ISSN: 1692-7257 - Volume 2 – Number 44 - 2024

RCTA
Revista Colombiana de Tecnologías de Avanzada
UNIPAMPLONA

Digital Object Identifier: 10.24054/rcta.v2i44.3069

Use of Machine Learning to detect P300-type brain signals by generating visual and auditory stimuli

Uso de Machine Learning para detectar señales cerebrales de tipo P300 generando estímulos visuales y auditivos

Alejandro Jesús Perdomo Cely[1], PhD(c). Camilo Ernesto Pardo Beainy[1]
MSc. Moshé Alonso Amarillo[2]

*[1] **Santo Tomás University,** Faculty of Electronic Engineering, GIDINT Investigation Group, Tunja, Boyacá, Colombia.*
*[2] **Los Andes University,** Faculty of Engineering, Biomedical Engineering Department, Master in Biomedical Engineering, Bogotá D.C, Cundinamarca, Colombia.*

***Correspondence:** alejandro.perdomo@usantoto.edu.co*

***Received:** february 16, 2024. **Accepted:** july 15, 2024. **Published:** august 10, 2024.*

***How to cite:** A. J. Perdomo Cely, C. E. Pardo Beainy, and M. Alonso Amarillo, "Use of Machine Learning to detect P300-type brain signals by generating visual and auditory stimuli", RCTA, vol. 2, no. 44, pp. 170–176, Aug. 2024. Recovered from https://ojs.unipamplona.edu.co/index.php/rcta/article/view/3069*

Abstract: The P300 signal is an evoked potential that occurs in the occipital region of the brain when an unexpected visual or auditory change to a light or sound pattern is presented. This pulse is commonly studied in the field of biomedicine, used in partial recovery of mobility in quadriplegic patients through a screen with different commands, in which the patient moves his eyes towards the desired command, and generating the P300 is performed. the desired command. It is from here that the Machine Learning models are used, being Logistic Regression, Decision Tree, Support Vector Machine and K Nearest Neighbors, to recognize characteristics of electroencephalographic signals with the presence and absence of P300 and an increase in data is applied to them by improving the training, in order to obtain the analysis of the best predictors of the P300 brain signal.

Keywords: Machine Learning, P300, electroencephalography, data augmetation.

Resumen: La señal P300 es un potencial evocado que se produce en la región occipital del cerebro cuando se presenta un cambio visual o auditivo inesperado a un patrón lumínico o sonoro. Este pulso es comúnmente estudiado en el campo de la biomedicina, usado en recuperación parcial de movilidad de pacientes cuadripléjicos por medio de una pantalla con diferentes comandos, en el que el paciente mueve los ojos hacia el comando que desea, y generando la P300 se realiza el comando deseado. Es a partir de aquí, que se le da uso a modelos de aprendizaje de Machine Learning, siendo Regresión Logística, Árbol de Decisión, Máquina de Soporte Vectorial y K Vecinos Más Cercanos, para reconocer características de señales electroencefalográficas con presencia y ausencia de P300 y se les aplica un aumento de datos mejorando los entrenamientos, para así obtener el análisis de los mejores predicadores de la señal cerebral P300.

Palabras clave: Machine Learning, P300, electroencefalografía, aumento de datos.

1. INTRODUCTION

The P300 electroencephalographic signal is an evoked potential that occurs in the brain [1], more specifically in the occipital region [2], which corresponds to the back of the head, when a random visual or auditory stimulus is perceived [3]. This signal is a positive voltage pulse generated approximately 300 milliseconds after the stimulus (see Fig. 1), although it can be found in ranges of 250 to 400 milliseconds [4].

