



# Ciberataques basados en ruido que generan ondas P300 falsas en interfaces cerebro-computadora

Enrique Tomás Martínez Beltrán • Mario Quiles Pérez<sup>1</sup> • Sergio López Bernal<sup>1</sup> • Alberto Huertas Celdra<sup>n2</sup> • Gregorio Martínez Pérez<sup>1</sup>

Recibido: 5 de marzo de 2021 / Revisado: 14 de abril de 2021 / Aceptado: 31 de mayo de 2021 / Publicado en línea: 10 de julio de 2021  
El autor(es) 2021, publicación corregida 2021

## Resumen

La mayoría de los escenarios actuales de aplicación de las Interfaces Cerebro-Computadora (BCI) utilizan señales electroencefalográficas (EEG) que contienen información del sujeto. Esto significa que si las EEG se manipularan maliciosamente, el correcto funcionamiento de los marcos BCI podría estar en riesgo. Desafortunadamente, esto sucede en marcos sensibles a ciberataques basados en ruido, y se necesitan más esfuerzos para medir el impacto de estos ataques. Este trabajo presenta y analiza el impacto de cuatro ciberataques basados en ruido que intentan generar ondas P300 falsas en dos fases diferentes de un marco BCI. Un conjunto de experimentos muestra que cuanto mayor es el conocimiento del atacante con respecto a las ondas P300, los procesos y los datos del marco BCI, mayor es el impacto del ataque. En este sentido, el atacante con menor conocimiento impacta un 1% en la fase de adquisición y un 4% en la fase de procesamiento, mientras que el atacante con mayor conocimiento impacta un 22% y un 74%, respectivamente.

**Palabras clave** Interfaces cerebro-computadora Ciberseguridad Ciberataques basados en ruido Integridad de datos Señal electroencefalográfica P300

## 1 Introducción

Las interfaces cerebro-computadora (BCI) presentan un canal de comunicación bidireccional entre el cerebro y el sistema externo. dispositivos. El ciclo de vida de la BCI es bidireccional, ya que puede adquirir la actividad neuronal producida por un sujeto y estimular

o inhibir neuronas. La Figura 1 muestra una vista reducida del ciclo completo de la BCI presentado en nuestro trabajo previo [15] con los procesos y comunicaciones realizados en ambas direcciones. Dado que este trabajo se centra en la adquisición de datos neuronales (representada por el flujo más oscuro en la Fig. 1), prestaremos más atención a esa dirección. En este sentido, las señales cerebrales producidas por la actividad cerebral son adquiridas y procesadas por la BCI. Finalmente, se transforman en un comando que las aplicaciones de la BCI pueden ejecutar. En ocasiones, este comando genera retroalimentación visual, auditiva o somatosensorial para el usuario, cerrando el ciclo. En la dirección opuesta, en gris en la Fig. 1, la estimulación neuronal también es posible para estimular áreas específicas del cerebro.

En el ámbito médico, las BCI proporcionan un sistema de comunicación alternativo que facilita la rehabilitación, la mejora de las habilidades motoras y el control de prótesis robóticas [3]. Las BCI también se utilizan para tratar la disfunción cognitiva [20], trastornos neurológicos como la esclerosis lateral amiotrófica (ELA) [4], o incluso para identificar y aliviar el dolor provocado por el síndrome del miembro fantasma [21]. Por otro lado, estos sistemas también permiten predecir una convulsión antes de que se produzca, lo que permite que los pacientes reciban la atención necesaria [12]. En situaciones de conducción, también se ha incrementado el uso de

y Enrique Tomás Martínez Beltrán  
enriquetomas.martinezb@um.es

Mario Quiles Pérez  
mario.quilesp@um.es

Sergio López  
Bernallopez@um.es

Alberto Huertas Celdra  
huertas@ifi.uzh.ch

Gregorio Martínez Pérez  
gregorio@um.es

<sup>1</sup> Departamento de Ingeniería de la Información y las Comunicaciones, Universidad de Murcia, 30100 Murcia, España

<sup>2</sup> Grupo de Sistemas de Comunicación (CSG), Departamento de Informática (Ifi), Universidad de Zurich UZH, 8050 Zurich, Suiza

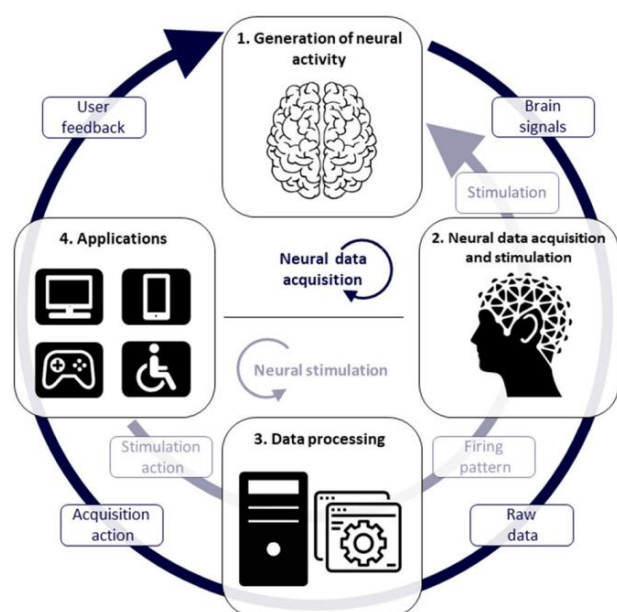


Fig. 1. Resumen de las fases del ciclo BCI. El flujo más oscuro en sentido horario muestra la monitorización de la señal neuronal. El flujo más claro en sentido antihorario indica estimulación neuronal.

Estos dispositivos detectan la embriaguez [28] o la somnolencia [11] en la carretera. Ambas perspectivas se complementan para promover una conducción óptima y reducir la posibilidad de accidentes. En otros sectores, como el entretenimiento, las BCI mejoran la interacción, la inmersión y, en resumen, la experiencia de juego de los jugadores [1]. Las BCI han llegado incluso al sector militar, donde se utilizan en el control mental de drones remotos [2] o exoesqueletos [5].

La mayoría de los escenarios de aplicación anteriores utilizan la señal electroencefalográfica, EEG [18] y potenciales evocados [7] como un medio para obtener información neuronal. Los potenciales evocados son patrones de señales generados automáticamente por el cerebro cuando se presentan estímulos al individuo. Dependiendo del tipo de estímulo, existen diferentes tipos de potenciales: visuales, auditivos, somatosensoriales o cognitivos. El potencial evocado P300 (o P3) [22] es uno de los más estudiados y conocidos para el registro cerebral. Esta respuesta originada por el cerebro lleva un pico de señal positivo apreciable en la señal EEG a 250-500 ms después de que se presenta un estímulo al individuo [22]. Diferentes procedimientos pueden permitir que el P300 aparezca en la señal EEG, como el Paradigma Oddball [8]. Este paradigma consiste en presentar una serie de estímulos conocidos, mezclados aleatoriamente, a un sujeto de los cuales el 10-20% son conocidos o familiares. Los estímulos visuales y auditivos pueden activar el P300, pero este trabajo se centra en los visuales.

Entre los potenciales evocados existentes, el P300 es uno de los más utilizados en aplicaciones de uso final. Este potencial tiene la capacidad de representar información neuronal considerable del sujeto, lo que lo convierte en una fuente de datos prometedora.

Información para el dispositivo final. Por esta razón, actualmente se utiliza en numerosas aplicaciones, como el control de sillas de ruedas, exoesqueletos militares y deletreadores. A pesar de las numerosas ventajas que ofrece el P300, la relevancia y el valor de los datos neuronales obtenidos aumentan la criticidad de los dispositivos BCI. En los últimos años, numerosos artículos se han centrado en la falta de medidas de seguridad tanto en el software como en el hardware de BCI. En este sentido, se han publicado investigaciones que ofrecen una perspectiva de ciberseguridad sobre los dispositivos BCI y la señal EEG adquirida.

Más específicamente, algunos autores detallaron varios ciberataques dirigidos contra la confidencialidad de los datos y la privacidad del usuario [13, 17], mientras que otros se centraron en afectar la integridad de la señal EEG atenuando los potenciales evocados [30].

En este contexto, a pesar de la cantidad de artículos que abordan los potenciales evocados, se requieren más esfuerzos para medir el impacto que los ciberataques tienen sobre ellos. Más específicamente, existe una falta de literatura sobre cómo puede verse comprometida la integridad de los datos gestionados por los marcos BCI.

Esta debilidad se complementa con un análisis limitado del impacto de los ciberataques en las diferentes fases del ciclo BCI. En este sentido, este artículo propone un estudio del impacto de los ciberataques centrados en la generación maliciosa de P300 en la señal EEG para determinar su impacto en los dispositivos BCI y, en consecuencia, en las aplicaciones finales. La investigación busca mostrar el impacto real de los ciberataques que afectan la integridad y la preocupación real por mantener la seguridad de los dispositivos en un mundo donde las BCI están adquiriendo un papel relevante en la forma en que los sujetos se comunican con el entorno.

Para mejorar algunas de las limitaciones anteriores, este trabajo presenta las siguientes contribuciones principales:

La selección de cuatro perfiles de ataque basados en ruido con conocimiento incremental para generar artificialmente potenciales P300 en señales de EEG. Por lo tanto, la variación entre los perfiles depende del conocimiento existente sobre el dispositivo BCI, aspectos de la señal EEG y el marco. Más detalladamente, el primer atacante conoce la presencia de comunicación inalámbrica entre el auricular BCI y el marco BCI; el segundo conoce conceptos teóricos de la señal EEG y el potencial P300, como su amplitud o intervalo de generación; el tercero conoce lo mismo que el segundo, así como la naturaleza y el procesamiento de los datos intercambiados; y el cuarto conoce lo mismo que el tercero, además de los modelos de clasificación utilizados para detectar el P300 y sus predicciones.

