applied sciences

MDPI

Article

Unsupervised Detection of Covariate Shift Due to Changes in EEG Headset Position: Towards an Effective Out-of-Lab Use of Passive Brain–Computer Interface

Daniele Germano [1,2,*], Nicolina Sciaraffa [3], Vincenzo Ronca [2,3], Andrea Giorgi [2,3], Giacomo Trulli [2], Gianluca Borghini [2,3], Gianluca Di Flumeri [2,3], Fabio Babiloni [2,3,4] and Pietro Aricò [1,3]

[1] Department of Computer, Control and Management Engineering Antonio Ruberti, Sapienza University of Rome, 00185 Rome, Italy; pietro.arico@uniroma1.it

[2] Laboratory of Industrial Neuroscience, Department of Molecular Medicine, Sapienza University of Rome, 00185 Rome, Italy; vincenzo.ronca@uniroma1.it (V.R.); andrea.giorgi@uniroma1.it (A.G.); giac.trulli@gmail.com (G.T.); gianluca.borghini@uniroma1.it (G.B.); gianluca.diflumeri@uniroma1.it (G.D.F.); fabio.babiloni@uniroma1.it (F.B.)

[3] BrainSigns srl, Lungotevere Michelangelo 9, 00198 Rome, Italy; nicolina.sciaraffa@gmail.com

[4] School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China

* Correspondence: daniele.germano@uniroma1.it

Abstract: In the field of passive Brain–computer Interfaces (BCI), the need to develop systems that require rapid setup, suitable for use outside of laboratories is a fundamental challenge, especially now, that the market is flooded with novel EEG headsets with a good quality. However, the lack of control in operational conditions can compromise the performance of the machine learning model behind the BCI system. First, this study focuses on evaluating the performance loss of the BCI system, induced by a different positioning of the EEG headset (and of course sensors), so generating a variation in the control features used to calibrate the machine learning algorithm. This phenomenon is called covariate shift. Detecting covariate shift occurrences in advance allows for preventive measures, such as informing the user to adjust the position of the headset or applying specific corrections in new coming data. We used in this study an unsupervised Machine Learning model, the Isolation Forest, to detect covariate shift occurrence in new coming data. We tested the method on two different datasets, one in a controlled setting (9 participants), and the other in a more realistic setting (10 participants). In the controlled dataset, we simulated the movement of the EEG cap using different channel and reference configurations. For each test configuration, we selected a set of electrodes near the control electrodes. Regarding the realistic dataset, we aimed to simulate the use of the cap outside the laboratory, mimicking the removal and repositioning of the cap by a non-expert user. In both datasets, we recorded multiple test sessions for each configuration while executing a set of Workload tasks. The results obtained using the Isolation Forest model allowed the identification of covariate shift in the data, even with a 15-s recording sample. Moreover, the results showed a strong and significant negative correlation between the percentage of covariate shift detected by the method, and the accuracy of the passive BCI system (p-value < 0.01). This novel approach opens new perspectives for developing more robust and flexible BCI systems, with the potential to move these technologies towards out-of-the-lab use, without the need for supervision for use by a non-expert user.

Keywords: passive brain–computer interface; electroencephalography; machine learning; covariate shift

check for updates

Citation: Germano, D.; Sciaraffa, N.; Ronca, V.; Giorgi, A.; Trulli, G.; Borghini, G.; Di Flumeri, G.; Babiloni, F.; Aricò, P. Unsupervised Detection of Covariate Shift Due to Changes in EEG Headset Position: Towards an Effective Out-of-Lab Use of Passive Brain–Computer Interface. *Appl. Sci.* **2023**, *13*, 12800. https://doi.org/10.3390/app132312800

Academic Editors: Sung Bum Pan and EunSang Bak

Received: 27 October 2023
Revised: 23 November 2023
Accepted: 27 November 2023
Published: 29 November 2023

1. Introduction

Industry 4.0, often referred to as the fourth industrial revolution, embodies a landscape where digital technologies intertwine with production processes, redefining the very nature of enterprises and human interaction with the surrounding environment [1]. Within this context, the integration of passive Brain–computer Interface (BCI) technologies may enable direct engagement of the user in the productive and/or industrial process [2,3]. Unlike standard and subjective methods (e.g., questionnaires), these technologies allow for a covert estimation (i.e., without asking anything to the user) of mental and emotional states of the user while engaged in the operational task, allowing for continuous feedback with the surrounding environment to put the user in the loop.

One of the most relevant mental states to be mentioned for out-of-the-lab applications, especially in safety-critical environments, is undoubtedly the mental workload [4–6]. In fact, it has been widely demonstrated that an operator's mental workload level (overload) that is too high could lead to degraded performance and/or an increased chance of committing errors [7].

The concept of mental workload cannot be viewed as singular; rather, it arises from various interrelated aspects. Numerous mental processes, including alertness, vigilance, mental effort, attention, mental fatigue, drowsiness, and more, may come into play during task execution. Moreover, these processes can be influenced at any given moment by the specific demands of the task [8–10].

The primary purpose of workload measurement is to assess the mental effort involved in task performance, aiming to predict potential declines in operator performance during operational tasks. This approach allows us to gather information about when a task might become overly complex for a participant or when a participant is experiencing cognitive strain [11–13]. This measure is, therefore, particularly important in safety-critical contexts, where human error (e.g., induced using overload or lack of attention) could induce disastrous consequences (e.g., piloting an airplane or air traffic management) [14,15].

Numerous studies have highlighted a significant correlation between the amplitude of Electroencephalographic (EEG) signals recorded on frontal and parietal EEG channels and variations in workload levels. Specifically, a significant EEG amplitude difference is observed in the Theta band on frontal channels and the Alpha band on parietal channels. For example, research conducted first by Gevins and later by Puma and Raufi [16–18] has demonstrated that variations in EEG signals detected in these cerebral regions are closely linked to changes in mental workload during the execution of working memory tasks. For example, Puma et al. conducted an experiment on 20 participants who were asked to perform the Priority Management Task, currently used by the Italian Civil Aviation Authority (ENAC) for the selection of airline pilots. These findings have been confirmed by a large number of subsequent studies performed in realistic settings [8,19], paving the way for an out-of-the-lab use of such systems in the near future.

In the current scenario, many studies have focused on creating pBCI devices that are more portable and comfortable but, at the same time, reliable [20–22]. As a result, there could be potential use not only within the confines of the laboratory but also in real-world settings, driving the use of passive BCI models in more practical contexts [23–27].

However, while every variable can be fully controlled in a laboratory environment, such control cannot be guaranteed in real operational situations [24,28].

In this regard, the amplitude of the EEG signal, and so of the EEG power spectrum, used often as a control feature of the passive BCI system for the mental states measurement, is dependent on many external variables, apart from the variations in user's mental states, for example, the impedance values of the EEG electrodes, and the specific positioning of the EEG sensors on the scalp [29]. Usually, good practice in a controlled setting is to keep impedance values below a certain threshold (e.g., 5 kΩ, [30]) and to position the EEG cap in a standardized way by following the standard 10–20 [31] (Figure 1). After placing the EEG cap on the user's head, the technician measures the distance between the Nasion and Inion, as well as the distance between the participant's left and right preauricular points. The vertex Cz is defined to be positioned on the cap at the midpoint of both distances. Consequently, the electrode Cz is aligned with the midpoint of the two distances. Once the alignment is deemed satisfactory, the EEG cap can be confirmed to be in the correct position [32]. All these procedural steps would not obviously be performed in a real setting, where the wearable EEG headset should instead be worn as soon as possible without wasting time for the upcoming operational activities. So, keeping the boundary conditions of the measuring system stable between one session and another could not be easily feasible. Anyhow, while impedance stability among repeated sessions seems to be a minor issue, since wearable sensors are often completely dry or wet with a saline solution, ensuring that the EEG headset is always worn in the same position between sessions seems to be a dubious assumption. Therefore, the change in sensors position along repeated sessions could compromise the effectiveness of the passive BCI system, inducing changes in the related control features (e.g., EEG spectrum amplitude), with the consequent decreasing in overall accuracy of mental states detection (e.g., mental workload).

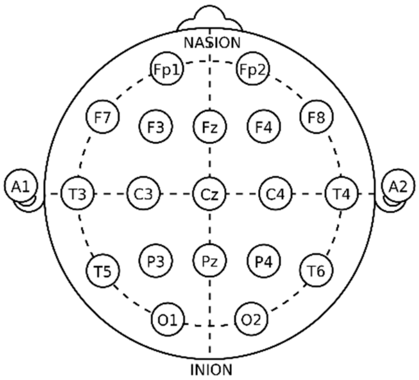


Figure 1. Electrode locations of International 10–20 system for EEG recording (source: Wikipedia).

This variation in control features, known as "covariate shift", can undermine the stability and efficacy of the entire passive BCI system. Although our study primarily focuses on detecting covariate shifts resulting from changes in the position of the EEG headset, it is essential to note that there are various sources of possible shifts in real-world environments outside of laboratories [33].

Among these, the potential for variations in EEG recordings due to involuntary or voluntary subject movements during task execution is relevant. While this scenario is not the primary focus of our paper, we believe that our approach can still be applied to identify covariate shifts following such movements, provided they result in a lasting change in the headset's position.

It is crucial to emphasize that currently, subject muscle movements that do not involve a physical displacement of the headset from its initial position are treated as artifacts and addressed via resizing or elimination.

However, this shift in the EEG headset position can result in a variation in the EEG amplitude or a distortion of the signal recorded by the electrodes in the new position. Therefore, even if the participant is experiencing the same level of mental state (e.g., workload) across different recording sessions, control features (e.g., amplitude of EEG power spectrum) can significantly change. This variability, caused by the covariate shift phenomenon, introduces a challenging element in the generation of processing models that can detect and ideally fix the negative effect that covariate shift may induce in passive BCI systems accuracy.

In the field of Artificial Intelligence, the covariate shift problem has been extensively investigated, as it poses a significant challenge across various domains [34,35]. This work focuses on the domain of Brain–computer Interfaces, where covariate shift may play a notably significant role.

Numerous studies have demonstrated how the non-stationarity of EEG signals can lead to variations in the distribution of data collected in different recording sessions [36–39]. This non-stationarity can be attributed to several factors, including not only fluctuations in user attention levels of fatigue but also electrode placement [36,40].

Some solutions proposed in the literature require the labeling of new data before starting the classification process [36,37,41]. This requirement limits the applicability of such approaches in real-world contexts, where there may be multiple data that exhibit covariate shifts, and this would uncontrollably increase the number of models that need to be trained each time and thus the waiting time to correctly identify the new coming recorded data.

Other approaches have focused on correcting covariate shifts using processed data samples via parallel processes. However, this methodology necessitates a substantial amount of new data and entails a temporal shift between the start of a new recording session and the identification of Covariate Shift [42,43].

However, in the current literature, the most widely explored approach to addressing covariate shift primarily leans toward anomaly detection. Some studies [36,37] relied on methods that use moving averages to identify anomalous records, i.e., those subject to covariate shift. These approaches prove effective in scenarios where the recorded trace remains relatively stable, with few anomalous values, compared to the EEG trace used for training. In the context of this work, we instead focus on scenarios where the presence of records with covariate shifts is assumed to be constant throughout the new EEG trace. For example, this might occur when the positioning of the recording headset is significantly different from the previous placement.

The primary contribution of this work, on the other hand, centers on the timely detection of the covariate shift caused by the incorrect positioning of the headset at the beginning of the new passive BCI session. In this context, it is crucial to promptly identify covariate shifts to take corrective actions on the data or at least inform the user to fix the headset's position. Furthermore, with the aim of simplifying and making BCI applications more accessible beyond the labs, we seek to minimize model training time and make it compatible with online use, which is desirable for real-context use.

This study has a twofold objective. On the one hand, to investigate to what extent variations in the sensor position of a commercial wearable EEG headset could induce a covariate shift in the recorded data, with an eventual decrease in the accuracy and reliability of the passive BCI system for mental workload evaluation. On the other hand, to develop and validate a methodology able to detect the occurrence of such covariate shift phenomenon in a timely manner, with the purpose of alerting the user about the issue, or ideally to automatically correct the shift.

Finally, a covariate shift correction function has also been tested, appearing to be able to mitigate the covariate shift effect under certain conditions of use.

In the following chapter, we will examine the two datasets used for our analysis. We will elaborate on the methodologies employed for data acquisition, the adopted protocol, and the various configurations used to simulate covariate shifts. The presentation of the Task, Setup, and Experimental Design will be distinct for both the Laboratory and Realistic datasets. Additionally, we will provide an overview of the methods used for signal processing, the Isolation Forest method, and describe the statistical approaches used to validate our analyses.

The third chapter will be dedicated to presenting the achieved results, organized by the two datasets and each conducted test (e.g., detection or correction of covariate shift). The final two chapters will be reserved for an in-depth discussion of the emerging results and the conclusions drawn from our research.

2. Materials and Methods

Two experimental datasets have been used in this study. The first one (i.e., Laboratory setting—Dataset) aimed to study the effectiveness of the proposed methodology in a controlled way. The second experimental dataset (i.e., Realistic setting—Dataset) aimed to test the methodology in a realistic setting, specifically by replicating the practical conditions that could induce the covariate shift phenomenon during real use. The following paragraphs will report the two experimental protocols, tasks, and participants' details.

2.1. Laboratory Setting—Dataset

In the following paragraphs, a detailed description of the dataset used to replicate laboratory conditions is provided. The paragraphs are organized to clearly separate the description of participants, work tasks, EEG setup used, and the conducted experiment. To perform this particular analysis, we utilized a dataset previously created and validated in earlier studies. Its original design makes it suitable for the current context.

