sensors

MDPI

Article

A Cybersecure P300-Based Brain-to-Computer Interface against Noise-Based and Fake P300 Cyberattacks

Giovanni Mezzina [1,†], Valerio F. Annese [2,†] and Daniela De Venuto [1,*]

[1] Department of Electrical and Information Engineering, Politecnico di Bari, 70125 Bari, Italy; giovanni.mezzina@poliba.it (G.M.)
[2] James Watt School of Engineering, University of Glasgow, Glasgow G12 8LT, UK; Valerio.Annese@glasgow.ac.uk (V.F.A.)
* Correspondence: daniela.devenuto@poliba.it (D.D.V.)
† Both authors contributed equally to this work.

Abstract: In a progressively interconnected world where the internet of things (IoT), ubiquitous computing, and artificial intelligence are leading to groundbreaking technology, cybersecurity remains an underdeveloped aspect. This is particularly alarming for brain-to-computer interfaces (BCIs), where hackers can threaten the user's physical and psychological safety. In fact, standard algorithms currently employed in BCI systems are inadequate to deal with cyberattacks. In this paper, we propose a solution to improve the cybersecurity of BCI systems. As a case study, we focus on P300-based BCI systems using support vector machine (SVM) algorithms and EEG data. First, we verified that SVM algorithms are incapable of identifying hacking by simulating a set of cyberattacks using fake P300 signals and noise-based attacks. This was achieved by comparing the performance of several models when validated using real and hacked P300 datasets. Then, we implemented our solution to improve the cybersecurity of the system. The proposed solution is based on an EEG channel mixing approach to identify anomalies in the transmission channel due to hacking. Our study demonstrates that the proposed architecture can successfully identify 99.996% of simulated cyberattacks, implementing a dedicated counteraction that preserves most of BCI functions.

Keywords: P300; brain-to-computer interface; BCI; EEG; machine learning; classification; cybersecurity; hacking; brainjacking

Citation: Mezzina, G.; Annese, V.F.; De Venuto, D. A Cybersecure P300-Based Brain-to-Computer Interface against Noise-Based and Fake P300 Cyberattacks. *Sensors* **2021**, *21*, 8280. https://doi.org/10.3390/s21248280

Academic Editor: Chang-Hwan Im

Received: 15 November 2021
Accepted: 8 December 2021
Published: 10 December 2021

1. Introduction

A brain-to-computer interface (BCI) is a direct communication channel between the user's neural activity and electronic devices [1]. Typically, BCIs are based on the recognition of a neural pattern during a specific mental task [1]. The neural activity can be recorded with both invasive and non-invasive equipment. Electroencephalography (EEG) is the most commonly used non-invasive technique in BCI systems [2]. Machine learning and classification algorithms are then used to analyze the neural signals, recognize the target neural pattern, and interpret the user's intention.

BCIs are growing progressively popular, as they are considered one of the most promising assistive technologies. As such, in the next decade, the global BCI market is expected to increase at a compound annual growth rate (CAGR) of 13.9%, from $1.5 million in 2020 to $5.5 million in 2030 [3]. To date, a variety of neural interfaces have been developed to aid people who have severe neuromuscular disorders [4], including brain-driven spellers [5], cars [1,6], wheelchairs [7], drones [8], and rehabilitative platforms [9].

As BCI technology approaches the market, concerns about cybersecurity have been raised [10]. With BCI hacking episodes, we refer to the possibility that a third party may read a patient's brain states or take control of the interface without the patient's or healthcare provider's consent. Currently, the eventuality of hacking is more plausible

than ever considering that recent BCI solutions have internet connections [11]. Taking control of a BCI can potentially lead to taking control of the user's intentions and actions. This raises the terrifying possibility of breaking into systems that control human cognition, feeling, and action and raises ethical concerns pertaining to privacy and physical or psychological harm [12]. The scenario is even more alarming after confirming that most of the current BCI solutions are seriously flawed due to the lack of control against hacking events. BCI uses well-established algorithms and neural patterns, which therefore are hackable. For instance, P300—arguably the most common neural pattern in BCIs—is easily reproducible, as it is widely understood and characterized in the literature. Despite the vulnerability of the current BCI frameworks, no definitive cybersecure solution has been established.

In this paper, we address the cybersecurity challenge by designing, implementing, and validating a novel architecture to monitor EEG headset and BCI framework communication channels. As a proof of concept, we focus on a P300-based interface trained using support vector machine (SVM) algorithms, but the proposed method is versatile and can be applied to other classifiers for BCI. In our study, we initially demonstrated that standard SVM-based P300 BCI interfaces are unable to discriminate real EEG data from synthetic signals. For this aim, we simulated cyberattacks using fake P300 signals. The testing dataset with real P300 data is composed of 7200 trial recordings from 5 different healthy subjects [13]. The testing dataset with fake P300 signals is composed of 2000 trials. Fake P300 signals were synthetically generated using modulated pink noise with a median filter or standard noise-based attacks. Thereafter, we designed and implemented a brain hacking recognizer (BHR), which uses an EEG channel mixing approach to identify anomalies in the transmission channel due to hacking events. We tested the BHR using both real and fake P300 datasets. We demonstrate that the BHR was able to identify 99.996% of attempted cyberattacks.

2. Related Works

Multiple independent scientific studies have demonstrated that BCIs are hackable [10,14,15], regardless of the hardware (e.g., EEG, ECOG, implants [10]), neural pattern (e.g., P300, movement-related potentials), and application (e.g., speller, wheelchair) [14]. In fact, most machine learning algorithms, including SVMs, decision trees, deep belief networks, artificial neural networks, random forest, and naïve Bayes, are vulnerable to cyber threats [15]. In the specific case of P300 classifiers, cyberattacks can cause a relative reduction in the area under the curve (AUC) of up to 74%, depending on the hacker's knowledge of the user's data [14].

Several solutions, summarized in Table 1, have been proposed to improve the security of a neural interface [16–30]. The proposed methods typically use one or more solutions to supervise the BCI session [17,18], authenticate the user [19–23], encrypt the data [24–27], and ultimately detect cyberattacks [28–30].

Supervised BCI systems identify cyberattacks by means of additional sensors and artificial intelligence monitoring the user's choices. For instance, the use of cameras [17], sensors, and actuators [23] in BCI navigation systems can prevent decisions that are dangerous or the result of hacking. However, additional hardware usually increases the cost and the complexity of the BCI system.

Authentication-based solutions establish a cybersecure communication channel between the user/sensors and the BCI framework. Authentication can involve the use of additional hardware, including user-specific radio frequency identification (RFID) technology [24], near-field or directional communication [28], and cameras for facial recognition [27]. Brainprint biometric authentication systems with no additional hardware using user-specific signatures in the EEG data have also been developed [20,21]. Nevertheless, the effect of P300 hacking events on authentication systems based on brain signatures is still debatable.

Table 1. State-of-the-art cybersecure frameworks.

