`, for  $xx \in \{1, \dots, 16\}$  a number identifying the subject. The `TEST` column identifies the variant of the N-back test (1 for the position 1-back, 2 for the arithmetic 1-back, and 3 for the dual position and arithmetic 2-back). The `PHASE` column identifies the phase in the session (1 for the baseline, 2 for the task, and 3 for the recovery).

The dataset also provides the performance of the subject's game during the N-back test. Each subject has three measurements—one for each task difficulty. Since the task phase lasts 20 min and the trial of the game is almost two minutes, the subject played the game many times, so scores are provided as a list of punctuation. The `.PARQUET` file also includes the fields `SUBJECT` and `TEST` defined as before for the identification of the subject and N-back test variant.

The answers to the TLX questionnaires were also collected for all games and participants and stored in a `.PARQUET` file including the fields `SUBJECT` and `TEST`.

The directory tree for the dataset of the N-back test is the following:

```

data_n_back_test
├── eeg
│   └── eeg.parquet
├── game_performance
│   └── game_scores.parquet
└── subjective_performance
    └── tlx_answers.parquet

```

where the file `eeg.parquet` stores the EEG signals for all participants, the file `game_scores.parquet` the game scores per subject and `tlx_answers.parquet` their answers to the TLXs questionnaires.

### 3.2. Heat-the-Chair Game

The EEG raw data stored in CSV files for each session was processed to remove parts outside the baseline and the task phase. The data of all participants was stored in a single `parquet` file, including metadata for the identification of the subject, game type, and phase. The dataset also provides the performance of the subject during the game, and the self-perceived workload for each task reported in the TLXs. The directory tree for the dataset for the Heat-the-Chair game is the following:

```

data_heat_the_chair
├── eeg
│   └── eeg.parquet
└── game_performance

```

```

├── subject_01_with.csv
├── subject_01_without.csv
├── ...
├── subject_26_with.csv
├── subject_26_without.csv
├── subjective_performance
└── tlx_answers.parquet

```

Each of the folders contains the following information:

1. The folder `eeg` contains the file `eeg.parquet` storing a pandas dataframe with all the EEG data and three extra columns (`SUBJECT`, `TEST`, `PHASE`) of metadata. The field `SUBJECT` identifies the volunteer as 'subject\_xx\_', for 'xx' a two-digit number. The field `TEST` identifies the game mode: 1 for a game without interruptions and 2 for a game with interruptions. The field `PHASE` identifies the stage of the experiment: 1 for the baseline and 2 for the game.
2. The folder `game_performance` contains the game scores for each subject and game into separated CSV. The name of these files follows the pattern 'subject\_xx\_mode', where xx is a two-digit number identifying the volunteer, and mode is the game type: 'with' for a game with interruptions and 'without', otherwise.
3. The folder `subjective_performance` contains the file `tlx_answers.parquet` with the answers to TLX questionnaires for all participants and games.

From the set of 17 volunteers, the seven subjects compressed between 1 and 16 subjects have also participated in the N-back test; the rest of them, from 17 to 26, were new participants.

### 3.3. Flight Simulator

The data recorded from the flight simulator is in the folder `data_flight_simulator`. To make the processing of data easier, the original 'csv' files provided by the sensors were prepared by adding additional columns and saved into a single one, 'parquet'. Five columns were added for both EEG and ECG. The column `SUBJECT` identifies the pilot who is flying (number 1 identifies pilot 1, whereas number 2 identifies pilot 2). The column `FLIGHT` indicates the flight experiment performed, ranging from 1 to 5. The column `PHASE` indicates the stage of the flight. Values are 'baseline' and 'flight' (see Figure 11). The column `THEORETICAL DIFFICULTY` represents the expected theoretical workload induced in the pilot, and the values range from -1 to 4 to indicate easy to hard. Each flight has its own theoretical difficulty. Finally, the `PERCEIVED DIFFICULTY` columns provide the perceived difficulty reported by the pilot about the complexity of the assigned mission. During the simulation, at every certain interval, the pilot scores the complexity of the scenario from 0 to 3. We encoded the perceived difficulty of the 'baseline' stage as -1.

Finally, the flight simulator data are organized into three sub-folders:

```

data_flight_simulator
├── eeg
│   └── eeg.parquet
├── perceived_difficulty
│   ├── flight_1.json
│   ├── flight_2_4.json
│   └── flight_3_5.json

```

Each of the folders contains the following information:

1. The folder `eeg` contains a single file `eeg.parquet` with the EEG data.
2. The folder `perceived_difficulty` contains the perceived difficulty of the pilots during the flight. It contains three 'json' files: `flight1.json`, `flight2_4.json`, and `flight3_5.json`.

The flight simulator experiment Wasim contains the collected data from two professional pilots who performed five simulation flights in an A320 cockpit. The expected induced degree of workload and the self-perceived workload reported by the pilots themselves are also registered.

#### 4. Data Validation

In this section, we present the technical validation of the proposed data. To show its usefulness for training and evaluation of AI models, we have conducted 3 different experiments:

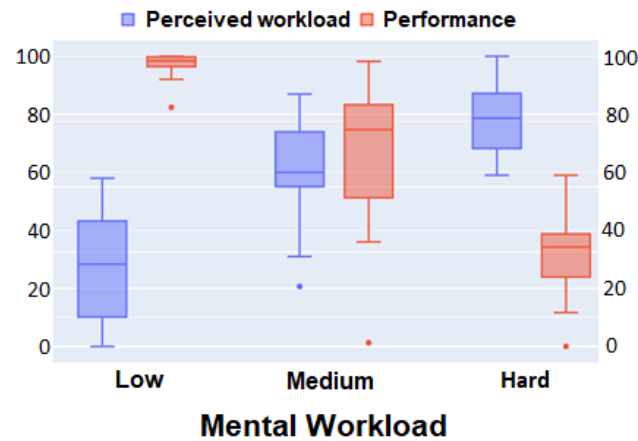
1. **Correlation to the Perceived Difficulty.** The goal of this experiment was to assess that the theoretical difficulty of each experimental scenario can be used as a valid annotation defining a ground truth for AI models. This assessment checks that the evolution of the difficulty perceived by participants increases along with the theoretical one. For the two games, we also include the performance to check if it is decreasing as theoretical performance increases.
2. **Analysis of Differences in Temporal Patterns.** The goal of this experiment was to show that the temporal waveforms of signals recorded under different WL conditions had different patterns. In particular, we have analyzed the number of spikes extracted from EEG recordings to assess whether their distribution is different across increasing levels of WL (N-back-test dataset), number of interruptions (serious game), and flight difficulty.
3. **Usefulness for Training AI models.** The goal of this experiment was to assess the usefulness of the presented dataset by training a DL model using the N-back test data and testing it on the Heat-the-Chair and flight simulator data to show its transfer task capability.

In the next sections, we report the experimental setup and results obtained for each of the experiments.

##### 4.1. Correlation to the Perceived Difficulty

To evaluate the technical quality of the collected data in the N-back test, we analyzed the answers to TLX questionnaires. Since TLX reports the self-perceived degree of workload enforced by the tasks, we put them in correspondence with the performance of the players. Figure 12 shows boxplots of the perceived difficulty given by the TLX questionnaire and game performance given by the percentage of correct operations for the 3 levels of difficulty of the game. On the one hand, the performance of subjects is decreasing with the theoretical difficulty of the game, as expected. On the other hand, the perceived workload of participants also has the expected increasing correlation with the theoretical difficulty. Finally, the perceived workload increases as the performance decreases, which is also consistent with the hypothesis that each test offers a challenge according to its difficulty level.