2.1.1. Experimental Participants

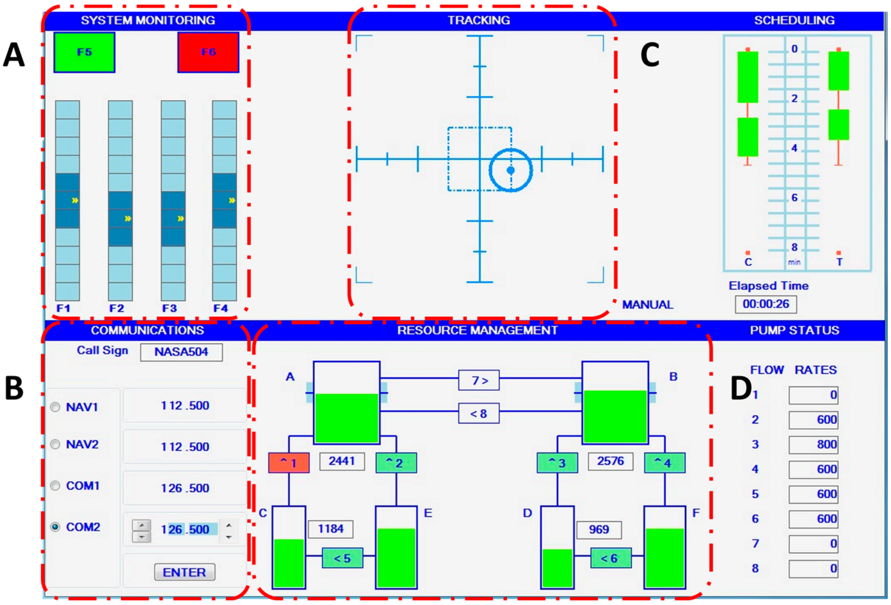
Ten participants, with a mean age of 25 ± 3, participated in this study. All participants were either students or staff members of the National University of Singapore (NUS). They were exclusively male, skilled in video games, right-handed, and in good health without any psychological or pharmaceutical issues. The study protocol received approval from the local Ethics Committee, and all participants provided written informed consent. Additionally, participants received compensation of $200 for their involvement in the experimental protocol. One of the 10 available participants had been excluded from the analysis for technical issues in the recorded data.

2.1.2. Experimental Task

The Multi-Attribute Task Battery (MATB, [44], Figure 2) provides a benchmark set of subtasks for trimming and controlling the task workload demand. Furthermore, the MATB can be used to simulate the activities inside the cockpit of an airplane and to provide a high degree of experimental task control in terms of complexity and difficulty. In this study, we set the MATB to simulate two possible scenarios at different difficulty levels (Easy, Hard) that could happen within a flight. We hypothesized that these conditions could induce different workload demands in the participants. This hypothesis has been tested using subjective assessment, reported in [45]. In particular, in the Easy condition, participants had to maintain the cursor in the center of the screen by manipulating the joystick to maintain the flight level of an airplane. In the hard condition, participants had to perform all the MATB sub-tasks at the same time to simulate a flight emergency. Before the beginning of the protocol, participants were trained to use the MATB software (version R2020b) until their performance was saturated with minimal errors commission. Throughout the experimental sessions, the MATB performance was estimated to ensure that participants would keep the same level of performance and that no learning effects would appear. The assessment of the training, in terms of task performance, cognitive resources, and task difficulty perception, was performed using the metric described in [45].

2.1.3. Experimental Setup

Scalp EEG has been recorded using the Waveguard© amplifier from ANT Neuro (Hengelo, The Netherlands), with a sample rate of 256 Hz from 64 Ag/AgCl EEG electrodes by following the 10–10 International System [31] referenced to the right mastoid and grounded to the aFz electrode, positioned on the head of the participant by a standard EEG cap in elastic textile. Not all electrodes were used for this study, but a subgroup of electrodes in the Frontal (Fpz, AF3, AF4, AF7, AF8) and Parietal (Pz, P3, P4) areas were selected to mimic the sensors' locations on the wearable headset employed in the second study (Mindtooth EEG system, Brain Products GmbH, Gilching, Germany, www.mindtooth-eeg.com accessed on 18 August 2023). In addition, the theta band on frontal channels and the alpha band on parietal channels' related features have been taken into account, with the aim of catching variations in the mental workload of the user while performing the experimental task [46,47].



Elapsed Time 00:00:26

Call Sign NASA504

○ NAV1 112.500
○ NAV2 112.500
○ COM1 126.500
● COM2 126.500

FLOW	RATES
1	0
2	600
3	800
4	600
5	600
6	600
7	0
8	0

Figure 2. Image of the Multi-Attribute Task Battery (MATB) interface: In the top left corner (**A**), is the emergency lights sub-task; in the bottom left corner (**B**), is the radio communication task; at the top, in the center (**C**), is the cursor tracking task, and finally, at the center bottom (**D**), is the fuel managing task.

Surrounding electrodes to those mentioned above were considered to mimic shift occurrences with different seriousness. To induce this shift, two changes were made: the positioning of the electrodes, i.e., the placement of the headset on the head, and the selection of a different electrode as the reference, i.e., the electrode subsequently utilized during the cleaning phase of the EEG trace. In particular:

- Control data:
  - ○ Channels: AF3-THETA, AF4-THETA, AF7-THETA, AF8-THETA, FPZ-THETA, P3-ALPHA, P4-ALPHA, PZ-ALPHA
  - ○ Reference: TP8
- Test data called *Test same Chs diff Ref*:
  - ○ Channels: AF3-THETA, AF4-THETA, AF7-THETA, AF8-THETA, FPZ-THETA, P3-ALPHA, P4-ALPHA, PZ-ALPHA
  - ○ Reference: TP8
- Test data called *Test same Ref diff Chs*:
  - ○ Channels: F3-THETA, F4-THETA, F7-THETA, F8-THETA, FZ-THETA, PO3-ALPHA, PO4-ALPHA, POZ-ALPHA
  - ○ Reference: TP8
- Test data called *Test diff Ref diff Chs*:
  - ○ Channels: F3-THETA, F4-THETA, F7-THETA, F8-THETA, FZ-THETA, PO3-ALPHA, PO4-ALPHA, POZ-ALPHA
  - ○ Reference: TP8

The three “Test” configurations (i.e., *Test same Chs diff Ref*, *Test same Ref diff Chs*, and *Test diff Ref diff Chs*) were generated to replicate what might happen during a recording session in which EEG headset placement occurs in a position different from an original session (i.e., Control session, used to train the passive BCI classification model).

The two central configurations (i.e., “*Test same Chs diff Ref*” and “*Test same Ref diff Chs*”) also allowed us to investigate whether a variation in channel positioning had more or less impact than a variation in the reference.

2.1.4. Experimental Design

The experimental protocol in controlled settings (Figure 3) comprised six recording sessions, with two sessions held per day—one in the morning and one in the afternoon. The initial four sessions took place on two consecutive days, referred to as *Day 1* and *Day 2*. The remaining two sessions occurred one week later, on *Day 9*. Each session included four conditions, during which participants randomly performed the two MATB conditions (Easy and Hard) twice. To prevent habituation or expectation effects, certain task parameters were randomly altered across the experimental sessions, including the order of stimulus presentation, radio frequencies, and the sequence of emergency lights. Each condition had a duration of 2.5 min. In summary, the complete dataset consisted of twelve pairs of conditions (four pairs of Easy and Hard conditions for each of the three experimental days, with two pairs conducted in the morning and two in the afternoon). In conclusion, before each Easy and Hard repetition, a 1 min baseline was recorded by asking the participants to fix the MATB task interface without reacting.

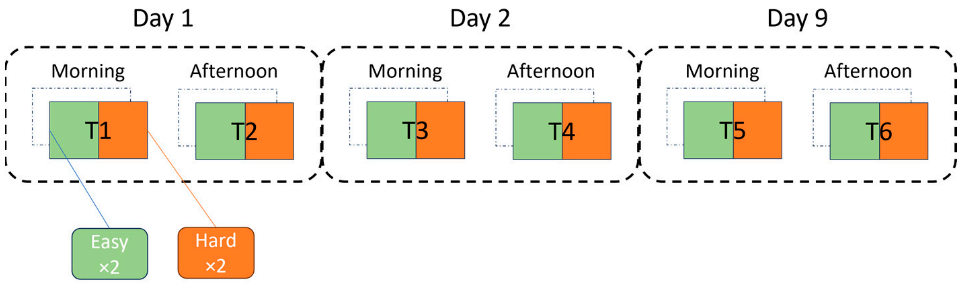


Figure 3. Experimental Design of Laboratory Setting. It has been composed of six recording sessions (T1–T6), two sessions per day, one in the morning and one in the afternoon. Each session consisted of four conditions.

Due to recording issues, participant 10 was excluded from the analyses in this study. Furthermore, during data cleaning, similar recording issues were detected in two sessions of participant 9, which were consequently excluded from the analysis.

Since we had only 9 participants available, in order to perform more accurate statistical analyses, we considered the sessions as independent variables for statistics rather than individual participants. This approach is justified by the fact that the sessions were recorded at different times, such as different parts of the day or on different days, and can thus be considered events independent from each other.

2.2. Realistic Setting—Dataset

In the following paragraphs, a detailed description of the dataset used to replicate realistic conditions is provided. The paragraphs are structured to keep separate the description of participants, work task, EEG setup used, and the conducted experiment. The dataset was specifically created for this experiment, allowing for the customization of tasks and configurations to replicate real-world use cases as closely as possible.

2.2.1. Experimental Participants

Ten participants (age = 33 ± 11) have been involved in this protocol: four women and six men were recruited on a voluntary basis. Each participant provided informed consent, and all data were pseudorandomized to safeguard against any association with individual identities. The experiments were conducted following the principles outlined in the 1975 Declaration of Helsinki, as revised in 2022 [48]. Experiments were approved by the Ethical Committee of the Sapienza University of Rome.

2.2.2. Experimental Task

As an experimental task, a multitasking application has been used, consisting of a set of four concurrent cognitive tasks of varying difficulty presented via split-screen (Figure 4), [28].

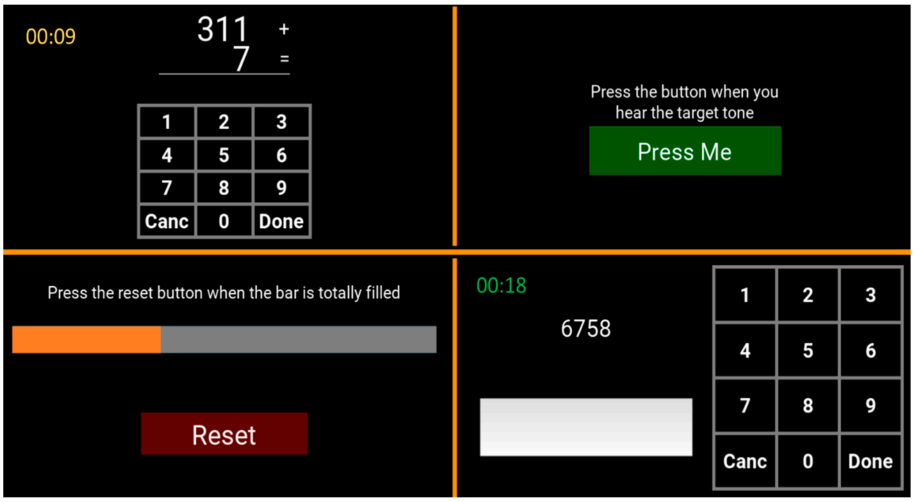

00:09
311 +
7 =
1 2 3
4 5 6
7 8 9
Canc 0 Done
Press the button when you hear the target tone
Press Me
Press the reset button when the bar is totally filled
Reset
00:18
6758
1 2 3
4 5 6
7 8 9
Canc 0 Done

Figure 4. Multitasking application screen. In the top left corner is the Mental Arithmetic task; in the top right corner is the auditory monitoring; in the bottom left is the visual monitoring; and in the bottom right is the Phone Number Entry Task.

The four chosen tasks are:

1. Mental arithmetic (left-up): Participants must input the results of additions into the numeric keypad. The difficulty escalates with an increase in the number of digits (ranging from 1 to 3) and carryover digits (ranging from 0 to 2).
2. Auditory monitoring (right-up): Participants are required to identify a target tone among two tons of different frequencies emitted at regular intervals. The task becomes more challenging as the similarity between the target tone and the distractor tone increases.
3. Visual monitoring (left-down): Participants need to reset a horizontal fill bar as soon as it becomes full. The difficulty level rises with an increase in the fill rate.
4. Phone number entry task (right-down): Participants must enter a number on a keypad. The task becomes more difficult with an increase in the number of digits to be entered, ranging from 4 to 10.

In accordance with the literature, carrying out multiple simultaneous tasks compared to the single-task approach leads to an increase in mental workload [28,44,49]. Therefore, the participants were asked to perform the four tasks individually to induce a low workload level (i.e., Easy workload). In this case, a 30-s condition was performed for the auditory monitoring, visual monitoring, phone number entry task, easy mental arithmetic task, and hard mental arithmetic task, for a total of 2:30 min of tasks. Instead, to induce an increasing in mental workload (i.e., Hard workload), participants performed the four concurrent tasks (multitasking phase) at the same time for 2:30 min. The ability of this experimental task to modulate the level of experienced user's workload has already been validated in [23].

2.2.3. Experimental Setup

Scalp EEG has been recorded using the wearable Mindtooth EEG Headset (Figure 5, Brain Products GmbH, Germany, www.mindtooth-eeg.com accessed on 18 August 2023), consisting of 8 EEG water-based electrodes of the 10–20 international system (AFz, AF3, AF4, AF7, AF8, Pz, P3, P4), referenced to the right mastoid, and grounded to the left mastoid. The water-based electrodes of the Mindtooth headset consisted of open-celled, hydrophilic, and highly absorbent cylindrical sponges. The material undergoes a hardening process when dry, transforming into a soft, expandable state when wet. This porous material is designed to absorb aqueous electrolyte solutions, with 1–2% sodium chloride solutions commonly used as electrolytes. The electrolyte solution facilitates a direct electrical connection to the participant's skin. Before starting the recording, the impedances have been brought below 100 kΩ [23,50].

Figure 5. Mindtooth EEG Headset.

2.2.4. Experimental Design

The experimental protocol (Figure 6) was composed of three recording sessions, and each session aimed to mimic the potential occurrence of the covariate shift phenomenon. In particular, during the first configuration, the headset was positioned following the 10–20 standard. In the second configuration, the headset was taken off and again put on, trying to maintain the same position as the first condition through visual inspection. During the third configuration, the headset was taken off and put on in a different position with respect to the first condition (i.e., all the electrodes, as well as the reference and ground, were moved two centimeters down). The second and third conditions have been randomized among the participants. Additionally, for each condition, two repetitions of both the experimental tasks (e.g., Easy workload and Hard workload) have been performed, in a randomized way, in order to not induce any habituation or expectation effect. Finally, before each condition, they have been recorded two rest conditions: the participants were asked to stay 1 min with closed eyes and 1 min with open eyes looking at a white cross on a screen. Before the beginning of the protocol, participants have been trained to use the multitasking software until their performance saturated with minimal errors commission.

