INSTITUTO TECNOLÓGICO DE BUENOS AIRES – ITBA

ESCUELA DE INGENIERÍA Y GESTIÓN

Análisis del rendimiento en el algoritmo de machine learning que emula ondas ERP P300 usado en un experimento con interfaces cerebro computador en pacientes con ELA (Esclerosis Lateral Amiotrófica).

AUTOR: Chavez Montaño, Alexander. (Leg. Nº 506218)

TUTOR: Ramele, Rodrigo

Trabajo final presentado para la obtención del título de especialista en ciencia de datos.

BUENOS AIRES

Segundo cuatrimestre, 2023

A mis dos emes: Macu y Marco.

Tabla de contenido

Abstract

Este trabajo complementa la investigación previa realizada en los experimentos descritos en el artículo *EEG Waveform Analysis of P300 ERP with Applications to Brain Computer Interfaces* [3] en pacientes con ELA (Esclerosis Lateral Amiotrófica): parte del resultado de dicho trabajo fue la de investigar algoritmos de machine learning que emulan ondas ERP P300: tomamos ese trabajo como punto de partida para realizar un análisis exploratorio tanto del objeto de estudio como de las herramientas computacionales disponibles, para luego realizar un abanico de pruebas que arrojen distinta *performance* y nos permitan proponer mejoras en la preconfiguración de dicho algoritmo.

1. Introducción

La electroencefalografía es una de las herramientas clínicas que, a lo largo de las últimas décadas, se ha convertido en uno de los principales métodos para obtener imágenes en tiempo real del comportamiento cerebral de manera no invasiva, portátil y móvil más usado en el ambiente médico [1]. Dentro de la electroencefalografía tenemos un conjunto de ondas con distintas características que varían en sus propiedades físicas como amplitud o frecuencia, como también en el origen y la ubicación en las distintas zonas del cerebro. La onda P300 se obtiene de ubicar un canal en el lóbulo parietal y su comportamiento es reactivo debido a estímulos esperados pero infrecuentes relacionados con actos cognitivos.

La electroencefalografía , sin embargo, está expuesta a alteraciones no deseadas en sus resultados, ya que, por más controlado, preciso y consistente que sea el ambiente donde se realiza el experimento o la toma de muestra, estaremos sujetos a factores fuera de nuestro control. Esta problemática se suele afrontar generando ambientes de pruebas donde se pueda recrear la situación con la mayor fidelidad posible. Los experimentos pasados y éste trabajo integrador se cimentan en la base de datasets sintéticos, artificiales, con los que se simulan respuestas de ondas ERP P300 a partir de electroencefalogramas reales, con resultados previamente conocidos, a fin de trabajar en la performance de algoritmos que logren resultados con mejoras en el tiempo.

Los métodos y los procedimientos cuantitativos para automatizar la decodificación de ondas EEG como la P300 se basa en EEG no invasivo [2]. Sin embargo, los métodos de la decodificación de señales, basadas en detección de formas de onda, y además con algoritmos de machine learning son relativamente escasos.

Se pretende entonces, a través de este trabajo, darle continuidad a la investigación previa realizada en los experimentos descritos en el artículo EEG Waveform Analysis of P300 ERP with Applications to Brain Computer Interfaces [3] en pacientes con ELA (Esclerosis Lateral Amiotrófica): en éste trabajo se realiza un análisis exploratorio de electroencefalogramas llamados pasivos; pacientes que participaron del experimento pero desconociendo las reglas de interacción con el *speller*, explicado más adelante en profundidad. En una etapa posterior se “inyectan” potenciales P300 en los lugares donde sabemos de antemano suceden los eventos, y realizamos modificaciones en las propiedades de las ondas que arrojan distinta *performance* para permitirnos obtener mejoras en la preconfiguración del algoritmo.

2. Estado del arte y Marco conceptual

DIRECTAMENTE VOY A INCLUIR EL EXPERIMENTO.

El estado del arte es el estado de la cuestión, dónde está parada la investigación hasta ése momento.