For the validation of the Heat-the-Chair game, we observe that, unlike the N-back test single memory task, this game includes simultaneous tasks (memory, perception, manual operations, and decision making) triggered by interruptions. Thus, the validation of these datasets is based on the correlation between perceived workload, participant performance, and the two modes of operation: with or without interruptions. As before, the perceived workload is given by the TLX questionnaire, while performance is given by the average time (in seconds) required to obtain pieces during the game. In this case, as the metric chosen to measure performance is the average time it takes the subject to obtain a piece, an increase in this time reflects a decrease in performance. Figure 13 shows boxplots of the perceived workload and performance obtained for games with and without interruptions. Since, in this case, the two quantities have different ranges, we show their boxplots with a standardized y-axis. The y-axis range shown in red on the left-hand side corresponds to the time needed to obtain pieces, while the range shown in blue on the right-hand side corresponds to the range of punctuation for TLX results, scaled between 0 and 50. As expected, both quantities are clearly increasing with the number of interruptions.

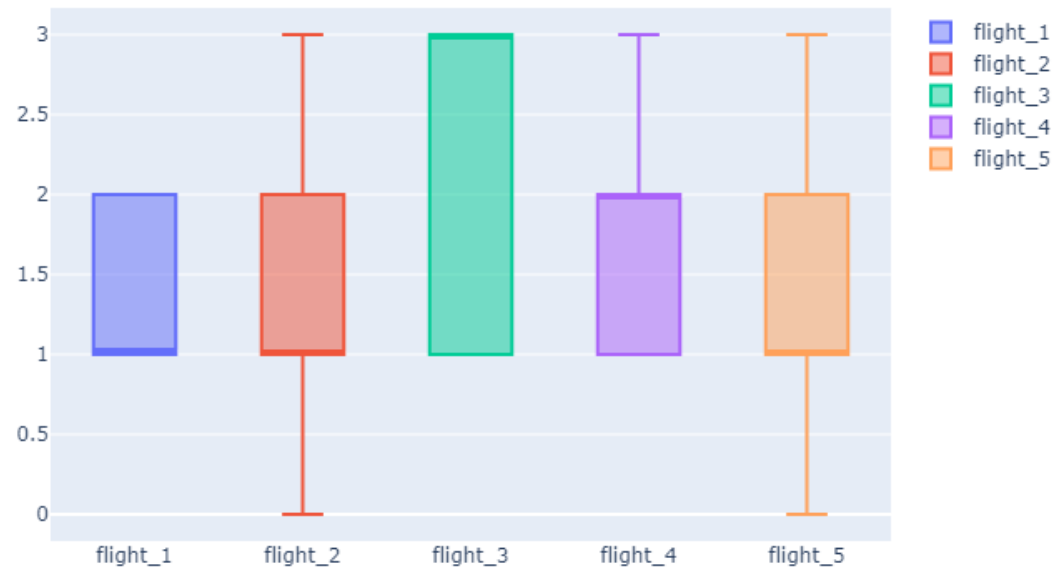


**Figure 12.** TLX analysis in the N-back test. Perceived workload and game performance across theoretical difficulty.



**Figure 13.** TLX analysis in the Heat-the-Chair test. Perceived workload and game performance across interruptions.

Given that we do not have any metric of pilot performance, for the validation of the flight simulations, we have analyzed the correlation between the distribution of the perceived difficulty (given by the dynamic TLX collected during the flight) and the global difficulty of the different flying scenarios. We recall that in this case, the perceived difficulty rates the complexity of the flight at time stamps according to 0—low, 1—mid—low, 2—mid—high, and 3—high. Figure 14 shows a boxplot of the perceived workload for the 5 flying scenarios with the median line in bold for better visualization. The theoretically easiest Flight 1 also has the lowest perceived difficulty, with values below 2 and 50% of the time stamps rated with mid—low difficulty. The twin flights, Flight 2 and Flight 5, have an identical distribution of values, with 50% of the flying time considered to be mid—low difficulty and only 25% rated as high. Flight 3, with the highest theoretical complexity, is also the one with the highest values of perceived WL, with 50% of the flight considered high-complexity by the monitoring pilot. Finally, Flight 4, of medium theoretical difficulty but with an unexpected event, is perceived as being of low—mid difficulty 50% of the time but with a 25% of high perceived complexity. This peak of complexity coincides with the time of the unexpected malfunction. This match between perceived and theoretical complexities validates as ground truth annotations both the global theoretical difficulty as well as the perceived complexity annotated by experts along the flying time.



**Figure 14.** TLX analysis in Simulated flights. The perceived workload in the flight scenarios.

#### 4.2. Analysis of Differences in Temporal Patterns

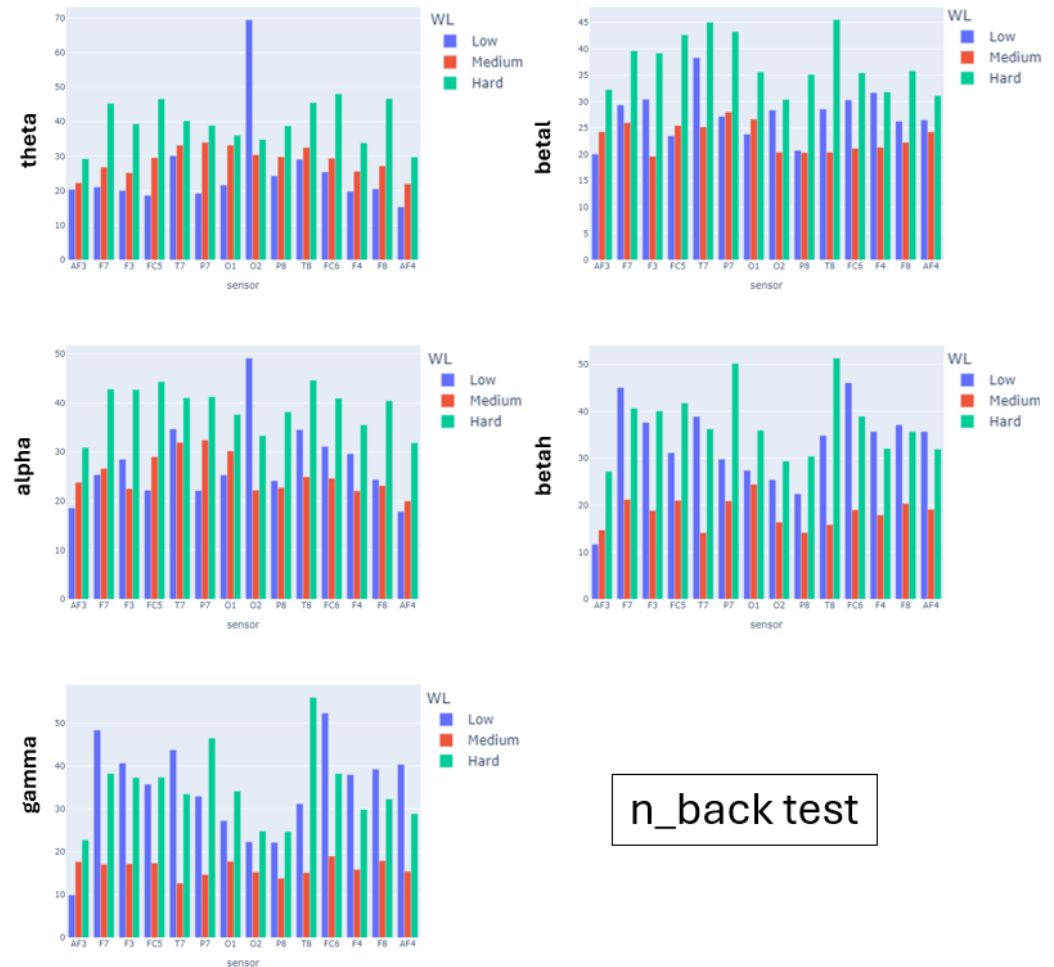
To analyze the differences in the temporal patterns of EEG recorded at different levels of task complexity, we have decomposed the signal of each EEG node into their power spectra waves ( $\theta$ ,  $\alpha$ ,  $\beta$ , and  $\gamma$ ). These signals are obtained by filtering each sensor signal at the following 4 frequency bands clinically related to main brain processes [11]:

1.  $\theta$  (4–8 Hz). The  $\theta$  activity is seen in drowsiness, arousal, and often during meditation. Dominant  $\theta$  activity is associated with relaxed, meditative, and creative states, memory recall, and ‘flow’ states. It is reported that an increase in theta, particularly frontal theta, is often associated with an increase in working memory load, especially in single-task contexts.
2.  $\alpha$  (8–12 Hz). The  $\alpha$  waves are the default ‘relaxed and alert’ mode of the brain. High  $\alpha$  levels appear in the frontal lobes during relaxation and are suppressed when other activities (like linguistic, abstract spatial thinking, or muscular) take place. It is reported that the alpha power during high-workload tasks might be lower than the alpha power during low-workload conditions.
3.  $\beta$  (12–25 Hz). This band is often associated with active, task-oriented, busy, or anxious thinking and active concentration. It is reported that beta power during a high-workload task is moderately greater than beta power in low-workload conditions. Numerous studies have established the involvement of this frequency in a variety of cognitive processes such as working memory [23], language processing [24], and decision making [25]. Since the Emotiv API provides access to two sub-bands in the  $\beta$  zone (12–18 Hz, labeled  $\beta_l$  and 18–25 Hz, labeled  $\beta_h$ ), we have analyzed both.
4.  $\gamma$  (greater than 25 Hz). The  $\gamma$  band activates when different populations of neurons network together to carry out demanding cognitive or motor functions requiring fast. Coupled processing is required [26]. It is reported that  $\gamma$  activation is related to emotions, perception, and attention. However, there are no conclusive studies of any relationship between  $\gamma$  power and WL.

For each power band, we computed the peaks of each node waveform as an indicator of temporal variability in brain activity associated with WL. Peaks were computed as local maxima with a value above 95% percentile of the node power spectra band.

Figures 15–17 show the barplots for the average number of peaks of the recordings for all participants of a given experiment. We show a different barplot for each band (rows) and experimental scenario (columns). For each barplot, we have grouped by node the number of peaks obtained under the different complexities of the experimental scenarios. For the N-back test and serious game, we have used the global theoretical difficulty, while for the

simulator, given the variety of complexity along a single flight, we have used the perceived workload grouped into easy-mild (scores 0, 1) and mild-high (scores 2, 3). The expected pattern is that the number of peaks is a monotonous function of the level of complexity for some of the EEG nodes.



**Figure 15.** Number of Peaks of EEG Nodes Power Bands. Barplots for the n\_back test Experiment.

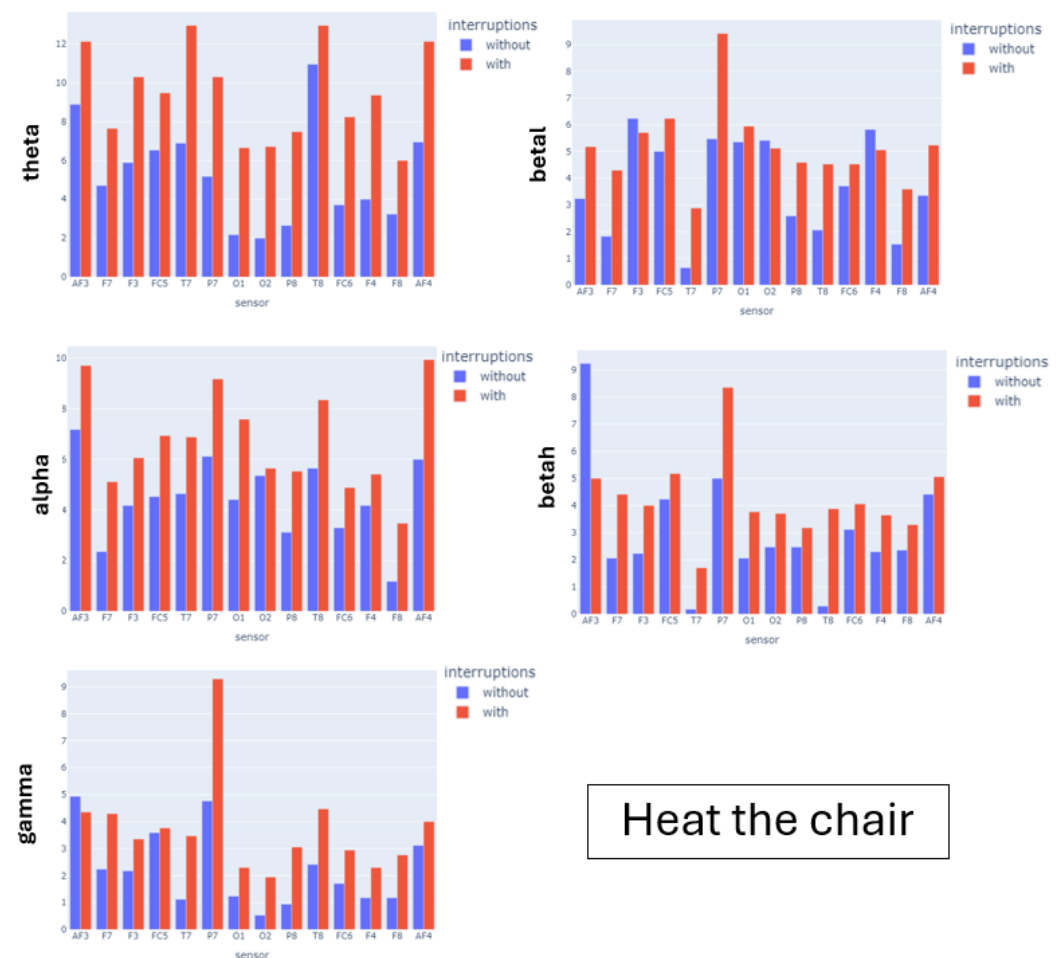
For the N-back test, the peaks of the  $\theta$  band are increasing in game complexity for all sensors except the occipital O7. This is aligned with the fact that the N-back test is a memory-demanding activity. The increasing pattern is also observed in most of the left sensors of the  $\alpha$  band associated with spatial thinking. The  $\beta$  and  $\gamma$  bands do not present a monotonous pattern, which is not surprising given that they are not associated with memory items. For the Heat-the-Chair game,  $\theta$  and  $\alpha$  are again the bands with a more prominent increasing pattern, which is followed for all nodes. This reflects the increased complexity in memory items and spatial coordination required to play the game. In this case, the  $\beta$  and  $\gamma$  bands also have an increasing pattern for most nodes, which indicates that the game requires multitask solving ( $\gamma$ ) and provokes anxious thinking ( $\beta$ ). For the flight simulator, the distribution of peaks is different. The  $\theta$  wave has a small decreasing pattern, although with very similar values (see ranges in Table 3) which are the highest ones. This indicates that flying is a knowledge-based task that requires recalling learned concepts. The peaks of the  $\alpha$  wave do not follow a well-defined pattern and are missing for some sensors. In fact, it is the band that has the lowest values (see ranges in Table 3). This suppression of the  $\alpha$  band indicates that even for easy flights, pilots flying an aircraft is a highly demanding mental task. The  $\beta_{ah}$  band is the only one presenting a clear increasing pattern for almost all nodes, which can be attributed to the fact that during hard flying

conditions, the pilot needs to make a lot of quick decisions. This is also reflected in the increasing pattern of the  $\gamma$  wave for both parietal nodes.