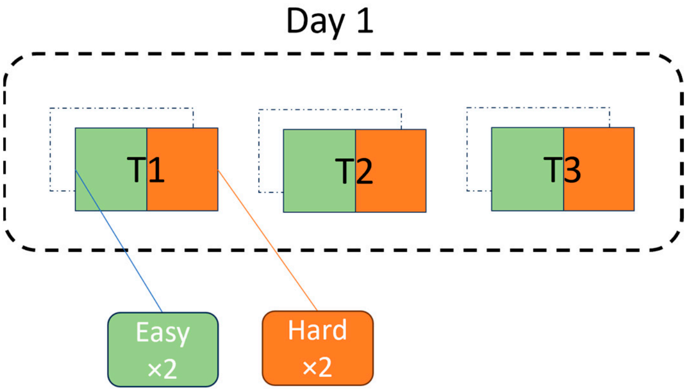


Figure 6. Experimental Design of Realistic Setting. It has been composed of three recording sessions (T1–T3). Each session consisted of four conditions (two Easy and two Hard). This design is maintained for all three configurations (Standard Positioning, Similar Positioning, and Different Positioning).

As the three recordings exhibit independence from one another, a systematic approach was devised. At each iteration, one recording was designated as the training dataset, while the remaining two were allocated for testing purposes. This procedure facilitated the creation of a Cross-Validation framework, where the training dataset was constituted by an individual recording, leaving the two remaining recordings to serve as distinct and independent test datasets.

In the following, we will name "Standard Positioning" each of the single configurations (first, second, or third) used to calibrate the machine learning model for workload classification. We will name "Similar Positioning" and "Different Positioning", respectively, the most similar configuration in terms of position and the most different configuration in relation to the current "Standard Positioning".

Table 1 displays the data (i.e., configurations) utilized as training data for Machine Learning models (i.e., Random Forest Classifier and Isolation Forest), along with their respective usage as test configurations, Similar Positioning, and Different Positioning. As observed in the table, Configuration 3, characterized by the widest movement of the headset from the reference position (i.e., standard 10–20), lacks a "similar" counterpart. Thus, when this configuration was used to train the models (Control configuration), predictions on the remaining two configurations were both considered as predictions on a "Different Configuration".

Table 1. Configurations utilized as training along with their usage as test for ML models.

Train Data/Standard Positioning	Similar Positioning	Different Positioning
Configuration 1	Configuration 2	Configuration 3
Configuration 2	Configuration 1	Configuration 3
Configuration 3	-	Configuration 1 and Configuration 2

2.3. Data Processing and Features Extraction

The signal processing and Feature Extraction steps were performed in the same way on both datasets described above. The initial step involved applying a band-pass filter to the EEG signal using a fifth-order Butterworth filter within the 2–30 Hz interval. Detection of blink artifacts employed the Reblinca method [51], and the correction was implemented by utilizing the ocular component estimated through a multi-channel Wiener Filter (MWF) [52]. Segmentation of EEG signals into 1-s epochs occurred, and artifact rejection was applied based on a threshold criterion of ± 80 µV [53]. This conservative threshold value was chosen over the default value of 100 µV suggested by the EEGlab toolbox, as a single criterion was adopted [54].

Both datasets were treated and processed in the same way. The only difference is represented by the AFz channel instead of the FPz channel for the Realistic Dataset because the AFz channel was not available in the Controlled dataset.

2.4. Machine Learning-Based Workload Index

For classifying the workload variations induced by the two different difficulty level conditions (i.e., Easy, Hard) for each of the two datasets, a specific implementation of the Random Forest Classifier [55] has been used. It is implemented in Python's scikit-learn library [56] and is based on the ensemble learning technique, combining multiple decision trees to make predictions [57].

Each tree is built on a randomly selected subset of the training data and a random subset of the features. The predictions from multiple trees are then combined to make the final prediction.

The algorithm follows these steps:

1. Random Sampling: Randomly select a subset of the training samples (with replacement) from the original dataset. This random sampling is called bootstrapping.
2. Random Feature Selection: Randomly select a subset of features (without replacement) from the available features. This subset is usually smaller than the total number of features.

3. Decision Tree Construction: Build a decision tree using the selected subset of samples and features. At each node of the tree, the algorithm selects the best split based on a criterion (e.g., Gini impurity for classification or mean squared error for regression).
4. Ensemble Creation: Repeat steps a-c to create a specified number of decision trees, each constructed on different subsets of samples and features.
5. Prediction Aggregation: The Random Forest combines the predictions of all decision trees using majority voting. The class that receives the most votes becomes the final prediction. For regression tasks, the algorithm averages the predictions of all decision trees to obtain the final prediction.

This function allows us to define many input parameters, but only the number of estimators has been optimized in the range from 50 to 500, as demonstrated in [23,55].

2.5. Covariate Shift Detection

The Isolation Forest [58] was the method investigated to detect the presence of covariate shift in new coming EEG data. In particular, it is a machine learning algorithm used for anomaly detection, which means it is designed to identify unusual or anomalous data points within a dataset. It is based on the concept of isolating anomalies rather than explicitly modeling normal data points.

The isolation forest algorithm works by randomly selecting a feature and then randomly selecting a split value within the range of that feature. By repeating this process recursively, the algorithm partitions the data points into individual trees called isolation trees. Anomalies are expected to require fewer splits to be isolated compared to normal data points.

The algorithm follows these steps:

1. Random Selection: Randomly select a feature from the d available features.
2. Random Split Point: Randomly select a split point between the minimum and maximum values of the selected feature.
3. Recursive Partitioning: Split the data based on the selected feature and split point, such that data points with feature values less than the split point go to the left branch, and those with values greater than the split point go to the right branch. Repeat this step recursively until all data points are isolated, or a predefined maximum tree depth is reached.
4. Tree Construction: Construct the isolation tree by repeating steps a–c until a specified number of isolation trees are created.
5. Anomaly Score Calculation: For a new data point, calculate its average path length (number of edges traversed) across all the isolation trees. The anomaly score is determined by comparing the average path length with the expected average path length of normal data points. Smaller average path lengths indicate anomalies.

We have used the isolation forest method since we hypothesized that the presence of a covariate shift could represent an anomaly with respect to the data recorded since that moment. So, we expected that the method could well fit the experimental problem.

In the current study, we employed the Isolation Forest model from the Sklearn [56] library. We tuned the “contamination” parameter of this method to adjust the expected proportion of anomalous data within the training condition. Multiple simulations were conducted each time using a different value (1%, 5%, 10%, 15%, 20%, 30%, 40%), with the aim of investigating the relationship between this percentage and the identified covariate shift occurrences in the test condition.

Covariate Shift Correction

As anticipated at the end of the introduction, although the covariate shift correction was outside of the scope of this paper, we would like to employ a simple correction method to investigate if it was possible to mitigate the negative effect induced by the covariate shift phenomenon. The outcome of this space transformation would be to modify the data distribution affected by covariate shift to make it similar to the control data (i.e., the same used to calibrate the machine learning model for workload evaluation).

To achieve this goal, the assumption was to make as references the resting recordings taken during the different conditions (i.e., OA recordings), with the assumption that the only differences within the distributions (i.e., control and testing) were induced by the different position of the headset, while the EEG patterns did not change. If this assumption works, it could be possible to modify the distribution of data with a covariate shift that closely approximates the distribution of the training dataset (i.e., with no covariate shift).

The following function was used to describe the relationship between rest and test conditions:

$$f(x) = m_2 + (x - m_1) \times \left(\frac{std_2}{std_1} \right) \quad (1)$$

m_2: mean of OA Control
x: value to be normalized
m_1: mean of OA Test
std_2: standard deviation of OA Control
std_1: standard deviation of OA Test

This correction is performed independently on each channel considered for the analysis.

2.6. Performance Analysis

On both the datasets, two variables have been analyzed: (I) the performance of the passive BCI system in classifying the two workload levels (i.e., easy vs. hard), and (II) the percentage of data affected by covariate shift phenomenon by using the isolation forest technique over all the conditions (i.e., Control data, Testing data with covariate shift, Testing data corrected). To determine whether a *shift* causes a decrease in classification accuracy, we initially investigated to what extent the performance of a Machine Learning model decreases when test data exhibits a shift with respect to training data.

Since the investigated variable (i.e., easy and hard conditions) is balanced between Easy and Hard values, to quantify the degradation in the performance of the Machine Learning classification model, we employed the accuracy metric [59].

To perform analyses involving the comparison of classification accuracy values, we computed a moving average of the classifier's predicted values (i.e., 0 or 1) by grouping the data into variable-length windows ranging from 1 s to 60 s. Consequently, at each time point (t), we obtained the accuracy value associated with the method by calculating the moving average over a window of length t records with a unitary shift.

$$Accuracy_{dec_t} = \frac{Number\ of\ correct\ assessments_{dec_t}}{Number\ of\ all\ assessments_{dec_t}} \quad \forall\, t = 1,\, 2, \ldots 60 \quad (2)$$

where at each decimation, the value of $Number\ of\ correct\ assessments_{dec_t}$ is computed by comparing the list of predicted values and correct values created by applying a moving average with window size t and unit step.

The second analysis was performed to investigate the effectiveness of the method employed to detect the presence of covariate shifts in recorded data, in particular, using the Isolation Forest [60] method. In this case, the model was trained using the control data, and then the percentage of data in the test conditions that were identified as outliers was evaluated. These outliers represent data that presents a *shift* with respect to the distribution of the control data [60].

As mentioned before, we tried to test whether a linear transformation of shift-affected data could improve classification accuracy. For this reason, the analyses previously described were performed both before and after the application of the linear transformation. The results obtained were then compared to determine if and under what conditions the application of the correction method could yield improvements in workload classification.

In order to establish the minimum number of test data samples required to optimize the detection of the covariate shift phenomenon, several simulations were conducted, each involving different sample sizes of recorded data. Specifically, the test data underwent an analytical process to identify the presence of shift by considering only the initial n seconds, where n ranged from 10, 15, 20, 30, 45, to 60 s. This approach was adopted because the type of shift of interest is associated with an incorrect positioning of the EEG electrode cap. Therefore, if this discrepancy is detectable in the initial moments of recording, it is presumed to persist in subsequent phases. Consequently, the objective was to determine the minimum length of the data subset to be analyzed for the detection of covariate shift.

Both analyses were conducted separately on both available datasets to ensure a comprehensive and accurate evaluation of the performance of the covariate shift detection and correction method.

2.7. Statistical Analysis

We performed non-parametric statistical tests since the obtained distributions were not normally distributed. In this regard, the Wilcoxon signed-rank test [61] (α = 0.05) was used to compare the Accuracy values and the percentage of anomalies detected using the isolation forest method obtained in the Control condition with the values obtained in each Test configuration. Moreover, the same test has been used to compare those variables before and after the application of the correction method. The Wilcoxon *p*-values was corrected using a multiple comparison correction method, specifically, the Bonferroni correction method [62], to ensure accurate statistical significance. Furthermore, we used Cohen's D test [63] to assess the effect size of the difference in accuracy performance between the Control/Standard configuration and the Test configurations and between pre- and post-correction results. The test result was evaluated using the interpretation made by Sawilowsky [64].

To compute the correlation between the accuracy values achieved using the classification model with the percentage of records identified with covariate shift, a repeated measures [65] test was conducted, considering individual participants in the different configurations of channels and references.

3. Results

In this chapter, the results obtained from the two datasets will be presented. The results will be categorized based on the type of analysis conducted (e.g., detection, correction of Covariate Shift, and correlation between Covariate Shift and accuracy) and the dataset used (Laboratory and Realistic).

3.1. Detection

In the following paragraphs, the graphs and results achieved in detecting covariate shifts using the Isolation Forest clustering method will be presented. As mentioned, the results will be categorized based on the dataset used.

3.1.1. Laboratory Setting—Dataset

Figure 7 shows the results achieved for each experimental session, training the model on the control data and classifying the four different configurations (Control Data, *Test same Chs diff Ref*, *Test same Ref diff Chs*, and *Test diff Ref diff Chs*). The four curves represent the four conditions on which the analysis was performed. Specifically, the blue curve identifies the classification results in the control condition used to train the classification model. The graph shows that except for a few experimental sessions, it is quite evident that the configuration "*Test diff Ref diff Chs*", which inherently represents a greater shift in the data compared to the control configuration, seems to achieve very low accuracy values different from those achieved by the control configuration. This difference is also noticeable in many experimental sessions for the other two configurations, which are also affected by data shifts.