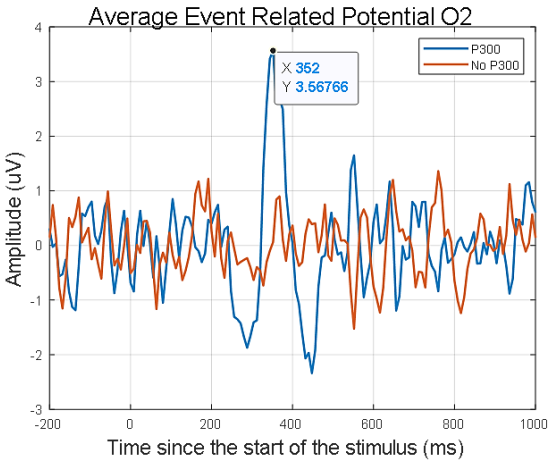


Fig. 1*. Comparison between presence and absence of the Event Related Potential in the electrode O2.*
Source: *own elaboration.*

The P300 is an electroencephalographic signal commonly used in the field of biomedicine for partial recovery of movement in paraplegic or quadriplegic patients [5]. Studies that support this concept correspond to the following:

•Brain-Computer Interfaces (BCI): BCIs based on the P300 signal have been developed to help people with physical disabilities communicate and control devices [6]. These applications range from controlling electric wheelchairs to writing text [7]. Studies such as evoked potentials are included, with which characteristics are extracted from the EEG signals after having applied pre-processing and filtering stages [8].

•Machine Learning Algorithms: Researchers have been applying a variety of Machine Learning algorithms, such as SVM systems, neural networks such as Deep Learning [9], and signal processing techniques to detect and decode P300 signals more accurately and efficiently. These studies begin with the communication between the electrode hardware, with the software that creates the stimuli, as well as the software that interweaves the EEG readings [10].

•Clinical Applications: Clinical applications have been explored, such as rehabilitating patients with spinal cord injuries [11] and improving the quality of life of people with paraplegia and quadriplegia. In addition, studies related to the creation of non-invasive electrodes and effective noise reduction are mentioned [12].

•Optimizing Visual Stimuli: Scientists have been investigating creating effective visual stimuli that could trigger reliable P300 responses, which is crucial for the accuracy of BCI systems, performing analysis on time series data, and explaining the difficulties and corrections to the main problems that occur in the capture of EEG signals [13].

•Interdisciplinary Collaborations: And last but not least, researchers have worked on creating effective visual stimuli that could trigger reliable P300 responses, which is crucial for the accuracy of BCI systems, performing analyzes on time series data, and explaining the difficulties and corrections to the main problems that occur in the detection of EEG signals [14].

From these ideas, the research was carried out aimed at predicting the P300 signal using supervised learning models from Machine Learning [15]. This has the practical purpose of being able to implement the most favorable models for electroencephalographic studies, where circumstances such as noise and high distance between the reading region and the study region can lead to the use of Machine Learning models with greater efficiency in detecting brain patterns, difficult to detect with the naked eye, along with efficient supervised learning models for the detection of brain signals.

The Logistic Regression [16], Decision Tree [17], Support Vector Machine, and K-Nearest-Neighbor [18] models were used, to which the exact same datasets were submitted, the models were trained, and each one returned the predicted data, thereby calculating the error that each one presented. The information provided to the models were the characteristics extracted from the electroencephalographic signals with the presence and absence of the P300, with their respective labels, with which the data was separated into training, validation and tests.

2. DEVELOPMENT

The study began with the assembly of the helmet that would hold the electrodes that will measure the voltages around the scalp, located between the occipital, temporal and central regions of the head. Through the OpenBCI program, the electrical variations in each of the 16 electrodes were captured (see Fig. 2), and transmitted through LSL communication, which allows sending data frames, up to 3 frames of different data simultaneously.