Tables 3–5 report descriptive statistics for the number of peaks summarized as mean  $\pm$  standard deviation computed for all nodes. The expanded descriptive statistics for each sensor can be found in Appendix A. For the N-back test, the ranges of all bands are higher during the tests compared to the ones of the baseline. However, the only band that has increasing ranges along test complexity is  $\theta$ . For serious games, the ranges of all bands are higher in the games with interruptions. Finally, for the simulator, only  $\beta_h$  and  $\gamma$  bands have higher ranges for mild-hard phases of the flight.

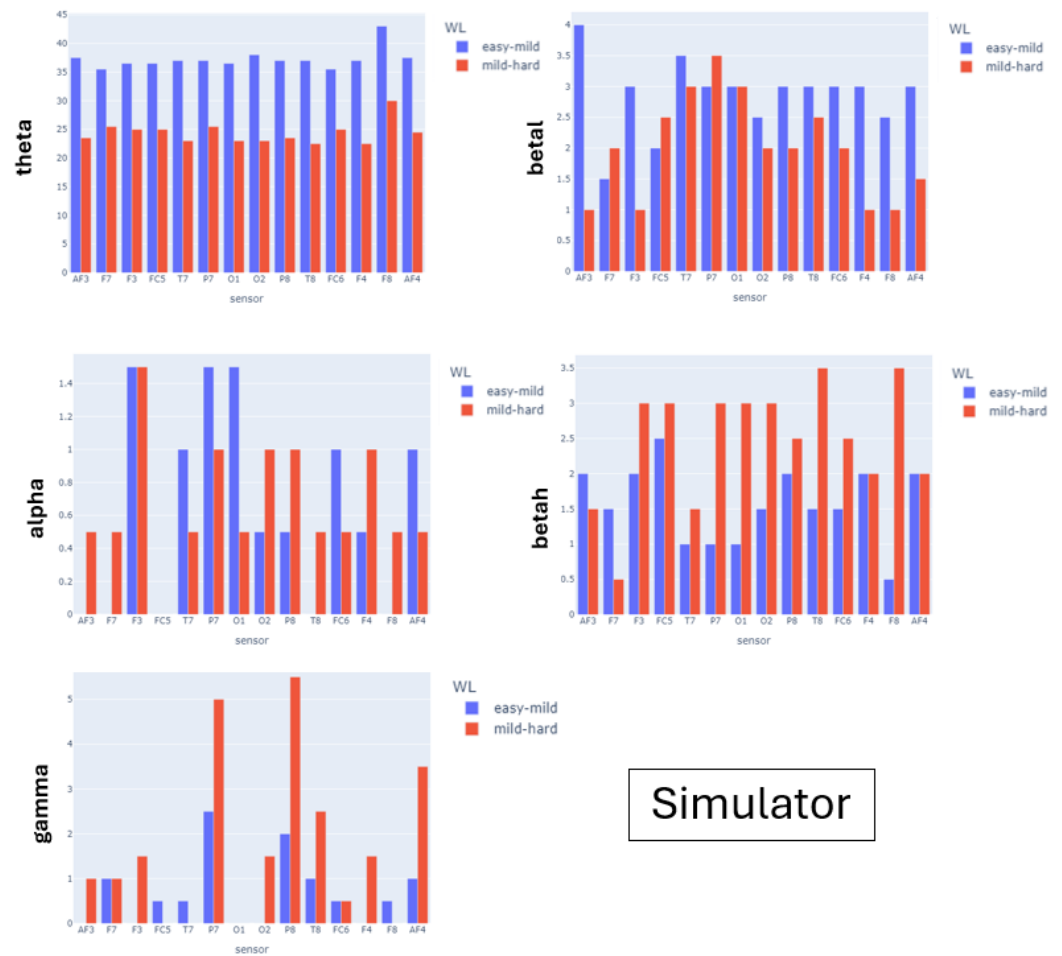
**Table 3.** Number of peaks of EEG node power bands. Summary of the descriptive statistics for the simulator experiment.

	Easy-Mild	Mild-Hard
theta	$37.25 \pm 1.73$	$24.39 \pm 1.88$
alpha	$0.64 \pm 0.58$	$0.68 \pm 0.36$
beta_l	$2.86 \pm 0.58$	$2.00 \pm 0.80$
beta_h	$1.57 \pm 0.52$	$2.46 \pm 0.83$
gamma	$0.68 \pm 0.75$	$1.68 \pm 1.76$



**Figure 16.** Number of peaks of EEG node power bands. Barplots for the Heat-the-Chair experiment.





**Figure 17.** Number of peaks of EEG node power bands. Barplots for the simulator experiment.

**Table 4.** Number of peaks of EEG node power bands. Summary of the descriptive statistics for the  $n\_back$  test.

	<b>BL</b>	<b>Low</b>	<b>Medium</b>	<b>Hard</b>
theta	95.43 ± 30.52	405.93 ± 213.24	458.57 ± 63.02	632.14 ± 100.31
alpha	125.79 ± 56.91	442.93 ± 129.18	407.07 ± 63.41	623.43 ± 72.52
beta_l	101.43 ± 25.68	440.14 ± 75.51	371.64 ± 44.39	597.43 ± 83.51
beta_h	122.49 ± 42.75	524.50 ± 145.39	294.86 ± 49.20	596.29 ± 115.51
gamma	143.57 ± 49.29	553.57 ± 181.67	259.29 ± 28.64	553.86 ± 142.94

**Table 5.** Number of peaks of EEG node power bands. Summary of the descriptive statistics for the Heat-the-Chair game.

	<b>Without</b>	<b>With</b>
theta	89.50 ± 44.29	160.57 ± 41.10
alpha	75.50 ± 27.20	115.00 ± 33.34
beta low	63.43 ± 31.00	87.79 ± 25.69
beta high	51.50 ± 38.17	71.93 ± 25.21
gamma	37.79 ± 23.91	63.50 ± 30.49

To detect if the differences in ranges were significant, we adjusted a generalized regression model [27] for the number of peaks with the complexity as a fixed factor and the sensor as a random effect to account for differences between them and correct for repeated measures. A different model was adjusted for each power band and experimental scenario. For the detection of significant differences, all models compare the number of peaks of each complexity level with the lowest-complexity ones. For each regression model, we report model parameters,  $p$ -values for significance in fixed effects, 95% CI for their mean values. In the case of a change in scale or transformation of data required to satisfy model assumptions, the CIs were back-transformed to the original scale. For all statistical analysis, a  $p$ -value  $< 0.05$  was considered significant. Statistical analyses were conducted using R version 4.3.2.

Tables 6–8 report a summary of the regression models for each of the experimental scenarios. For the N-back test, the complexity factor was significant for all bands except  $\gamma$ . For the  $\theta$  band, peaks were significantly increasing with complexity. This is not the case for the remaining bands, where models detect a significant decrease in the number of peaks of the Medium complexity. The peaks of the high-complexity waves are significantly higher than the low-complexity waves, except for  $\gamma$ . The models for the Heat-the-Chair game detect a significant increase in the number of peaks of games with interruptions for all bands. Finally, for the flight simulator, models detect a significant increase in the number of peaks of the  $\beta$  waves and a significant decrease for  $\theta$ . There are no significant differences for  $\alpha$  and  $\theta$ .