Method	Solution	Application	Ref.
Supervision	Navigation unit (camera, laser odometry)	P300-based navigation system	[17]
	Navigation unit (wheel encoders, proximity sensors, gyro sensor, and an RGBD sensor)	P300-driven wheelchair	[23]
Authentication	RFID technology	Bidirectional BCI	[24]
	Near-field communication	User framework	[28]
	Facial recognition	IoT framework	[27]
	Brainprint biometric authentication	P300 speller	[25]
		BCI framework	[26]
Encryption	BCI anonymizer	BCI Framework	[29]
	Tensor-based data representation	EEG data	[30]
	Chaotic encryption	EEG data	[18]
	Randomization	BCI Framework	[19]
Cyberattack identification software	User-specific action profile	User framework	[20]
	User-specific EEG data	EEG data	[21]
		P300 BCI	[22]

Encryption techniques protect the communication between the sensors and the BCI framework, arguably the weakest point in BCI systems. Encryption methods for BCI applications include the use of an anonymizer [29], unconventional tensor-based data representation [30], standard encryption algorithms [18], and randomization [19]. Cyberattack identification software has also been developed to identify threats in a timely manner. The dedicated identification software uses the user-specific EEG data [29,30] and profile [20] to identify operational anomalies using artificial intelligence. However, software-based encryption and cyberattack identification solutions typically use complex unconventional algorithms and require high consumption of computational resources.

These barriers represent a gap to be bridged for implementing cybersecure BCI frameworks in a real-life scenario. As such, we propose a novel approach that uses a low-resource, reproducible, and versatile encryption method together with minimal hardware modifications and therefore is suitable for real-life applications.

3. Simulated Cyberattacks

Our study focuses on SVMs, which are widely used for BCI implementation. Five different SVM-based BCI models were trained using real P300 data. Then, the models were tested using real P300 data to quantify the performance of the algorithms in identifying P300 for BCI applications. Thereafter, several cyberattacks were simulated. We focused on the specific hacking profile shown in Figure 1, where the attacker hacks the transmission step preceding data processing and classification. We assume that the attackers have knowledge of the hardware (EEG headset), the wireless communication system, and the BCI framework. The attacker also has an external database of EEG signals ("Ext. DB" in Figure 1a) collected with the same settings and equipment. In the simulated cyberattacks, the real P300 dataset was modified by introducing fake signals. The introduction of fake signals was used to demonstrate that SVM-based models fail to recognize that the signal is not legitimate. The weakness of the SVM models was quantified by comparing relevant metrics calculated with real and fake signals.

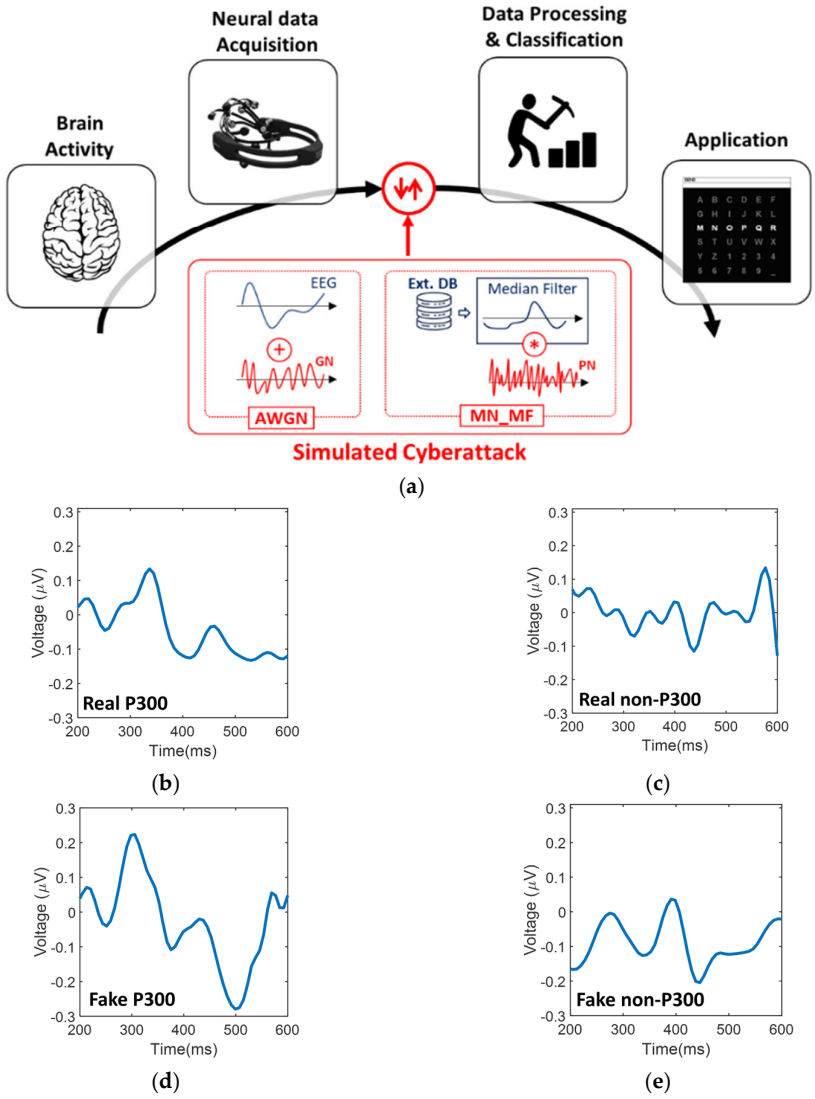

Figure 1. (**a**) BCI processing chain from acquisition to application with simulated cyberattacks on raw data between the neural acquisition and transmission and the BCI framework. (**b**) Example of real P300 and (**c**) real non-P300 signals (subject 1, channel P4, averaged over 200 acquisitions). (**d**) Example of fake P300 and (**e**) fake non-P300 signals (MF3,001, averaged over 200 signals).

3.1. Materials and Methods

Training dataset (real P300). A public EEG dataset was used in this work [13]. The training dataset is composed of data from 5 different men who are non-smokers and aged 21–41. None of the participants had any prior BCI experience or a history of neurological disorders. The dataset was produced using the standard 6 × 6 Donchin and Farewell P300 Speller Matrix. The training dataset includes 240 P300 (Figure 1b) and 1200 non-P300 trials (Figure 1c). Data were recorded using the EMOTIV EPOC + EEG wireless headset with 14 channels and a 128 Hz sampling frequency [31]. Data were filtered using two notch filters (50 Hz and 60 Hz) and a bandpass filter (0.2–45 Hz). Additional information on the training dataset can be found in [13].

EEG recordings were pre-processed before training the models according to best practices in the P300 recognition application [32–35]. Pre-processing was identical regardless of the SVM model to be trained. First, the EEG stream was time windowed between 50 ms and 600 ms after each stimulation onset due to the intrinsic nature of P300 patterns.

Indeed, P300 deflection typically occurs in a range from 250 ms to 500 ms, even in the presence of neurocognitive disease [34]. The resulting trials were low-pass filtered with a zero-phase 8th-order Butterworth filter with a cutoff frequency of 15 Hz. This interval was determined to be one of the most discriminative according to a batch of 27 works collected at the 5th Graz BCI Conference [33]. The filtered signals were detrended through a global average approach [33]. Next, the trials were downsampled by a factor of 4 through an average-based technique, achieving 32 Sa/s, reducing the random variations in EEG signals. The last step consisted of a zero-mean and unity standard deviation normalization [32–35]. The resulting data subset constitutes the features used as input for training the SVM models.

Testing dataset (real P300). The testing set is composed of EEG recordings from the same database as above [13]. The testing dataset includes 240 P300 and 1200 non-P300 trials for each of the five participants. Data were acquired using the same equipment and settings as in the training dataset. The pre-processing and feature extraction steps were also replicated from the training dataset.