La definición y el despliegue de un escenario realista para ejecutar los ataques anteriores y demostrar su viabilidad en dos fases o procesos de un marco BCI: adquisición y procesamiento de EEG. El escenario propuesto considera un vídeo con imágenes conocidas y desconocidas para el sujeto. Estos estímulos visuales generan una reacción en las ondas cerebrales del sujeto.

Se basa en el paradigma Oddball, donde se presentan estímulos visuales familiares (objetivo) dentro de un conjunto de estímulos desconocidos (no objetivo). El escenario también considera un auricular BCI para adquirir el EEG y un marco que implementa el ciclo BCI (véase la Fig. 1) para obtener las señales del EEG, procesarlas y detectar el P300.

Análisis del impacto de los cuatro perfiles de ataque basados en ruido que afectan al escenario propuesto. En este contexto, los resultados obtenidos demuestran que un mayor conocimiento de la BCI y del escenario aumenta el impacto de los ciberataques basados en ruido. Asimismo, se muestra que la puntuación AUC del mejor clasificador que detecta P300 se reduce un 1 %, un 3 %, un 12 % y un 22 % al atacar la fase de adquisición, y un 4 %, un 10 %, un 41 % y un 74 % cuando la fase de procesamiento de datos se ve afectada por cada uno de los cuatro perfiles, respectivamente.

El resto del artículo se estructura de la siguiente manera. La sección 2 analiza los problemas de seguridad en dispositivos BCI y las publicaciones más relevantes. También revisa artículos centrados en ciberataques basados en ruido y sus impactos.

Posteriormente, la Sección 3 se centra en los ciberataques basados en ruido que afectan a los marcos BCI, describiendo los detalles de los cuatro perfiles de ataque propuestos. La Sección 4 presenta el diseño y la implementación de un escenario realista compuesto por un caso de uso y un marco BCI. Posteriormente, la Sección 5 detalla los experimentos y el impacto de los cuatro ataques basados en ruido que afectan las fases de adquisición y procesamiento del marco BCI. Finalmente, la Sección 6 presenta algunas conclusiones y trabajos futuros.

## 2 Trabajos relacionados

Esta sección revisa el estado del arte en cuanto a problemas comunes de ciberseguridad en las BCI. Posteriormente, analiza trabajos que utilizan ciberataques basados en ruido para afectar la señal EEG adquirida.

### 2.1 Cuestiones de ciberseguridad en las BCI

En los últimos años, diversos trabajos han estudiado las implicaciones de las BCI en ciberseguridad. Sin embargo, estos estudios solo se centran en aspectos parciales, ignorando la totalidad de los problemas de ciberseguridad. Para abordar estas limitaciones, López Bernal et al. [15] analizaron el estado actual de la ciberseguridad en las BCI desde la perspectiva de la confidencialidad, la integridad y la disponibilidad de la información intercambiada. Finalmente, el estudio incluyó posibles contramedidas para los ataques analizados.

Estudios posteriores han clasificado los ciberataques según el tipo de escenario de aplicación: aplicaciones médicas, entretenimiento, autenticación y basadas en teléfonos inteligentes.

Aplicaciones. En este sentido, Li y Conti [14] detallaron que los atacantes pueden generar comandos ilícitos y lograr el mal funcionamiento de las prótesis o crear acciones incorrectas. Por otro lado, destacan la generación de patrones en la señal EEG para vulnerar los sistemas de autenticación. Rushanan et al. [25] se centraron en los problemas de ciberseguridad en la primera y la última fase del ciclo BCI (véase la Fig. 1). Los autores demostraron que la comunicación con la BCI y con

Las aplicaciones finales pueden ser capturadas o espiadas, en algunos casos incluso modificando los datos transmitidos.

Los dispositivos BCI basados en EEG han ganado popularidad en los últimos años debido a su versatilidad y bajo costo, convirtiéndolos en un objetivo atractivo para posibles ciberataques. Uno de los usos de estas tecnologías es adquirir información neuronal a partir de estímulos. En este contexto, Martinovic et al. [17] realizaron algunos experimentos para robar información crítica del sujeto, como el código PIN de 4 dígitos, información bancaria e incluso el lugar de residencia de la persona. Los autores utilizaron un BCI comercial, el auricular Emotiv EPOC, y muestrearon estímulos visuales durante 250 ms con un intervalo de 2 s entre imágenes. Lange et al. [13] ampliaron la investigación de Martinovic con la recuperación total o parcial del código PIN propuesto, agregando diferentes escenarios que vulneran la privacidad del individuo. De manera similar, Rosenfeld [24] reafirmó la preocupación por la extracción de información y presentó aplicaciones en escenarios forenses y antiterroristas.

Otros ataques, realizados por Frank et al. [6], reducen los intervalos entre estímulos visuales haciéndolos subliminales.

La literatura también ha estudiado el impacto de los ciberataques en la fase de procesamiento del ciclo BCI. La mayoría de los dispositivos BCI cuentan con un módulo de clasificación encargado de interpretar la señal adquirida. Por lo tanto, estos ataques corrompen los modelos con muestras adversarias, lo que tiene un impacto significativo en la BCI y en las acciones previstas por el usuario.

En este sentido, Zhang y Wu [29] definieron un método de señal de gradiente rápido no supervisado (UFGSM) para atacar tres redes neuronales convolucionales (CNN) populares en BCI, demostrando su eficacia. En otros casos, la densidad y la alta frecuencia de la señal EEG dificultan su procesamiento local. Juhasz [10] analizó la posibilidad de migrar clústeres locales a una infraestructura en la nube, lo que reduce significativamente el tiempo de ejecución y garantiza la seguridad de los datos.

### 2.2 Ciberataques basados en ruido

Otros trabajos en la literatura estudian ciberataques que afectan la integridad y disponibilidad de los datos transmitidos. Más específicamente, los ciberataques se han diseñado para afectar directamente la señal capturada en las fases de adquisición o procesamiento del ciclo BCI. Estas amenazas buscan ocultar segmentos de señales neuronales, principalmente asociados con eventos.

Potenciales Relacionados (ERP). En otros casos, su objetivo es incitar al atacante a generarlos deliberadamente.

Estas alteraciones de datos constituyen un problema importante en muchos escenarios de aplicación.

En los dispositivos de EEG, los problemas se agravan al adquirir una señal muy susceptible al ruido. Por lo tanto, los ciberataques utilizan la técnica de generación de ruido para afectar los datos adquiridos. En este contexto, Zhang et al. [30] implementaron un sistema de deletreo basado en EEG con P300. Los autores generan perturbaciones adversarias demasiado pequeñas para ser percibidas al añadirlas a las señales de EEG, pero que pueden inducir al sistema a deletrear cualquier palabra que el atacante desee.

Asimismo, solo consideran un escenario de caja blanca donde el atacante conoce todo sobre el modelo utilizado, ajustando los parámetros al escenario desplegado. A pesar de ser el primer trabajo que demuestra el impacto del ruido en la toma de decisiones, estos ataques se limitan a afectar al P300 y no permiten su generación en segmentos específicos de EEG. Otros estudios, como el realizado por Jiang et al. [9], consideraron ataques de caja negra basados en la transferibilidad. Para lograr este propósito, el atacante entrenó un modelo para replicar el modelo legítimo. Posteriormente, generó ejemplos adversarios empleando mecanismos de ruido dinámico con el modelo entrenado, utilizándolos para atacar el modelo legítimo. Por el contrario, Meng et al. [19] consideraron ataques de caja blanca para problemas de regresión donde se conoce toda la información sobre el algoritmo de aprendizaje. Esta suposición permite generar perturbaciones en la señal de EEG de entrada para variar el resultado en una cantidad específica. Además, los autores consideraron la transferibilidad del procedimiento a escenarios de caja negra donde se desconocen los modelos.

### 3 ciberataques basados en ruido para generar P300 falsos en marcos BCI

Esta sección presenta cuatro perfiles de ataque diferentes que utilizan ciberataques basados en ruido para afectar la detección de ondas P300 por parte del marco BCI que detecta P300. La selección de cuatro perfiles está determinada por el número de fases del ciclo BCI implementado: generación de actividad neuronal, adquisición de señal EEG, procesamiento y detección de P300. Los ciberataques propuestos apuntan a generar P300 falsos en señales EEG que previamente estaban ausentes. Este procedimiento se realiza en dos fases del ciclo BCI: adquisición y procesamiento. Sin embargo, estos no son los únicos tipos de amenazas enfocadas en violar la integridad de los datos. Existen otras modalidades de ciberataques basados en ruido en la literatura, donde en lugar de generar señales artificialmente, la amenaza causa una atenuación o eliminación de P300 en la señal EEG [30]. Estos ataques están más allá del alcance de este artículo, aunque es un buen punto de partida para trabajos futuros.

Los perfiles del estudio se ordenan de forma incremental según el conocimiento que el atacante posee sobre el marco BCI y el escenario de aplicación. Este conocimiento incremental implica que un perfil particular presenta las características y funcionalidades de los anteriores, lo que resulta en técnicas de ataque más robustas para vulnerar el marco BCI. La Figura 2 resume gráficamente las características de cada perfil de atacante, mostrando en color más oscuro los datos, procesos y antecedentes que el atacante conoce. Por lo tanto, los perfiles de atacante poseen conocimiento asociado con las cuatro fases del ciclo BCI implementadas en este trabajo.