**Table 6.** Number of peaks of EEG node power bands. Regression model for the N-back test.

		Coefficient	$p$ -Value	95% CI
$\theta$	Low	$5.93 \times 10^0$	-	(326.66, 425.48)
	Medium	$1.89 \times 10^{-1}$	0.02	(394.68, 514.09)
	Hard	$5.07 \times 10^{-1}$	$<0.001$	(542.43, 706.53)
$\alpha$	Low	$6.06 \times 10^0$	-	(383.55, 471.91)
	Medium	$2.81 \times 10^2$	$<0.001$	(361.16, 452.99)
	Hard	$4.98 \times 10^2$	$<0.001$	(577.51, 669.34)
$\beta_l$	Low	$6.07 \times 10^0$	-	(400.17, 468.23)
	Medium	$-1.62 \times 10^{-1}$	$<0.001$	(340.27, 398.14)
	Hard	$3.10 \times 10^{-1}$	$<0.001$	(545.68, 638.48)
$\beta_h$	Low	$6.21 \times 10^0$	-	(432.24, 568.05)
	Medium	$-5.42 \times 10^{-1}$	$<0.001$	(251.50, 330.53)
	Hard	$1.59 \times 10^{-1}$	0.03	(506.76, 665.99)
$\gamma$	Low	$6.25 \times 10^0$	-	(434.64, 598.50)
	Medium	$-6.95 \times 10^{-1}$	$<0.001$	(216.88, 298.64)
	Hard	$4.06 \times 10^{-2}$	0.67	(452.64, 623.28)

**Table 7.** Number of peaks of EEG power bands. Regression model for the Heat-the-Chair game.

		Coefficient	$p$ -Value	95% CI
$\theta$	without	$4.38 \times 10^0$	-	(61.19, 97.85)
	with	$6.71 \times 10^{-1}$	$<0.001$	(119.76, 191.50)
$\alpha$	without	$4.24 \times 10^0$	-	(54.09, 84.98)
	with	$4.63 \times 10^{-1}$	$<0.001$	(85.90, 134.95)
$\beta_l$	without	$3.99 \times 10^0$	-	(39.07, 69.27)
	with	$4.47 \times 10^{-1}$	$<0.001$	(61.11, 108.35)
$\beta_h$	without	$3.60 \times 10^0$	-	(20.72, 52.21)
	with	$6.25 \times 10^{-1}$	0.02	(38.71, 97.56)
$\gamma$	without	$3.44 \times 10^0$	-	(21.80, 40.69)
	with	$6.31 \times 10^{-1}$	$<0.001$	(40.98, 76.50)

**Table 8.** Number of peaks of EEG power bands. Regression model for the flight simulator.

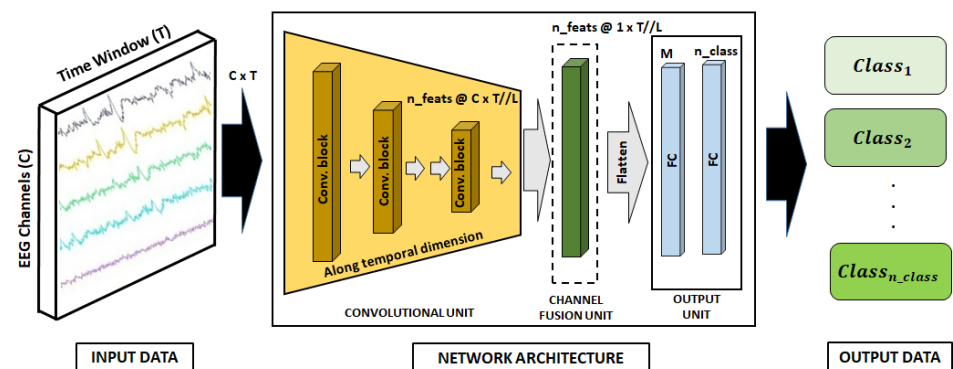
		Coefficient	<i>p</i> -Value	95% CI
$\theta$	easy-mild	$3.62 \times 10^0$	-	(35.92, 38.51)
	mild-hard	$-4.25 \times 10^{-1}$	<0.001	(23.48, 25.17)
$\alpha$	easy-mild	$6.43 \times 10^{-1}$	-	(0.37, 0.92)
	mild-hard	$3.57 \times 10^{-2}$	0.85	(0.40, 0.95)
$\beta_l$	easy-mild	$2.86 \times 10^0$	-	(2.46, 3.26)
	mild-hard	$-8.57 \times 10^{-1}$	<0.001	(1.60, 2.40)
$\beta_h$	easy-mild	$1.57 \times 10^0$	-	(1.17, 1.97)
	mild-hard	$8.93 \times 10^{-1}$	<0.001	(2.07, 2.86)
$\gamma$	easy-mild	$6.79 \times 10^{-1}$	-	(-0.09, 1.45)
	mild-hard	$1.00 \times 10^0$	0.06	(0.91, 2.45)

#### 4.3. Usefulness for Training AI Models

In this section, we illustrate the usefulness of the EEG dataset by training AI models. We report partial results of previous work on DL models for WL detection, published in [22]. In that work, several architectures for EEG channel fusion were presented and validated on the *n*\_back test data using a leave-one-subject-out cross-validation scheme. The best performers trained on the whole *n*\_back set were also validated on the Heat-the-Chair set to assess the task transfer of models. We summarize the main findings and report preliminary results obtained on the flight simulator by models trained on the *n*back test.

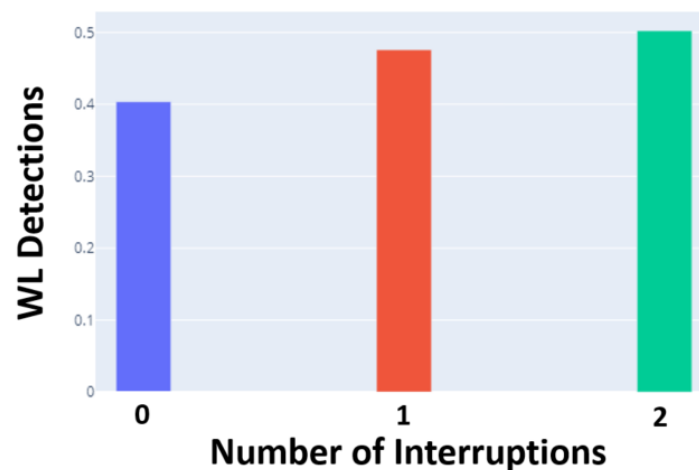
Our best-performing architecture was a convolutional neural network (CNN) that fuses the channels of the features obtained after the convolutional block and before the Fully Connected (FC) layers. Concretely, signals are processed as follows: input data are taken by the input data module and feed the convolutional module, which performs feature extraction. The number of channels remains along convolutions so that a channel fusion unit transforms them into a single signal channel for each feature. The outcome of the previous unit is flattened and processed by two FC layers to combine features and perform output predictions.

