MCST: a Python framework for multi-channel signals feature extraction for Machine Learning

Victor Ocampo-Villalobos [a], *Daniel E. Hernandez* [*b], *Leonardo Trujillo* [a].

[a] Departamento de Ingeniería Eléctrica y Electrónica, vocampo009@gmail.com , leonardo.trujillo@tectijuana.edu.mx.
[b] Departamento de Ingeniería Industrial, daniel.hernandezm@tectijuana.edu.mx.
Tecnológico Nacional de México/Instituto Tecnológico de Tijuana, Calzada Tecnológico S/N, Fraccionamiento Tomás Aquino, Tijuana, Baja California, C.P. 22414
[*] Corresponding author

Resumen

En la era contemporánea de los datos, la integración de sensores, dispositivos inteligentes y maquinaria para capturar datos relacionados con la producción se ha vuelto cada vez más prevalente entre las organizaciones. Si bien estos datos tienen el potencial de impulsar la modelización predictiva y mejorar los procesos de toma de decisiones, el volumen abrumador a menudo plantea desafíos en la construcción de modelos. Para abordar esto, la extracción de rasgos es una etapa fundamental en el pipeline de aprendizaje de máquina. Este artículo presenta Multi-channel Signal Tools, una paqueteria basada en Python diseñada para mejorar la extracción de rasgos en señales multi-canal dentro de los pipelines de aprendizaje de máquina. Para evaluar las características extraídas utilizando la herramienta propuesta, se procesaron datos del conjunto de datos EEG SEED_IV [1] y se desarrolló un clasificador de bosques aleatorios, en un escenario multi-clase, para el reconocimiento del estado emocional. A pesar de utilizar datos de solo 12 canales EEG, el clasificador logra resultados comparables a modelos más complejos que utilizan 62 canales y datos de movimiento ocular. Nuestros hallazgos subrayan la efectividad de las herramientas de programación propuestas para optimizar el procesamiento de señales multi-canal, mejorar la escalabilidad y facilitar el análisis de datos. Los resultados muestran el papel de la herramienta propuesta en el avance de las técnicas de extracción de rasgos y su impacto potencial en diferentes dominios.

Palabras clave— Aprendizaje de máquina, Clasificación, EEG, Extracción de rasgos, Señales multi-canal.

Abstract

In the contemporary era of data, the integration of sensors, smart devices, and machinery for capturing production-related data has become increasingly prevalent among organizations. While this data holds the potential to fuel predictive modeling and enhance decision-making processes, the sheer volume often poses challenges in model construction. Addressing this, feature extraction emerges as a pivotal stage in the machine learning pipeline. This paper introduces Multi-channel Signal Tools, a Python-based framework aimed at enhancing feature extraction in multi-channel signals within machine learning pipelines. To evaluate the features extracted using the proposed framework, we process data from the SEED_IV [1] *EEG dataset and develop a random forest classifier, in a multi-class scenario, for emotional state recognition. Despite utilizing data from only 12 EEG channels, the classifier achieves results comparable to more complex models using 62 channels and eye movement data. Our findings underscore the effectiveness of the proposed programming framework in streamlining multi-channel signal processing, improving scalability, and facilitating insightful data analysis. The result show Multi-channel Signal Tools' role in advancing feature extraction techniques and its potential impact across different domains.*

Keywords— *Classification, EEG, Feature Extraction, Machine Learning, Multi-channel signals.*

1. INTRODUCTION

Currently, we find ourselves in the so-called era of data, largely due to digital transformation and the development of new data capture and transfer technologies. Therefore, the paradigm of Industry 4.0 promotes the use of sensors, smart devices, and machines that enable organizations to continuously collect production-related data. This, in turn, allows for better management of their systems. Thus, organizations face both new opportunities and challenges, one of which is predictive analysis using computational tools capable of detecting patterns in data and making informed decisions. Moreover, constructing models that enable predictions and, consequently, better decision-making. It is in these needs that Artificial Intelligence (AI) and Machine Learning (ML) come into play. The objective of implementing ML tools in the industry is to automate processes to maximize efficiency, increase sustainability, improve supply chain management, and identify system barriers even before they occur [2, 3].