#### 3.1 Primer perfil: el atacante conoce la existencia de una comunicación inalámbrica

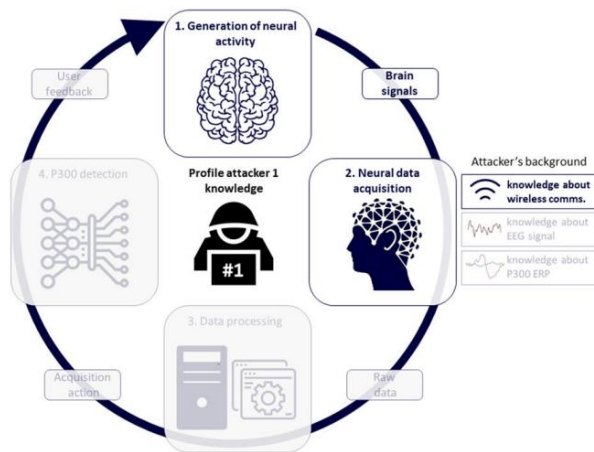
En este perfil, el atacante conoce la comunicación inalámbrica entre el auricular BCI y el marco BCI. Sin embargo, desconoce los datos intercambiados, las fases del ciclo BCI implementado por el marco BCI (detalladas en la Sección 1), el formato de los datos transmitidos por el auricular BCI ni las estructuras de almacenamiento de información implementadas por el marco. Asimismo, el atacante no posee los conocimientos necesarios para comprender las señales de EEG ni la generación de P300 para realizar un ataque preciso. La Figura 2a muestra el conocimiento del atacante sobre las fases del marco BCI y el intercambio de datos, así como su experiencia en EEG y P300.

Basándose en el supuesto anterior, el atacante genera una serie de ruidos aleatorios. Este ruido pertenece a un rango determinado por el atacante y se aplica de forma pseudoaleatoria durante la comunicación inalámbrica de datos entre los auriculares BCI y el marco BCI.

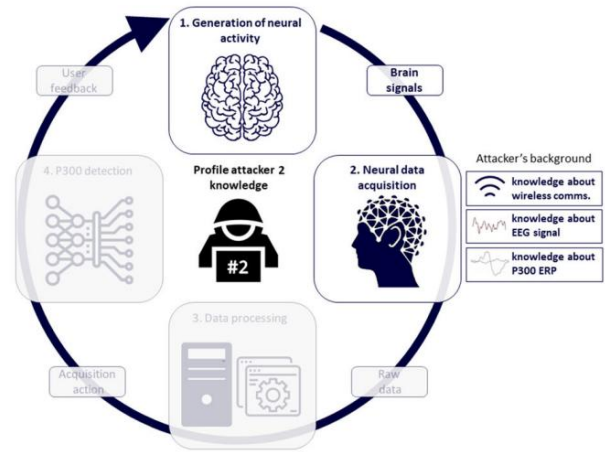
#### 3.2 Segundo perfil: el atacante tiene antecedentes con respecto a las ondas P300

Este atacante tiene cierto conocimiento del marco BCI utilizado en el escenario. En particular, conoce los mecanismos más comunes para la adquisición de señales cerebrales (fase 1 de la Fig. 2b) y las debilidades de cada uno. La debilidad del EEG reside en la alta sensibilidad al ruido externo y la necesidad de procesar los datos para obtener información relevante (véase la Sección 2.2). Asimismo, el atacante conoce la información general de P300 y las técnicas para favorecer su generación o atenuación (latencia, polaridad, amplitud o los estímulos que la desencadenan).

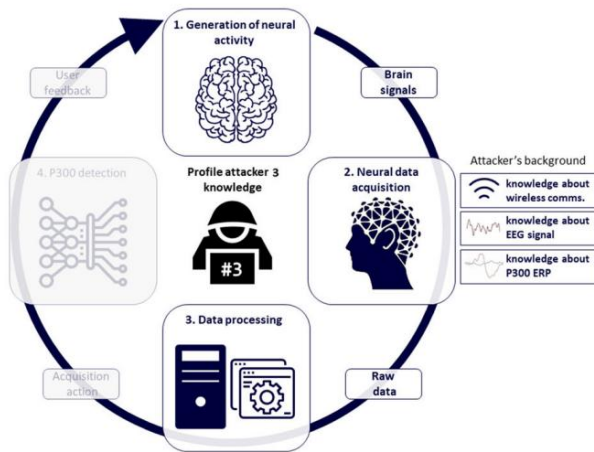
Según la información previa, el atacante genera una plantilla de ruido con una forma similar a un potencial P300 o un ruido pseudoaleatorio para dificultar la detección de un potencial P300. Los diferentes ruidos se aplican aleatoriamente a la señal EEG durante las fases de adquisición y procesamiento del marco BCI.



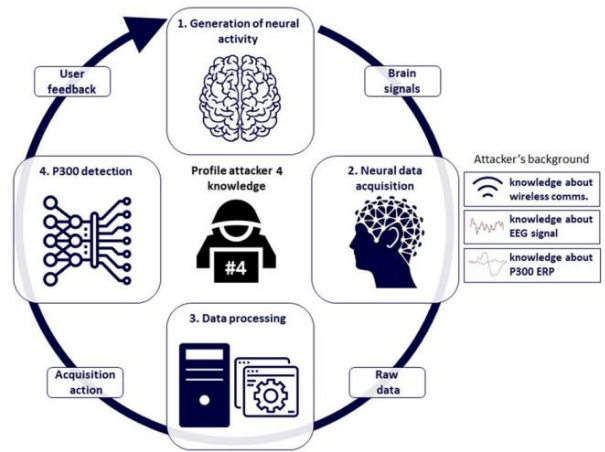
(a) First profile: the attacker knows the existence of a wireless communication



(b) Second profile: the attacker has background regarding P300 waves



(c) Third profile: the attacker has knowledge about the BCI framework and P300 waves



(d) Fourth profile: the attacker knows not only the same as the third, but also the P300 detection model details and outputs

Fig. 2 Perfiles del atacante. Cada subfigura representa un perfil de atacante diferente: los componentes en rojo describen los datos, procesos y antecedentes que conoce el atacante.

### 3.3 Tercer perfil: el atacante tiene conocimiento sobre el marco BCI y la onda P300

Este atacante conoce los detalles de las fases de adquisición y procesamiento de datos del marco BCI. Por un lado, conoce los datos transmitidos entre el auricular BCI y el marco BCI. En concreto, tiene acceso permanente al voltaje medido por cada electrodo del auricular BCI. Por lo tanto, conoce la frecuencia de muestreo del BCI y la posición de los electrodos en el cuero cabelludo. Por otro lado, conoce las técnicas de procesamiento aplicadas a los datos de la señal (más detalles en la sección 4.2). Esto significa que el ciberataque basado en ruido puede afectar las frecuencias no filtradas por los filtros paso banda (3-17 Hz en

El caso del procesamiento P300). También conoce los parámetros de rechazo basados en la amplitud pico a pico aplicada a cada electrodo. Por lo tanto, el atacante puede generar ruido con una amplitud dinámica adaptada al valor de voltaje previo. La Figura 2c muestra el conocimiento del atacante (en color más oscuro), así como los aspectos desconocidos (en color más claro).

Según los supuestos anteriores, el atacante puede generar ruido dinámico, variando sus características según los datos adquiridos por la BCI. Este atacante tiene mayor control sobre el funcionamiento de la BCI, modificando exactamente los datos que considera relevantes para el ataque. En el caso del P300, la modificación pretende afectar los datos de las diferentes épocas para generar ondas P300, lo que afecta a las aplicaciones externas que utilizan este ERP como...



Un transmisor de información neuronal. En resumen, el atacante intenta vulnerar la BCI adaptando las condiciones del ciberataque a las fases de adquisición y procesamiento del marco de la BCI.

3.4 Cuarto perfil: el atacante conoce no solo lo mismo que el tercero, sino también los detalles y resultados del modelo de detección P300

El último atacante conoce todo el marco BCI, incluyendo su implementación y los datos intercambiados en cada fase. La principal diferencia con el atacante anterior reside en el conocimiento de los detalles del módulo de clasificación basado en aprendizaje automático o profundo capaz de detectar ondas P300, detallados en la sección 4.3. Este atacante conoce la salida del modelo, por lo que puede adaptar el ciberataque en función de este valor obtenido durante la evaluación. En otras palabras, el atacante aplica ruido a la señal de EEG y, en función de la salida del modelo, adapta el ataque para evaluaciones sucesivas (véase la figura 2d).

En este caso, la generación de ruido se basa en la creación automática de plantillas según el ajuste del modelo a los datos. El uso de plantillas de ruido puede abordar algunos problemas: (1) adaptar el ruido al escenario implementado, independientemente de la funcionalidad que se esté realizando, y (2) adaptar el ruido a las condiciones externas o fisiológicas del usuario; por ejemplo, un paciente con mayor latencia en el P300 debido a ELA.

4 Configuración del escenario

Esta sección detalla el escenario implementado para obtener señales de EEG y detectar potenciales P300. El escenario se divide en tres componentes: (1) un monitor donde se presentan estímulos visuales al sujeto siguiendo el paradigma Oddball; (2) un auricular BCI no invasivo para adquirir la señal de EEG mientras el sujeto visualiza los estímulos; y (3) un sistema BCI que obtiene la señal de EEG, la sincroniza con los estímulos visuales mostrados en el monitor y procesa los datos para detectar potenciales P300.

4.1 Caso de uso

El caso de uso propuesto busca presentar estímulos visuales a un sujeto, que forman parte de un video, y generar potenciales P300. Se empleó el paradigma Oddball para activar la generación de este potencial evocado. Se seleccionó un conjunto de imágenes, donde el 20% eran familiares para el usuario (imágenes objetivo) y el resto eran desconocidas (imágenes no objetivo). El experimento comienza con 30 s de actividad EEG basal. A continuación, se aplican los estímulos visuales.

Se muestran aleatoriamente en la pantalla con un intervalo de 0,250 s (véase la Fig. 3). El experimento finaliza cuando se muestran al usuario todas las imágenes del conjunto inicial. La Tabla 1 incluye todos los parámetros utilizados en la implementación del marco y utilizados durante los experimentos.