Figure 18 depicts the proposed neural architecture. Notice that *C* corresponds to the number of channels, *T* to the time window, *L* to the convolution, and  $n_{feats}$  to the number of features extracted by the convolutional block  $L^{th}$ . Hence, after the convolutional process  $L$ ,  $x^{C \times T}$  becomes  $x^{n_{feats} \times C \times (T//L)}$ , and  $T//L$  is due to the pooling operation in each convolutional block. During the channel fusion unit, the input  $x^{n_{feats} \times C \times (T//L)}$  becomes  $x^{n_{feats} \times 1 \times (T//L)}$ . The convolution unit has 3 blocks consisting of one convolutional layer with max pooling and with 16, 32, and 64 neurons for each convolutional layer, respectively. The classification layer has 256 neurons. The output unit has 2 blocks consisting of one convolutional layer before the classification layer. The first one has 64 neurons, and the second one projects convolutional features also using 64 neurons.

**Figure 18.** Neural network architecture for mid-fusion.

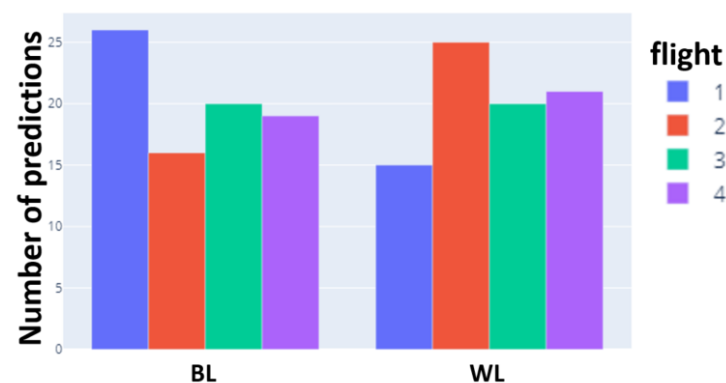
For the validation of the models, the N-back test data were split using a leave-one-subject-out scheme, and 16 independent models were trained on 15 participants and tested in the remaining one. The models were trained for the classification of BL and medium phases using a weighted cross-entropy loss to compensate for imbalances between baseline and workload phases. We used a batch size of 750, Adam as the optimization method, 100 epochs, and a learning rate of 0.0001. The quality metrics were the sensitivity and specificity, considering BL the positive class.

Results show a 76.25% sensitivity and 87.81% specificity in WL detection for a leave-one-out subject evaluation in the N-back-test data and good task transfer with the detected WL increasing with the number of interruptions in the Heat-the-Chair game (see Figure 19).



**Figure 19.** Correspondence between interruptions in the Heat-the-Chair game and the number of WL detection.

Regarding the capability of task transfer, Figure 20 shows the barplots for the number of predictions in BL and WL for 4 of the 5 flights (the last has not been checked, as it is very similar to the second). As expected, the highest number of BL detections is for Flight 1, while Flight 2 and Flight 4 show more WL detections since they are characterized as more difficult. Flight 3 was not as expected and could be attributed to the discrepancies in waveform between synthetic games and the simulator, detected in Section 4.2.



**Figure 20.** Correspondence between the number of interruptions in the Heat-the-Chair game and the number of WL detections by the neural network.

## 5. Conclusions

This paper provides a complete dataset of physiological recordings from electroencephalogram (EEG) and electrocardiogram (ECG) devices, which are useful for testing methods that recognize mental workload. Three different experiments have been presented in whose participants were induced to different levels of workload:

1. on the well-known dual N-back game
2. on a specifically designed serious game mimicking the increase of workload an aircraft pilot can suffer
3. on a flight simulator

Technical validation at three different levels shows the correlation between objective measures from the experiments and the corresponding subjective self-perceived complexity from subjects. Our games are specifically designed to target mental activities and, thus, can be used to assess the capability of a physiological sensor to detect mental WL or any other specific mental effort. Moreover, we have shown that they could be powerful means for collecting unambiguous annotated data valid for training AI models.

However, there is room for improvement. The results obtained on the transfer to the simulator are sub-optimal for an AI system deployable in the cockpit. The analysis of the power band patterns of Section 4.2 shows that the synthetic games have a mental demand different from that of flying pilots. Therefore, more specific serious games should be designed to guarantee a fully successful transfer to flight-deck scenarios.

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

**Funding:** This work was partially supported by the Cleansky grant number 831993, MCIN/AEI/10.13039/501100011033 and “ERDF A way of making Europe” (grant PID2021-126776OB-C21), and the European Union NextGenerationEU/PRTR (grant TED2021-132802B-I00), the Department of Research and Universities of the Generalitat of Catalonia in the SGR 2021 CICS Research Group (Code: 2021 SGR 01623) and CERCA Programme/Generalitat de Catalunya.

**Institutional Review Board Statement:** The Ethics Committee on Animal and Human Research (CEEA) of the Universitat Autònoma de Barcelona, Spain, provided us with an approval letter to collect neurophysiological data in the Project “E-PILOTS (H2020)” Grant agreement ID: 831993. The letter was signed on 2 September 2022, by Núria Pérez Pastor, Secretary of the CEEA.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** The dataset is publicly available at the Digital Document Deposit of the Universitat Autònoma de Barcelona, accessible at <https://doi.org/10.5565/ddd.uab.cat/259591> (accessed on 26 December 2023). No registration is required for anyone who would like to download and use these data.

**Conflicts of Interest:** Author Daniel Álvarez was employed by the company Aslogic. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Appendix A. Extended Descriptive Statistics

In this appendix, we show the tables of expanded descriptive statistics for each experiment. Each cell reports the mean of the number of peaks for all subjects in a specific sensor and WL. Rows group the sensor signals based on their power spectra, while columns correspond to the different levels of WL.