La electroencefalografía es una de las herramientas clínicas que, a lo largo de las últimas décadas, se ha convertido en uno de los principales métodos para obtener imágenes en tiempo real del comportamiento cerebral de manera no invasiva, portátil y móvil más usado en el ambiente médico ***[8]****. Sin embargo, está expuesta a alteraciones no deseadas en sus resultados, ya que, por más controlado, preciso y consistente que sea el ambiente donde se realiza el experimento o la toma de muestra, estaremos sujetos a que el objeto de estudio, en este caso es el ser humano, incurrirá en desconcentración o desenfoque al momento de hacer las pruebas y esto modificar la respuesta esperada.*

Dentro de la electroencefalografía tenemos un conjunto de ondas con distintas características que varían en sus propiedades físicas como amplitud o frecuencia, como también en el origen y la ubicación en las distintas zonas del cerebro. La onda P300 se obtiene de ubicar un canal en el lóbulo parietal y su comportamiento es reactivo debido a estímulos esperados pero infrecuentes relacionados con actos cognitivos.

Los métodos y los procedimientos cuantitativos para automatizar la decodificación de ondas EEG como la P300 se basa en EEG no invasivo ***[2]****. Sin embargo, los métodos de la decodificación de señales, basadas en detección de formas de onda, y además con algoritmos de machine learning, es relativamente escaso.*

Son ocho participantes sanos: saludables, visión normal o corregida en la normalidad, sin antecedentes de trastornos neurológicos, entre 20 y 40 años de edad, y los datos de EEG se recopilan en una sola sesión de grabación. Cada sujeto está sentado en una silla cómoda, con su vista alineada con una pantalla de computadora ubicada a un metro frente a él/ella. El manejo y procesamiento de los datos y estímulos se realiza mediante la plataforma OpenVibe.

Marco conceptual
: define con precisión los conceptos centrales
del dominio del problema (Palabras Clave).
Marco teórico
: explica las teorías sobre el
Estado del Arte
: sintetiza el estado actual que se encuentre en
otras investigaciones o situaciones similares a la estudiada.
Antecedentes
: describe aquellos antecedentes que contemplen
diversas formas de resolución de ese problema , que se hayan
construido en forma previa a la investigación.
---

3. Definición del problema

La continuidad de los proyectos de investigación es vital para obtener avances y mejoras en los resultados de los experimentos. Particularmente, las investigaciones en el tratamiento de señales electroencefalográficas con modelos de machine learning son un campo de estudio relativamente nuevo y con poca disponibilidad de datos, lo que genera obstáculos que impiden realizar experimentos comparativos a gran escala. En el caso de los experimentos descritos en el artículo *EEG Waveform Analysis of P300 ERP with Applications to Brain Computer Interfaces* [3] en pacientes con ELA (Esclerosis Lateral Amiotrófica), su continuidad se encontraba pausada por razones ajenas a este documento. Es posible que, una vez reanudada estas investigaciones, se puedan dar saltos posteriores con experimentos comparativos a gran escala.

Dentro del artículo descrito, se usó un algoritmo de machine learning que ensambla electroencefalogramas de pacientes pasivos; pacientes que participaron del experimento desconociendo las reglas de interacción con los equipos, con potenciales P300 en los lugares donde sabemos de antemano que suceden los eventos evocados. Es necesario analizar y testear dicho algoritmo para que la investigación continúe.

El EDA (Análisis Exploratorio de Datos, por sus siglas en inglés) y modificaciones en las propiedades de las ondas obtenidas de ese ensamble nos permitirán ampliar el abanico de resultados que arrojan distinta *performance* en los resultados. Esto permitirá proponer mejoras en la preconfiguración del algoritmo.

4. Justificación del estudio

Se pretende darle continuidad al proyecto de investigación basados en los experimentos descritos en el artículo *EEG Waveform Analysis of P300 ERP with Applications to Brain Computer Interfaces* [3] en pacientes con ELA (Esclerosis Lateral Amiotrófica).

Este trabajo final integrador permitirá ampliar la gama de resultados del algoritmo mencionado en este documento. La salida del algoritmo es una onda compuesta por un electroencefalograma y potenciales ERP. Se pretende, modificando las propiedades de dicha onda, ofrecer una gama de resultados que sirvan de complemento para evaluar y proponer mejoras en la *performance* del algoritmo. A priori, y debido a las particularidades del experimento, las nuevas ondas generadas tendrán efectos distintos. No se espera verificar si el comportamiento final sigue siendo el mismo.