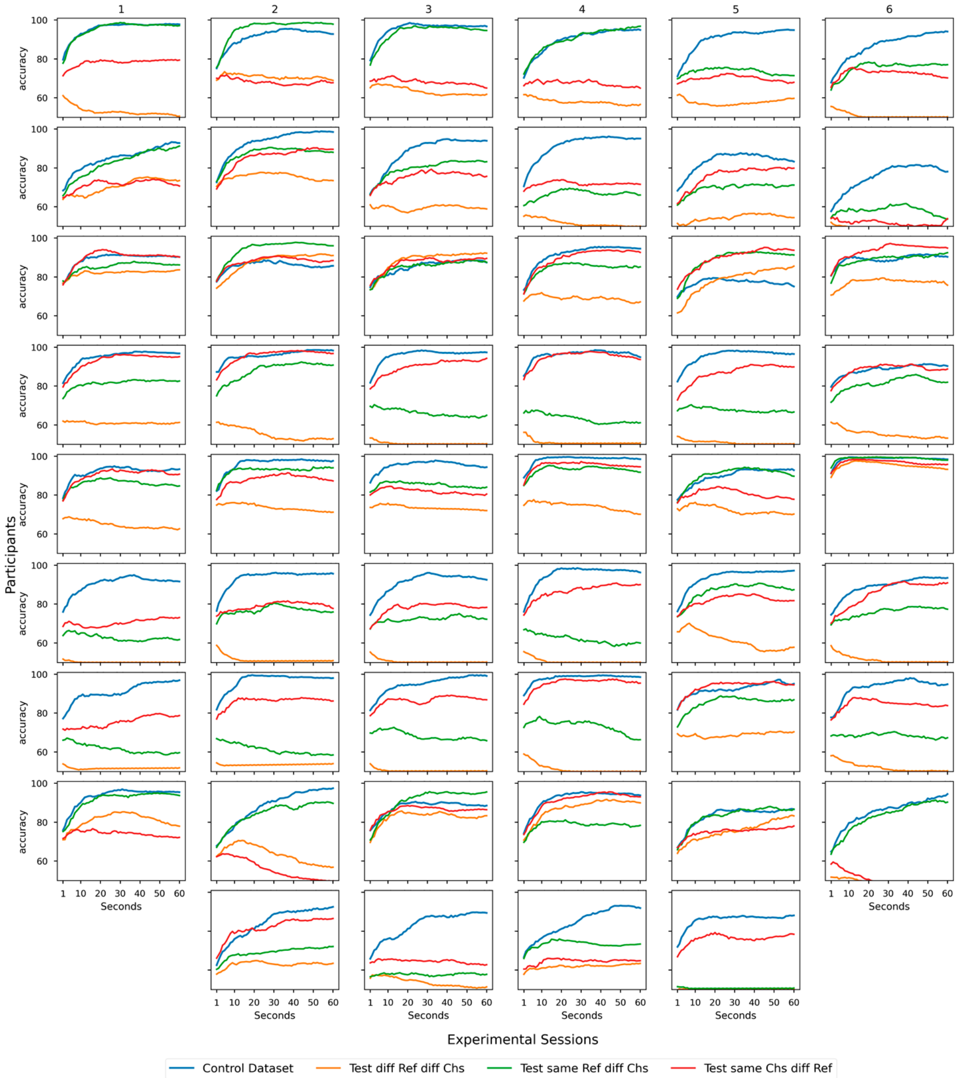


Figure 7. Accuracy was achieved in classifying four configurations per experimental session. Each line displays model training accuracy on control data and classification of the four configurations. The Y-axis ranges from 50 to 100; if a configuration's curve is not visible, its accuracy is below 50. Numbers 1 to 6 identify the registration sessions present for each participant.

Figure 8 represents the average accuracy curve over 52 sessions (i.e., all available sessions considering the nine subjects described above) on each configuration and the corresponding Wilcoxon test performed, comparing the values achieved in the control condition with those achieved in the remaining 3 test configurations. It is quite evident how the "*Test diff Ref diff Chs*" configuration, even when averaging all the values from the 52 experimental sessions, does not achieve good levels of classification accuracy compared to what one would obtain by training and using the classification model with the control configuration. The other configurations also differ significantly from the values achieved by the control configuration. The Wilcoxon test shows that these differences are significant ($p < 0.05$) for all three configurations (the three dashed curves appear to overlap, and therefore only one seems to be visible). Comparing the average accuracy values achieved by the three test configurations with respect to the control configuration, using the Cohen test, we note that the difference in the means is very significant, having for all three comparisons a D value greater than 2 (in particular 9.11 for Control vs. *Test diff Ref diff Chs*; 3.43 for Control vs. *Test same Ref diff Chs*; 2.94 for Control vs. *Test same Chs diff Ref*).

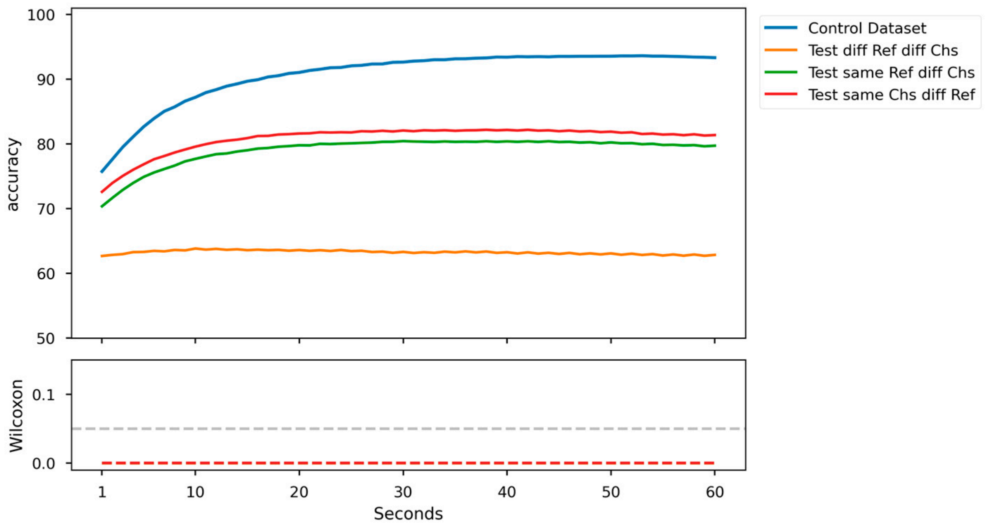


Figure 8. Mean accuracy across all tests for each configuration and corresponding Wilcoxon test. Dashed lines represent *p*-values for Wilcoxon tests comparing Control accuracy with that obtained in "*Test same Chs diff Ref*", "*Test same Ref diff Chs*", and "*Test diff Ref diff Chs*" configurations (same color legend). The last two lines are not visible due to overlap. All *p*-values are below 10^{-8}.

Table 2 displays classification accuracy values and corresponding *p*-values achieved at specific levels of decimation (in seconds) across different channel and reference configurations. The *p*-value is computed between the control configuration and each of the test configurations; hence, the value is not present in the column related to the control data.

Table 2. Classification accuracy values and *p*-values achieved across different configurations.

	Control		*Test Diff Ref Diff Chs*		*Test Same Ref Diff Chs*		*Test Same Chs Diff Ref*	
Second	**Accuracy**	***p*-Value**	**Accuracy**	***p*-Value**	**Accuracy**	***p*-Value**	**Accuracy**	***p*-Value**
10	0.87	-	0.64	4×10^{-10}	0.78	9×10^{-8}	0.8	3×10^{-8}
30	0.93	-	0.63	5×10^{-10}	0.8	1×10^{-7}	0.82	2×10^{-7}
60	0.93	-	0.63	9×10^{-10}	0.8	1×10^{-7}	0.81	3×10^{-8}

Figure 9 shows the percentage of anomalies (%CS) for each experimental configuration at different values of the *contamination* parameter, split per difficulty level (Easy—Hard). Each point on the curve represents the percentage of records within the subset (i.e., the value on the X-axis) identified as having a shift in the data compared to control data (i.e., each point of the curves is the average value across all experimental sessions). The Wilcoxon test showed that all the performance levels for each condition differed significantly from the control one ($p < 0.05$). Furthermore, it can be observed how the data from the 'Hard' condition seem to be more inclined to be identified as shifts, as they have a higher number of records identified as shifts, even with the same percentage of anomaly and configuration. Additionally, as seen in the classification performance, in this case, the worst configuration appears to be the "*Test diff Ref diff Chs*" configuration.

The following four tables (Tables 3–6) display the percentage values of records identified as covariate shift, using a contamination level of 30% and dividing the data into Hard and Easy, as presented graphically in Figure 9. In the tables, alongside the information regarding the percentage of records, the Wilcoxon *p*-value is provided. This *p*-value is calculated by comparing the values of the test configuration with those of the control configuration.

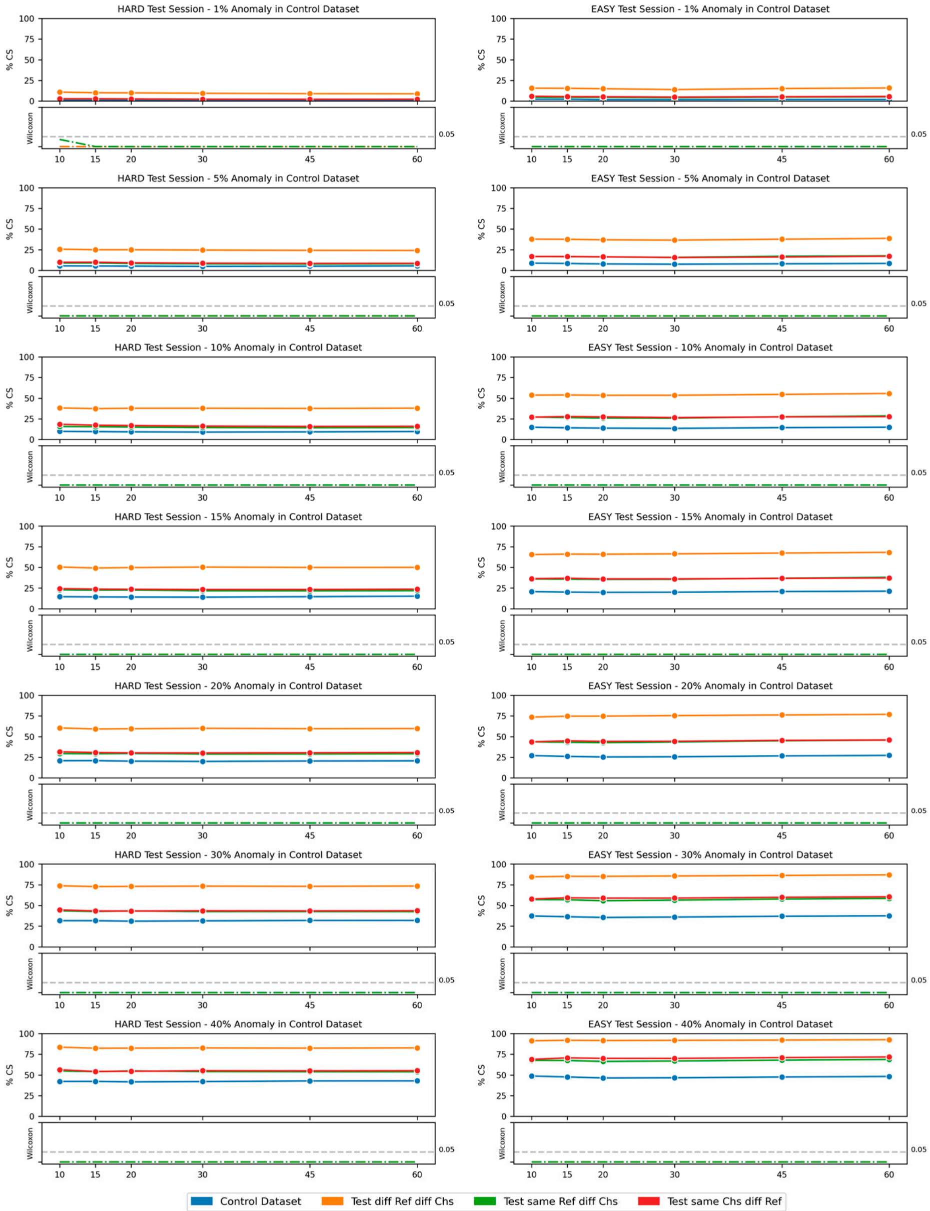


Figure 9. Each graph pair on a row represents a distinct isolation forest model trained with a different contamination level, using the Hard and Easy condition data. Each curve point signifies the percentage of records within the subset (X-axis value) identified with a data shift compared to control data. At each graph's bottom is the Wilcoxon test, with dashed curves indicating p-values by comparing %CS values of the control configuration with the other three configurations separately. All p-values are below 10^{-5} except one (dashed green line, first box on the left).

Table 3. Control Data.

	Hard		Easy	
N Record	**%CS**	***p*-Value**	**%CS**	***p*-Value**
10	0.42	-	0.49	-
15	0.42	-	0.48	-
20	0.42	-	0.46	-
30	0.42	-	0.47	-
45	0.43	-	0.48	-
60	0.43	-	0.48	-

Table 4. *Test diff Ref diff Chs.*

	Hard		Easy	
N Record	**%CS**	***p*-Value**	**%CS**	***p*-Value**
10	0.84	1×10^{-49}	0.91	4×10^{-51}
15	0.82	1×10^{-50}	0.92	1×10^{-51}
20	0.82	3×10^{-51}	0.92	1×10^{-51}
30	0.83	1×10^{-51}	0.92	1×10^{-51}
45	0.82	1×10^{-51}	0.92	1×10^{-51}
60	0.83	1×10^{-51}	0.93	1×10^{-51}

Table 5. *Test same Ref diff Chs.*

	Hard		Easy	
N Record	**%CS**	***p*-Value**	**%CS**	***p*-Value**
10	0.56	2×10^{-17}	0.69	3×10^{-33}
15	0.54	5×10^{-20}	0.71	2×10^{-38}
20	0.54	2×10^{-26}	0.70	1×10^{-40}
30	0.55	1×10^{-28}	0.70	1×10^{-45}
45	0.55	1×10^{-34}	0.71	1×10^{-48}
60	0.55	1×10^{-34}	0.72	1×10^{-50}

Table 6. *Test same Chs diff Ref.*

	Hard		Easy	
N Record	**%CS**	***p*-Value**	**%CS**	***p*-Value**
10	0.55	5×10^{-16}	0.68	2×10^{-32}
15	0.54	1×10^{-16}	0.68	8×10^{-40}
20	0.55	1×10^{-17}	0.66	7×10^{-44}
30	0.54	3×10^{-20}	0.67	1×10^{-47}
45	0.54	3×10^{-23}	0.68	4×10^{-50}
60	0.54	4×10^{-23}	0.69	1×10^{-50}

Figure 10 shows the percentage of values identified with covariate shift (%CS) in the four test conditions, using recordings obtained under the resting condition (i.e., open-eyes). The graph presents participants on the X-axis and the percentage of records in CS on the Y-axis. Each data point represents the mean of values across the six recording sessions (four sessions for subject 9). This graph also includes a fifth curve (i.e., the purple curve) related to the cross-session result, wherein one session is used for training the Isolation Forest model, and n-1 sessions are used for testing.

3.1.2. Realistic Setting—Dataset

As in the case of the Laboratory Setting, Figure 11 shows the results achieved in each test of the classification model trained using the control data (i.e., Standard Positioning) and used to classify the data for the three different configurations. The three curves refer to the three different configurations. Specifically, the blue curve identifies the average (decimated) accuracy values obtained by averaging the accuracy results from considering the three configurations one at a time as the training data set of the classification model. Additionally, using this dataset, the graph shows that except for a few participants, it is quite evident that the configuration "Different Positioning", which inherently represents a greater shift in the data compared to the standard one, seems to achieve very low accuracy values different from those achieved by the Standard Positioning.