Los experimentos se aplicaron a dos sujetos diferentes con características físicas similares. Tenían 22 y 23 años, respectivamente, ambos de aproximadamente 1,80 m de altura y no presentaban problemas cognitivos ni neurológicos. La postura mantenida durante el experimento fue perpendicular al suelo, con el monitor frente a los ojos del sujeto, evitando movimientos involuntarios y, por lo tanto, ruido adicional en la señal EEG. Además, el repositorio oficial del proyecto [16] contiene los scripts necesarios para el despliegue del escenario y las directrices para su personalización.

4.2 Adquisición y procesamiento de EEG

La fase de adquisición es el proceso mediante el cual el marco BCI obtiene la actividad neuronal generada por el cerebro del usuario. Este estudio realiza la adquisición de EEG utilizando una BCI no invasiva, el auricular EEG OpenBCI Ultracortex Mark IV [27]. Durante la monitorización, se utilizan ocho electrodos (Fp1, Fp2, C3, C4, P7, P8, O1, O2). Los electrodos se distribuyen según el sistema internacional 10-20 [23], mientras que la frecuencia de muestreo del proceso de registro es de 250 Hz. Simultáneamente, se sincronizan los estímulos visuales mostrados al usuario y la señal monitorizada. Este ajuste de tiempo es esencial para determinar la forma de onda generada en relación con la imagen objetivo mostrada.

Las señales de EEG adquiridas pueden verse alteradas por el ruido causado por algunos artefactos, como el parpadeo, los movimientos musculares, los movimientos oculares o la respiración. El ruido puede afectar el rendimiento de la BCI al sobrecargarla con datos adicionales. Por este motivo, los datos se procesan antes de continuar con el ciclo de la BCI. En primer lugar, la señal se reduce con una relación de muestreo de 5, pasando de 250 muestras por segundo (250 Hz) a 50 muestras por segundo (50 Hz). Posteriormente, se aplica un filtro Notch mediante el método de superposición-adición FIR con fase cero. Este filtro atenúa la frecuencia a 50 Hz y sus múltiplos debido al ruido causado por el cableado eléctrico del sistema BCI en Europa. Tras eliminar la frecuencia específica, los datos de EEG se filtran con un filtro paso banda con el filtro Butterworth de octavo orden.



Fig. 3 Distribución temporal de la presentación de diferentes estímulos visuales. El símbolo "T" denota una imagen objetivo, mientras que "NT" denota una imagen no objetivo.

Tabla 1 Parámetros utilizados en el experimento

Parámetro del experimento	Valor
Tamaño del monitor externo	1920 1080
Separación entre individuo y monitor	60 centímetros
Número de imágenes	180
% Imágenes de destino	20
% Imágenes no objetivo	80
Intervalo de tiempo entre imágenes	0,250 s
Tiempo de fluctuación variable	0.2
Línea base inicial	30 segundos

Rango de frecuencia entre 3 y 17 Hz para eliminar otros ruidos de alta frecuencia. El objetivo es mantener las frecuencias dentro del rango especificado y rechazar el resto.

Finalmente, se emplea el Análisis de Componentes Independientes (ICA) en el procesamiento, una potente técnica para reducir el ruido mediante la separación de fuentes independientes mezcladas linealmente en múltiples electrodos. Al final de la fase de procesamiento, la señal de EEG se divide en épocas, segmentos de EEG clasificados según la eventualidad producida. Cada época comienza 0,1 s antes del evento y 0,8 s después.

A los segmentos correspondientes a eventos objetivo se les asigna una etiqueta con el valor "2", y a los eventos no objetivo, un valor "1". La Figura 4 muestra un conjunto de épocas del escenario y los segmentos de la señal EEG correspondientes a cada electrodo. Las épocas se asocian con un identificador de evento, como se mencionó anteriormente: "1" para el objetivo y "2" para el no objetivo. Las líneas verticales verdes marcan el inicio del evento.

### 4.3 Detección P300

La última fase del marco BCI implementado busca detectar las ondas P300 en la señal de EEG capturada y procesada. Para ello, el marco BCI implementado utiliza clasificadores, que son elementos del aprendizaje supervisado que intentan predecir el resultado basándose en modelos entrenados.

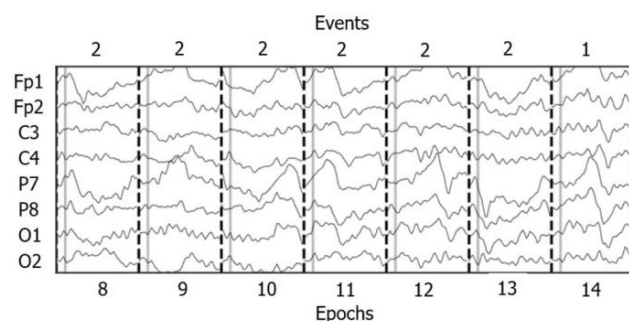


Fig. 4 Conjunto de épocas de la señal EEG marcada

El framework utiliza los siguientes clasificadores para la detección de potenciales P300: Clasificador I, que emplea algoritmos de estandarización escalar y regresiones; Clasificador II, que consiste en un modelo con un borde de decisión lineal, generado mediante el ajuste de densidades de clases condicionales a los datos y utilizando la regla de Bayes; Clasificador III, la misma operación que el clasificador II pero añadiendo xDAWN como filtro espacial; Clasificador IV, estimación de la matriz de covarianza de los posibles potenciales, proyección espacial de la tangente y regresiones; y Clasificador V, con una estimación de la matriz de covarianza y clasificación por Distancia Mínima a la Media.

Antes de entrenar los modelos de clasificación, se analiza la actividad de la señal EEG para comprobar la calidad de los datos obtenidos en la fase anterior. La Figura 5 muestra la actividad cerebral en seis instantes diferentes, utilizando el promedio de todas las eventualidades producidas durante la visualización de la imagen objetivo.

Durante el primer segundo, cuando se muestra el estímulo visual, se observa actividad cerebral en el área occipital relacionada con el procesamiento de estímulos visuales (primera representación de la Fig. 5). Sin embargo, la actividad cerebral aumenta considerablemente alrededor de los 217 ms, seguido de una disminución de esta actividad a valores negativos en un intervalo de 90 ms. Como afirma la literatura, se produce un aumento de la actividad eléctrica en el área occipital cuando se genera el P300 [26]. Asimismo, describe el P300 como una disminución de voltaje en la señal, que puede alcanzar valores negativos (segunda representación en la Fig. 5), luego aumenta el voltaje a un pico de 20–40  $\mu V$  (tercera representación) y finalmente, una ligera disminución de voltaje (cuarta representación) [22]. Finalmente, en la quinta y sexta representaciones, la actividad cerebral no muestra patrones característicos relacionados con las áreas neutrales del cerebro.

Una vez relacionada la actividad cerebral con el posible P300, los clasificadores se entrenan con cada uno de los segmentos etiquetados obtenidos en la fase anterior. Los datos se dividen manualmente en dos conjuntos diferentes: datos de entrenamiento y datos de prueba, con proporciones del 75% y el 25%, respectivamente. Se ha aplicado un proceso de validación cruzada y estratificada (debido a la naturaleza no balanceada del conjunto de datos) al conjunto de datos de entrenamiento. La estrategia implementada, StratifiedShuffleSplit, permite 10 particiones de los datos de entrada, generando diez combinaciones diferentes. Cada combinación se divide nuevamente en dos conjuntos de datos: datos de entrenamiento y datos de prueba, con las mismas proporciones que el subconjunto anterior. Mientras que los primeros se utilizan para entrenar los clasificadores, los segundos se utilizan para evaluar la precisión de las predicciones dadas. Usando la validación cruzada

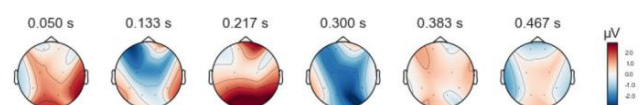


Fig. 5 Actividad cerebral con el promedio de valores capturados durante la visualización del objetivo.

La validación detecta una situación de sobreajuste cuando el modelo entrenado  
El modelo no se generaliza bien con nuevos datos de prueba.

4.4 Generación de ruido

En esta subsección se presenta el procedimiento de generación de ruido.  
utilizado por los atacantes y las consideraciones matemáticas de  
los conceptos necesarios para generarlos.  
El objetivo principal de la generación de ruido es alterar la  
Señal EEG original. Esta técnica debe utilizar ruidos de  
Diferentes relaciones señal-ruido para evaluar el rendimiento del clasificador  
P300 en diversas condiciones de ruido.  
La relación señal-ruido (SNR) se puede definir de la siguiente manera:

$$\text{Relación señal-ruido} \approx 10 \log_{10} \frac{\text{RMS}^2_{\text{señal}}}{\text{RMS}^2_{\text{ruido}}}$$

01p

donde RMSsignal es el valor de la raíz cuadrada media (RMS) de  
la señal y RMSnoise es el valor RMS del ruido.  
Se crea generación de ruido, basada en señales aleatorias.  
A través de un modelo básico de ruido llamado Ruido Gaussiano Blanco  
Aditivo (AWGN). En primer lugar, es aditivo, por lo que el ruido generado...  
Se añade ruido a la señal. En segundo lugar, el ruido tiene la  
misma distribución de potencia en cada frecuencia, siendo la potencia  
constante de densidad espectral. Finalmente, es gaussiana porque  
utiliza un modelo matemático para calcular la probabilidad de  
los eventos generados.  
El modelo AWGN agrega una variable aleatoria gaussiana de media cero.  
variable a su señal original. La varianza de esa variable aleatoria  
La variable afectará la potencia de ruido promedio. Para una variable  
aleatoria gaussiana X, la potencia promedio es  
 $E[X^2] = \sigma^2$

En la generación de ruido blanco  $\sigma^2 = 0$ , por lo que la potencia promedio es  
Entonces es igual a la varianza  $\sigma^2$ . En resumen, la implementación de  
AWGN puede realizarse de dos maneras diferentes: (1)  
Calcular la varianza en función de la relación señal-ruido  
relación señal/ruido (SNR) o (2) seleccionar una potencia de ruido específica y  
Aplicándolo a la señal EEG. En este trabajo, la segunda forma...  
Se implementa. La Tabla 2 compara los diferentes niveles de ruido.  
generaciones utilizadas en este trabajo, siendo cada tipo de ruido  
diferenciado por el nivel de ruido RMS (dB) y por el dinamismo en su  
aplicación en la señal EEG.