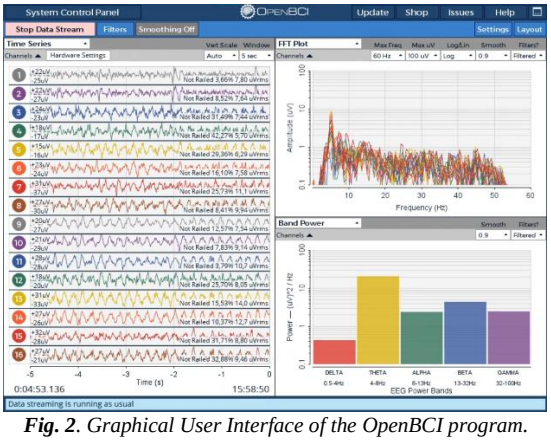


***Fig. 2**. Graphical User Interface of the OpenBCI program.*
***Source:** own elaboration.*

Once the LSL communication was initiated, the data frame was linked to the markers created by Psychopy, corresponding to the presence (1) or absence (0) of the P300 signal, through the random generation of visual and auditory stimuli. Having a total of 4 people in the study (see Fig. 3), each one of them was tasked with looking at a screen, counting how many times a sudden change of a constant blue square appeared, and pressing a button at each appearance of the stimulus [19]. This generated markers that indicated when a random stimulus had appeared and when it had not, providing the information to be intertwined with the OpenBCI electroencephalographic signals.

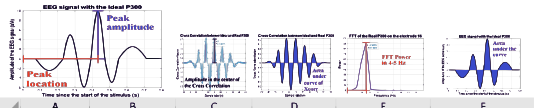
***Fig. 3**. Participants during the P300 generation tests: 1/A (left), 2/C (upper right), 3/Ma (middle right), and 4/Mo (lower right).*
***Source:** own elaboration.*

LabRecorder was the program in charge of unifying the LSL information that came from OpenBCI, with the LSL information that Psychopy returned, creating a final .xdf format file which was opened with a program created in Matlab called EEGLab [20]. Once imported into Matlab, the windows containing the presence or absence of the P300 signal were created, 6 main characteristics were extracted from each electrode, which were the following:

- The maximum peak of the signal: The highest value that the signal window reached.
- The location of the maximum peak: The position in time of the peak of the signal window.
- The center of the cross-correlation: At the time of performing the cross-correlation between an artificial signal that resembles the P300 signal, with the signal window to extract features, the value of the center of the correlation is obtained, where the higher this data is, the more similar it will be to the P300 signal.
- The area under the cross-correlation curve: If the area is located with a value greater than the reference zero, it will indicate that it is closer to being a P300 signal.
- The average power of frequencies from 4 to 5Hz: The set of signals were transformed from the time domain to the frequency domain, and the average of the powers at the frequencies from 4 to 5Hz was calculated, and if the average power is high, it is an indication of being a P300.
- The area under the signal curve: The area under the curve of the original signal was

calculated, which was useful information for the supervised learning models.

Everything was saved in a spreadsheet (see Fig. 4), obtaining a total of 96 characteristics, due to the fact that the 6 characteristics were calculated for each electrode of the 16 that were used, and an additional column for the presence or absence marker of P300.



	A	B	C	D	E	F
1	Peak_Value_1	Peak_location_1	XCorr_center_1	Mean_Xcorr_1	FFT_Power_4-5Hz_1	Average_Amplitude_1
2	2.40541458	14	-0.358593078	0.00290447297	27.1703823	8.99764151
3	2.42888829	63	0.0315264642	0.00351330664	35.8149077	12.48126
4	1.32545018	101	-0.159215866	-0.00130093566	14.5573871	-2.74642101
5	0.9939062	101	0.26765662	0.000837720821	2.10517432	0.175643081
6	1.3435719	95	0.00114547253	0.0011958884	0.702492435	1.80206277
7	1.04066467	8	0.209712277	0.00759315697	4.14364935	9.93909715
8	1.79652774	101	0.0299912751	-0.00221292161	10.2765798	-4.76529517
9	2.14355516	31	0.404461402	0.00152610008	35.5921641	3.68989308
10	2.80185294	5	0.0797767019	0.00385656685	16.24231	10.5695438

***Fig. 4**. Dataset with its corresponding extracted characteristics.*
***Source:** own elaboration.*

Finally, data augmentation was applied for the datasets, maintaining the values within the standard deviation ranges, with an increase of 10 times the original dataset (see Fig. 5), in order to amplify the training of the Machine Learning models.