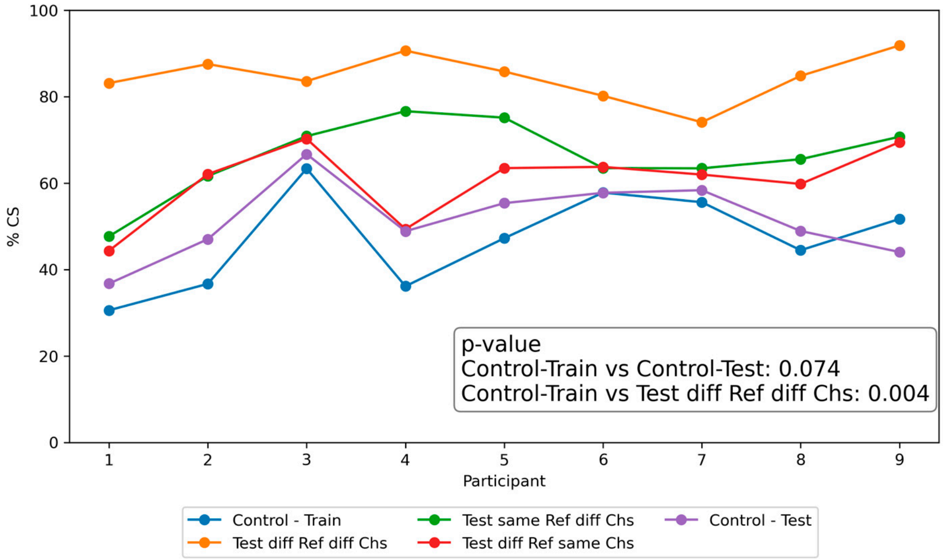


Figure 10. Percentage of records showing Covariate Shift with rest condition data. Each curve point represents the participant's percentage of records with covariate shifts across different configurations using rest condition data. The box displays Wilcoxon test results comparing the blue curve with the purple curve and the orange curve. This highlights the increase in covariate shift as configurations change, even during the resting condition.

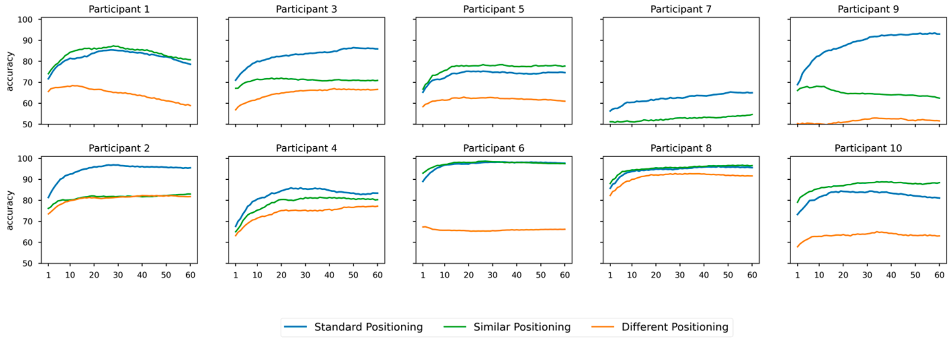


Figure 11. Accuracy was achieved in the classification of the three configurations. Each line shows the accuracy reached by training the model on the control data (Standard Positioning) and classifying the three different configurations. The Y axis of the graph includes values from 50 to 100; therefore, when the curve of some configuration is not visible in the graph, it is because the accuracy value achieved is not greater than 50. Each graph corresponds to the results achieved in one participant.

Figure 12 represents the mean of the accuracy curve over the 10 participants on each configuration and the corresponding Wilcoxon test performed by comparing the values achieved in the control configuration (Standard Positioning) and those achieved in the remaining 2 test configurations. In particular, the dashed orange curve shows how the *p*-values of the Wilcoxon test are always significant (<0.05). Specifically, the Similar Positioning (green curve) test shows not significant *p*-values (it does not appear to be present in the graph because its value, for each decimation point, exceeds the upper limit of the graph). Cohen's test calculated between Standard and Different Positioning shows a value exceeding 8 (i.e., 8.26), indicating a strong and significant difference between the means of the two distributions. Instead, the difference between the means of the Standard Positioning values and those of Similar Positioning is of lesser impact, with a Cohen's D value of 2.24.

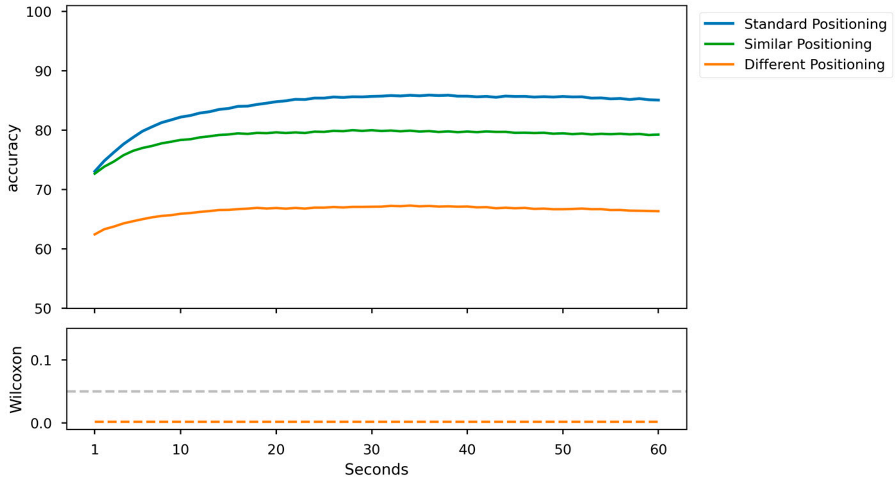


Figure 12. Mean Accuracy across all tests for each configuration and the corresponding Wilcoxon test. The test compares control condition (Standard Positioning) values with the other two configurations. The dashed orange line represents the *p*-value of the Wilcoxon test comparing accuracy in the control and 'Different Positioning' configurations. The green dashed line for 'Standard Positioning' versus 'Similar Positioning' is not visible, exceeding the 0.15 Y-axis limit for every decimation value. All *p*-values are below 10^{-7}. The dashed grey line indicates the significance level chosen for the test (i.e., 0.05).

Table 7 displays classification accuracy values and corresponding *p*-values achieved at specific levels of decimation (in seconds) across configurations. The *p*-value is computed between the Standard Positioning and each of the other positions; hence, the value is not present in the column related to the Standard Positioning.

Table 7. Classification accuracy values and *p*-values achieved across configurations.

	Standard Positioning		**Similar Positioning**		**Different Positioning**	
Second	**Accuracy**	***p*-Value**	**Accuracy**	***p*-Value**	**Accuracy**	***p*-Value**
10	0.82	-	0.78	0.23	0.66	1×10^{-3}
30	0.86	-	0.8	0.23	0.67	1×10^{-3}
60	0.85	-	0.79	0.28	0.66	1×10^{-3}

Figure 13 shows the percentage of anomalies for each experimental configuration at different values of the *contamination* parameter, split per difficulty level (Easy—Hard). Each point on the curve represents the percentage of records within the subset (i.e., the value on the X-axis) identified as having a shift in the data compared to control data. The statistical analysis revealed that when comparing the percentages of values with covariate shift in the standard position and in a different position, the *p*-values are consistently significant, except for a few cases where the contamination level in training is 1%. However, a different scenario emerges when comparing the values obtained in the standard position with those in a similar position, as achieving significant results requires a high level of contamination in training, specifically exceeding 30%.

The following three tables (Tables 8–10) display the percentage values of records identified as covariate shifts, using a contamination level of 30% and dividing the data into Hard and Easy, as presented graphically in Figure 13. In the tables, alongside the information regarding the percentage of records, the Wilcoxon *p*-value is provided. This *p*-value is calculated by comparing the values of the test configuration with those of the control configuration.

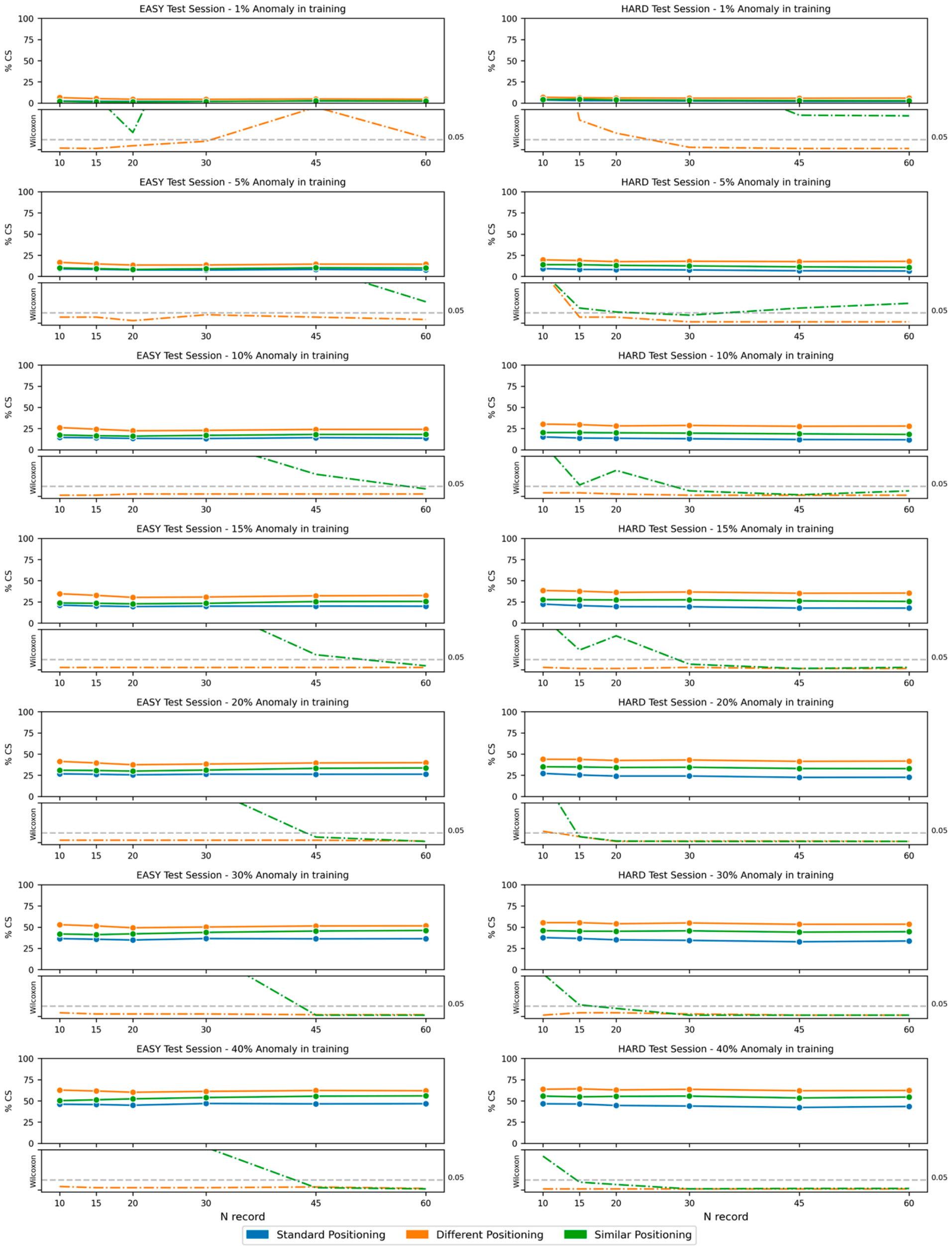


Figure 13. Each graph pair on a row represents a distinct isolation forest model trained with a different contamination level, using the Hard and Easy condition data. Each curve point indicates the percentage of records within the subset (X-axis) identified with data shift compared to the control. Wilcoxon tests at the graph bottoms show *p*-values, comparing %CS values of the control with the other two configurations. The dashed grey line indicates the significance level chosen for the test (i.e., 0.05).

Table 8. Standard Positioning.

	Hard		Easy	
N Record	**%CS**	***p*-Value**	**%CS**	***p*-Value**
10	0.38	-	0.37	-
15	0.37	-	0.36	-
20	0.35	-	0.35	-
30	0.34	-	0.37	-
45	0.33	-	0.36	-
60	0.34	-	0.36	-

Table 9. Similar Positioning.

	Hard		Easy	
N Record	**%CS**	***p*-Value**	**%CS**	***p*-Value**
10	0.46	0.1	0.42	0.37
15	0.45	0.02	0.41	0.69
20	0.45	0.01	0.42	0.43
30	0.46	1×10^{-3}	0.44	0.16
45	0.44	1×10^{-3}	0.45	1×10^{-3}
60	0.45	1×10^{-3}	0.46	1×10^{-3}

Table 10. Similar Positioning.

	Hard		Easy	
N Record	**%CS**	***p*-Value**	**%CS**	***p*-Value**
10	0.55	1×10^{-3}	0.53	5×10^{-3}
15	0.55	5×10^{-3}	0.51	3×10^{-3}
20	0.54	5×10^{-3}	0.49	3×10^{-3}
30	0.55	5×10^{-3}	0.50	3×10^{-3}
45	0.53	1×10^{-3}	0.51	3×10^{-3}
60	0.53	1×10^{-3}	0.52	3×10^{-3}

3.2. Correction

In the following paragraphs, the graphs and results achieved in correcting covariate shifts using the linear correction method will be presented. As mentioned, the results will be categorized based on the dataset used.

3.2.1. Laboratory Setting—Dataset

Figure 14 shows the Accuracy values achieved using the RF model trained on the control data, pre and post-application of the correction method. In particular, the dashed line is referred to as the accuracy pre-shift correction, and the continuous line is referred to as the post-shift correction. It can be easily observed that, for the "*Test diff Ref diff Chs*" configuration, the accuracy values curve increases at each decimation step after data correction through the transformation function. However, this transformation does not appear to have yielded improvements in the other configurations affected by shifts. The statistical analysis reveals that all the comparisons made appear to be statistically significant, as for all decimation levels and all three Wilcoxon curves, the *p*-values fall below the 0.05 threshold. The Cohen's test calculated between pre- and post-correction values using the test configuration "*Test diff Ref diff Chs*" demonstrates a value exceeding 6, indicating a strong and significant difference between the means of the two distributions. Additionally, notable differences are observed in the distributions of the other two tests, with values of 4.26 ("*Test diff Ref same Chs*") and 3.57 ("*Test same Ref diff Chs*"), respectively.

Table 11 compares the accuracy values achieved pre- and post-correction for some selected decimation values for all configurations except the Control one. For each comparison, the Wilcoxon *p*-value is also reported.