When a signal is generated by a single sensor or source, it is referred to as a single-channel signal. On the other hand, if a signal involves various sensors or sources, it is described as a multi-channel signal. An example of this type of signal is Electroencephalography (EEG) signals, composed of data from various electrodes according to the capturing device being used; in this case, the sensors are the same, capturing electrical activity in different regions of the brain [4, 5]. However, there are signals formed by different sensors, such as in the case of mechanical systems where data from vibration, temperature, humidity, noise sensors, among other relevant physical factors for the system under study, may be available [2, 6]. Due to the complexity of multi-channel signals, primarily stemming from the vast amount of data they contain, one of the fundamental stages in building a ML model is the feature extraction process.

The process of constructing a ML model typically includes stages such as data acquisition, data cleaning, feature engineering, model selection, training, validation, and testing [7]. Feature engineering involves the process of transforming raw data, sensor data, into a suitable format for building ML model, aiming to enhance their performance and predictive accuracy. It includes the extraction, modification, or selection of features, to provide an algorithm with more relevant and informative input. This process often requires domain knowledge and creativity to identify meaningful patterns, relationships, or interactions within the data. Feature engineering plays a crucial role in improving model interpretability, addressing data quality issues, handling non-linear relationships, and adapting the data to meet the specific requirements of different ML algorithms. Feature extraction, in particular, plays a vital role in ML, involving the transformation of raw data into a condensed and relevant set of features. This technique reduces dimensionality, improving computational efficiency, mitigating noise, and enhancing model performance. By focusing on essential patterns, feature extraction contributes to better generalization, interpretability, and adaptability to algorithm requirements [8]. Overall, it is a vital step in optimizing data for effective modeling, leading to more efficient and accurate ML applications.

In this context, this paper introduces Multi-channel Signal Tools (MCST), a Python-based programming framework designed to facilitate the processing of multi-channel signals and extraction of features in both the time and frequency domains. The aim is to streamline the feature extraction process within the ML pipeline. These tools were employed to preprocess data from the SEED_IV EEG dataset [1], and the resulting features were utilized to develop a random forest classifier, yielding results comparable to those in the state-of-the-art.

2. METHODS

The MCST framework encompasses three primary stages for feature extraction. Firstly, signal splitting enables users to segment the signal into time windows for subsequent processing. Secondly, in the core stage of feature extraction, MCST offers methods for extracting features in both the time and frequency domains. Lastly, for storing the features, the framework provides tools to save the resulting data into files for future use in the ML pipeline.

2.1 Singal split

This method separates an input signal into time windows, it is common for the data capture of multi-channel signals to occur continuously over extended periods of time—minutes, hours, days. However, it's also typical to aim for building learning models that operate on considerably smaller samples of these signals. In this regard, the *split_cvs* tool allows for the segmentation of an input signal into smaller time windows. For this purpose, the input signal, the frequency at which it was captured, and the desired size in seconds of the time windows for feature extraction are required.

2.2 Feature extraction

The primary functionality of the framework is feature extraction in both the time and frequency domains. MCST enables the extraction of features from a single channel and features that combine data from two channels using asymmetry. The features extraction implements the operators describe in [5]. We define $\xi(t) \in \mathbb{R}^T$ as the vector containing the time series from a single channel, where T is the number of samples in ξ. The time derivative is represented as $\dot{\xi}(t)$.