Hacked datasets (fake signals). For each subject, 400 out of 1440 real trials were substituted with fake signals. Real trials substituted with fake ones were randomly selected using the Mersenne Twister 19937 method (seed: 4) [36]. Therefore, fake signals constituted 27.8% of the dataset, mimicking hacking attempts during the normal use of the BCI. Fake P300 (Figure 1d) and non-P300 (Figure 1e) signals were synthetically generated via MATLAB using 5 different methods. As such, five hacked datasets for each of the five participants were generated. The methods used to generate fake signals were:

(i) Addition of white Gaussian noise with SNR = 20 dB (AWGN 20);
(ii) Addition of white Gaussian noise with SNR = 40 dB (AWGN 40);
(iii) Modulated noise with 3rd-order median filter and low-amplitude pink noise (MF3, 001);
(iv) Modulated noise with 9th-order median filter and low-amplitude pink noise (MF9, 001);
(v) Modulated noise with 9th-order median filter and high-amplitude pink noise (MF9, 05).

Methods (i) and (ii) implement a standard noise-based attack with the main objective of altering the acquired EEG signal information content. These methods use a noise model known as additive white Gaussian noise (AGWN). The technique adds Gaussian noise (GN in Figure 1) to the selected EEG signal. By manipulating the signal-to-noise ratio (SNR), it is possible to create various noise scenarios. Specifically, we implemented AWGN by deriving variance from SNR and measured EEG signal power.

Methods (iii)–(v) implement a technique to reproduce synthetic EEG trials containing P300 features in both the time and spectral domains. These methods consider the rejection parameters typically used in the winsorizing process to identify and prevent abnormal voltage amplitudes due to physiological and non-physiological artifacts. However, winsorizing can also be used as a discriminative procedure in noise-based attacks and must be addressed in a fake P300 generation context. The proposed method is schematized in Pseudocode 1.

Real P300 and non-P300 signals were acquired using the same settings and equipment (line 1 of Pseudocode 1). Data were acquired from subjects other than the one under attack that is unknown to the attacker. Next, P300 and non-P300 data from each channel were averaged, resulting in two bidimensional matrices: mP3 for P300 trials and mNP3 for non-P300 ones (lines 3, 4). To emphasize only those features common to all P300/non-P300 trials over all analyzed subjects, a channel-specific median filter was applied (lines 5, 6).

Different nth-order unidimensional median filters were investigated. Median filtering is widely applied in image processing. The median filter replaces the center value in a selected window with the median value of all of the points within the window:

$$\mathbf{y}(\mathrm{m}) = \mathrm{median}\left(\mathbf{x}\left(\mathrm{m} - \frac{\mathrm{n}-1}{2} : \mathrm{m} + \frac{\mathrm{n}-1}{2}\right)\right) \quad (1)$$

where $\mathbf{y}$ and $\mathbf{x}$ are the filter output and input signals, respectively, m is the considered sample index, and n is the filter order.

To generalize the attack, invalidating all of the main performance metrics of the BCI, both targets and non-targets are independently cyberattacked. A random selection of the median signal to be used for the attack generation is carried out at line 7 of Pseudocode 1.

Pseudocode 1. Routine for fake P300 generation based on modulated noise with a median filter.

```
1. load P300_Trials, NotP300_Trials
2. # Generate attack
3.        mP3      ←  mean(P300_trials)
4.        mNP3     ←  mean(NotP300_trials)
5.        MF_mP3   ←  medianfilter(mP3)
6.        MF_mNP3  ←  medianfilter(mNP3)
7.        atck     ←  random_select (MF_mP3, MF_mNP3)
8. for (channel):
9.        EEG_lims    ←  extract EEG(channel) streaming limits
10.       atck_lims   ←  extract atck(channel) streaming limits
11.       k           ←  adapt_factor(EEG_lims, atck_lims)
12.       pn          ←  generatePinkNoise
13.       →generate MN_MF fake P300(channel)
```

Next, for each channel, the attacker simulator extracts the EEG amplitude limits from the signal streaming and the same parameters from the median-filtered waveforms. This process allows the system to generate a correction factor for the waveform to prevent its amplitude from approaching the winsorizing limits, preventing amplitude-based attack identification (line 11).

Finally, the attack waveform is generated as:

$$\mathbf{MN\_MF} = ((\alpha_1 \cdot \mathbf{pn}) * \mathbf{MF\_sig}) \cdot \mathrm{k} \quad (2)$$

where $\mathbf{MN\_MF}$ is a generated attack (P300 or non-P300), and α_1 is a hyperparameter that manages the amplitude of the locally generated pink noise vector $\mathbf{pn}$. $\mathbf{MF\_sig}$ represents the median-filtered signal regardless of whether it is P300 or non-P300, while k represents the correction function for protection against winsorizing. The dataset composition is summarized in Table 2.

SVM Training. For each subject, five different SVM-based BCI models were trained using the respective training set. The implemented SVMs have (i) linear (L), (ii) quadratic (Q), (iii) cubic (C), (iv) medium Gaussian (MG), and (v) coarse Gaussian (CG) kernels. The models were trained using the MATLAB Classification Learner tool by MathWorks with a k-fold cross-validation (k = 5) approach. Table 2 summarizes the datasets and methods used. For each subject and each model, we calculated accuracy, precision, recall, and F1 (defined in Table 2). For each of the above metrics, a comparative parameter named "cyberattack effect" was introduced in this study. It consists of the difference between the same metrics calculated with real and fake P300 data.

Table 2. Datasets used in this study and metrics of interest.

DATASETS			
Dataset	**Data Type**	**Method**	**Description**
Training	Real data	EEG acquisition	5 subjects, 1440 trials per subject
Testing	Real data	EEG acquisition	5 subjects, 1440 trials per subject
	Hacked data	EEG + AWGN 20	5 subjects, 1040 real trials per subject 400 fake trials per subject
		EEG + AWGN 40	
		EEG + MF3,001	
		EEG + MF9,001	
		EEG + MF9,05	

METRICS		
Parameter	**Relation** [1]	
Accuracy (A)	$\mathrm{A} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$	(3)
Precision or positive predicted value (PPV)	$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$	(4)
Recall or true positive rate (TPR)	$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$	(5)
F1 score (F1)	$\mathrm{F1} = 2\frac{\mathrm{PPV} \cdot \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}}$	(6)
Cyberattack impact (Δ)	$\Delta = \mathrm{V}_{real} - \mathrm{V}_{fake}$	(7)

[1] TP: true positives; TN: true negatives; FP: false positives; FN: false negatives; V_{real}: generic metric calculated with real P300 dataset; V_{fake}: same metric calculated with fake P300 dataset.

3.2. Results of the Simulated Cyberattack

Each trained model was individually tested using all of the testing datasets in Table 2. The results of the simulated cyberattack are summarized in Figure 2. Data are presented as averages and standard deviations over the five subjects. Subject-specific data are reported in Tables S1–S5 of the Supplementary Materials of this work. Figure 2 shows that the presence of cyberattacks generated using (MF9, 001), (AWGN20), (AWGN40), and (MF9, 05) leads to a reduction in precision and F1-score parameters compared to the real P300 dataset regardless of the used SVM model. On average, all of the parameters are affected by the introduction of fake signals.