Tabla 2 Características del ruido  
generado

Tipo de ruido	Nivel de potencia	Nivel de ruido RMS (dB)
Gaussiana con rango estático	Bajo	0:8
Gaussiana con rango estático	Alto	5
Gaussiano con rango dinámico	Adaptado	Variación dentro del rango de 0,8 a 5

5 Resultados y discusión

En esta sección se resumen los resultados de la aplicación de ciberataques  
basados en ruido en el EEG para cada uno de los atacantes.  
perfiles definidos en la Secc. 3. Estos ciberataques afectan a dos  
diferentes fases del ciclo BCI: (1) fase de adquisición,  
donde se aplica el ruido durante la adquisición de la  
ondas cerebrales mediante electrodos colocados en el cuero cabelludo, y (2)  
fase de procesamiento, en la que el ruido se aplica una vez que  
Los datos se encuentran en el marco BCI y han sido procesados por el  
Tercera fase. La Figura 8 muestra los ciberataques realizados por  
cada perfil para el mismo segmento de señal EEG y proporciona una  
comparación visual entre las técnicas de ataque  
y el impacto resultante en la señal.  
Para realizar los ataques se deben tener en cuenta varias consideraciones  
deben tenerse en cuenta. Por un lado, la  
El ruido físico (analógico) utilizado para atacar esta fase se simula  
digitalmente en la señal adquirida. Por lo tanto, un ruido similar  
El impacto se obtiene sin utilizar equipo adicional para  
generación de ruido. Por otro lado, la aplicación de  
El ruido en la fase de procesamiento representa malware que afecta  
el marco BCI, que genera un impacto en los datos  
intercambiados entre las fases tres y cuatro del BCI  
marco. El malware se comporta de manera similar al físico  
ataque para establecer una comparación entre el atacante  
perfiles.  
Los perfiles de ataque descritos en las siguientes subsecciones  
comparten las mismas técnicas de generación de ruido descritas  
en la Sección 4.4. A pesar de generar ambos comportamientos de ruido  
(físicos y malware) con las mismas técnicas, ellos  
Varían en tiempo y forma dependiendo del atacante.  
conocimiento del marco, adaptando y enfocando la  
objetivo de generación de ruido al provocar la aparición de la  
P300, aumentando así el impacto global de la propuesta  
estructura.  
Se mide el impacto generado por cada perfil de ataque  
del marco BCI. En particular, el marco utiliza  
los clasificadores descritos en la Sección 4.3 para proporcionar  
una métrica agregada del ataque de rendimiento utilizando el Área Bajo  
La métrica Curve (AUC). Dado que el objetivo de los ataques es...  
generar ondas P300 en la señal EEG que no  
los contienen, el valor AUC se obtiene evaluando  
solo épocas no objetivo del EEG. Finalmente, una relación  
se establece entre las métricas obtenidas al afectar la  
señal legítima por el ruido y el conocimiento del atacante.



## 5.1 Señal EEG legítima

Esta sección describe la señal de EEG adquirida durante el estudio sin la perturbación causada por ciberataques basados en ruido. La Figura 6 muestra un fragmento de la señal de EEG legítima durante la fase de adquisición, más oscura (en azul), y después del procesamiento, más clara (en naranja). Más específicamente, la figura representa un segmento de 10 s de la señal de EEG capturada durante el estudio. Mientras que la señal de EEG sin procesar muestra el ruido habitual causado por algunos artefactos, la señal procesada proporciona más información al estrechar la frecuencia de las ondas cerebrales y reducir el ruido mediante técnicas de procesamiento. Además, la figura presenta el inicio de cada época con su etiqueta correspondiente según el tipo de evento producido (imagen objetivo o no objetivo).

Posteriormente, la señal de EEG procesada se introduce en los clasificadores entrenados de la fase de detección P300. La Figura 7 muestra los valores de AUC obtenidos por los cinco clasificadores utilizados en este trabajo. Como se puede observar, el marco puede clasificar aproximadamente entre el 50 % y el 80 % de las épocas no objetivo. Entre todos los clasificadores, los clasificadores I y V (con valores de AUC de 0,746 y 0,792, respectivamente) destacan como los más prometedores.

## 5.2 Perfil del primer atacante

El primer atacante genera dos tipos diferentes de ruido: (1) Ruido gaussiano con un rango estático de 0,8 dB y (2) ruido gaussiano con un rango estático de 5 dB (véase la figura rectangular en gris en la Fig. 8a). El atacante solo conoce la comunicación inalámbrica que se produce, por lo que el objetivo es alterar la señal en la fase de adquisición. La generación de los ruidos se prolonga durante toda la fase de adquisición, donde ambos ruidos se intercalan con un intervalo de 2 a 3 s.

Del mismo modo, el ataque está dirigido a todos los canales BCI, aplicando la misma cantidad e intervalo de ruido a todos ellos.

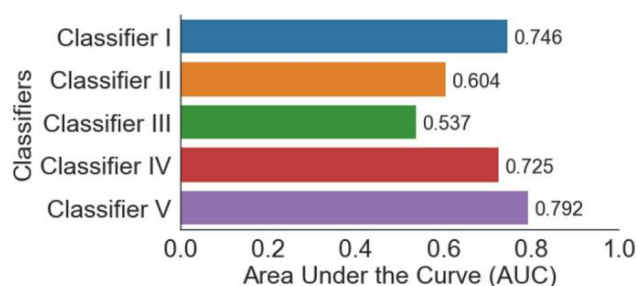


Fig. 7 Valores de AUC obtenidos por el clasificador al evaluar un no objetivo utilizando una señal de EEG legítima

Además, la figura 8a incluye marcas de cruces y marcas de verificación para indicar si el ruido aplicado es detectado como P300 por el clasificador con mejor rendimiento.

La Tabla 3 muestra los valores de AUC para cada clasificador y el comportamiento del ruido. Los resultados obtenidos en la Tabla 3 y Las siguientes tablas son el resultado de evaluar únicamente las épocas no objetivo de la señal EEG, como se explicó al principio de la sección 5. Por lo tanto, los valores de AUC determinan el impacto de los ataques para generar ondas P300 y, en consecuencia, las épocas etiquetadas como objetivo. A partir de los resultados obtenidos, se puede concluir que tanto el ruido de malware como el ruido físico obtienen una reducción similar de los valores de AUC con respecto a la señal EEG legítima. Estos resultados se deben a que ambos ruidos se aplican arbitrariamente a toda la señal EEG. La aplicación del ruido afecta a un conjunto de muestras aleatorias desconocidas para el atacante, extendiendo el ataque a toda la señal EEG adquirida sin ninguna adaptación.

Por lo tanto, el ruido no considera los parámetros de adquisición de la señal EEG, como la frecuencia de muestreo, la división de épocas o las características de la onda P300. Este procedimiento afecta tanto a las épocas objetivo como a las no objetivo, y las últimas son evaluadas por los clasificadores. Los valores de AUC obtenidos indican