Time domain features

The first set of time domain features are statistical features: defined as follows:

- **Power**:
$$P_\xi = \frac{1}{T}\Sigma_{-\infty}^{\infty} \mid \xi(t)|^2 \quad [1]$$
- **Mean**:
$$\mu_\xi = \frac{1}{T}\Sigma_{t=1}^{T}\, \xi(t) \quad [2]$$
- **Standard Deviation**:
$$\sigma_\xi = \sqrt{\frac{1}{T}\Sigma_{t=1}^{T}\big(\xi(t) - \mu_\xi\big)^2} \quad [3]$$
- **1st difference**:
$$\delta_\xi = \frac{1}{T-1}\Sigma_{t=1}^{T-1}\, |\xi(t+1) - \xi(t)| \quad [4]$$
- **Normalized 1st difference**:
$$\bar{\delta}_\xi = \frac{\delta_\xi}{\sigma_\xi} \quad [5]$$
- **2nd difference**:
$$\gamma_\xi = \frac{1}{T-2}\Sigma_{t=1}^{T-2}|\xi(t+2) - \xi(t)| \quad [6]$$
- **Normalized 2nd difference**:
$$\gamma_\xi = \frac{\gamma_\xi}{\sigma_\xi} \quad [7]$$

The Power P_ξ features represent the strength of the signal or consumed energy per unit of time. The mean μ_ξ and standard deviation σ_ξ features are statistical moments of the signal. The $1st$ and $2nd$ differences describe the changes of a signal over time and the normalized $1st$ difference is used to quantify the self-similarities contained in a signal.

MCST considers additional time domain features such as Hjorth features described in [9]: Activity, which describes the variance of the signal; Mobility, is the standard deviation of the slope of the signal using the standard deviation of the amplitude as reference; and finally, Complexity, which measures the variation of the signal using a smooth curve as reference. These features are calculated as follows:

- Activity: $A_\xi = \frac{\Sigma_{t=1}^{T}\left(\xi(t)-\mu_\xi\right)^2}{T}$ [8]
- Mobility: $M_\xi = \sqrt{\frac{var\left(\dot{\xi}(t)\right)}{var(\xi(t))}}$ [9]
- Complexity: $C_\xi = \frac{M\left(\dot{\xi}(t)\right)}{M(\xi(t))}$ [10]

where $\dot{\xi}(t)$ represents the first derivative of $\xi(t)$ with respect to time.

Another time feature in MCST is the Non-Stationary Index (NSI) proposed in [10]. The NSI is based on the variation of different segments of the signal over time. A stationary signal is one whose statistical properties, such as mean, variance, and autocorrelation, remain constant over time. In contrast, a non-stationary signal exhibits variations in its statistical properties over time. This feature is calculated by dividing the signal ξ into small segments and the mean μ_i for each segment is calculated. The NSI is defined as the standard deviation of the segments' means. In this way, as this index increases, the signal is considered as "less stationary".

The last time domain features included in MCST are the Higher Order Crossings (HOC) proposed in [11]. The objective of this features is to characterize the oscillatory nature of a signal. They are calculated by a sequence of high-pass filters applied over a zero-mean time series $Z(t)$ as follows:

$$\Im_k\{Z(t)\} = \nabla^{k-1}Z(t) \quad [11]$$

where ∇^k is the backward difference operator with order $k = 1,..,10$. The resulting D_k features are defined by counting the number of sign changes in the processed signal $Z(t)$. There is a total of 10 HOC features, that together with the Hjorth features, the NSI and the statistical features previously described, it results in a total of 21 time domain features that can be extracted from a single channel. Note that these features can be extracted from each of the channels independently.

Frequency features

Fourier analysis stands out as a widely adopted method in signal processing. This is because it excels in pinpointing narrowband elements within a signal, enabling a detailed characterization of its components. The prevalent frequency characteristics typically involve power features obtained from various frequency ranges. These features are derived from the power spectrum or band power, denoted as $P(u)$, of the signal $\xi(t)$. The band power is calculated using the Fourier Transform, where $P(u) = |F(u)|^2$ is determined by the square of the magnitude of the frequency spectrum $|F(u)|$. The frequency bands considered in MCST, listed in Table 1, are taken from the analysis of electroencephalogram (EEG) signals such as the one described in [4, 12].