Figure 3 shows the cyberattack impact, as defined in Table 2. Averaging parameter values over the five subjects and the five SVM kernels results in a reduction in the accuracy parameter, which ranges between 3% (MF3, 001) and 7% (AWGN 20). Precision is reduced by about 5% with all of the cyberattacks except for AWGN 40, which reduces the precision by 4%. Recall shows a reduction between 2% (AWGN 40) and 5% (MF3, 001). The F1-score decreases by 4% in most cases, except for AWGN 20, which achieves a 6% decrease. Considering the overall cyberattack impact, as shown in Figure 3, the attack using fake signals generated with (MF3, 001) is demonstrated to be the most effective hacking method. This method simultaneously reduces all parameters relative to those calculated on the real P300 dataset, regardless of the used SVM model.

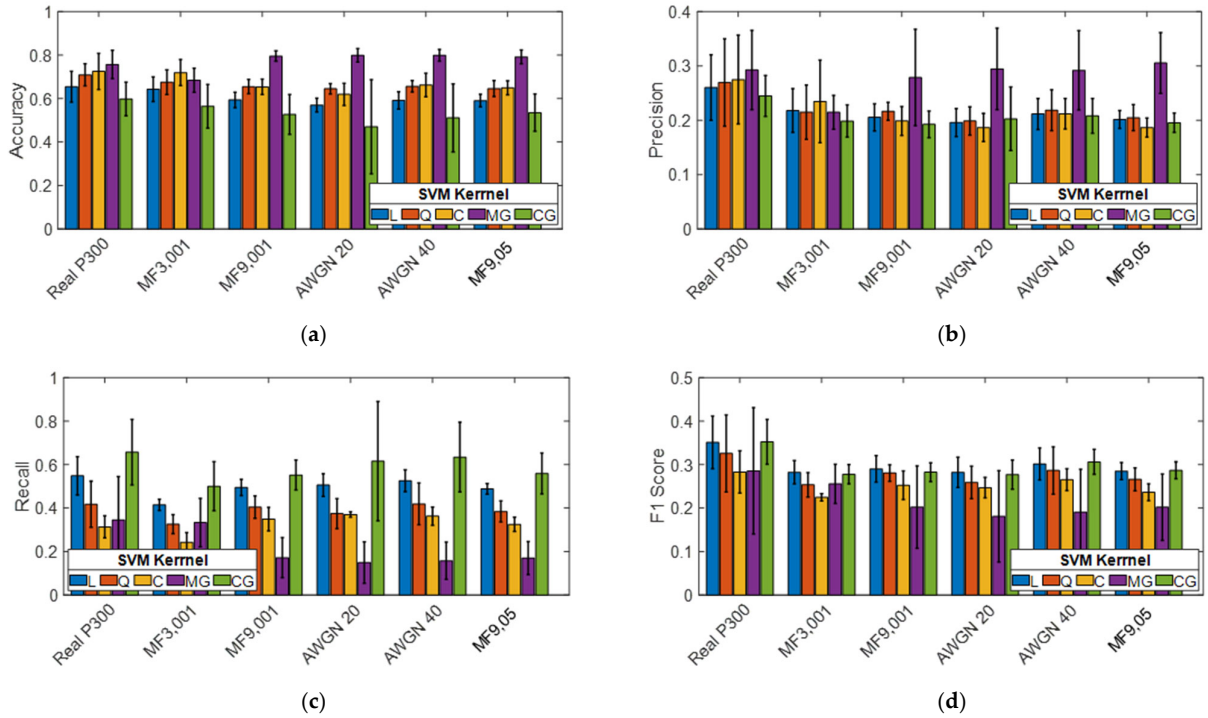

Figure 2. Test of the SVM models using the real P300 dataset (real P300) vs. hacked P300 datasets. Performance metrics are (**a**) accuracy, (**b**) precision, (**c**) recall, and (**d**) F1 score. Data are represented as averages and standard deviations over the five subjects.

The simulated cyberattacks primarily show that standard classification algorithms are incapable of identifying cyberattacks, consequently losing their overall accuracy in discriminating P300 trials because of wrong classification outcomes. The fake P300 data eluded all of the SVM models used in this study. This performance is intolerable in a real-life application. These simulation results are in line with findings from [10,14,15], raising serious concerns about the security of most BCIs.

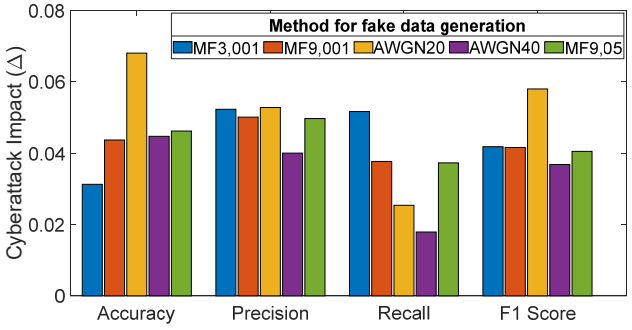

Figure 3. Metric reduction versus BCI performance metrics for each cyberattack method.

4. Architecture of the Brain Hacking Recognizer (BHR) for the Cybersecure BCI

The proposed cybersecure architecture is based on encrypted multi-channel communication between the EEG headset and the BCI framework. We designed and implemented a brain hacking recognizer (BHR) architecture that supervises the communication between the EEG headset and the BCI framework. The BHR aims to (i) detect communication anomalies and potential cyberattacks, (ii) supply the BCI framework with possible

actions to inhibit the attack, and (iii) lead the BCI to a secure status. Figure 4 depicts the proposed architecture.

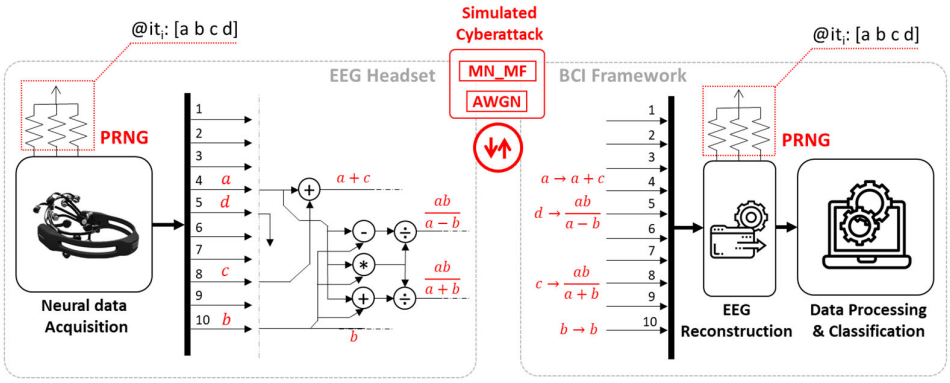

Figure 4. Architecture of a 10-channel BHR.

4.1. Brain Hacking Recognizer

The working principle of the BHR is based on a pseudo-random number generator (PRNG). A PRNG consists of a set of pull-up resistors identifying a hardware-specific random seed replicated on both the EEG headset and the BCI framework [37]. As such, the PRNG supplies known random numbers to the EEG headset and BCI framework. The random numbers are deterministic if the seed and iteration numbers are known (as in the presented application).