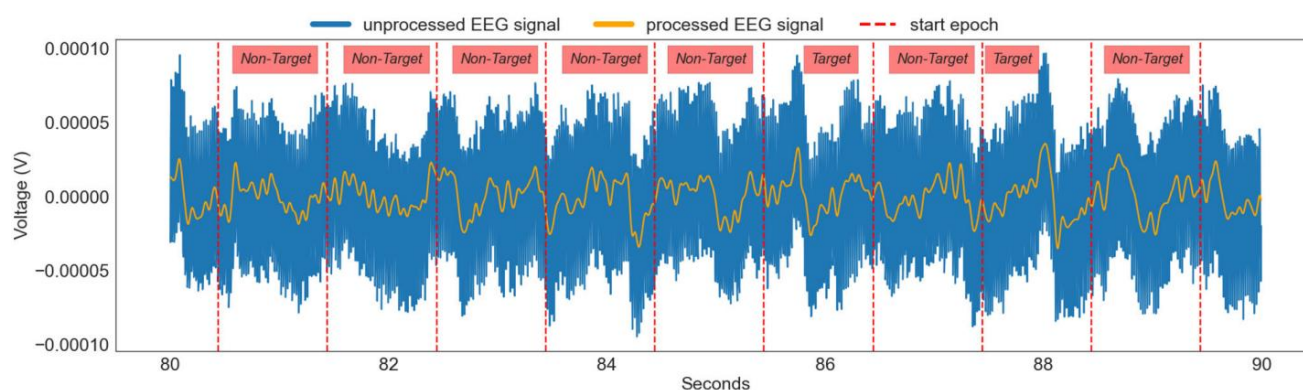
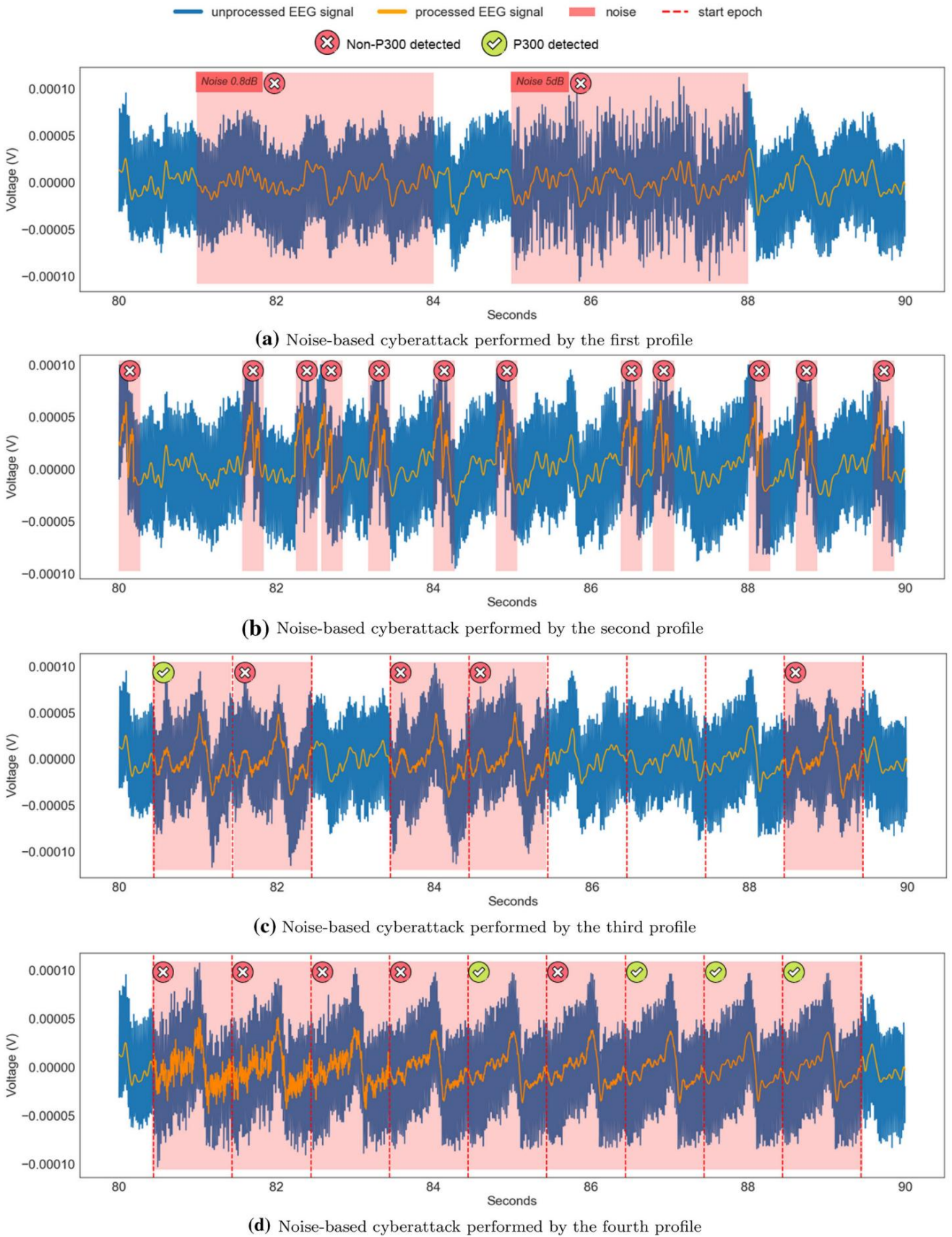


Fig. 6 Señal EEG legítima (Figura en color en línea)



b Fig. 8 Impacto de los ciberataques basados en ruido en la señal EEG dependiendo del perfil del atacante

que el ataque no fue suficiente para desencadenar la generación de P300 en la señal EEG.

5.3 Perfil del segundo atacante

El segundo atacante genera ruidos específicos en función de la Onda P300. Por lo tanto, genera ruidos siguiendo la Características y comportamiento del P300, compuesto por (1) incremento variable del ruido al inicio de la ataque en los primeros 100 a 300 ms, intentando que la señal EEG no está en valores negativos, (2) una disminución del ruido, apuntando que la señal puede alcanzar valores de voltaje negativos cercanos a cero, (3) un incremento del ruido pretendiendo que el la señal alcanza 20–40 IV y (4) una disminución de la variable ruido entre generaciones (ver figura rectangular con gris color en la Fig. 8b).

Este perfil de atacante presenta una ligera disminución de la Valores de AUC para todos los clasificadores (ver Tabla 4) en comparación con los obtenidos con la señal legítima. Esta generalizada La disminución se debe a que no se utilizan fragmentos largos con ruido y

acentuándolos en áreas específicas de la señal EEG. Aunque el ruido se genera simulando el P300 ola, el atacante no conoce la fase de adquisición variables, como la frecuencia de muestreo, el voltaje rango, o la sincronización con los estímulos.

5.4 Perfil del tercer atacante

El tercer atacante realiza un proceso similar al del perfil anterior. La principal diferencia es que ahora el atacante... genera el ruido en los segmentos de señal EEG que afectan diferentes épocas individualmente (ver Fig. 8c). Negativamente impacta a los clasificadores generando ruido más específico en aquellos segmentos de EEG en los que se produce un evento no objetivo. La figura 8c muestra un ejemplo de ello, donde un clasificador predice la inexistencia de P300 en las cinco épocas atacados, excepto el primero. El objetivo del atacante es para alterar el segmento de señal EEG en relación con el no objetivo evento y generar una onda P300. La generación de ruido implica Monitorizar los datos transmitidos y, más concretamente, la Voltaje medido por cada electrodo. Una vez que la información Se sabe que el atacante genera un ruido similar al P300 onda pero con la frecuencia y amplitud adaptadas a la Resto de la señal EEG. Además, limita la potencia del ruido.

Tabla 3 Valores de AUC por Clasificador y comportamiento del ruido en El primer perfil del atacante


Comportamiento del ruido				
		ruido físico	Ruido de malware	Señal legítima
Clasificadores	Clasificador I	0.738	0.721	0.746
	Clasificador II	0.587	0.583	0.604
	Clasificador III	0.536	0.525	0.537
	Clasificador IV	0.716	0.701	0.725
	Clasificador V	0.783	0.759	0.792

Tabla 4 Valores del AUC por Clasificador y comportamiento del ruido en El segundo perfil del atacante



Comportamiento del ruido				
		ruido físico	Ruido de malware	Señal legítima
Clasificadores	Clasificador I	0.737	0.701	0.746
	Clasificador II	0.587	0.581	0.604
	Clasificador III	0.521	0.517	0.537
	Clasificador IV	0.718	0.689	0.725
	Clasificador V	0.774	0.710	0.792

Tabla 5 Valores de AUC por Clasificador y comportamiento del ruido en El tercer perfil del atacante

Comportamiento del ruido				
		ruido físico	Ruido de malware	Señal legítima
Clasificador	Clasificador I	0.678	0.445	0.746
	Clasificador II	0.489	0.412	0.604
	Clasificador III	0.501	0.313	0.537
	Clasificador IV	0.679	0.389	0.725
	Clasificador V	0.695	0.468	0.792

nivel, adaptando la señal a los parámetros de la fase de procesamiento.

Asimismo, tanto el ruido físico como el malware...

Se generan ruidos en los electrodos O1 y O2 (colocados en la región occipital del cerebro) ya que el atacante sabe los voltajes de cada electrodo del cuero cabelludo y cuáles son involucrado en la generación del P300 a través de la visualización estímulos.

La Tabla 5 muestra los valores de AUC obtenidos en cada ruido.

Ciberataque basado en la nube. En este perfil de atacante, se observan cifras sustanciales Se observan cambios en los valores de AUC obtenidos En la señal legítima, en particular el ruido de malware. A diferencia de los perfiles anteriores, el conocimiento de la señal EEG... procesamiento y sincronización respecto a lo mostrado Los estímulos generan un entorno beneficioso para el atacante. De igual forma, el ruido se genera sin sobrepasar el rechazo de artefactos del marco, evitando así una tensión reducción en la onda P300 generada. Esta conclusión tiene Su representación en (1) una disminución del 6–12% en los valores del AUC con ruido físico y (2) una disminución del 35-40% con malware ruido, tanto en relación con la señal legítima.

5.5 Perfil del cuarto atacante

El cuarto perfil se centra en atacar la adquisición y fases de procesamiento del ciclo BCI pero que tienen información sobre el clasificador y sus predicciones. Este tipo de ataque tiene similitudes con los ataques adversarios a las máquinas y Aprendizaje profundo en la literatura, como el método FGSM (Fast Gradient Signed Method). La diferencia con el FGSM es que este último necesita calcular los gradientes utilizando modificaciones del  $\epsilon$ , mientras que el ataque propuesto utiliza el Características del P300 para aplicarlo en forma de ruido adaptativo a La señal EEG. El objetivo es modificar la señal EEG, maximizando la probabilidad de que los clasificadores predigan P300. La Figura 8d muestra la adaptación del ruido a lo largo de la señal EEG. Si bien los clasificadores predicen inicialmente la modificación de las épocas como No-P300, la adaptación continua al ruido conduce a la generación del P300 en la señal EEG (quinta época en la figura). La adaptación al ruido es continua hasta Todas las épocas se clasifican como P300 (séptima época en adelante). En resumen, el atacante utiliza la retroalimentación recibida por el clasificadores para refinar el ruido generado, disminuyendo la impacto del ataque y potenciar la generación de la P300.