Figure 15 shows the values (in %) of records identified with covariate shift at different values of contamination, using a subsample of the first 15 records in each test condition. The values achieved pre and post-application of the correction function are also compared. The Wilcoxon statistical test is calculated by comparing the values in the test conditions achieved pre- and post-correction. The curves related to the comparison between pre- and post-correction values indicate that, concerning the "*Test diff Ref diff Chs*" configuration, the *p*-values consistently remain significant (<0.05). In contrast, for the other two comparisons in the different configurations, statistically significant results are only achieved when using a contamination level of 1% or 40%.

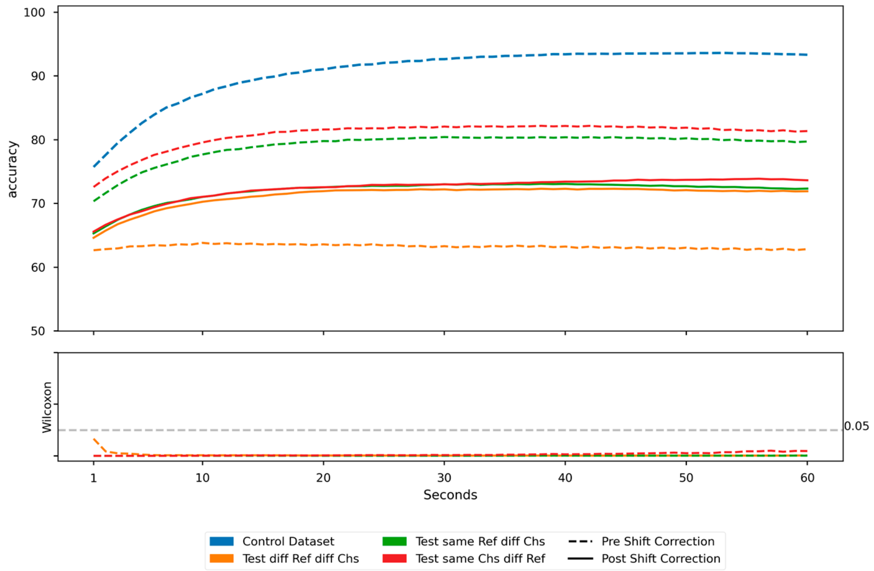


Figure 14. Comparison of classification accuracy before and after correction. Dashed curves show pre-correction accuracy using registration data (see Figure 8), while solid curves represent post-correction accuracy. Colors indicate test data configurations. The bottom box displays p-values of accuracy comparison at each decimation point, evaluating the impact of the correction function on the same configuration's data. All p-values are below 10^{-4}, except for the first value of *Test diff Ref diff Chs* comparison.

Table 11. Accuracy values achieved pre- and post-correction in each configuration.

	Test Diff Ref Diff Chs			*Test Diff Ref Diff Chs*			*Test Same Chs Diff Ref*		
Second	**Acc Pre**	**Acc Post**	***p*-Value**	**Acc Pre**	**Acc Post**	***p*-Value**	**Acc Pre**	**Acc Post**	***p*-Value**
10	0.64	0.7	1×10^{-3}	0.78	0.71	8×10^{-5}	0.8	0.71	3×10^{-4}
30	0.63	0.72	4×10^{-4}	0.8	0.73	2×10^{-4}	0.82	0.73	1×10^{-3}
60	0.63	0.72	7×10^{-4}	0.8	0.72	1×10^{-9}	0.81	0.74	1×10^{-8}

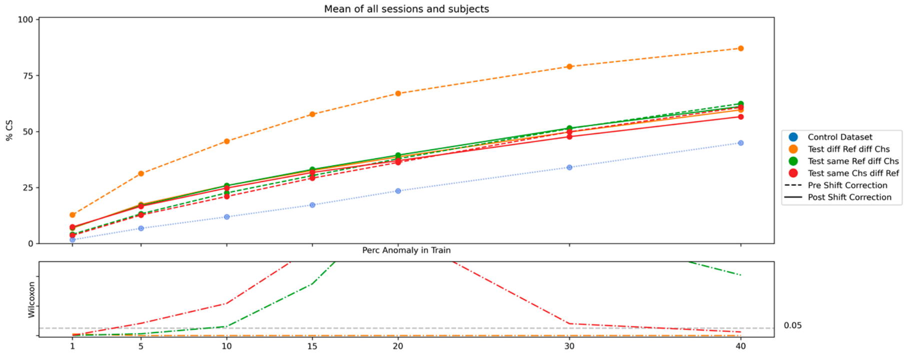


Figure 15. Comparison of anomaly detection results before and after correction. Dashed curves show the percentage of anomalies identified using the registration data, while solid curves represent post-correction results. The bottom box displays Wilcoxon p-values of the comparison of anomaly percentages before and after applying the correction function on the same configuration's data based on different contamination levels in the isolation forest method.

3.2.2. Realistic Setting—Dataset

Figure 16 shows the Accuracy values achieved during the classification process using the RF model trained on the control data (i.e., Standard Positioning), pre- and post-application of the correction method. The dashed curves represent the accuracy values achieved by the model in classification using the registration data before the application of the correction function (as also depicted in Figure 12). Conversely, the solid curves depict the accuracy values achieved by the model in classification using the data after the application of the correction function. For both types of curves, the color denotes the configuration used for the test data. At the bottom, there is a box displaying *p*-value values for each decimation point. The comparison is made between the accuracy values before and after the application of the transformation function on data from the same configuration. The statistical analysis reveals *p*-values all exceeding the chosen level of significance (0.05). When comparing the average accuracy achieved pre- and post-correction using Cohen's test, we observe that, unlike what we observed in the laboratory dataset, the difference is not as pronounced for both configurations. Specifically, Cohen's D value when comparing the values recorded with the "Different Positioning" configuration is 0.63, whereas in the other configuration (i.e., "Similar Positioning"), it is 1.58.

Table 12 compares the accuracy values achieved pre- and post-correction for some selected decimation values for all positioning configurations except the Standard one. For each comparison, the Wilcoxon *p*-value is also reported.

Figure 17 shows the values (in %) of records identified with covariate shift at different values of contamination, using a subsample of the first 15 records in each test condition. The values achieved pre- and post-application of the correction function are also compared. The dashed curves represent the percentage of values identified with shift by isolation forest using the registration data before the application of the correction function. Conversely, the solid curves depict the percentage of values identified with shift using the data after the application of the correction function. For both types of curves, the color denotes the configuration used for the test data. At the bottom, there is a box displaying *p*-value values for each value of the percentage of anomaly in the training dataset (i.e., contamination level of isolation forest method). The Wilcoxon statistical test is calculated by comparing the values in the test conditions achieved pre- and post-correction.

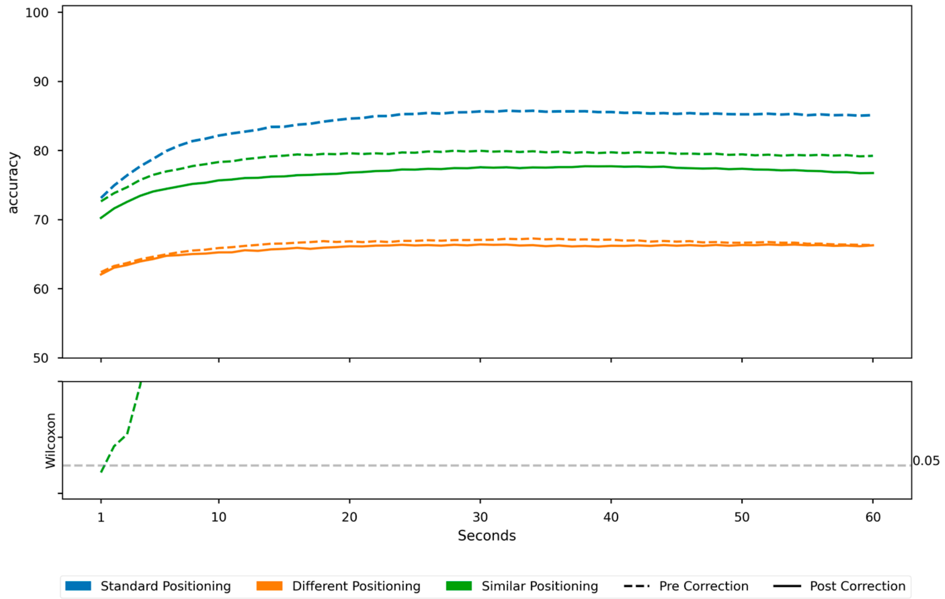


Figure 16. Comparison of pre- and post-correction accuracy curves using registration data. Dashed curves depict accuracy values before applying correction, while solid curves show accuracy after correction. The bottom box displays *p*-values of accuracy comparison at each decimation point. The Wilcoxon graph box is limited to 0.2, with *p*-values mostly exceeding the upper bound (except for initial values of the green curve), making them not visible in the graph.

Table 12. Accuracy values achieved pre- and post-correction in each configuration.

	Similar Positioning			Different Positioning		
Second	**Acc Pre**	**Acc Post**	***p*-Value**	**Acc Pre**	**Acc Post**	***p*-Value**
10	0.78	0.76	0.32	0.66	0.65	0.62
30	0.8	0.78	0.49	0.67	0.66	0.77
60	0.79	0.77	0.62	0.66	0.66	0.92

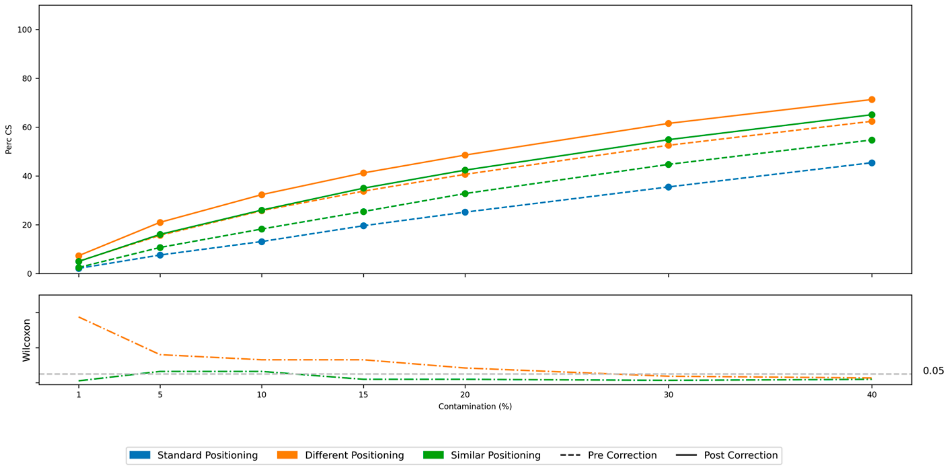


Figure 17. Comparison percentage of data with shift using a 15-record subsample and various contamination levels. Dashed curves show values identified before correction, while solid curves depict values after correction. The bottom box displays *p*-values for comparing percentages before and after correction at different contamination levels. Curve colors indicate the configurations being compared.

The statistical analysis shows statistically significant *p*-values in the comparison between pre- and post-correction values for the Similar Positioning configuration. However, significant values are observed only when considering a contamination level of 30 or 40 for the Different Positioning configuration.

3.3. Correlation between Covariate Shift and Accuracy

The following paragraphs present the results and graphs obtained in each dataset regarding the correlation analysis between the percentage of covariate shift in the data and the classification accuracy achieved using the Random Forest model.

3.3.1. Laboratory Setting—Dataset

Figure 18 shows the correlation between the accuracy values achieved by the classification model and the percentage of records identified with covariate shifts in the four analyzed configurations. Each individual point represents the percentage of records identified with a shift and the corresponding accuracy achieved using the classification method in a specific experimental session of a particular participant. For each participant, there are 24 points that correspond to the six experimental sessions in the four configurations (for participant 9, there are only 16 points since there are only four experimental sessions). Each line indicates the linear correlation between the points for each participant. The correlation has been calculated by considering the six sessions (4 sessions for participant 9) of each subject. The analysis showed a correlation of -0.622, with a p value of 8.06×10^{-23}.

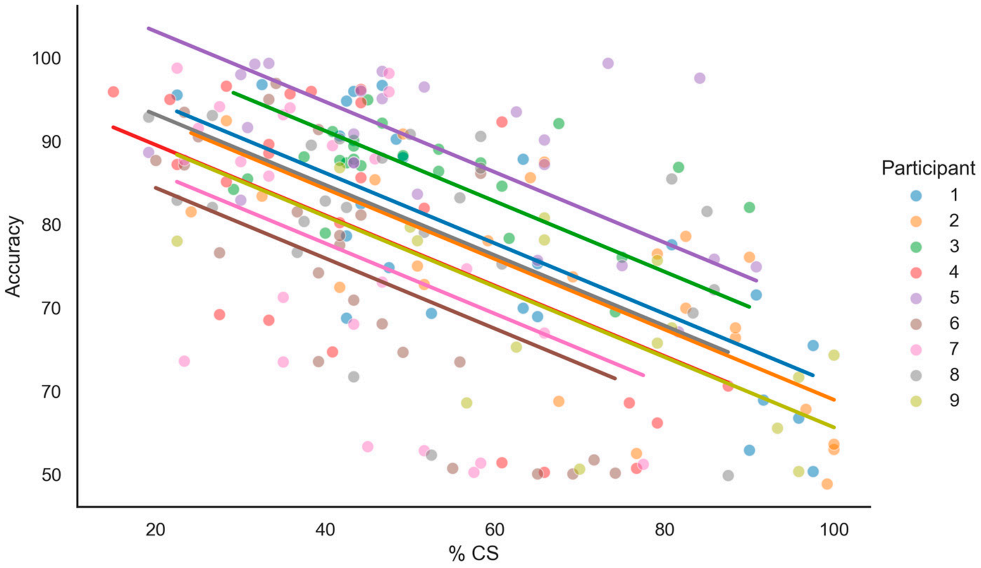


Figure 18. Correlation between Accuracy and Covariate Shift Percentage in each participant. Each point represents the record shift percentage and corresponding accuracy in a specific experimental session. Each line indicates the linear correlation between the points for each participant.

3.3.2. Realistic Setting—Dataset

Figure 19 correlates the accuracy values achieved by the classification model with the percentage of records identified with covariate shifts in the four analyzed configurations. Each individual point represents the percentage of records identified with a shift and the corresponding accuracy achieved using the classification method in a specific experimental session of a particular participant. For each participant, there are 3 points that correspond to the three configurations. The results are presented considering each participant. The analysis showed a correlation of -0.807, with a p value of 9×10^{-6}.