Table 1 EEG Frequency band for feature extraction

Bandwidth (Hz)	Band name
1-4	δ
4-8	θ
8-10	$slow\ \alpha$
8-12	α
12-30	β
30-64	γ

These ranges may vary depending on the specific field. However, it is assumed that the features remain constant over time. Following the methodology outlined in [12], these features are derived using the Short Time Fourier Transform (STFT), employing a one-second Hamming window without any overlap. From the resulting segments, the mean power, minimum, maximum, and variance are extracted as features. This is done for each frequency bands listed in Table 1. Then, the alpha-beta ratio is calculated with the band power ration between bata and alpha β/α. Finally, the Differential Entropy feature proposed in [13] is computed for each frequency bands. A total of 31 frequency features are extracted from a single channel. Similar to the time domain, these features can be extracted for each channel independently.

Asymmetry features

The features outlined previously are computed using a single channel. However, examining two or more channels simultaneously may reveal descriptive features that elucidate the relationship between data captured by distinct sensors, potentially from diverse regions within the system under study. To prevent an explosion of potential features, MCST focuses on implementing binary features, utilizing signals from two channels simultaneously to derive them. Hence, asymmetry features are defined by combining information from pairs of channels. The features are evaluated using these channel pairs include differential asymmetry and rational asymmetry. The former is calculated as the difference between features derived from the signals of independent channels 'r' and 'l', given by:

$$\delta x = x_l - x_r \quad [12]$$

where x_u is a feature, computed from the signal of channel u. Similarly, rational asymmetry is computed as:

$$\Delta x = \frac{x_l}{x_r} \quad [13]$$

Asymmetry features are derived from both statistical features in the time domain and all frequency features, resulting in a comprehensive set of 74 asymmetry features per pair of channels.

In this context, MCST enables the extraction of a total of 52 features from a single channel and 74 features from a pair of channels. This process can be applied selectively to a subset of channels within the entire input.

3. EXPERIMENTS AND RESULTS

To demonstrate the effectiveness of the proposed features in constructing a predictive model, MCST was applied to the data from the SEED_IV dataset. This dataset is a collection of signal recordings, captured during a visual emotion induction experiment. It consists of EEG signals captured from participants while they were exposed to various emotional stimuli, mainly videos, designed to provoke a certain emotion in the individuals. The dataset includes annotations indicating the emotional states experienced by the participants during the task. It is commonly used in research for developing ML models to classify emotional states based on EEG data.

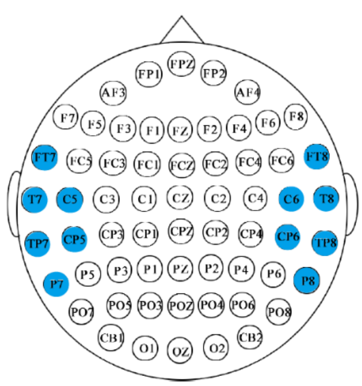
3.1 SEED_IV dataset

The SEED_IV dataset, as described in [1], comprises signals recorded from 62 electrodes and emotional neurofeedback tests. These tests aim to establish correlations between the recorded signals and specific emotional states. The dataset captures emotional responses from 15 participants exposed to 24 videos, of approximately 120 second, carefully selected to elicit neutral, happy, sad, and fearful emotional reactions. This protocol was repeated across three separate sessions, resulting in a total of 1,080 sets of signals collected from the 62 electrodes. To ensure consistency, EEG signals were down sampled to 200Hz and filtered within the range of 1 to 75Hz. Each trial of the evoked potential was categorized based on the elicited emotional state: neutral, happy, sad, or fear.

3.2 Processing with MCST

Initially, the signals underwent segmentation into 10-second intervals to try to identify emotions within that timeframe. It is postulated that each segment of the signal portrays the same emotional expression observed in the complete signals. This resulted in approximately 80 observations of each emotion per subject per session. Subsequently, drawing on findings from [5] demonstrating that signals from just 12 electrodes suffice to distinguish between EEG signals of different emotions, features were extracted from these specific electrodes, as illustrated in Figure *1* Selected electrodes for feature extractionFigure 1. Consequently, for each set of EEG signals associated with its respective emotion, a total of 1,068 features were extracted. This comprised 21 time-domain and 31 frequency-domain features for each of the 12 individual channels, along with 74 features from each of the 6 pairs of asymmetrical electrodes.