Tabla 6 Valores de AUC por Clasificador y comportamiento del ruido en El cuarto perfil del atacante



Comportamiento del ruido				
		ruido físico	Ruido de malware	Señal legítima
Clasificadores	Clasificador I	0.603	0.201	0.746
	Clasificador II	0.441	0.112	0.604
	Clasificador III	0.467	0.104	0.537
	Clasificador IV	0.621	0.155	0.725
	Clasificador V	0.618	0.212	0.792

Tabla 7 Valores AUC del clasificador V según perfil de atacante y ruido comportamiento

Comportamiento del ruido			
			
		Ruido físico	Ruido de malware
Perfiles de atacantes	Perfil I	0.783	0.759
	Perfil II	0.774	0.710
	Perfil III	0.695	0.468
	Perfil IV 0.618		0.212

Los valores de AUC obtenidos en este perfil (véase la Tabla 6) son los más bajos del estudio, incluyendo los obtenidos por los perfiles anteriores y la señal legítima. Estos valores abarcan todos los vectores de ataque realizados, incluyendo el periodo de aprendizaje para generar el ruido ideal que favorezca la generación del P300. Este enfoque resulta en valores de AUC mayores que cero. Cuando el clasificador detecta principalmente P300 con una plantilla de ruido específica, se aplica en ataques sucesivos. Por lo tanto, el clasificador siempre es engañado, lo que significa que el ataque puede generar un P300 donde no lo había la mayoría de las veces. Mientras que los valores de AUC...

Los obtenidos con el ruido físico han disminuido en un 20-28% respecto a la señal legítima, los del ruido de malware han disminuido en una proporción más significativa, obteniendo valores entre 0,104 y 0,212.

Finalmente, la Tabla 7 compara los valores de AUC obtenidos por el Clasificador V, el que presenta el mejor rendimiento de predicción, con la señal legítima (véase la Fig. 7), según el comportamiento del ruido y el perfil del atacante. Por un lado, los valores muestran una ligera disminución progresiva del AUC con el ruido físico en los diferentes perfiles de ataque.

La disminución se sitúa entre el 1 % y el 22 % con respecto a la señal legítima, siendo del 1 % para el primer perfil y del 22 % para el cuarto. Por otro lado, los valores de AUC del ruido de malware presentan una disminución significativa entre el segundo y el tercer perfil, con una reducción del 34 %, y entre el tercer y el cuarto perfil, del 55 %. De igual forma, la aplicación de ruido de malware en el cuarto perfil tiene un impacto del 74 % en comparación con el obtenido en la señal legítima. Por lo tanto, la aplicación de ruido de malware en la fase de procesamiento por el cuarto perfil de atacante tiene el mayor impacto en el AUC, lo que se traduce en una alta generación de potenciales P300 en la señal EEG.

## 6 Conclusión

Este trabajo presenta cuatro perfiles de atacantes incrementales que generan ciberataques basados en ruido que afectan a los marcos BCI inteligentes que detectan ondas P300. El primer perfil conoce la comunicación inalámbrica entre los auriculares BCI y el marco BCI. El segundo tiene información sobre...

Ondas P300. El tercero conoce el marco BCI y el cuarto también conoce los detalles del modelo de detección P300.

y salidas. Para cada perfil, se consideran dos tipos de ruido: (1) físico, que afecta la fase de adquisición de la señal EEG de los marcos BCI, y (2) basado en malware, que impacta la fase de procesamiento. Para medir el impacto de los ataques, hemos implementado un escenario realista para la adquisición de señales EEG compuesto por (1) un video que muestra estímulos visuales conocidos y desconocidos, (2) un auricular BCI no invasivo, y (3) un marco BCI que implementa las fases de adquisición, procesamiento y detección de P300 del ciclo de vida de la BCI. Los experimentos realizados han demostrado que un mayor conocimiento sobre el ciclo BCI permite a un atacante realizar ataques más sofisticados para generar ondas P300. Asimismo, hemos observado que los ataques que afectan la fase de procesamiento tienen un impacto más significativo en la generación de P300. En particular, la puntuación AUC del mejor clasificador que detecta P300 se reduce en un 1%, 3%, 12% y 22% cuando se ataca la fase de adquisición, y en un 4%, 10%, 41% y 74% cuando la fase de procesamiento de datos se ve afectada por cada uno de los cuatro perfiles, respectivamente.

Como trabajo futuro, planeamos estudiar el impacto de nuevas técnicas y objetivos de aplicación de ruido, creando nuevos vectores de ataque. Asimismo, la materialización de diferentes vectores de ataque puede dar lugar a nuevos perfiles de atacantes con un impacto diferente al descrito en este trabajo. Una de las líneas futuras podría profundizar en un perfil de atacante centrado en el hardware BCI, a un nivel de abstracción más bajo o desde la perspectiva de la estimulación cerebral. Sería interesante comparar los perfiles y determinar el impacto o la criticidad originada en futuras investigaciones. Del mismo modo, consideramos un estudio con un número mayor de muestras de señales de EEG etiquetadas para el entrenamiento de clasificadores, así como algoritmos más sofisticados para detectar el P300. Finalmente, proponemos utilizar un mayor número de electrodos en la monitorización de señales de EEG, ya que la interpolación podría reducir el impacto de estos ciberataques.

Contribuciones de los autores Conceptualización: ETMB y MQP; Metodología: ETMB; Software: ETMB; Validación: ETMB, SLB y AHC; Análisis formal: ETMB y SLB; Investigación: ETMB y MQP; Recursos: AHC y GMP; Curación de datos: ETMB y MQP; Redacción—preparación del borrador original: ETMB; Redacción—revisión y edición: ETMB, MQP, AHC, SLB y GMP; Visualización: ETMB y MQP; Supervisión: AHC y GMP; Proyecto



Administración: BPM. Todos los autores han leído y aceptado la versión publicada del manuscrito.

Financiación: Financiación de acceso abierto proporcionada gracias al acuerdo CRUE-CSIC con Springer Nature. Este trabajo ha sido financiado parcialmente por (a) Bit and Brain Technologies SL en el marco del proyecto CyberBrain: Ciberseguridad en BCI para Asistencia Avanzada a la Conducción, asociado a la Universidad de Murcia (España), (b) la Oficina Federal Suiza para la Contratación de Defensa (armasuisse) con el proyecto CyberSpec (CYD-C-2020003), y (c) la Universidad de Zúrich (UZH).

Disponibilidad de datos No aplicable.

Disponibilidad del código Todo el código utilizado en este trabajo está disponible públicamente en [16].

## Declaraciones

Conflicto de intereses No aplicable.

Aprobación ética No aplicable.

Acceso abierto Este artículo está licenciado bajo una Licencia Creative Commons Atribución 4.0 Internacional, que permite el uso, intercambio, adaptación, distribución y reproducción en cualquier medio o formato, siempre y cuando se otorgue el crédito correspondiente al autor original y a la fuente, se proporcione un enlace a la licencia Creative Commons y se indique si se realizaron cambios. Las imágenes u otro material de terceros en este artículo están incluidos en la licencia Creative Commons del artículo, a menos que se indique lo contrario en una línea de crédito al material. Si el material no está incluido en la licencia Creative Commons del artículo y el uso que pretende darle no está permitido por la regulación legal o excede el uso permitido, deberá obtener permiso directamente del titular de los derechos de autor. Para ver una copia de esta licencia, visite <http://creativecommons.org/licenses/by/4.0/>.