¿Porqué se lleva a cabo esta investigación?

¿Cómo se ubica en relación con las demás? ¿Qué blancos pretende llenar o qué correcciones intenta hacer a los aportes de otros autores? ¿Busca acaso modificar las condiciones en que se produce un hecho ya estudiado para observar si se repite de la misma manera?

---

5. Alcance del trabajo y limitaciones

Los experimentos realizados en el artículo *EEG Waveform Analysis of P300 ERP with Applications to Brain Computer Interfaces* [3] en pacientes con ELA (Esclerosis Lateral Amiotrófica) permiten un sinnúmero de estudios y experimentos posteriores. El presente trabajo está enfocado únicamente al testeo del algoritmo destinado a generar una onda compuesta de manera sintética. Se espera que, con los resultados obtenidos al modificar deliberadamente las propiedades de la onda, podamos tener un espectro de respuestas que nos permitan porponer mejoras para trabajos posteriores.

6. Hipótesis

NO LLEVO NADA

---

7. Objetivos

General.

Darle continuidad a los experimentos e investigaciones previas realizadas en el Instituto Tecnológico de Buenos Aires ITBA, analizando y obteniendo resultados del rendimiento en el algoritmo de machine learning que emula ondas ERP P300 usado en un experimento con interfaces cerebro computador en pacientes con ELA (Esclerosis Lateral Amiotrófica).

Específicos.

- Realizar un recorrido en la mayor cantidad de información disponible sobre los objetos de estudio: por un lado, el potencial de eventos evocado P300. Por otro, electroencefalogramas de pacientes sanos y pacientes con ELA (Esclerosis Lateral Amiotrófica), y por último todo el conjunto de señales que se obtienen producto de una interfaz cerebro computador (BCI por sus siglas en inglés) vinculado a un instrumento de comunicación llamado P300 speller.

- Completar un análisis exploratorio de datos (EDA por sus siglas en inglés) con los datasets usados en los experimentos, a través de las herramientas computacionales disponibles: librerías de python especializadas en el manejo de datos y electroencefalografía.

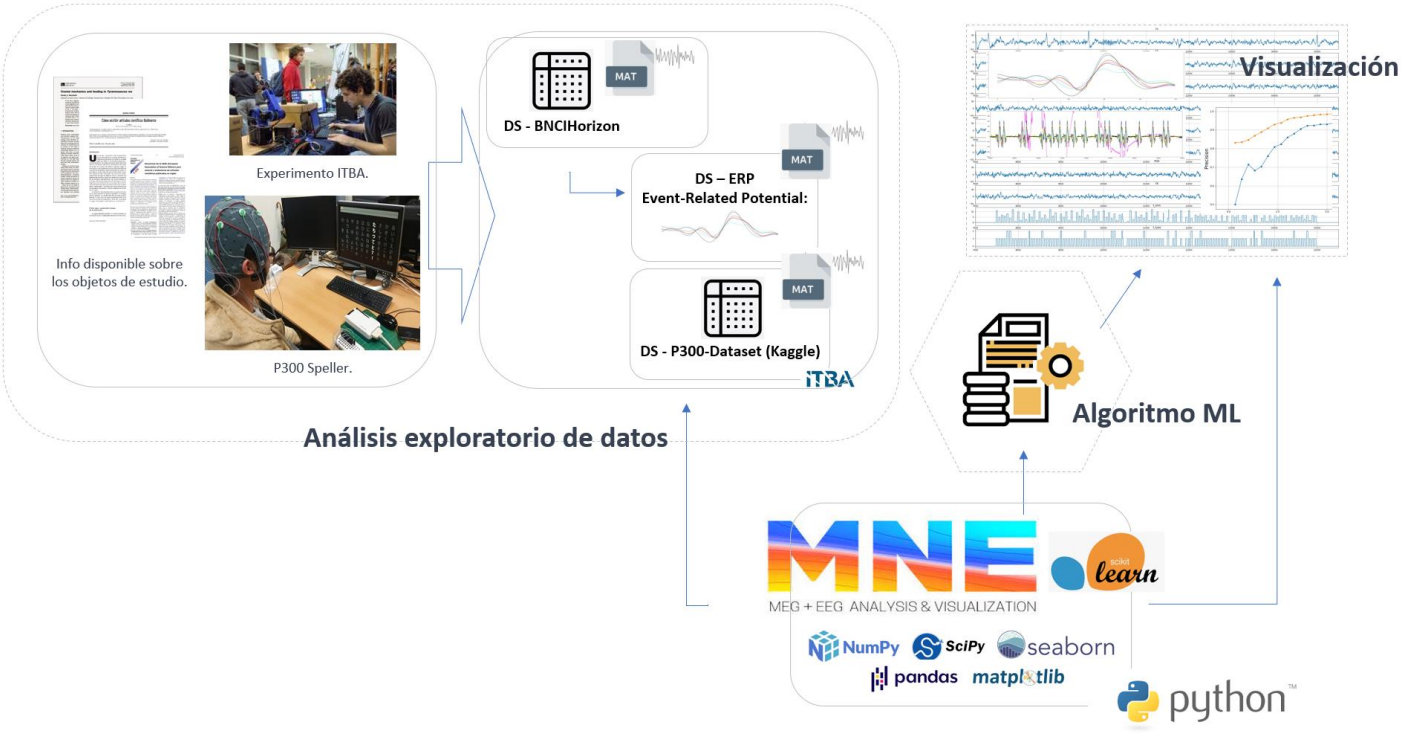
  BNCI Horizon 2020: 8. Speller P300 with ALS patients (008-2014).
  ITBA. P300 dataset of 8 healthy subjects.