Figure 1 Selected electrodes for feature extraction



Source: imagen taken from [14]

3.3 Classification

To assess the efficacy of the extracted features in constructing machine learning models, we followed the experimental setup outlined in [1]. This involved establishing a 4-class classification problem within a subject- and session-dependent context, resulting in 45 distinct experiments across the 15 subjects and their 3 sessions each. We employed a Random Forest (RF) classifier using scikit-learn, utilizing 500 estimators and default configurations for the remaining parameters, and incorporating all 1068 extracted features. To evaluate the classifier's performance, we conducted a 10-fold cross-validation procedure. The average classification accuracy per session is presented in Table *2*.

Table 2 Classification accuracy (%) average per session

Session	**1**	**2**	**3**	**All**
Mean	72.99	73.64	79.61	75.41
Std	8.55	9.84	8.58	9.49

Figures 2-4 display the confusion matrices corresponding to each session. These matrices represent the cumulative testing results across the 10 folds in the evaluation process.

Figure 2 Confusion matrix for session 1

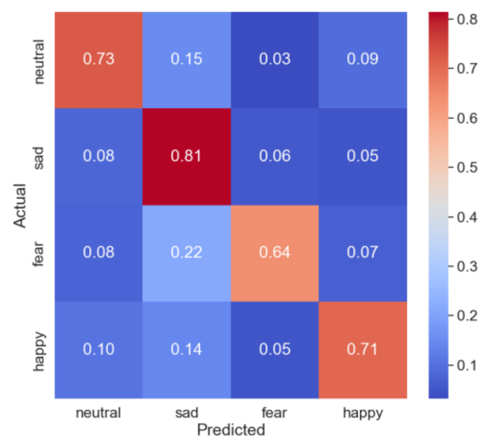


Figure 3 Confusion matrix for session 2

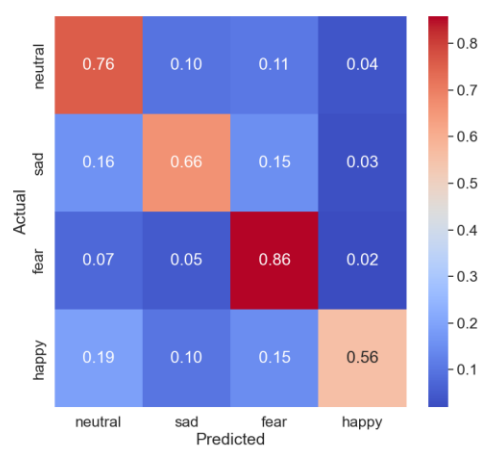
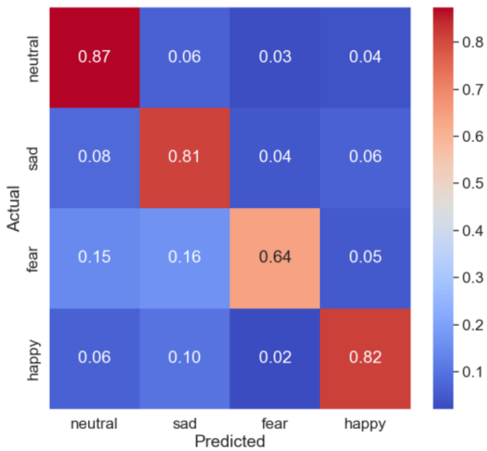
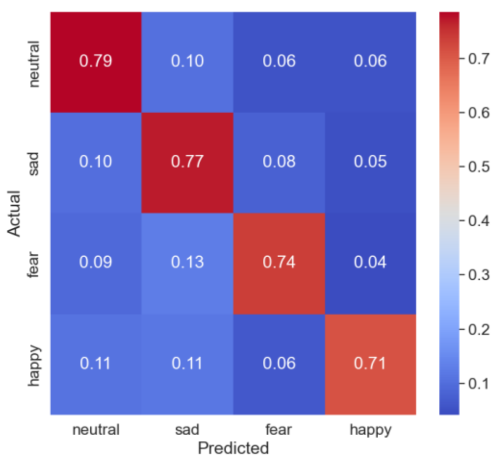


Figure 4 Confusion matrix for session 3



Then, Figure *5* illustrates the confusion matrix across the 45 experiments. It is evident that the RF classifier finds it easier to recognize the neutral emotion, while signals captured during instances of induced happiness prove to be more challenging to differentiate.