At a generic ith iteration, the PRNG on the EEG headset side generates four indexes (a, b , c , and d in Figure 4). As per the demonstrative BHR overview on 10 channels in Figure 4, these indexes are used to randomly select four specific EEG channels (e.g., in the ith iteration, index a corresponds to EEG channel 4). Data on the selected channels are mixed using algebraic operations in Figure 4, with channel b transmitted unaltered. The new channel composition is buffered and transmitted to the BCI framework.

Thereafter, the BCI framework receives the data. The BCI framework, having the same PRNG setting and iteration number, is aware of the four indexes a, b , c , and d. Data on the channel indexed as b represent a reference pattern. The same signal can also be estimated using the remaining channels a, c , and d using appropriate inverse operations, as shown in Figure 4. Let us use $\hat{b}$ to denote the estimation of the signal on channel b using only data from channels a, c , and d . The BHR estimates $\hat{b}$ using the following inverse operations:

$$\frac{ab}{a+b} = x \qquad \frac{ab}{a-b} = y \tag{8}$$

$$\frac{1}{b} + \frac{1}{a} = \frac{1}{x} \qquad \frac{1}{b} - \frac{1}{a} = \frac{1}{y} \tag{9}$$

Let us define:

$$\frac{1}{b} = e \qquad \frac{1}{a} = f \tag{10}$$

Considering the sum among the equations composing the system defined by Equations (9) and (10) results in:

$$e = \frac{\left(\frac{1}{c} + \frac{1}{d}\right)}{2} \rightarrow \hat{b} \tag{11}$$

where $\hat{b}$ is the estimation of the signal on the channel indexed as b.

If $\hat{b}$ coincides with data received on channel b, the BHR marks the transmission as legitimate and proceeds with the estimation of a and the reconstruction of signal on channel c, feeding the data processing and classification block discussed in Section 2. In detail, if the transmission is attack-free, the reconstruction returns, ideally, $\hat{b} = b$ because channels indexed as c and d are unaltered. In this case, the signal on the channel indexed as a can be estimated from $\hat{b}$ according to:

$$f = \frac{1}{x} - e \rightarrow \hat{a} \tag{12}$$

Obtaining the estimation â, it is also possible to derive the signal on channel *c* on the EEG headset side.

If $\hat{b}$ does not match data received on channel b, the BHR marks the transmission as "compromised" and plans a counteraction to inhibit the attack. In the presence of a cyberattack, we assume that the attacker is unaware of the seed (a, b , c , and d) and iteration number. Therefore, even if an attack based on the P300 trial with physiological characteristics is generated, the attacker will modify the channels, making channel b reconstruction impossible on the BCI framework side. Moreover, even if the attacker is aware of the channel mixing methodology, it has a very low probability of finding the right disposition among the available ones. A detailed analysis of this probability is provided in Section 3.2.

In any case, an erroneous attack trial would lead to $\hat{b} \neq b$. In this case, a cyberattack inhibition step follows the BHR alert (i.e., BCI under attack). It is beyond the scope of this work to define the best counterattack, as this is application-dependent. However, as a proof of concept, let us consider a P300 speller. The unsafe situation in a P300 speller is the presence of observations erroneously considered to be positive (i.e., false positive). This situation would lead to incorrect word spelling. In this case, false negatives do not affect the proper functioning of the system. For cyberattacks on P300-based spellers, we, therefore, recommend forcing the BCI to classify the corrupted trial as non-P300, blocking its operation with an attack warning for the user. The concept can be easily expanded to a wheelchair or car that drives through a BCI [38].

4.2. Results of the BHR Rejection Capabilities

To demonstrate that the BHR is effective in rejecting cyberattacks, a testbench composed of a total of 144 million attack trials was tested. The testbench was realized on the basis of 100,000 sessions. Each session consisted of 1440 attacks (i.e., fake signals), which represent the whole dataset. In this study, we assumed that the hacker is aware of the presence of the BHR and has knowledge of the PRNG hardware/algorithm, but the seed and iteration number are unknown. This represents the worst-case scenario.

Figure 5 depicts the occurrence of successfully hacked trials per session. The results show that in 5780 sessions out of 100,000 (5.78%), a single trial out of 1440 available was successfully hacked. In the same context, 160/100,000 sessions (0.16%) recorded the successful corruption of two trials. Similarly, 4/100,000 sessions (0.004%) presented three affected trials. Overall, 6112 attacks out of 144 million trials (0.0042%) overcame the BHR protection system. Therefore, BHR was successful in rejecting 99.996% of the cyberattack trials regardless of the hacking type (e.g., noise-based or physiological-like attack).

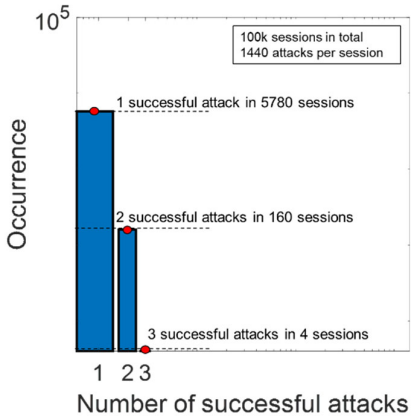

Figure 5. Frequency of successfully hacked trials of 100,000 sessions composed of 1440 attacks (whole dataset).

To provide a complete overview of the proposed brainjacking recognition method, the computational time of the overall processing chain on the BCI framework receiver side was assessed. The test was conducted on a laptop PC with Intel Core i5-10210U and 8 GB RAM. Starting from the stimulus onset, the whole processing chain on the receiver side required a total of 612.7 ms. Specifically, the EEG data buffer filling for the trial definition required 600 ms (97.9% of the total time), while 11.58 ± 0.38 ms (2% of the total time) was required in the classification stage. The introduction of the BHR increased the processing time by 0.826 ms ± 0.69 ms (0.1% of the total time) on average. Therefore, the adoption of the implemented architecture has a negligible impact on the computational time of the system, allowing it to readily detect and inhibit a cyberattack.

4.3. Results of the Simulated Cyberattack with BHR

The five previously trained SVM models were supplied with the proposed BHR. Each new BCI was tested on the hacked P300 datasets reported in Table 2. The nature of the BHR working principle leads the system to react in the same way regardless of the hacking type. Indeed, according to the inhibition procedure, the BHR forces the BCI to label the trial as non-P300 as soon as it detects discrepancies between signals on specific channels. This happens for any considered hacking type, resulting in a unique BCI-dependent outcome.

Figure 6 shows the performance metric behavior when the real P300 dataset (reference) or the hacked P300 datasets are used to test the BCIs with and without the BHR. The first and last five blocks of histograms are repurposed from Figure 3. The second block of histograms reports the mean value and standard deviation of the specific parameter computed on the five subjects with different SVM kernels when BCIs are supplied with the BHR. The mean value over the five SVM models is also highlighted with a dotted red line in Figure 6.