## Referencias

- Ahn, M., Lee, M., Choi, J., Jun, S.: Una revisión de juegos con interfaz cerebro-computadora y una encuesta de opinión a investigadores, desarrolladores y usuarios. *Sensors* (Basel, Suiza) 14, 14601–14633 (2014). <https://doi.org/10.3390/s140814601>
- Al-Nuaimi, FA, Al-Nuaimi, RJ, Al-Dhaheri, SS, Ouhbi, S., Belkacem, AN: Persecución de drones mentales mediante una interfaz cerebro-computadora basada en EEG. En: 16.ª Conferencia Internacional sobre Entornos Inteligentes (IE), 2020, págs. 74–79 (2020). <https://doi.org/10.1109/IE49459.2020.9154926>
- Asociación de Fisiatras Académicos: Controlar una prótesis con el cerebro. Asociación de Fisiatras Académicos (2017). Recuperado el 27 de enero de 2021 de <https://www.science.dailymail.com/releases/2017/02/170206084904.htm>
- Birbaumer, N., Hochberg, LR: Una comunicación útil en las interfaces cerebro-computadora. *Neurología* 91(3), 109–110 (2018). <https://doi.org/10.1212/WNL.0000000000005804>
- Crea, S., Nann, M., Trigili, E., Cordella, F., Baldoni, A., Badesa, F., Catalán, J.M., Zollo, L., Vitiello, N., Aracil, N., Soekadar, S.: Viabilidad y seguridad del control autónomo de exoesqueleto de brazo completo guiado por visión y EEG/EOG compartido para realizar actividades de la vida diaria. *Sci. Rep.* (2018). <https://doi.org/10.1038/s41598-018-29091-5>
- Frank, M., Hwu, T., Jain, S., Knight, RT, Martinovic, I., Mittal, P., Perito, D., Sluganovic, I., Song, D.: (2017) Uso de EEG basado Dispositivos BCI para sondear subliminalmente información privada. En: *Actas del Taller sobre Privacidad en la Sociedad Electrónica de 2017*, WPES '17, págs. 133–136. Asociación para la Maquinaria de Computación, Nueva York (2017). <https://doi.org/10.1145/3139550.3139559>
- Friganovic, K., Medved, M., Cifrek, M.: Interfaz cerebro-computadora basada en potenciales evocados visuales de estado estacionario. En: 39.ª Convención Internacional sobre Información y Comunicación de 2016 *Tecnología*, Electrónica y Microelectrónica (MIPRO), págs. 391–395 (2016). <https://doi.org/10.1109/MIPRO.2016.7522174>
- Jang, YS, Ryu, SA, Park, KC: Análisis de la elección de objetivo relacionada con P300 en el paradigma Oddball. *J. Inf. Commun. Converg. Eng.* (2011). <https://doi.org/10.6109/jicce.2011.9.2.125>
- Jiang, X., Zhang, X., Wu, D.: Aprendizaje activo para ataques adversarios de caja negra en interfaces cerebro-computadora basadas en EEG. En: *Serie de Simposios IEEE sobre Inteligencia Computacional (SSCI) de 2019*, pp. 361–368 (2019). <https://doi.org/10.1109/SSCI44817.2019.9002719>
- Juhász, Z.: Comparación cuantitativa de costos del procesamiento de datos de EEG en infraestructura local y en la nube. *Clust. Comput.* (2020). <https://doi.org/10.1007/s10586-020-03141-y>
- Kanna, RK, Vasuki, R.: Aplicaciones avanzadas de BCI para la detección del estado de somnolencia mediante ondas de EEG. *Mater. Today Proc.* (2021). <https://doi.org/10.1016/j.matpr.2021.01.784>
- Kumar, STS, Kasthuri, N.: Clasificación de convulsiones EEG basada en la explotación de la reconstrucción del espacio de fase y el aprendizaje extremo. *Clúster. Computadora*. 22(5), 11477–11487 (2019). <https://doi.org/10.1007/s10586-017-1409-z>
- Lange, J., Massart, C., Mouraux, A., Standaert, FX: Ataques de canal lateral contra el cerebro humano: el caso del código PIN (versión extendida). *Brain Inform.* 5, 12 (2018). <https://doi.org/10.1186/s40708-018-0090-1>
- Li, QQ, Ding, D., Conti, M.: Aplicaciones de interfaz cerebro-computadora: desafíos de seguridad y privacidad. En: *Conferencia IEEE sobre Comunicaciones y Seguridad de Redes (CNS) de 2015*, pp. 663–666 (2015). <https://doi.org/10.1109/CNS.2015.7346884>
- Lo'pez Bernal, S., Huertas Celdra'n, A., Martí'nez Pe'rez, G., Bar-ros, MT, Balasubramaniam, S.: Seguridad en interfaces cerebro-computadora: estado del arte, oportunidades y desafíos futuros. *ACM Comput. Surv.* (2021). <https://doi.org/10.1145/3427376>
- Martí'nez Beltrán, ET: *enriquetomasmb/bci* (2021). Recuperado el 21 de febrero de 2021 de <https://github.com/enriquetomasmb/bci>
- Martinovic, I., Davies, D., Frank, M., Perito, D., Ros, T., Song, D.: Sobre la viabilidad de ataques de canal lateral con interfaces cerebro-computadora. En: 21.º Simposio de Seguridad USENIX (USENIX Security 12), págs. 143–158. Asociación USENIX, Bellevue (2012). Recuperado el 15 de enero de 2021 de <https://www.usenix.org/conference/usenixsecurity12/technical-sessions/presentation/martinovic>
- McFarland, D., Wolpaw, J.: Interfaces cerebro-computadora basadas en EEG. *Curr. Opin. Biomed. Eng.* 4, 194–200 (2017). <https://doi.org/10.1016/j.cobme.2017.11.004>
- Meng, L., Lin, C., Jung, T., Wu, D.: Ataque de objetivo de caja blanca para problemas de regresión BCI basados en EEG. En: Gedeon, T., Wong, KW, Lee, M. (eds.) *Procesamiento de la Información Neural — 26.ª Conferencia Internacional, ICONIP 2019, Sidney, NSW, Australia, 12–15 de diciembre de 2019*, Actas, Parte I, Lecture Notes in Computer Science, vol. 11953, pp. 476–488. Springer (2019). [https://doi.org/10.1007/978-3-030-36708-4\\_39](https://doi.org/10.1007/978-3-030-36708-4_39)
- Monaco, A., Sforza, G., Amoroso, N., Antonacci, M., Bellotti, R., de Tommaso, M., Di Bitonto, P., Di Sciascio, E., Diacono, D., Gentile, E., Montemurno, A., Ruta, M., Ulloa, A., Tangaro, S.: El proyecto PERSON: un juego serio de interfaz cerebro-computadora para el tratamiento del deterioro cognitivo. *Tecnología de la salud*. 9(2), 123–133 (2019). <https://doi.org/10.1007/s12553-018-0258-y>

21. Peña, A., Arango, J., Mazo, J.: Sistema para rehabilitación del síndrome del miembro fantasma utilizando interfaz cerebro-computador y realidad aumentada. *Rev. Iber. Hermana. Tecnol. inf.* (2013). <https://doi.org/10.4304/risti.11.93-106>
22. Picton, T.: La onda P300 del potencial humano relacionado con eventos. *J. Clin. Neurophysiol. Off. Publ. Am. Electroencephalogr. Soc.* 9, 456–79 (1992). <https://doi.org/10.1097/00004691-199210000-00002>
23. Rojas, G., Alvarez, C., Montoya, C., de la Iglesia-Vaya, M., Cisternas, J., Ga'lvez, M.: Estudio de redes de conectividad funcional en estado de reposo utilizando la posición de electrodos de EEG como semilla. *Portada. Neurociencia* (2018). <https://doi.org/10.3389/fnins.2018.00235> 24.
- Rosenfeld, JP: P300 en la detección de información oculta y engaño: una revisión. *Psicofisiología* 57(7), e13362 (2020). <https://doi.org/10.1111/psyp.13362>
25. Rushanan, M., Rubin, AD, Kune, DF, Swanson, CM: SoK: seguridad y privacidad en dispositivos médicos implantables y redes de área corporal. En: Simposio IEEE sobre Seguridad y Privacidad de 2014, pp. 524-539 (2014). <https://doi.org/10.1109/SP.2014.40> 26. Takano, K., Ora, H., Sekihara, K., Iwaki, S., Kansaku, K.: Actividad coherente en las cortezas parietooccipital bilateral durante la operación P300-BCI. *Front. Neurol.* 5, 74 (2014). <https://doi.org/10.3389/fneur.2014.0007427>. La GUI de OpenBCI: Documentación de OpenBCI (2021). Recuperado el 19 de febrero de 2021 de <https://docs.openbci.com/docs/06Software/01-OpenBCISoftware/GUIDocs>
28. Vinothraj, T., Alfred, DD, Amarakeerthi, S., Ekanayake, J.: Detección de pacientes alcohólicos basada en BCI. En: 17.º Congreso Mundial Conjunto de la Asociación Internacional de Sistemas Difusos y 9.ª Conferencia Internacional sobre Computación Suave y Sistemas Inteligentes (IFSAS-SCIS), 2017, págs. 1-6 (2017). <https://doi.org/10.1109/IFSAS-SCIS.2017.8305564> 29. Zhang, X., Wu, D.: Sobre la vulnerabilidad de los clasificadores
- CNN en BCI basados en EEG. *IEEE Trans. Neural Syst. Rehabil. Eng.* (2019). <https://doi.org/10.1109/TNSRE.2019.2908955>
30. Zhang, X., Wu, D., Ding, L., Luo, H., Lin, CT, Jung, TP, Chavarriaga, R.: Ruido pequeño, grandes errores: perturbaciones adversariales inducen errores en los delectreadores de la interfaz cerebro-computadora. *Natl Sci. Rev.* (2020). <https://doi.org/10.1093/nsr/nwaa233>

Nota del editor: Springer Nature permanece neutral con respecto a los reclamos jurisdiccionales en los mapas publicados y las afiliaciones institucionales.



Enrique Tomás Martínez Beltrán es estudiante de Máster en Nuevas Tecnologías en la Universidad de Murcia, especializado en redes y telemática. Actualmente trabaja en su Proyecto Fin de Máster sobre ciberseguridad e Interfaces Cerebro-Computadora. Simultáneamente, investiga la automatización de ataques y defensas en diferentes escenarios con el equipo de CyberData-Lab. Sus intereses incluyen las tecnologías de ciberseguridad.

y nuevo



Mario Quiles Pérez, estudiante de Ingeniería Informática en la Universidad de Murcia (UMU). Mención en Tecnologías de la Información y la Comunicación.

Gies. Actualmente, en su último año académico, investiga los posibles ataques a las interfaces cerebro-máquina. Interesado e iniciado en ciberseguridad, desarrolla escenarios de entrenamiento para la detección de ataques adversarios en redes internas. En el futuro, espera dedicarse de lleno al mundo de la ciberseguridad.



Sergio López Bernal es licenciado y máster en Informática por la Universidad de Murcia, y máster en Arquitectura e Ingeniería para el IoT por IMT Atlantique, Francia.

Actualmente cursa el doctorado en la Universidad de Murcia. Sus líneas de investigación incluyen la seguridad de las TIC en interfaces cerebro-computadora y la seguridad de redes y de la información.



Alberto Huertas Celdra'n recibió el título de M.Sc. y Ph.D.

Licenciado en Informática por la Universidad de Murcia, España. Actualmente es investigador postdoctoral asociado al Grupo de Sistemas de Comunicación (CSG) de la Universidad de Zúrich (UZH). Sus intereses científicos incluyen los sistemas ciberfísicos médicos (MCPS), las interfaces cerebro-computadora (BCI), la ciberseguridad, la privacidad de datos, la autenticación continua, la tecnología semántica y el análisis contextual.

sistemas conscientes y redes de computadoras.



Gregorio Martínez Per'ez Catedrático de Universidad del Departamento de Ingeniería de la Información y las Comunicaciones de la Universidad de Murcia, España.

Su actividad científica se centra principalmente en la ciberseguridad y las redes, trabajando también en el diseño y la monitorización autónoma de aplicaciones y sistemas críticos en tiempo real. Trabaja en diversos proyectos de investigación IST nacionales (14 en la última década) y europeos (11 en la última década) relacionados con estos temas, siendo

investigador principal en la mayoría de ellos.

Ha publicado 160 artículos en actas de congresos, revistas y periódicos nacionales e internacionales.