- Modificar, en distintos rangos, las propiedades de la onda compuesta de manera sintética y verificar el rendimiento del algoritmo *drugsignal.py*

Presentar todos los resultados obtenidos de las modificaciones en las propiedades de la onda compuesta y proponer mejoras en la preconfiguración del algoritmo.

Documentar todo este trabajo final integrador en un paper.

8. Metodologías a usar



La estructura de este trabajo de integrador está pensada en función de los objetivos en el punto anterior. Primero se ofrecerá un análisis exploratorio de datos, no solamente de los datos en sí, sino también del contexto del experimento realizado el ITBA mencionado anteriormente. Este análisis exploratorio contendrá los datasets usados en los experimentos:

- BNCI Horizon 2020: 8. Speller P300 with ALS patients (008-2014).
- ITBA. P300 dataset of 8 healthy subjects.

En segundo lugar, habrá una descripción de la preparación de los datos para ser modelados, complementado con las herramientas computacionales disponibles: librerías de python especializadas en el manejo de datos y electroencefalografía.

En tercer lugar, se aplicará el modelo con las distintas variaciones en las propiedades de la onda compuesta de manera sintética. Y en cuarto y último lugar, se mostrarán los gráficos de los resultados obtenidos de las modificaciones más representativas en las propiedades de la onda compuesta, acompañados con una propuesta de mejoras en la preconfiguración del algoritmo.

Las herramientas y/o librerías usadas en este proyecto se pueden clasificar en dos; matemáticas y de electroencefalografía, todas concentradas en el lenguaje de programación Python. Los electroencefalogramas que fueron usados son *Matlab files*: archivos de extensión .mat en versiones con funcionalidades de almacenamiento de *arrays* de n dimensiones de hasta 100.000.000 elementos por arreglo y 2^{31} bytes por variable.

Dentro de las librerías matemáticas se encuentran ***NumPy*** para permitir el manejo de arreglos grandes y multidimensionales, ***SciPy*** con módulos para optimización , álgebra lineal , integración , interpolación , funciones especiales , FFT , procesamiento de señales e imágenes, entre otros, ***Matplotlib*** y ***Seaborn*** para visualización de los datos, y ***Pandas*** para la manipulación y análisis tanto de los archivos usados como fuentes de datos como para los distintos procesos intermedios en el análisis exploratorio. La librería destinada al machine learning es ***Scikit-learn***: dispone de algoritmos de clasificación, regresión y agrupamiento , que incluyen support vector machine, random forest, k -means y DBSCAN. Tiene la versatilidad para interactuar con el resto de librerías mencionadas anteriormente.