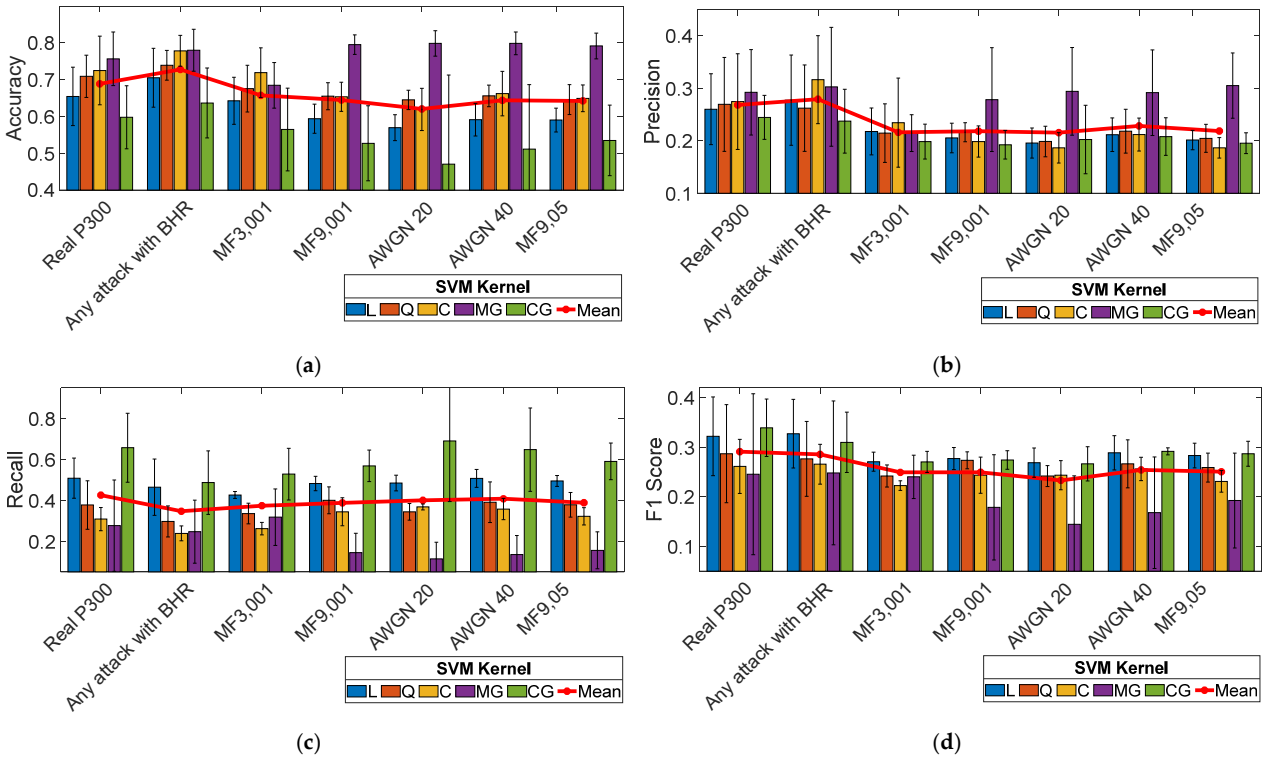

Figure 6. Test of the SVM models using the real P300 dataset (real P300) vs. hacked datasets. Performance metrics are (**a**) accuracy, (**b**) precision, (**c**) recall, and (**d**) F1-score. Data are presented as averages and standard deviations over the five subjects. The black line with markers reports the average value of the metric calculated on the five SVM kernels.

The results in Figure 7 show that, overall, the BCIs supplied with the proposed BHR improve their accuracy, precision, and F1-score parameters at the expense of the recall metric. Indeed, as per Equation (5), the increase in FN, with a consequent reduction in TP, due to the attack on P300 trials (constrained with the label non-P300 by the inhibition procedure) hardly affected the BCI recall. Figure 7 illustrates the impact on BCI performance through the cyberattack impact parameter, as defined in Table 2. The results demonstrate that accuracy is improved relative to the real P300 dataset reference by about 4% and by 7.1% for the less effective attack (i.e., the MF3, 001). Similarly, the precision increases by 1.1% relative to the reference and 5.1% for the less effective attack (i.e., AWGN 40). The F1-score is still reduced relative to the reference performance by 3.3%; however, the BHR introduction improves the F1-score by 0.3% if compared with the AWGN 40 dataset result (less effective attack in terms of F1-score). Conversely, the recall parameter shows a large decrease of 10.9% considering the reference value and 5.8% relative to the most effective attack source (i.e., MF3, 001). Subject-specific data are reported in Tables S6–S10 of the Supplementary Materials of this work.

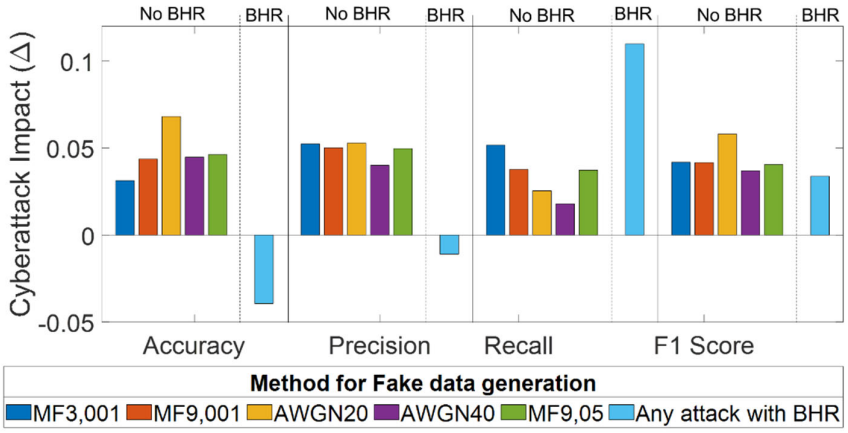

Figure 7. Metric reduction versus BCI performance metrics for each cyberattack method.

4.4. Discussion

The approach proposed in this work applies to a specific scenario where (i) the hacker attacks the communication channel between sensors and the BCI framework; (ii) the hacker has knowledge of the hardware, (iii) wireless communication system, and (iv) BCI framework; (v) the attacker uses EEG signals from other subjects other than the user under attack; and (vi) fake P300 signals are synthetically created using noise-based and median filter techniques. Arguably, the scenario proposed in this work is one of the most plausible. However, there are many other scenarios where the hacker operates differently or has different degrees of knowledge. For instance, the attacker might be able to access user-specific data or might have direct access to the BCI framework. The assessment of the BHR performance in different hacking scenarios is beyond the scope of this work and will be investigated in future works. Nevertheless, since the working principle of the BHR is based on a low-resource pseudo-random algebraic combination of the data to find anomalies in communication before the BCI framework operation, the nature and the provenience of data (i.e., P300 from the same subject, different subjects, synthetic or simply noise) are irrelevant. The minimum requirement for this architecture to work is to have at least four synchronized communication channels from the sensors to the BCI framework. Although we used EEG data to demonstrate the architecture, the data type is irrelevant to the proper functioning of the cybersecure framework. Therefore, we envisage that the proposed architecture can be applied to other multi-channel systems for measuring neural potentials, including electrocorticography (ECOG) [39] and local field potentials (LFP).

Furthermore, the BHR reconstructs data on the BCI framework end before any classification or machine learning is performed. As such, the features of the neural pattern and the algorithm used for its classification do not affect the BHR operation. Although we demonstrated the proposed solution for a P300 SVM-based BCI, the BHR is modular and can be applied to recognize hacking attempts on BCI systems based on other patterns (e.g., as movement or pre-movement brain potentials) and algorithms (e.g., neural networks, decision tree, random forest) [40].