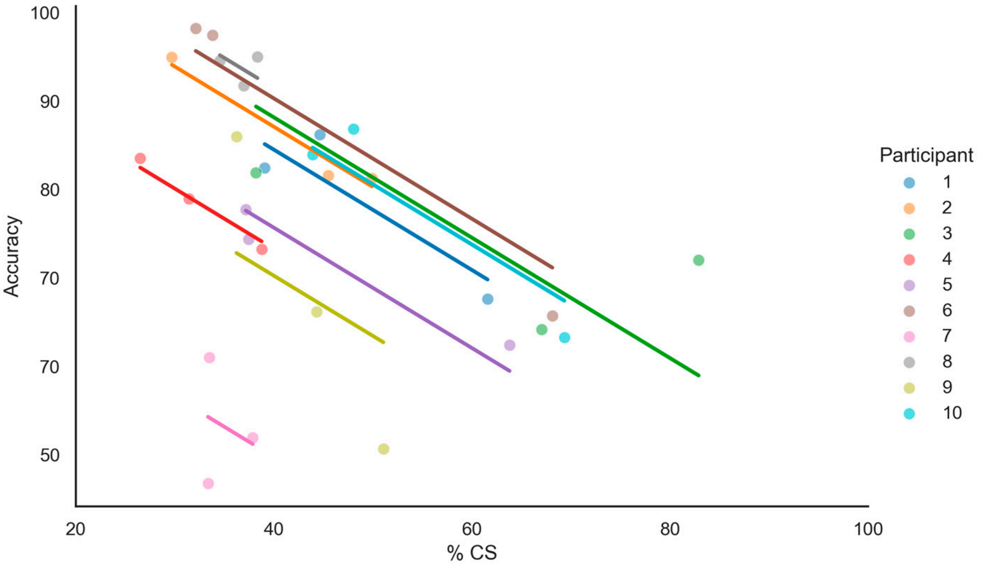


Figure 19. Correlation between Accuracy and Percentage of Covariate Shift achieved in each participant. Each point represents the record shift percentage and corresponding accuracy in a specific experimental session. Each line indicates the linear correlation between the points for each participant.

4. Discussion

Wearable EEG headsets are reaching a level of signal quality high enough to enter the market via out-of-the-lab passive BCI applications, such as for training purposes or monitoring during critical operational activities, by maintaining both wearability and comfort. In laboratory settings, standard EEG systems are usually employed, and it is good practice and even feasible to control all the possible variables and prevent any confounding effect, such as the quality of sensor contact and positioning. In contrast, in more real settings, and by using wearable devices, it would not be possible to maintain the same level of control; above all, it cannot be guaranteed that the position of the headset will be the same among different recording sessions. This could induce possible variations in the control features (e.g., EEG power spectrum) used in input to the machine learning model of the passive BCI system, not related to physiological changes of the user's state, causing a covariate shift phenomenon in newly recorded data.

In the current study, we investigated the effectiveness of an algorithm for anomaly detection, i.e., the isolation forest, in detecting the covariate shift phenomenon due to a change in the headset's sensors' positioning among different sessions.

4.1. Detection

By looking at the results achieved in both the laboratory and realistic datasets, it is quite evident that the covariate shift phenomenon significantly impacts the performance of the machine learning algorithm behind the passive BCI system, inducing decrements of even 20% in accuracy (p-value < 0.05), in the worst condition. In particular, the covariate shift phenomenon is reflected in a degradation of classification performance, quantified through accuracy, which becomes more pronounced with an increase in the shift. The phenomenon seems to be more prominent, as expected, to a variation in both recording and reference sensors' positioning, with respect to the only recording or only reference sensors' positioning. In fact, no statistical difference was found between the latter two configurations.

The significant difference in accuracy values is also highlighted by Cohen's D values, which, in both the laboratory and realistic datasets, consistently exceed 2 for each comparison between control configuration (or similar positioning) and test configuration.

It is worth noting to highlight that the phenomenon of covariate shift persists regardless of the amount of new testing data, as depicted in Figures 9 and 13. This observation implies the persistent nature of data shifts over time, induced by the change in sensor positioning. Hence, the need to detect the occurrence of the shift as soon as possible.

In this regard, the isolation forest method was able to significantly detect the covariate shift phenomenon, on average, using just the first 15 s of new coming data, allowing the possibility to alert the user to adjust the headset position or eventually to run correction algorithms able to properly correct the data distribution accordingly.

4.2. Correction

Regarding the correction of the shift through the implementation of the proposed linear transformation, it appears to be effective just in the case of the laboratory dataset and only under specific conditions. In particular, Figure 14 highlights that the application of this transformation has led to a significant enhancement in classification accuracy (and a consequent significant decreasing of the detected covariate shift on corrected data) only for the "*Test diff Ref diff Chs*" test. When comparing accuracy before and after the correction, a significant increasing emerges, with a p-value < 0.05. Nevertheless, the same behavior is not observed in the case of the realistic dataset, where Figure 16 does not reveal any significant difference in accuracy levels, with and without correction.

In the laboratory dataset, the application of the method resulted in a significant difference between accuracy values before and after, notably highlighted by Cohen's D measure. This difference is particularly noticeable in the configuration where we expected more variation in data (i.e., *Test diff Ref diff Chs*), where the Cohen's D value reaches 6.48. Conversely, in the realistic dataset, there is no Cohen's D value indicating a significant average difference in achieved accuracy values.

In a realistic context, as described, the application of the correction function seems to worsen the situation, compared to the control distribution, as depicted in Figure 17, in contrast to the laboratory scenario (Figure 15).

This behavior could be associated with the lack of stationarity of the EEG signal in similar conditions. In particular, even if participants have been asked to keep their eyes open during the three repetitions, it cannot be guaranteed that the elicited features used to normalize the new data with anomalies were the same, apart from the difference in sensors' position.

On the contrary, the analysis performed over the repeated sessions of the laboratory dataset (MATB) showed stability in EEG traces during rest sessions (i.e., eyes-opened), maybe because the participant was asked in that specific case to look at the interface without reacting. To assess this stability, we employed an index based on the percentage of data affected by covariate shift, using the same methodologies as described in this study. Specifically, for each session of each participant, we built an Isolation Forest model and tested it against the remaining 11 sessions, considering the four configurations described in this work. The results, presented in Figure 10, reveal that the blue and violet curves exhibit similar trends (p-value > 0.05). However, as we move to conditions with different sensor positioning, we observe an increase in the percentage of data affected by covariate shifts. The worst condition (depicted by the orange curve) occurs when both recording and reference sensor changes are introduced with respect to the original sensor's positioning (Control data).

4.3. Correlation between Covariate Shift and Accuracy

In conclusion, the analysis conducted on the correlation between the percentage of data with covariate shift and the accuracy of the classification model revealed a strong negative and significant correlation in both datasets. To perform this analysis, we applied the repeated measures correlation method [65], averaging the values for each participant. In the laboratory dataset, a correlation with a coefficient of -0.62 (p-value < 0.001) was observed, while in the realistic dataset, we found an even stronger correlation with a coefficient of -0.81 (p-value < 0.001).

An evident limitation of this initial exploratory study lies in the limited number of subjects involved in both datasets. While we were able to treat sessions as independent tests in the laboratory dataset, thereby increasing the number of tests for more statistically robust results, this approach was not feasible for the realistic dataset. Another limitation is the homogeneity of certain variables (e.g., age, tasks) that could pose challenges for drawing definitive inferences from the results. However, this study represents an initial exploration of a new method for detecting Covariate Shift under specific conditions, which will undoubtedly be further investigated using subsequent research. For example, we aim to expand the number of subjects involved and consider different working configurations in future research. Additionally, we plan to explore alternative correction methods that take into account the lack of stationarity of the control features, compensate for the data distortion induced by a different position, and mitigate the covariate shift phenomenon. In addition, more repeatable features than the EEG features in resting state, e.g., eyes blinks, will be investigated since they could be used as a more stable reference to correct the new coming data affected by covariate shift.

5. Conclusions

This work has introduced and validated a method that is able to detect the covariate shift occurrence induced by variations in the positioning of the EEG headset during multiple experimental sessions. The method has been tested both in a controlled setting, in which the differences in headset positioning were simulated by considering near electrodes (e.g., Fz with FPz, Pz with POz), and in a more realistic setting, in which a commercial wearable headset has been used, and it was mimic the situation in which the user was expected to wear the headset in a similar position, or in a completely different position, with respect to an initial one. Results showed that the isolation forest method proposed in this study was able to significantly detect the presence of the covariate shift phenomenon by using just 15 s of new coming data, independent of the severity of the anomaly, and in a completely not supervised way (i.e., the methods needs just original data, but does not need any knowledge about the kind of anomaly that could happen on new recorded data).

Our proposed solution differs significantly from existing solutions and studies in the literature for several reasons. Primarily, we utilize a wearable device to mimic what would probably happen in an out-of-the-lab situation. This approach simulates, in fact, the use of passive BCI applications outside of controlled laboratory environments.

Furthermore, our method stands out as it does not rely on using an already collected test dataset, whether labelled or unlabelled. It solely calibrates on the data from one session (assumed to be devoid of data affected by covariate shift) and can be used in real-time to predict, for instance, the presence of Covariate Shift during the recording of new sessions. This implies its capability to identify these variations as they occur, without the need for pre-existing data.

Additionally, the application field on which our method was developed and tested involves headset movements during different recording sessions in passive BCI applications for neurometric analysis, particularly for analyzing cognitive workload.

In addition, applying a simple linear normalization to data does not allow compensation for the negative effect induced by the different positions, and this behavior could be related to the lack of stationarity hypothesis in the EEG signal. We believe that the proposed normalization-based correction method performs well in the laboratory task but not in the realistic one due to the stability of the EEG signal during the recording of the resting-state sessions (i.e., open-eyes).

Author Contributions: Conceptualization, D.G., F.B. and P.A.; Data curation, D.G. and G.T.; Formal analysis, D.G. and A.G.; Funding acquisition, F.B.; Investigation, D.G. and P.A.; Methodology, D.G. and N.S.; Project administration, P.A.; Resources, D.G. and P.A.; Software, D.G.; Supervision, P.A.; Validation, D.G., V.R., G.B. and G.D.F.; Visualization, D.G.; Writing—original draft, D.G.; Writing—review and editing, D.G., N.S. and P.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was co-funded by the European Commission through H2020 projects "MINDTOOTH: Wearable device to decode human mind by neurometrics for a new concept of smart interaction with the surrounding environment" (GA n. 950998) and "FITDRIVE: Monitoring devices for overall FITness of Drivers" (GA n. 953432). The individual grants "CHALLENGES: CompreHensive frAmework to aLLEge andaNalyse surGEons' Stress" (Bando Ateneo Medio 2021), "BRAINORCHESTRA: Multimodal teamwork assessment through hyperscanning technique" (Bando Ateneo Medio 2022) provided by Sapienza University of Rome to Gianluca Borghini, "REMES–Remote tool for emotional states evaluation" provided to Vincenzo Ronca, and "HFAUX-Aviation: Advanced tool for Human Factors evaluation for the AUXiliary systems assessment inAviation", provided by Sapienza University of Rome to Vincenzo Ronca are also acknowledged. This work has also been carried out within the framework of the GURU (Sviluppo di un sistema multisensoriale a realtà mista per l'addestramento dinamico di lavoratori in ambienti ad alto rischio), co-financed by INAIL institute within the call BRIC2021; and GR-2019-12369824 "Detecting "windows of responsiveness" in Minimally Conscious State patients: a neurophysiological study to provide a multimodal-passive Brain–computer Interface", funded by Italian Ministry of Health.

Institutional Review Board Statement: The experiment was conducted following the principles outlined in the Declaration of Helsinki of 1975, as revised in 2008, and it received the approval of the Sapienza University of Rome ethical committee (nr. 2507/2020 approved on 4 August 2020).

Informed Consent Statement: All the participants signed the informed consent and the related information sheet, in which the study was explained, before participating in the experiment.

Data Availability Statement: For privacy reasons, given the sensitive nature of the data, the aggregated data analyzed in this study will not be publicly disclosed but might be available on request from the corresponding author.