Figure 5 Overall confusion matrix



In our pursuit of enhancing the classification accuracy of the model, we opted to integrate a feature selection step as a filtering approach. This involved implementing a Kernel Principal Component Analysis (KPCA) operator with a Radial Basis Function (RBF) kernel. Utilizing scikit-learn, we conducted the operation to identify the optimal 50 components for classification. Once more, we utilized a RF classifier, evaluating it across the 45 experiments via a 10-fold cross-validation approach. We then compared the results of our proposed model with those described in [1], where the authors introduced various methods. These methods include a model constructed solely using eye movement to classify emotions, another utilizing EEG signals, a third leveraging data from both EEG and eye movement (referred to as Feature Fusion), and finally, a multimodal deep neural network (DNN) incorporating both data sources. The comprehensive comparison of overall classification accuracy between the proposed model and these methods is detailed in Table *3*.

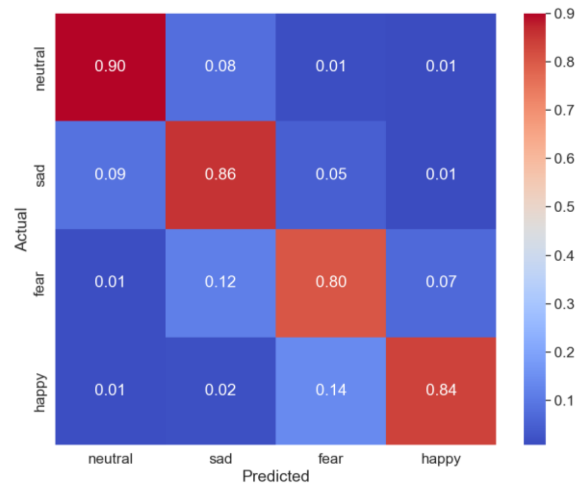
Table 3 Overall classification accuracy (%) performance comparison

Method	**Eye**	**EEG**	**Feat Fusion**	**DNN**	**MCST-RF**
Mean	67.82	70.33	75.88	85.11	**85.22**
Std	18.04	14.45	16.44	11.79	**5.66**

Source: results reported in [1] and the proposed method

Figure *6* illustrates the confusion matrix across the 45 experiments. These results indicate that the neutral emotion is the most easily distinguishable, whereas signals representing a fearful emotional reaction appear to be more challenging to differentiate. These findings underscore the importance of feature engineering, as they demonstrate that using the using the right set of features, a straightforward off-the-shelf model achieves comparable performance to more complex models such as deep neural networks.

Figure 6 Confusion matrix over the 45 experiments using KPCA for feature selection



4. CONCLUSIONS AND FUTURE WORK

This paper addresses the challenge of feature extraction in signals for constructing predictive models. By introducing the Multi-channel Signal Tools (MCST) framework and harnessing its capabilities, we have enhanced the feature extraction process for multi-channel signals within a machine learning pipeline. Through processing data from the SEED_IV EEG with MCST, we have successfully developed a random forest classifier that achieves results comparable to more sophisticated models in classifying emotional states, and using less data while we implement data from 12 channels, the DNN model uses the 62 channels and eye movement data.

Furthermore, our findings highlight the significance of leveraging computational tools such as MCST to facilitate the processing of multi-channel signals. This not only enhances the scalability and effectiveness of feature extraction but also empowers researchers and practitioners to extract meaningful insights from complex data sources.