**Table A1.** Number of Peaks of EEG Power Bands. Descriptive statistics for the n\_back test.

		BL	Low	Medium	Hard			BL	Low	Medium	Hard	
$\theta$		AF3	3.00	20.38	22.31	29.25	$\beta_t$	AF3	4.00	20.06	24.25	32.25
		F7	5.00	21.06	26.81	45.25		F7	5.00	29.38	26.00	39.63
		F3	4.00	20.06	25.19	39.38		F3	6.00	30.44	19.63	39.19
		FC5	5.00	18.63	29.56	46.56		FC5	7.00	23.50	25.44	42.63
		T7	6.00	30.13	33.19	40.25		T7	6.00	38.31	25.19	45.00
		P7	6.00	19.25	34.00	38.94		P7	6.00	27.19	28.00	43.25
		O1	5.00	21.63	33.19	36.06		O1	6.00	23.81	26.69	35.63
		O2	10.00	69.50	30.44	34.81		O2	7.00	28.38	20.38	30.38
		P8	7.00	24.31	29.81	38.75		P8	6.00	20.75	20.31	35.13
		T8	9.00	29.13	32.50	45.50		T8	10.00	28.56	20.38	45.50
		FC6	7.00	25.44	29.44	48.06		FC6	6.00	30.31	21.13	35.44
		F4	7.00	19.88	25.63	33.88		F4	9.00	31.69	21.31	31.81
		F8	5.00	20.50	27.19	46.69		F8	5.00	26.25	22.25	35.81
		AF4	4.00	15.31	22.00	29.75		AF4	5.00	26.50	24.25	31.13
		mean	<b>5.93</b>	<b>25.37</b>	<b>28.66</b>	<b>39.51</b>		mean	<b>6.29</b>	<b>27.51</b>	<b>23.23</b>	<b>37.34</b>
		std	<b>1.94</b>	<b>13.33</b>	<b>3.94</b>	<b>6.27</b>		std	<b>1.59</b>	<b>4.72</b>	<b>2.77</b>	<b>5.22</b>
$\alpha$		AF3	3.00	18.56	23.81	30.88	$\beta_h$	AF3	3.00	11.69	14.69	27.19
		F7	6.00	25.38	26.63	42.81		F7	8.00	45.06	21.19	40.63
		F3	6.00	28.50	22.50	42.69		F3	8.00	37.63	18.88	40.06
		FC5	7.00	22.19	29.00	44.31		FC5	7.00	31.19	21.00	41.75
		T7	5.00	34.69	31.94	41.06		T7	10.00	38.88	14.13	36.25
		P7	6.00	22.13	32.44	41.25		P7	6.00	29.81	20.88	50.19
		O1	6.00	25.31	30.19	37.63		O1	7.00	27.44	24.44	35.94
		O2	12.00	49.13	22.19	33.31		O2	5.00	25.44	16.38	29.38
		P8	9.00	24.13	22.75	38.19		P8	6.00	22.44	14.19	30.44
		T8	16.00	34.56	24.94	44.63		T8	14.00	34.81	15.81	51.31
		FC6	11.00	31.13	24.63	40.94		FC6	10.00	46.06	19.00	38.94
		F4	11.00	29.63	22.06	35.50		F4	10.00	35.69	17.94	32.06
		F8	7.00	24.38	23.13	40.44		F8	8.00	37.13	20.38	35.69
		AF4	5.00	17.88	20.00	31.88		AF4	6.00	35.69	19.13	31.94
		mean	<b>7.86</b>	<b>27.68</b>	<b>25.44</b>	<b>38.96</b>		mean	<b>7.71</b>	<b>32.78</b>	<b>18.43</b>	<b>37.27</b>
		std	<b>3.51</b>	<b>8.07</b>	<b>3.96</b>	<b>4.53</b>		std	<b>2.70</b>	<b>9.09</b>	<b>3.07</b>	<b>7.22</b>
$\gamma$		AF3	3.00	9.94	17.69	22.75						
		F7	9.00	48.38	17.06	38.25						
		F3	8.00	40.69	17.19	37.31						
		FC5	9.00	35.75	17.38	37.38						
		T7	10.00	43.75	12.69	33.50						
		P7	7.00	32.94	14.69	46.50						
		O1	8.00	27.25	17.75	34.13						
		O2	5.00	22.31	15.25	24.81						
		P8	6.00	22.19	13.81	24.69						
		T8	14.00	31.19	15.13	56.00						
		FC6	14.00	52.38	19.00	38.25						
		F4	13.00	38.00	15.88	29.88						
		F8	10.00	39.25	17.94	32.31						
		AF4	10.00	40.38	15.44	28.88						
		mean	<b>9.00</b>	<b>34.60</b>	<b>16.21</b>	<b>34.62</b>						
		std	<b>3.23</b>	<b>11.35</b>	<b>1.79</b>	<b>8.93</b>						

**Table A2.** Number of Peaks of EEG Power Bands. Descriptive statistics for the Heat-the-Chair game.

		Without	With		Without	With	
$\theta$	AF3	8.88	12.12	$\beta_l$	AF3	3.24	5.18
	F7	4.71	7.65		F7	1.82	4.29
	F3	5.88	10.29		F3	6.24	5.71
	FC5	6.53	9.47		FC5	5.00	6.24
	T7	6.88	12.94		T7	0.65	2.88
	P7	5.18	10.29		P7	5.47	9.41
	O1	2.18	6.65		O1	5.35	5.94
	O2	2.00	6.71		O2	5.41	5.12
	P8	2.65	7.47		P8	2.59	4.59
	T8	10.94	12.94		T8	2.06	4.53
	FC6	3.71	8.24		FC6	3.71	4.53
	F4	4.00	9.35		F4	5.82	5.06
	F8	3.24	6.00		F8	1.53	3.59
	AF4	6.94	12.12		AF4	3.35	5.24
	mean	5.26	9.45		mean	3.73	5.16
	std	2.51	2.33		std	1.76	1.46
	$\alpha$	AF3	7.18		9.71	$\beta_h$	AF3
F7		2.35	5.12	F7	2.06		4.41
F3		4.18	6.06	F3	2.24		4.00
FC5		4.53	6.94	FC5	4.24		5.18
T7		4.65	6.88	T7	0.18		1.71
P7		6.12	9.18	P7	5.00		8.35
O1		4.41	7.59	O1	2.06		3.76
O2		5.35	5.65	O2	2.47		3.71
P8		3.12	5.53	P8	2.47		3.18
T8		5.65	8.35	T8	0.29		3.88
FC6		3.29	4.88	FC6	3.12		4.06
F4		4.18	5.41	F4	2.29		3.65
F8		1.18	3.47	F8	2.35		3.29
AF4		6.00	9.94	AF4	4.41		5.06
mean		4.44	6.76	mean	3.03		4.23
std		1.54	1.89	std	2.16		1.43
$\gamma$		AF3	4.94	4.35			
	F7	2.24	4.29				
	F3	2.18	3.35				
	FC5	3.59	3.76				
	T7	1.12	3.47				
	P7	4.76	9.29				
	O1	1.24	2.29				
	O2	0.53	1.94				
	P8	0.94	3.06				
	T8	2.41	4.47				
	FC6	1.71	2.94				
	F4	1.18	2.29				
	F8	1.18	2.76				
	AF4	3.12	4.00				
	mean	2.22	3.74				
	std	1.36	1.73				

**Table A3.** Number of Peaks of EEG Power Bands. Descriptive statistics for the simulator experiment.

		Easy-Mild	Mild-Hard			Easy-Mild	Mild-Hard
$\theta$	AF3	37.5	23.5	$\beta_t$	AF3	4	1
	F7	35.5	25.5		F7	1.5	2
	F3	36.5	25		F3	3	1
	FC5	36.5	25		FC5	2	2.5
	T7	37	23		T7	3.5	3
	P7	37	25.5		P7	3	3.5
	O1	36.5	23		O1	3	3
	O2	38	23		O2	2.5	2
	P8	37	23.5		P8	3	2
	T8	37	22.5		T8	3	2.5
	FC6	35.5	25		FC6	3	2
	F4	37	22.5		F4	3	1
	F8	43	30		F8	2.5	1
	AF4	37.5	24.5		AF4	3	1.5
	mean	37.25	24.39286		mean	2.857143	2
std	1.729471	1.882045	std	0.580288	0.801784		
$\alpha$	AF3	0	0.5	$\beta_h$	AF3	2	1.5
	F7	0	0.5		F7	1.5	0.5
	F3	1.5	1.5		F3	2	3
	FC5	0	0		FC5	2.5	3
	T7	1	0.5		T7	1	1.5
	P7	1.5	1		P7	1	3
	O1	1.5	0.5		O1	1	3
	O2	0.5	1		O2	1.5	3
	P8	0.5	1		P8	2	2.5
	T8	0	0.5		T8	1.5	3.5
	FC6	1	0.5		FC6	1.5	2.5
	F4	0.5	1		F4	2	2
	F8	0	0.5		F8	0.5	3.5
	AF4	1	0.5		AF4	2	2
	mean	0.642857	0.678571		mean	1.571429	2.464286
std	0.580288	0.358924	std	0.529728	0.833758		
$\gamma$	AF3	0	1				
	F7	1	1				
	F3	0	1.5				
	FC5	0.5	0				
	T7	0.5	0				
	P7	2.5	5				
	O1	0	0				
	O2	0	1.5				
	P8	2	5.5				
	T8	1	2.5				
	FC6	0.5	0.5				
	F4	0	1.5				
	F8	0.5	0				
	AF4	1	3.5				
	mean	0.678571	1.678571				
std	0.746591	1.758725					