5. Conclusions

Cybersecurity is an aspect that has received far less attention when designing and implementing a BCI. In this work, a novel architecture to improve cybersecurity in BCI systems is presented. The working principle of the architecture is based on the transmission of a linear combination of EEG data on randomly selected and dynamically changing communication channels. This approach allows identifying breaches of the wireless communication between the EEG headset and the BCI framework. Wireless communication is indeed the weakest point in current BCI solutions, as it is exposed to external cyberattacks.

Our study shows that with no security features, current BCI systems are exposed to cyber threats. By using simulated cyberattacks, we demonstrate that SVM models, which are very popular in BCIs, are incapable of discriminating real EEG data from fake signals. After implementing the proposed cybersecurity solution, the BCI framework was able to identify 99.996% of simulated cyberattacks. As such, the BCI system can therefore implement routines to both inhibit further attacks and ensure the safe functioning of the user interface according to the specific application.

The architecture proposed in this study is versatile and can be easily implemented on a variety of BCI models. We believe that our study can also raise awareness of the risk of using exposed BCI systems so that designers will dedicate special attention to this aspect of the interface.

Supplementary Materials: The following are available online at www.mdpi.com/article/10.3390/s21248280/s1, Tables S1–S5: Test of the SVM models using the real P300 dataset (real P300) vs. hacked P300 datasets for Subject *number of subject from 1 to 5*. Tables S6–S10: Test of the SVM models using the real P300 dataset (real P300) vs. hacked P300 datasets for Subject *number of subject from 1 to 5* when BCI is supplied with the proposed BHR.

Author Contributions: Conceptualization, G.M., V.F.A. and D.D.V.; methodology, G.M. and V.F.A.; software, G.M.; validation, G.M., V.F.A. and D.D.V.; formal analysis, G.M.; investigation, G.M. and V.F.A.; resources, D.D.V.; data curation, G.M. and V.F.A.; writing—original draft preparation, V.F.A. and G.M.; writing—review and editing, G.M., V.F.A. and D.D.V.; visualization, G.M. and V.F.A.; supervision, D.D.V.; project administration, D.D.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: In this study, we used a publicly available EEG dataset [13]. The dataset has been modified to simulate cyberattacks, as described in Section 2.

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

References

1. De Venuto, D.; Annese, V.F.; Mezzina, G. Real-time P300-based BCI in mechatronic control by using a multi-dimensional approach. *IET Softw.* **2018**, *12*, 5, doi:10.1049/iet-sen.2017.0340.
2. Annese, V.F.; Mezzina, G.; De Venuto, D. Towards mobile health care: Neurocognitive impairment monitoring by BCI-based game. In Proceedings of the 2016 IEEE SENSORS, Orlando, FL, USA, 30 October–3 November 2016, doi:10.1109/ICSENS.2016.7808745.
3. Brain Computer Interface Market Size and Industry Trends|2030. Available online: https://www.alliedmarketresearch.com/brain-computer-interfaces-market (accessed on 16 September 2021).
4. Wolpaw, J.R.; McFarland, D.J.; Neat, G.W.; Forneris, C.A. An EEG-based brain-computer interface for cursor control. *Electroencephalogr. Clin. Neurophysiol.* **1991**, *78*, 252–259, doi:10.1016/0013-4694(91)90040-B.
5. Yin, E.; Zhou, Z.; Jiang, J.; Yu, Y.; Hu, D. A dynamically optimized SSVEP brain-computer interface (BCI) speller. *IEEE Trans. Biomed. Eng.* **2015**, *62*, 1447–1456, doi:10.1109/TBME.2014.2320948.
6. Navarro, J.; Reynaud, E.; Osiurak, F. Neuroergonomics of car driving: A critical meta-analysis of neuroimaging data on the human brain behind the wheel. *Neurosci. Biobehav. Rev.* **2018**, *95*, 464–479, doi:10.1016/J.NEUBIOREV.2018.10.016.
7. De Venuto, Daniela, Valerio Francesco Annese, Giovanni Mezzina, Michele Ruta, and Eugenio Di Sciascio. "Brain-computer interface using P300: a gaming approach for neurocognitive impairment diagnosis." In 2016 IEEE International High Level Design Validation and Test Workshop (HLDVT), pp. 93-99. IEEE, 2016
8. Christensen, S.M.; Holm, N.S.; Puthusserypady, S. An improved five class MI based BCI Scheme for Drone Control Using Filter Bank CSP. In Proceedings of the 7th International Winter Conference on Brain-Computer Interface, BCI 2019, Gangwon, Korea, 18–20 February 2019, doi:10.1109/IWW-BCI.2019.8737263.
9. De Venuto, D., Annese V. F., Mezzina G. and Defazio G. "FPGA-Based Embedded Cyber-Physical Platform to Assess Gait and Postural Stability in Parkinson's Disease," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, **2018**, vol. 8, no. 7, pp. 1167-1179, doi: 10.1109/TCPMT.2018.2810103.
10. Pugh, J.; Pycroft, L.; Sandberg, A.; Aziz, T.; Savulescu, J. Brainjacking in deep brain stimulation and autonomy. *Ethics Inf. Technol.* **2018**, *203*, 219–232, doi:10.1007/S10676-018-9466-4.
11. Hosseini, M.-P.; Pompili, D.; Elisevich, K.; Soltanian-Zadeh, H. Optimized Deep Learning for EEG Big Data and Seizure Prediction BCI via Internet of Things. *IEEE Trans. Big Data* **2017**, *3*, 392–404, doi:10.1109/TBDATA.2017.2769670.