Conflicts of Interest: Despite being partially employed by a private company (i.e., BrainSigns srl), the authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Yang, F.; Gu, S. Industry 4.0, a Revolution That Requires Technology and National Strategies. *Complex Intell. Syst.* **2021**, *7*, 1311–1325. [CrossRef]
2. Villalba-Diez, J.; Ordieres-Meré, J. Human–Machine Integration in Processes within Industry 4.0 Management. *Sensors* **2021**, *21*, 5928. [CrossRef] [PubMed]
3. Douibi, K.; Le Bars, S.; Lemontey, A.; Nag, L.; Balp, R.; Breda, G. Toward EEG-Based BCI Applications for Industry 4.0: Challenges and Possible Applications. *Front. Hum. Neurosci.* **2021**, *15*, 705064. [CrossRef] [PubMed]
4. Sciaraffa, N.; Germano, D.; Giorgi, A.; Ronca, V.; Vozzi, A.; Borghini, G.; Di Flumeri, G.; Babiloni, F.; Aricò, P. Mental Effort Estimation by Passive BCI: A Cross-Subject Analysis. In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Guadalajara, Jalisco, Mexico, 1–5 November 2021; pp. 906–909.
5. Midha, S.; Maior, H.A.; Wilson, M.L.; Sharples, S. Measuring Mental Workload Variations in Office Work Tasks Using fNIRS. *Int. J. Hum.-Comput. Stud.* **2021**, *147*, 102580. [CrossRef]
6. Novak, D.; Sigrist, R.; Gerig, N.J.; Wyss, D.; Bauer, R.; Götz, U.; Riener, R. Benchmarking Brain-Computer Interfaces Outside the Laboratory: The Cybathlon. *Front. Neurosci.* **2018**, *11*, 756.
7. Reason, J. Human Error: Models and Management. *BMJ* **2000**, *320*, 768–770. [CrossRef]
8. 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]
9. Mazur, L.M.; Mosaly, P.R.; Moore, C.; Comitz, E.; Yu, F.; Falchook, A.D.; Eblan, M.J.; Hoyle, L.M.; Tracton, G.; Chera, B.S.; et al. Toward a Better Understanding of Task Demands, Workload, and Performance during Physician-Computer Interactions. *J. Am. Med. Inform. Assoc.* **2016**, *23*, 1113–1120. [CrossRef]
10. Silva, F.P. Da Mental Workload, Task Demand and Driving Performance: What Relation? *Procedia Soc. Behav. Sci.* **2014**, *162*, 310–319. [CrossRef]
11. Marchand, C.; De Graaf, J.B.; Jarrassé, N. Measuring Mental Workload in Assistive Wearable Devices: A Review. *J. NeuroEngineering Rehabil.* **2021**, *18*, 160. [CrossRef]
12. Longo, L.; Wickens, C.D.; Hancock, G.; Hancock, P.A. Human Mental Workload: A Survey and a Novel Inclusive Definition. *Front. Psychol.* **2022**, *13*, 883321. [CrossRef] [PubMed]
13. Bagheri, M.; Power, S.D. Simultaneous Classification of Both Mental Workload and Stress Level Suitable for an Online Passive Brain–Computer Interface. *Sensors* **2022**, *22*, 535. [CrossRef]
14. Di Flumeri, G.; Giorgi, A.; Germano, D.; Ronca, V.; Vozzi, A.; Borghini, G.; Tamborra, L.; Simonetti, I.; Capotorto, R.; Ferrara, S.; et al. A Neuroergonomic Approach Fostered by Wearable EEG for the Multimodal Assessment of Drivers Trainees. *Sensors* **2023**, *23*, 8389. [CrossRef]
15. Zeng, H.; Li, X.; Borghini, G.; Zhao, Y.; Aricò, P.; Di Flumeri, G.; Sciaraffa, N.; Zakaria, W.; Kong, W.; Babiloni, F. An EEG-Based Transfer Learning Method for Cross-Subject Fatigue Mental State Prediction. *Sensors* **2021**, *21*, 2369. [CrossRef]
16. Gevins, A.; Smith, M.E.; McEvoy, L.; Yu, D. High-Resolution EEG Mapping of Cortical Activation Related to Working Memory: Effects of Task Difficulty, Type of Processing, and Practice. *Cereb. Cortex* **1997**, *7*, 374–385. [CrossRef]
17. Puma, S.; Matton, N.; Paubel, P.-V.; Raufaste, É.; El-Yagoubi, R. Using Theta and Alpha Band Power to Assess Cognitive Workload in Multitasking Environments. *Int. J. Psychophysiol.* **2018**, *123*, 111–120. [CrossRef]
18. Raufi, B.; Longo, L. An Evaluation of the EEG Alpha-to-Theta and Theta-to-Alpha Band Ratios as Indexes of Mental Workload. *Front. Neuroinform.* **2022**, *16*, 861967. [CrossRef]
19. Hamann, A.; Carstengerdes, N. Investigating Mental Workload-Induced Changes in Cortical Oxygenation and Frontal Theta Activity during Simulated Flights. *Sci. Rep.* **2022**, *12*, 6449. [CrossRef]
20. Craik, A.; González-España, J.J.; Alamir, A.; Edquilang, D.; Wong, S.; Sánchez Rodríguez, L.; Feng, J.; Francisco, G.E.; Contreras-Vidal, J.L. Design and Validation of a Low-Cost Mobile EEG-Based Brain–Computer Interface. *Sensors* **2023**, *23*, 5930. [CrossRef]
21. Bai, O.; Lin, P.; Huang, D.; Fei, D.-Y.; Floeter, M.K. Towards a User-Friendly Brain-Computer Interface: Initial Tests in ALS and PLS Patients. *Clin. Neurophysiol.* **2010**, *121*, 1293–1303. [CrossRef]
22. Park, S.; Han, C.-H.; Im, C.-H. Design of Wearable EEG Devices Specialized for Passive Brain—Computer Interface Applications. *Sensors* **2020**, *20*, 4572. [CrossRef]
23. Sciaraffa, N.; Di Flumeri, G.; Germano, D.; Giorgi, A.; Di Florio, A.; Borghini, G.; Vozzi, A.; Ronca, V.; Babiloni, F.; Aricò, P. Evaluation of a New Lightweight EEG Technology for Translational Applications of Passive Brain-Computer Interfaces. *Front. Hum. Neurosci.* **2022**, *16*, 901387. [CrossRef]
24. Giorgi, A.; Ronca, V.; Vozzi, A.; Sciaraffa, N.; di Florio, A.; Tamborra, L.; Simonetti, I.; Aricò, P.; Di Flumeri, G.; Rossi, D.; et al. Wearable Technologies for Mental Workload, Stress, and Emotional State Assessment during Working-like Tasks: A Comparison with Laboratory Technologies. *Sensors* **2021**, *21*, 2332. [CrossRef]

25. Singh, G.; Chanel, C.P.C.; Roy, R.N. Mental Workload Estimation Based on Physiological Features for Pilot-UAV Teaming Applications. *Front. Hum. Neurosci.* **2021**, *15*, 692878. [CrossRef]
26. Angrisani, L.; Arpaia, P.; De Benedetto, E.; Esposito, A.; Moccaldi, N.; Parvis, M. Brain-Computer Interfaces for Daily-Life Applications: A Five-Year Experience Report. In Proceedings of the 2021 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), Glasgow, UK, 17 May 2021; pp. 1–6.
27. Angrisani, L.; Arpaia, P.; Esposito, A.; Gargiulo, L.; Natalizio, A.; Mastrati, G.; Moccaldi, N.; Parvis, M. Passive and Active Brain-Computer Interfaces for Rehabilitation in Health 4.0. *Meas. Sens.* **2021**, *18*, 100246. [CrossRef]
28. Sciaraffa, N.; Di Flumeri, G.; Germano, D.; Giorgi, A.; Di Florio, A.; Borghini, G.; Vozzi, A.; Ronca, V.; Varga, R.; van Gasteren, M.; et al. Validation of a Light EEG-Based Measure for Real-Time Stress Monitoring during Realistic Driving. *Brain Sci.* **2022**, *12*, 304. [CrossRef]
29. Scrivener, C.L.; Reader, A.T. Variability of EEG Electrode Positions and Their Underlying Brain Regions: Visualizing Gel Artifacts from a Simultaneous EEG-fMRI Dataset. *Brain Behav.* **2022**, *12*, e2476. [CrossRef]
30. Hinrichs, H.; Scholz, M.; Baum, A.K.; Kam, J.W.Y.; Knight, R.T.; Heinze, H.-J. Comparison between a Wireless Dry Electrode EEG System with a Conventional Wired Wet Electrode EEG System for Clinical Applications. *Sci. Rep.* **2020**, *10*, 5218. [CrossRef]
31. Jasper, H.H. The Ten-Twenty Electrode System of the International Federation. *Electroencephalogr. Clin. Neurophysiol.* **1958**, *10*, 371–375. Available online: https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=1219144 (accessed on 20 September 2023).
32. Sazgar, M.; Young, M.G. Overview of EEG, Electrode Placement, and Montages. In *Absolute Epilepsy and EEG Rotation Review*; Springer International Publishing: Cham, Switzerland, 2019; pp. 117–125, ISBN 978-3-030-03510-5.
33. Domingos, C.; Marôco, J.L.; Miranda, M.; Silva, C.; Melo, X.; Borrego, C. Repeatability of Brain Activity as Measured by a 32-Channel EEG System during Resistance Exercise in Healthy Young Adults. Available online: https://www.mdpi.com/1660-4601/20/3/1992 (accessed on 10 November 2023).
34. Raza, H.; Prasad, G.; Li, Y. Adaptive Learning with Covariate Shift-Detection for Non-Stationary Environments. In Proceedings of the 2014 14th UK Workshop on Computational Intelligence (UKCI), Bradford, UK, 8–10 September 2014; pp. 1–8.
35. Dharani, Y.G.; Nair, N.G.; Satpathy, P.; Christopher, J. Covariate Shift: A Review and Analysis on Classifiers. In Proceedings of the 2019 Global Conference for Advancement in Technology (GCAT), Bangalore, India, 18–20 October 2019; pp. 1–6.
36. Raza, H.; Prasad, G.; Li, Y. EWMA Model Based Shift-Detection Methods for Detecting Covariate Shifts in Non-Stationary Environments. *Pattern Recognit.* **2015**, *48*, 659–669. [CrossRef]
37. Raza, H.; Samothrakis, S. Bagging Adversarial Neural Networks for Domain Adaptation in Non-Stationary EEG. In Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019; pp. 1–7.
38. Shenoy, P.; Krauledat, M.; Blankertz, B.; Rao, R.P.N.; Müller, K.-R. Towards Adaptive Classification for BCI. *J. Neural Eng.* **2006**, *3*, R13–R23. [CrossRef]
39. Satti, A.; Guan, C.; Coyle, D.; Prasad, G. A Covariate Shift Minimisation Method to Alleviate Non-Stationarity Effects for an Adaptive Brain-Computer Interface. In Proceedings of the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 105–108.
40. Blankertz, B.; Tomioka, R.; Lemm, S.; Kawanabe, M.; Muller, K. Optimizing Spatial Filters for Robust EEG Single-Trial Analysis. *IEEE Signal Process. Mag.* **2008**, *25*, 41–56. [CrossRef]
41. Sugiyama, M. Learning Under Non-Stationarity: Covariate Shift Adaptation by Importance Weighting. In *Handbook of Computational Statistics*; Gentle, J.E., Härdle, W.K., Mori, Y., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 927–952, ISBN 978-3-642-21550-6.
42. Jang, S.; Park, S.; Lee, I.; Bastani, O. Sequential Covariate Shift Detection Using Classifier Two-Sample Tests. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 28 June 2022; pp. 9845–9880.
43. Feutry, C.; Piantanida, P.; Alberge, F.; Duhamel, P. A Simple Statistical Method to Detect Covariate Shift. In Proceedings of the XXVIIème Colloque Francophone de Traitement du Signal et des Images (Gretsi 2019), Lille, France, 26 August 2019.
44. Comstock, J.R.; Arnegard, R.J. *The Multi-Attribute Task Battery for Human Operator Workload and Strategic Behavior Research*; NASA Langley Research Center: Hampton, VA, USA, 1992.
45. Borghini, G.; Aricò, P.; Di Flumeri, G.; Sciaraffa, N.; Colosimo, A.; Herrero, M.-T.; Bezerianos, A.; Thakor, N.V.; Babiloni, F. A New Perspective for the Training Assessment: Machine Learning-Based Neurometric for Augmented User's Evaluation. *Front. Neurosci.* **2017**, *11*, 325. [CrossRef]
46. Klimesch, W. EEG Alpha and Theta Oscillations Reflect Cognitive and Memory Performance: A Review and Analysis. *Brain Res. Rev.* **1999**, *29*, 169–195. [CrossRef]
47. Arico, P.; Borghini, G.; Di Flumeri, G.; Colosimo, A.; Graziani, I.; Imbert, J.-P.; Granger, G.; Benhacene, R.; Terenzi, M.; Pozzi, S.; et al. Reliability over Time of EEG-Based Mental Workload Evaluation during Air Traffic Management (ATM) Tasks. In Proceedings of the 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milano, Italy, 25–29 August 2015; Volume 2015, pp. 7242–7245. [CrossRef]
48. WMA—The World Medical Association-WMA Declaration of Helsinki—Ethical Principles for Medical Research Involving Human Subjects 2020. Available online: https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/ (accessed on 10 August 2023).

49. Wetherell, M.A.; Sidgreaves, M.C. Secretory Immunoglobulin-A Reactivity Following Increases in Workload Intensity Using the Defined Intensity Stressor Simulation (DISS). *Stress Health* **2005**, *21*, 99–106. [CrossRef]
50. Kappenman, E.S.; Luck, S.J. The Effects of Electrode Impedance on Data Quality and Statistical Significance in ERP Recordings. *Psychophysiology* **2010**, *47*, 888–904. [CrossRef]
51. Di Flumeri, G.; Arico, P.; Borghini, G.; Colosimo, A.; Babiloni, F. A New Regression-Based Method for the Eye Blinks Artifacts Correction in the EEG Signal, without Using Any EOG Channel. In Proceedings of the 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, 16–20 August 2016; Volume 2016, pp. 3187–3190. [CrossRef]
52. Somers, B.; Francart, T.; Bertrand, A. A Generic EEG Artifact Removal Algorithm Based on the Multi-Channel Wiener Filter. *J. Neural Eng.* **2018**, *15*, 036007. [CrossRef]
53. Hubbard, J.; Kikumoto, A.; Mayr, U. EEG Decoding Reveals the Strength and Temporal Dynamics of Goal-Relevant Representations. *Sci. Rep.* **2019**, *9*, 9051. [CrossRef]
54. Delorme, A.; Makeig, S. EEGLAB: An Open Source Toolbox for Analysis of Single-Trial EEG Dynamics Including Independent Component Analysis. *J. Neurosci. Methods* **2004**, *134*, 9–21. [CrossRef]
55. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
56. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
57. Breiman, L. *Classification and Regression Trees*; Routledge: New York, NY, USA, 2017; ISBN 978-1-315-13947-0.
58. Liu, F.T.; Ting, K.M.; Zhou, Z.-H. Isolation Forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008; pp. 413–422.
59. Thomas, E.; Dyson, M.; Clerc, M. An Analysis of Performance Evaluation for Motor-Imagery Based BCI. *J. Neural Eng.* **2013**, *10*, 031001. [CrossRef]
60. Liu, F.T.; Ting, K.M.; Zhou, Z.-H. Isolation-Based Anomaly Detection. *ACM Trans. Knowl. Discov. Data* **2012**, *6*, 1–39. [CrossRef]
61. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biom. Bull.* **1945**, *1*, 80–83. [CrossRef]
62. Lee, S.; Lee, D.K. What Is the Proper Way to Apply the Multiple Comparison Test? *Korean J. Anesth.* **2018**, *71*, 353–360. [CrossRef]
63. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Lawrence Erlbaum Associates: Hillsdale, NJ, USA, 1988; ISBN 978-0-8058-0283-2.
64. Sawilowsky, S.S. New Effect Size Rules of Thumb. *J. Mod. App. Stat. Meth.* **2009**, *8*, 597–599. [CrossRef]
65. Bakdash, J.Z.; Marusich, L.R. Repeated Measures Correlation. *Front. Psychol.* **2017**, *8*, 456. [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.