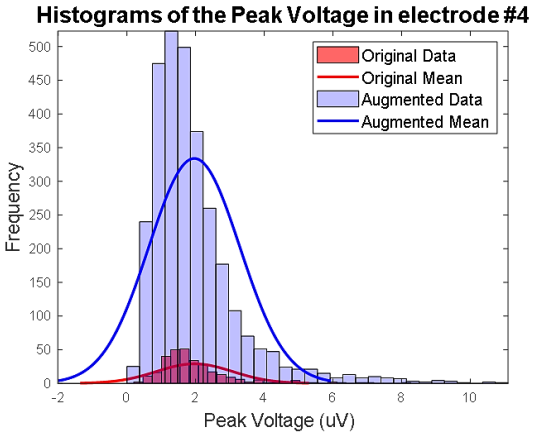


***Fig. 5**. Histograms comparing the data from the original dataset (red) with the augmented dataset (blue).*
***Source:** own elaboration.*

Afterwards, the information from the datasets was entered into the training models, and with this each of them was validated and tested. Each one provided the results of satisfactory and erroneous predictions, which were subjected to metric evaluation, to finalize the analysis of these results and establish which would be the best predictors of the P300 signal.

3. RESULTS

Once the confusion matrices were obtained from the participants, involving both the validation and the test ones, for the four types of models (see Fig. 6), it was observed that the confusion matrices related to SVM (Support Vector Machine) and KNN (K-Nearest-Neighbors) have a higher number of True Positives and True Negatives, compared to the LR (Logistic Regression) and Tree (Decision Tree) values, which had many more False Positive and False Negative values.

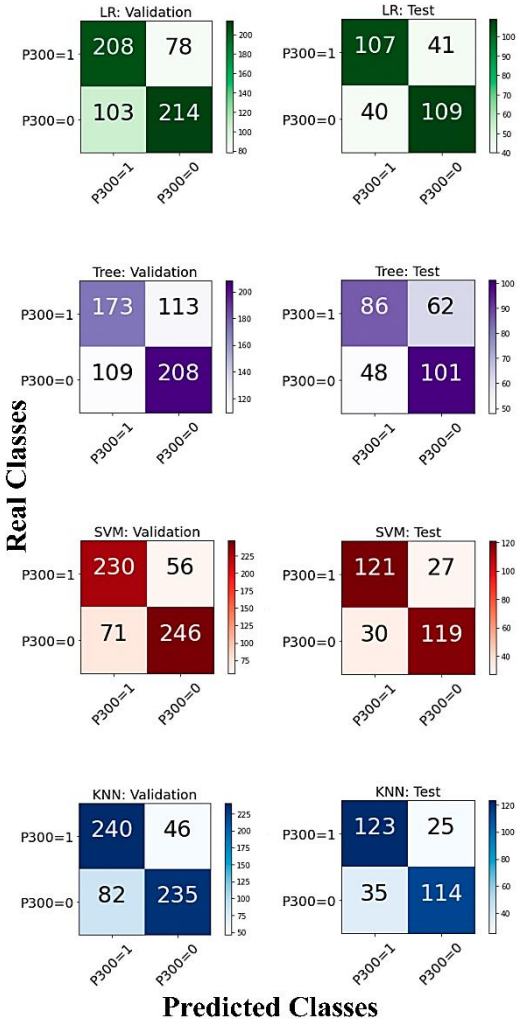


***Fig. 6**. Confusion matrices for the validation data (left) and the test data (right) with Logistic Regression (Green), Decision Tree (Purple), Support Vector Machine (Red) and K-Nearest-Neighbors (Blue) from the participant 1/A.*
***Source:** own elaboration*

The F1 score is a number between 0 and 1, which indicates the percentage of correct answers that the model made when it has been trained and has been presented with new data to predict. A value close to

zero indicates that it has not been satisfactorily trained, therefore, it does not predict correctly, while a value close to one indicates greater precision in the predictions it makes. This score was the metric used to establish the best predictors of the P300 signal.

3.1. Validation Data

Performing the calculations of the F1 score of the four participants, and of both datasets per participant, being the original and the augmented data, the following graph was obtained (see Fig. 7).

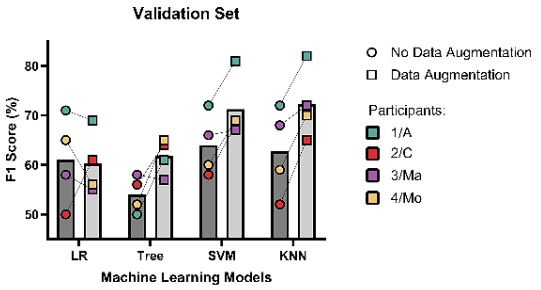


Fig. 7*. F1 Score percentage of the four participants for the validation data.*
Source: *own elaboration.*

This corresponds to 20% of the total data, since 70% of it is training data.