12. Attiah, M.A.; Farah, M.J. Minds, motherboards, and money: Futurism and realism in the neuroethics of BCI technologies. *Front. Syst. Neurosci.* **2014**, 86, doi:10.3389/FNSYS.2014.00086.
13. Fouad, I.A. A robust and reliable online P300-based BCI system using Emotiv EPOC + headset. *J. Med. Eng. Technol.* **2021**, *45*, 94–114, doi:10.1080/03091902.2020.1853840.
14. Beltrán, E.T.M.; Pérez, M.Q.; Bernal, S.L.; Celdrán, A.H.; Pérez, G.M. Noise-based cyberattacks generating fake P300 waves in brain–computer interfaces. *Clust. Comput.* **2021**, 1–16, doi:10.1007/S10586-021-03326-Z.
15. Shaukat, K.; Luo, S.; Varadharajan, V.; Hameed, I.A.; Chen, S.; Liu, D.; Li, J. Performance Comparison and Current Challenges of Using Machine Learning Techniques in Cybersecurity. *Energies* **2020**, *13*, 2509, doi:10.3390/EN13102509.
16. Sergio López Bernal, Alberto Huertas Celdrán, Gregorio Martínez Pérez, Michael Taynnan Barros, and Sasitharan Balasubramaniam. 2021. Security in Brain-Computer Interfaces: State-of-the-Art, Opportunities, and Future Challenges. ACM Comput. Surv. 54, 1, Article 11 (January 2022), 35 pages. DOI:https://doi.org/10.1145/3427376
17. Escolano, C.; Antelis, J.M.; Minguez, J. A telepresence mobile robot controlled with a noninvasive brain-computer interface. *IEEE Trans. Syst. Man Cybern. Part. B Cybern.* **2012**, *42*, 793–804, doi:10.1109/TSMCB.2011.2177968.
18. AIatropoulos, L.; Moysis, L.; Giakoumis, A.; Volos, C.; Ouannas, A.; Goudos, S. Medical Data Encryption based on a Modified Sinusoidal 1D Chaotic Map and Its Microcontroller Implementation. In Proceedings of the 2021 10th International Conference on Modern Circuits and Systems Technologies (MOCAST), Thessaloniki, Greece, 5–7 July 2021, doi:10.1109/MOCAST52088.2021.9493422.
19. Takabi, H.; Bhalotiya, A.; Alohaly, M. Brain Computer Interface (BCI) Applications: Privacy Threats and Countermeasures. In Proceedings of the 2016 IEEE 2nd International Conference on Collaboration and Internet Computing (CIC), Pittsburgh, PA, USA, 1–3 November 2016; pp. 102–111, doi:10.1109/CIC.2016.026.
20. Sasko, A.; Hillsgrove, T.; Gagneja, K.; Katugampola, U. System Usage Profiling Metrics for Notifications on Abnormal User Behavior. *Commun. Comput. Inf. Sci.* **2019**, *1113*, 149–160, doi:10.1007/978-3-030-34353-8_11.
21. Gui, Q.; Yang, W.; Jin, Z.; Ruiz-Blondet, M.V.; Laszlo, S. A residual feature-based replay attack detection approach for brainprint biometric systems. In Proceedings of the 2016 8th IEEE International Workshop on Information Forensics and Security (WIFS 2016), Location: Abu Dhabi, United Arab Emirates, 4–7 December 2016, doi:10.1109/WIFS.2016.7823907.
22. Belkacem, A.N. Cybersecurity Framework for P300-based Brain Computer Interface. In Proceedings of the 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, ON, Canada, 11–14 October 2020; pp. 1–6, doi:10.1109/SMC42975.2020.9283100.
23. Salhi, K.; Alimi, A.M.; Khelifa, M.M.B.; Gorce, P. Improved secure navigation of wheelchairs using multi-robot system and cloud computing technologies. In Proceedings of the 2015 11th International Conference on Information Assurance and Security (IAS), Marrakesh, Morocco, 14–16 December 2015; pp. 50–54, doi:10.1109/ISIAS.2015.7492744.
24. Ajrawi, S.; Rao, R.; Sarkar, M. Cybersecurity in Brain-Computer Interfaces: RFID-based design-theoretical framework. *Inform. Med. Unlocked* **2021**, *22*, 100489, doi:10.1016/J.IMU.2020.100489.
25. Rathi, N.; Singla, R.; Tiwari, S. Authentication framework for security application developed using a pictorial P300 speller. *Brain-Comput. Interfaces* **2020**, *7*, 70–89, doi:10.1080/2326263X.2020.1860520.
26. Borkotoky, C.; Galgate, S.; Nimbekar, S.B. Human computer interaction: Harnessing P300 potential brain waves for authentication of individuals. In Proceedings of the 1st Bangalore Annual Compute Conference, Compute 2008, Bangalore, India, 18–20 January 2008, doi:10.1145/1341771.1341797.
27. Munoz, C.M.B.; Cruz, F.G.; Valero, J.S.J. Software architecture for the application of facial recognition techniques through IoT devices. In Proceedings of the 2020 Congreso Internacional de Innovación y Tendencias en Ingeniería (CONIITI), Bogotá, Colombia, 30 September–2 October 2019, doi:10.1109/CONIITI51147.2020.9240416.
28. Zou, Y.; Zhu, J.; Wang, X.; Hanzo, L. A Survey on Wireless Security: Technical Challenges, Recent Advances, and Future Trends. *Proc. IEEE* **2016**, *104*, 1727–1765, doi:10.1109/JPROC.2016.2558521.
29. Bonaci, T.; Herron, J.; Matlack, C.; Chizeck, H.J. Securing the exocortex: A twenty-first century cybernetics challenge. In Proceedings of the IEEE Conference on Norbert Wiener in the 21st Century (21CW), Boston, MA, USA, 24–26 June 2014, doi:10.1109/NORBERT.2014.6893912.
30. Rahman, M.L.; Bardhan, S.; Neupane, A.; Papalexakis, E.; Song, C. Learning Tensor-Based Representations from Brain-Computer Interface Data for Cybersecurity. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 389–404, doi:10.1007/978-3-030-10997-4_24.
31. EMOTIV EPOC+ 14-Channel Wireless EEG Headset | EMOTIV. Available online: https://www.emotiv.com/epoc/ (accessed on 21 September 2021).
32. Hoffmann, U.; Vesin, J.M.; Ebrahimi, T.; Diserens, K. An efficient P300-based brain–computer interface for disabled subjects. *J. Neurosci. Methods* **2008**, *167*, 115–125, doi:10.1016/J.JNEUMETH.2007.03.005.
33. Bougrain, L.; Saavedra, C.; Ranta, R.; Bougrain, L.; Saavedra, C.; Ranta, R. Finally, What Is the Best Filter for P300 Detection?. 2012. Available online: https://hal.inria.fr/hal-00756669 (accessed on 2 December 2021).
34. Kaper, M.; Meinicke, P.; Grossekathoefer, U.; Lingner, T.; Ritter, H. BCI competition 2003—Data set IIb: Support vector machines for the P300 speller paradigm. *IEEE Trans. Biomed. Eng.* **2004**, *51*, 1073–1076, doi:10.1109/TBME.2004.826698.
35. Patrone, M.; Lecumberry, F.; Martín, Á.; Ramirez, I.; Seroussi, G. EEG Signal Pre-Processing for the P300 Speller. In *Iberoamerican Congress on Pattern Recognition;* Springer: Berlin/Heidelberg, Germany, 2015; pp. 559–566, doi:10.1007/978-3-319-25751-8_67.

36. Parisot, A.; Bento, L.M.S.; Machado, R.C.S. Testing and selecting lightweight pseudo-random number generators for IoT devices. In Proceedings of the 2021 IEEE International Workshop on Metrology for Industry 4.0 & IoT (MetroInd4. 0&IoT), Rome, Italy, 7–9 June 2021; pp. 715–720, doi:10.1109/METROIND4.0IOT51437.2021.9488454.
37. Baldanzi, L.; Crocetti, L.; Falaschi, F.; Bertolucci, M.; Belli, J.; Fanucci, L.; Saponara, S. Cryptographically Secure Pseudo-Random Number Generator IP-Core Based on SHA2 Algorithm. *Sensors* **2020**, *20*, 1869, doi:10.3390/S20071869.
38. De Venuto, D.; Annese, V.F.; Mezzina, G. An embedded system remotely driving mechanical devices by P300 brain activity. In Proceedings of the Design, Automation & Test in Europe Conference & Exhibition (DATE), Lausanne, Switzerland, 27–31 March 2017, doi:10.23919/DATE.2017.7927139.
39. De Venuto, Daniela, and Jan Rabaey. "RFID transceiver for wireless powering brain implanted microelectrodes and backscattered neural data collection." *Microelectronics Journal*, **2014**, 45, no. 12: 1585-1594.
40. De Venuto, D., and Ohletz M. J. "On-chip test for mixed-signal ASICs using two-mode comparators with bias-programmable reference voltages." *Journal of Electronic Testing*, **2001**, 17.3: 243-253.