Por otra parte se usó la librería **MNE**: permite la exploración, visualización y análisis de datos neurofisiológicos humanos: MEG, EEG, sEEG, ECoG, NIRS, entre otros.

9. Análisis Exploratorio de Datos (EDA)

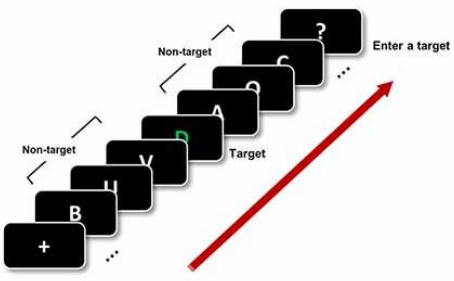
9.1 La enfermedad.

La esclerosis lateral amiotrófica o ELA, es una enfermedad degenerativa de las neuronas en el cerebro, el tronco cerebral y la médula espinal que controlan el movimiento de los músculos voluntarios. En la ELA, las células nerviosas (neuronas) motoras se desgastan o mueren y ya no pueden enviar mensajes a los músculos. Con el tiempo, esto lleva a debilitamiento muscular, espasmos e incapacidad para mover los brazos, las piernas y el cuerpo. La afección empeora lentamente. Cuando los músculos en la zona torácica dejan de trabajar, se vuelve difícil o imposible respirar.

En pacientes con ELA de etapas intermedias y avanzadas, es necesario el uso de dispositivos tecnológicos para la comunicación, como el P300 Speller.

9.2 El P300 Speller y el *oddball paradigm* (paradigma del bicho raro).

El P300 Speller es un dispositivo que conforma uno de las BCI (Brain Computer Interfaces) más usados en este tipo de aplicaciones. Su funcionamiento se acoge al paradigma del bicho raro: al usuario/paciente se le presenta una matriz de caracteres de 6 por 6 (ver figura) de manera intermitente, sucesiva y aleatoria. La tarea del usuario/paciente será enfocar su atención en los caracteres de una palabra prescrita por el investigador; es decir, un carácter a la vez. Cuando éstas contienen el carácter deseado (es decir, una fila particular y una columna determinada) se registra el potencial evocado P300 en el registro del EEC.

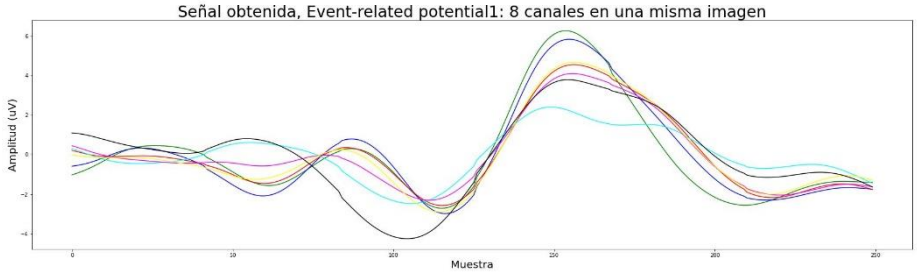


9.3 ¿Qué es un ERP (Event Related Potential)?

De forma paralela, es necesario explicar qué es una señal P300. La palabra *evocada* es clave: en medicina, se refiere a una actividad que puede ser detectada sincrónicamente después de una cantidad específica de tiempo después del inicio de un estímulo. Si estamos a la espera de que un computador nos dé una señal visual y nos la da, en nuestro cerebro ocurre un evento de este tipo. En términos médicos *es una actividad inducida* .

La onda P300 es entonces, una señal en el cerebro con amplitud positiva relacionada con eventos. Para esta investigación, los eventos serán aquellos provocados bajo el *paradigma del bicho raro*: El sujeto detecta un estímulo "objetivo" ocasional en un tren regular de estímulos estándar.