3.2. Test Data

This set corresponded to the test data with 10% of the total datasets, giving the following values as results (see Fig. 8).

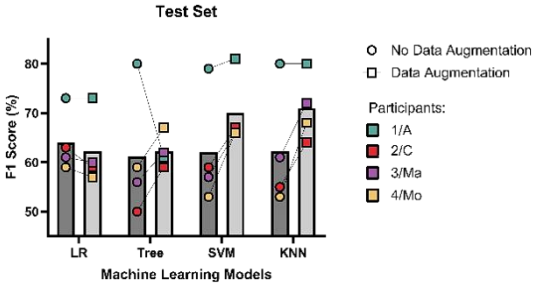


Fig. 8*. F1 Score percentage of the four participants for the test data.*
Source: *own elaboration.*

This allowed us to confirm that, in both validation and test cases, SVM and KNN were those that presented the best performance in predictions, and improved when the augmented data was entered, confirming the benefit of augmenting the data to train even better the Machine Learning models.

4. CONCLUSIONS

With these final results, it was possible to complete the project, where the following conclusions were drawn up:

•With the help of the OpenBCI, Psychopy and Matlab programs, it was possible to obtain the most favorable results in terms of hardware with the helmet, and software with the communication of information via LSL. Psychopy resulted very effective in terms of establishing markers coordinated with the appearance of random visual and auditory stimuli, and with the use of LabRecorder the EEG signals were successfully anchored with the markers, being evidenced by the evoked potential seen at the O1 and the O2 electrodes. Matlab allowed us to obtain the signals and markers to extract their characteristics and thus create the desired dataset for the study, along with data augmentation to improve the P300 prediction results.

•The electroencephalography helmet has a high performance as long as the electrodes can be adjusted correctly on the scalp, since it is affected by the amount of hair the person can have at the time of carrying out the study. If possible, for people with a lot of hair, it is recommended to adjust the electrodes as deep to the helmet as possible, and if possible, use conductive gel to increase the conductivity of the electrodes on the head.

•The characteristics extracted from each electrode during the study were sufficient to train the supervised learning models, since the analysis method was applied for each instant in which a stimulus appeared, allowing predictions to be made with each P300 marker that would appear in brain signals.

•The Machine Learning models mostly presented an improvement in terms of P300 signal predictions when the original dataset got applied the data augmentation, due to the fact that having more data to train allowed them to guess correctly with a greater precision the new data, information that was initially received without the classes.

•The supervised learning models “Support Vector Machine” and “K-Nearest-Neighbors” turned out to be the best predictors of the appearance of the P300 signal in this study, which was less affected by circumstances such as noise, while it was more favored by the data augmentation.

RECOGNITION

To the participants who collaborated in the data collection of the study.

And to the institutions that participated in this project, collaborating with the infrastructure and elements necessary for the implementation and development of this research.

REFERENCES

[1] B. Miner, Y. Pan, C. Burzynski, L. Iannone, M. Knauert, T. Gill and H. Yaggi, "0726 Agreement Between an Electroencephalography-Measuring Headband and Polysomnography in Older Adults with Sleep Disturbances," in *Sleep*, Oxford Academic, 2023, pp. A319-A320.

[2] O. A. Broggi Angulo, D. G. Koc Gonzáles y P. C. Martinez Esteban, «Guía de procedimiento de electroencefalografía y videoelectroencefalografía,» Ministerio de Salud de la República del Perú, San Borja, 2022.

[3] Y. Zhang, H. Xu, Y. Zhao, L. Zhang y Y. Zhang, «Application of the P300 potential in cognitive impairment assessments after transient ischemic attack or minor stroke,» *Neurological Research*, vol. 43, nº 4, pp. 336-341, 2021.