## References

1. 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. *Neurosci. Biobehav. Rev.* **2014**, *44*, 58–75. [[CrossRef](#)] [[PubMed](#)]
2. Lin, Y.; Mutz, J.; Clough, P.J.; Papageorgiou, K.A. Mental Toughness and Individual Differences in Learning, Educational and Work Performance, Psychological Well-being, and Personality: A Systematic Review. *Front. Psychol.* **2017**, *8*, 1345. [[CrossRef](#)]
3. Chin, Z.Y.; Zhang, X.; Wang, C.; Ang, K.K. EEG-based Discrimination of Different Cognitive Workload Levels from Mental Arithmetic. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, Honolulu, HI, USA, 18–21 July 2018; Volume 2018, pp. 1984–1987. [[CrossRef](#)]
4. Zhao, Y.; Tang, J.; Cao, Y.; Jiao, X.; Xu, M.; Zhou, P.; Ming, D.; Qi, H. Effects of Distracting Task with Different Mental Workload on Steady-State Visual Evoked Potential Based Brain Computer Interfaces—An Offline Study. *Front. Neurosci.* **2018**, *12*, 79. [[CrossRef](#)]
5. Alifah, S.K.; Widyantara, P.B.; Puspasari, M.A. The Effect of Mental Workload Towards Mental Fatigue on Customer Care Agent using Electroencephalogram. In Proceedings of the 5th International Conference on Industrial and Business Engineering, Hong Kong, China, 27–29 September 2019; pp. 173–177. [[CrossRef](#)]
6. Huang, J.; Pugh, Z.H.; Kim, S.; Nam, C.S. Brain dynamics of mental workload in a multitasking context: Evidence from dynamic causal modeling. *Comput. Hum. Behav.* **2024**, *152*, 108043. [[CrossRef](#)]
7. Gao, S.; Wang, L. Effects of mental workload and risk perception on pilots' safety performance in adverse weather contexts. In Proceedings of the Engineering Psychology and Cognitive Ergonomics. Cognition and Design: 17th International Conference, EPCE 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, 19–24 July 2020; pp. 278–291.
8. McDonnell, A.S.; Crabtree, K.W.; Cooper, J.M.; Strayer, D.L. This is your brain on Autopilot 2.0: The influence of practice on driver workload and engagement during on-road, partially automated driving. *Hum. Factors* **2023**. [[CrossRef](#)]
9. NASA Task Load Index (TLX), v. 1.0, Manual 1986. Available online: <https://humansystems.arc.nasa.gov/groups/TLX/downloads/TLX.pdf> (accessed on 26 December 2023)
10. Rim, B.; Sung, N.J.; Min, S.; Hong, M. Deep Learning in Physiological Signal Data: A Survey. *Sensors* **2020**, *20*, 969. [[CrossRef](#)] [[PubMed](#)]
11. Chikhi, S.; Matton, N.; Blanchet, S. EEG power spectral measures of cognitive workload: A meta-analysis. *Psychophysiology* **2022**, *59*, e14009. [[CrossRef](#)]
12. Di Flumeri, G.; Borghini, G.; Aricò, P.; Sciaraffa, N.; Lanzi, P.; Pozzi, S.; Vignali, V.; Lantieri, C.; Bichicchi, A.; Simone, A.; et al. EEG-based Mental Workload Neurometric to Evaluate the Impact of Different Traffic and Road Conditions in Real Driving Settings. *Front. Hum. Neurosci.* **2018**, *12*, 509. [[CrossRef](#)]
13. Chen, J.; Asce, A.M.; Taylor, J.E.; Asce, M.; Comu, S. Assessing Task Mental Workload in Construction Projects: A Novel Electroencephalography Approach. *J. Constr. Eng. Manag.* **2017**, *143*, 04017053. [[CrossRef](#)]
14. Jaeggi, S.M.; Buschkuhl, M.; Jonides, J.; Perrig, W.J. Improving Fluid Intelligence with Training on Working Memory. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 6829–6833. [[CrossRef](#)]
15. Sevchenko, N.; Ninaus, M.; Wortha, F.; Moeller, K.; Gerjets, P. Measuring Cognitive Load Using In-Game Metrics of a Serious Simulation Game. *Front. Psychol.* **2021**, *12*, 906. [[CrossRef](#)] [[PubMed](#)]
16. Lim, W.L.; Sourina, O.; Wang, L.P. STEW: Simultaneous Task EEG Workload Data Set. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2018**, *26*, 2106–2114. [[CrossRef](#)]
17. Beh, W.K.; Wu, Y.H.; Wu, A.Y.A. MAUS: A Dataset for Mental Workload Assessment on N-back task Using Wearable Sensor. 2021. Available online: <https://iee-dataport.org/open-access/maus-dataset-mental-workload-assessment-n-back-task-using-wearable-sensor> (accessed on 26 December 2023).
18. Qu, H.; Shan, Y.; Liu, Y.; Pang, L.; Fan, Z.; Zhang, J.; Wanyan, X. Mental Workload Classification Method Based on EEG Independent Component Features. *Appl. Sci.* **2020**, *10*, 3036. [[CrossRef](#)]
19. Han, S.Y.; Kwak, N.S.; Oh, T.; Lee, S.W. Classification of Pilots' Mental States Using a Multimodal Deep Learning Network. *Biocybern. Biomed. Eng.* **2020**, *40*, 324–336. [[CrossRef](#)]
20. Dehais, F.; Duprès, A.; Blum, S.; Drougard, N.; Scannella, S.; Roy, R.N.; Lotte, F. Monitoring Pilot's Mental Workload Using ERPs and Spectral Power with a Six-Dry-Electrode EEG System in Real Flight Conditions. *Sensors* **2019**, *19*, 1324. [[CrossRef](#)] [[PubMed](#)]
21. Ahmad, M.I.; Keller, I.; Robb, D.A.; Lohan, K.S. A Framework to Estimate Cognitive Load Using Physiological Data. *Pers. Ubiquitous Comput.* **2020**, *27*, 2027–2041. [[CrossRef](#)]
22. Hernández-Sabaté, A.; Yauri, J.; Folch, P.; Piera, M.À.; Gil, D. Recognition of the Mental Workloads of Pilots in the Cockpit Using EEG Signals. *Appl. Sci.* **2022**, *12*, 2298. [[CrossRef](#)]
23. Chen, Y.; Huang, X. Modulation of alpha and beta oscillations during an n-back task with varying temporal memory load. *Front. Psychol.* **2016**, *6*, 2031. [[CrossRef](#)]
24. Weiss, S.; Mueller, H.M. "Too many betas do not spoil the broth": The role of beta brain oscillations in language processing. *Front. Psychol.* **2012**, *3*, 201. [[CrossRef](#)]
25. Marco-Pallarés, J.; Münte, T.F.; Rodríguez-Fornells, A. The role of high-frequency oscillatory activity in reward processing and learning. *Neurosci. Biobehav. Rev.* **2015**, *49*, 1–7. [[CrossRef](#)]

- 
26. Amo, C.; De Santiago, L.; Barea, R.; López-Dorado, A.; Boquete, L. Analysis of gamma-band activity from human EEG using empirical mode decomposition. *Sensors* **2017**, *17*, 989. [[CrossRef](#)] [[PubMed](#)]
  27. Booth, J.G. Generalized Linear Models with Random Effects: Unified Analysis via H-Likelihood by Y. Lee, J. A. Nelder, and Y. Pawitan. *Biometrics* **2007**, *63*, 1296–1297 [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.