La onda P300 solo ocurre si el sujeto participa activamente en la tarea de detectar los objetivos. Su amplitud varía con la improbabilidad de los objetivos. Su latencia varía con la dificultad de discriminar el estímulo objetivo de los estímulos estándar. Una latencia pico típica cuando un sujeto adulto joven hace una discriminación simple es de 300 ms. En pacientes con capacidad cognitiva disminuida, el P300 es más pequeño y más tardío que en sujetos normales de la misma edad.



Se desconoce el origen intracerebral de la onda P300 y su papel en la cognición no se comprende con claridad. El P300 puede tener múltiples generadores intracerebrales, con el hipocampo y varias áreas de asociación de la neocorteza contribuyendo al potencial registrado en el cuero cabelludo. La onda P300 puede representar la transferencia de información a la conciencia, un proceso que involucra muchas regiones diferentes del cerebro [10].

9.4 Los datasets y las señales.

Ya mencionado anteriormente, los electroencefalogramas que fueron usados son *Matlab files*: archivos de extensión .mat en versiones con funcionalidades de almacenamiento de *arrays* de n dimensiones de hasta 100.000.000 elementos por arreglo y 2^{31} bytes por variable. El dataset de BNCI Horizon, el 008-2014, contiene un grupo completo de potenciales evocados P300 registrados con la interfaz cerebro computador BCI2000[11]. De éste dataset obtendremos el *ERPTemplate*. El otro, el P300-Dataset, está conformado por 8 EEGs de donde se extraerán algunos para realizar las pruebas del algoritmo. Todos los

datasets están basados en el paradigma Farwell y Donchin [12] mencionado en el punto 9.2. Las señales usadas en este trabajo serán descritas a continuación:

9.4.1 El ERPTemplate

Si bien la descripción de qué es un ERP está en el punto 9.3, es importante mencionar que éste ERP se extrae artificialmente del dataset de BNCI Horizon (008-2014) para ser "inyectado" a una señal EEG del P300-Dataset con el fin de crear una señal sintética que nos permita realizar las modificaciones de latencia y amplitud en donde quede empalmado dicho ERP. El proceso está descrito en el punto 9.5. En el archivo *a_analisis_ERPTemplate.ipynb* hay un análisis más en detalle de la estructura y propiedades de la onda.

(MIRAR EL VIDEO DE LA DISTRITAL Y COMPLETAR ALGUNOS DATOS FÍSICOS DE LA ONDA EN EL CÓDIGO)

9.4.2 P300-Dataset

El P300-Dataset está conformado por 8 EEGs distribuidos en dos grupos según la modalidad del experimento: pasivos P300S01,02,03,06 y activos P300S04, 05, 07 y 08. Este trabajo se enfoca en los pacientes pasivos: las trazas de EEG donde se superponen las pantillas ERP son de los pacientes que **no se enfocan en ninguna letra en particular**. Todo está allí, excepto el componente P300 ERP. Es por esto que se utiliza la información de marcadores para localizar los segmentos verdaderos donde se debería encontrar el P300, y esas ubicaciones de tiempo se utilizan para superponer la forma de onda de ERP extraída.

9.4.2.1 Estructura

(ACÁ VA LA SEÑAL EEG: LOS OCHO CANALES, Y LOS T-STIM Y GRAFICAS)

9.5 drugsignal, getstims y getlabels **(y las funciones que se me escapen).**

Hablar de las funciones y las alteraciones,

---

1. Introducción 100%
2. Estado del arte 30% (Párrafo corto de contexto y luego directamente el experimento)
3. Definición del problema 100%
4. Justificación del estudio 100%
5. Alcances del trabajo y limitaciones 100%
6. Hipótesis 0%
7. Objetivos 100%
8. Metodología 100%
    - Técnicas
    - Herramientas
9. Análisis exploratorio de datos 100%
    - 9.1 La enfermedad. 100%
    - 9.2 El P300 Speller y el oddball paradigm (paradigma del bicho raro). 100%
    - 9.3 ¿Qué es un ERP (Event Related Potential)? 100%
    - 9.4 Los datasets y las señales. 100%
        - 9.4.1 El ERPTemplate 90%
        - 9.4.2 P300-Dataset 100%
            - 9.4.2.1 Estructura 0%
    - 9.5 drugsignal, getstims y getlabels 0%