[4] J. M. Macías Macías, J. A. Ramirez Quintana, J. S. A. Méndez Aguirre, M. I. Chacón Murgia y A. D. Corral Sáenz, «Procesamiento Embebido de P300 Basado en Red Neuronal Convolucional para Interfaz Cerebro-Computadora Ubicua,» *ReCIBE. Revista electrónica de Computación, Informática, Biomédica y Electrónica, vol. 9, núm. 2*, pp. 1-24, 2020.

[5] «Electroencefalografía (EEG),» 2018. [En línea]. Available: https://brainsigns.com/es/science/s2/technologies/eeg.

[6] S. Silva Pereira, E. Ekin Özer and N. Sebastian-Galles, "Complexity of STG signals and linguistic rhythm: a methodological study for EEG data," *Cerebral Cortex*, vol. 34, no. 2, 2024.

[7] L. E. Morillo, «ANÁLISIS VISUAL DEL ELECTROENCEFALOGRAMA,» pp. 145-153.

[8] C. F. Blanco Díaz y A. F. Ruiz Olaya, «Caracterización de señales de EEG relacionadas a potenciales evocados visuales en estado estacionario,» *Ontare*, pp. 18-20, 2019.

[9] R. Chandra Poonia, V. Singh y S. Ranjan Nayak, Deep Learning for Sustainable Agriculture, A volume in Cognitive Data Science in Sustainable Computing, India: Elsevier, 2022.

[10] M. Razavi, V. Janfaza, T. Yamauchi, A. Leontyev, S. Longmire-Monford and J. Orr, "OpenSync: An opensource platform for synchronizing multiple measures in neuroscience experiments," pp. 3-7, 2021.

[11] T. Mo, W. Huang, W. Sun, Y. Hu, L. Mcdonald, Z. Hu, L. Chen, J. Liao, B. Hermann, V. Prabhakaran y H. Zeng, «Activation Map Reveals Language Impairment in Children with Benign Epilepsy with Centrotemporal Spikes (BECTS),» *Neuropsychiatric Disease and Treatment*, vol. 19, p. 1949–1957, 2023.

[12] C. Biarnés Rabella, «Diseño, caracterización y evaluación de electrodos capacitivos para la medida de ECG y EEG,» *Universitat Politécnica de Catalunya*, pp. 12-15, 2018.

[13] F. Wu, M. Gong, J. Ji, G. Peng, L. Yao, Y. Li and W. Zeng, "Interval and subinterval perturbation finite element-boundary element method for low-frequency uncertain analysis of structural-acoustic systems," *Journal of Sound and Vibration*, vol. 462, no. 114939, 2019.

[14] L. Bianchi, A. Antonietti, G. Bajwa, R. Ferrante, M. Mahmud y P. & Balachandran, «A functional BCI model by the IEEE P2731 working group: data storage and sharing,» *Brain-Computer Interfaces*, vol. 8, nº 3, p. 108–116, 2021.

[15] S. Gannouni, A. Aledaily, K. Belwafi and H. Aboalsamh, "Emotion detection using electroencephalography signals and a zero-time windowing-based epoch estimation and relevant electrode identifcation," *Nature Portfolio*, pp. 5-7, 2021.

[16] I. M. Hojas, «Regresión Logística en Python,» [En línea]. Available: https://www.statdeveloper.com/regresion-logistica-en-python/.

[17] R. Romo, «Árboles de Decisión / Decision Trees con python,» [En línea]. Available: https://rubenjromo.com/decision-trees/.

[18] L. Gonzales, «K Vecinos más Cercanos – Teoría,» 19 Julio 2019. [En línea]. Available: https://aprendeia.com/algoritmo-k-vecinos-mas-cercanos-teoria-machine-learning/.

[19] J. G. J. R. S. S. M. M. H. R. S. H. K. E. &. L. J. K. Peirce, «PsychoPy2: Experiments in behavior made easy.,» 2019. [En línea]. Available: https://doi.org/10.3758/s13428-018-01193-y.

[20] Z. Zhang, X. Liang, W. Qin, S. Yu and Y. Xie, "matFR: a MATLAB toolbox for feature ranking," *Bioinformatics*, vol. 36, no. 19, p. 4968–4969, 2020.