As future work, it is imperative to address more complex versions of the problem, such as session-independent and/or subject-independent formats. Additionally, considering other evaluation scenarios like leave-one-subject-out, which is a common practice for processing biomedical data, would be beneficial. Furthermore, there is a need to evaluate MCST over other signal sources, such as industrial sensors, for fault detection in machinery. These efforts will contribute to a deeper understanding of the capabilities of MCST and its potential applications across different domains.

5. ACKNOWLEDMENTS

This work was partially funded by CONAHCYT (Mexico) project CF-2023-I-724 and TecNM through the projects “Visualización del proceso de aprendizaje realizado por un sistema de programación genética” and “Clasificación de señales multicanal con técnicas de aprendizaje de máquina”.

6. REFERENCES

[1] W. -L. Zheng, W. Liu, Y. Lu, B. -L. Lu and A. Cichocki, "EmotionMeter: A Multimodal Framework for Recognizing Human Emotions," *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 1110-1122, 2019.

[2] M. Cerrada, L. Trujillo, D. Hernández, H. Correa Zevallos, J. Macancela, D. Cabrera and R. Vinicio Sánchez, "AutoML for Feature Selection and Model Tuning Applied to Fault Severity Diagnosis in Spur Gearboxes," *Mathematical and Computational Applications,* vol. 27, no. 1, 2022.

[3] A. Angelopoulos, E. Michailidis, N. Nomikos, P. Trakadas, A. Hatziefremidis, S. Voliotis and T. Zahariadis, "Tackling Faults in the Industry 4.0 Era—A Survey of Machine-Learning Solutions and Key Aspects," *Sensors,* vol. 20, no. 1, 2020.

[4] D. E. Hernández, L. Trujillo, E. Z. Flores, O. M. Villanueva and O. Romo-Fewell, "Detecting Epilepsy in EEG Signals Using Time, Frequency and Time-Frequency Domain Features," in *Computer Science and Engineering—Theory and Applications*, Cham, Springer, 2018, pp. 167-182.

[5] D. Hernandez, L. Trujillo and A. Rodriguez, "AutoML for Emotion Recognition in EEG signals," *Numero Especial de la Revista Aristas: Investigación Básica y Aplicada.* , vol. 8, no. 16, 2021.

[6] J. Enríquez Zárate, M. Gómez López, J. Carmona Troyo and L. Trujillo, "Analysis and Detection of Erosion in Wind Turbine Blades," *Mathematical and Computational Applications*, vol. 27, no. 1, 2022.

[7] R. Ashmore, R. Calinescu and C. Paterson, "Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges," *ACM Computing Survey,* vol. 54, no. 5, pp. 1-39, 2021.

[8] R.-C. Chen, C. Dewi, S.-W. Huang and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," *Journal of Big Data*, vol. 7, no. 52, 2020.

[9] B. Hjorth, "EEG analysis based on time domain properties," *Electroencephalography and Clinical Neurophysiology,* vol. 29, no. 3, pp. 306--310, 1970.

[10] J. Hausdorff, A. Lertratanakul, M. E. Cudkowicz, A. L. Peterson, D. Kaliton and A. L. Goldberger, "Dynamic markers of altered gait rhythm in amyotrophic lateral sclerosis," *Journal of applied physiology,* vol. 88, no. 6, pp. 2045--2053, 20000.

[11] P. C. Petrantonakis and L. J. Hadjileontiadis, "Emotion Recognition From EEG Using Higher Order Crossings," *IEEE Transactions on Information Technology in Biomedicine,* vol. 14, no. 2, pp. 186--197, 2010.

[12] R. Jenke, A. Peer and M. Buss, "Feature Extraction and Selection for Emotion Recognition from EEG," *IEEE Transactions on Affective Computing,* vol. 53, pp. 327--339, 2014.

[13] R.-N. Duan, J.-Y. Zhu and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, San Diego, CA, USA, 2013.

[14] W.-L. Zheng and B.-L. Lu, "Investigating Critical Frequency Bands and Channels for EEG-based Emotion Recognition with Deep Neural Networks," *IEEE Transactions on Autonomous Mental Development,* vol. 7, no. 3, pp. 162--175, 2015.

.