10. Testeo del algoritmo 0%
11. Resultados 0%
12. Referencias-Bibliografía 70%

10. Testeo del algoritmo

11. Resultados

12. Referencias / Bibliografía

(OJO QUE YA ESTÁN ORGANIZADAS Y CITADAS)

1. Hartman, A.L. Atlas of EEG Patterns; Lippincott Williams & Wilkins: Philadelphia, PA, USA, 2005.

2. Guger, C.; Allison, B.Z.; Lebedev, M.A. Introduction. In Brain Computer Interface Research: A State of the Art Summary 6; Springer: Cham, Switzerland, 2017; pp. 1–8.

3. Ramele, R.; Villar, A.J.; Santos, J.M.; EEG Waveform Analysis of P300 ERP with Applications to Brain Computer Interfaces: Computer Engineering Department, Instituto Tecnológico de Buenos Aires (ITBA), Published: 16 November 2018.

4. Harari, Y.N; Homo Deus: A Brief History of Tomorrow; Jerusalem University

5. Skoog, D.A.; West, D.M.; Holler, F.J.; Crouch, S.R. Analytical Chemistry: An Introduction; Saunders College, Publishing: Philadelphia, PA, USA, 2000.

6. Owens, T.J.; Zandt, G.; Taylor, S.R. Seismic evidence for an ancient rift beneath the Cumberland Plateau, Tennessee: A detailed analysis of broadband teleseismic P waveforms. J. Geophys. Res. Solid Earth 1984, 89, 7783–7795.

7. Stockman, G.; Kanal, L.; Kyle, M. Structural pattern recognition of carotid pulse waves using a general waveform parsing system. Commun. ACM 1976, 19, 688–695.

8. Hartman, A.L. Atlas of EEG Patterns; Lippincott Williams & Wilkins: Philadelphia, PA, USA, 2005.

9. Picton, T.W.; The P300 Wave of the Human Event-Related Potential; Journal of clinical neurophysiology, American electroencephalographic society, 1992.

(EL SUMMARY ESTÁ MUY BUENO)

27. Luck, S.J. An Introduction to the Event-Related Potential Technique; USA; MIT Press: Cambridge, MA, USA, 2005; Volume 78, p. 388.

10. The P300 wave of the human event-related potential

https://pubmed.ncbi.nlm.nih.gov/1464675/

[11] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, e J. R. Wolpaw, «BCI2000: a general-purpose brain-computer interface (BCI) system», IEEE Trans. Biomed. Eng., vol. 51, n. 6, pagg. 1034–1043, 2004

[12] L. A. Farwell e E. Donchin, «Talking off the top of your head: toward a mental prosthesis utilizing eventrelated brain potentials», Electroencephalogr. Clin. Neurophysiol., vol. 70, n. 6, pagg. 510–523, 1988.

Paper	Título
brainsci-08-00199.pdf	EEG Waveform Analysis of P300 ERP with Applications to Brain Computer Interfaces
P300 dataset of 8 healthy subjects.pdf	P300 dataset of 8 healthy subjects.pdf
fncom-13-00043.pdf	Histogram of Gradient Orientations of Signal Plots Applied to P300 Detection
UMA-BCI Speller.pdf	UMA-BCI SPELLER: PLATAFORMA DE COMUNICACIÓN DE FÁCIL CONFIGURACIÓN BASADA EN EL BCI2000 P300 SPELLER
P300 Speller with patients with ALS	P300 Speller with patients with ALS
Picton 1992	The P300 wave of the human Event-Related- Potential.
vucic2020.pdf	P300 jitter latency, brain-computer interface and amyotrophic lateral sclerosis
tesis_n3966_Gambini.pdf	Modelos de segmentación basados en regiones y contornos activos aplicados a imágenes de radar de apertura sintética
fnins-07-00267.pdf	MEGandEEGdataanalysiswithMNE-Python

Archivos .mat

https://la.mathworks.com/help/matlab/import_export/mat-file-versions.html