Pipeline to detect spike-and-wave EEG patterns based on polynomial regression modeling and Taylor series feature selection

Matias F. Adell[1], Javier Balda[1], Facundo Casas[1], Carlos D'Giano[3], and Antonio Quintero-Rincón[1,2,3][0000−0003−0186−4049]

[1] Department of Data Science, Data Science and AI Laboratory, Catholic University of Argentina (UCA), Buenos Aires, Argentina.
[2] Department of Computer Sciences, Catholic University of Argentina (UCA), Buenos Aires, Argentina.
[3] Epilepsy and Telemetry Integral Center, Foundation for the Fight against Pediatric Neurological Disease (FLENI), Buenos Aires, Argentina.
{matiasadell,javierbalda,facundocasas,antonioquintero}@uca.edu.ar

Abstract. Epilepsy is a common neurological disorder diagnosed and monitored through EEG recordings. Accurate spike-and-wave (SW) pattern classification is crucial for distinguishing this epileptic seizure disorder from normal brain wave activity (NW). However, mathematically modeling SW remains challenging, affecting classification accuracy. This study proposes a pipeline in two stages combining polynomial regression techniques, and data processing, in a machine-learning classification scheme. At the first stage of decision-making, the idea is to create a generalized waveform mother that represents all the waveforms of the EEG patterns, such as SW and NW. This waveform is derived from a polynomial regression model that is assessed by the truncation error of the Taylor series. In the second stage, a feature selection algorithm based on a vector that includes the coefficients from Taylor and the statistical properties of the SW and NW waveforms was designed for the machine learning classifier. This algorithm uses the confidence interval to extract the Taylor series points that do not represent the generalized mother equation. This yields a dimensional reduction of this vector, which can be used in a classification and detection scheme. Three polynomial regression models, such as Fourier, Gaussian, and sums-of-sines were evaluated using the pipeline methodology. The best model was the Fourier regression, which achieved an accuracy of 96.2% using the SVM classifier with a Gaussian kernel to detect spike-and-wave patterns.

Keywords: Spike-and-wave · Polynomial regression · Taylor series · Feature selection

1 Introduction

Epilepsy is one of the most common neurological diseases, affecting approximately 50 million people worldwide [1]. This condition is characterized by the

occurrence of epileptic seizures, which are the result of abnormal and excessive electrical activity in the brain. The electroencephalogram (EEG) is a crucial biomedical tool used for the diagnosis and treatment of epilepsy because it allows for the recording and analyzing of brain waves to detect epileptiform activity, such as the spike-and-wave (SW) waveform pattern [2]. In the healthcare industry, EEG signals are widely used for detecting and classifying epileptiform waveform patterns, essential for an accurate diagnosis [3]. However, signal interpretation remains challenging due to the complexity of the waveforms and the need to differentiate between epileptiform activity and normal brain waves (NW) [4]. Automating this process using machine learning techniques and advanced signal processing has gained ground in the last decade, improving the accuracy and efficiency of diagnosis [5, 6]. However, difficulties persist in the precise mathematical representation of SW patterns and in assessing the impact of errors in this representation on classification models [7]. The problem lies in the need to develop mathematical models in EEG analysis that accurately capture the shape of SW waves, allowing them to be differentiated from NW and improving the classification algorithms' accuracy [8]. Today, the lack of precision in waveform representation can lead to significant errors in classification, directly affecting the ability of automatic systems to make reliable diagnoses [9, 10, 11, 12, 13]. This study addresses this problem by implementing a comprehensive pipeline that combines regression techniques, data processing, and classification models to analyze SW and NW waveforms. This pipeline, the main contribution of this work, consists of two stages. The first stage is for decision-making, and the second is for feature selection and classification. The decision-making goal is to use polynomial regression modeling to create a generalized mother equation based on the EEG waveforms. This generalized waveform is asses through the truncation error of the Taylor series. This stage produces a polynomial regression function and the Taylor coefficients at each point of this function. In this study, Fourier regression was the best model regarding the other two models studied, such as Gaussian, and Sum-of-Sines. The second stage applies the optimal results from the first stage to detect and classify spike-and-wave epileptiform patterns in EEG signals based on the feature selection algorithm. The input of this algorithm is the feature vector given by mean, median, standard deviation, kurtosis, and skewness from the SW and NW waveforms, and the Taylor series coefficients in each point. This algorithm focuses on extracting the Taylor series points that are not representative from the generalized mother equation using the confidence interval, yielding a dimensional reduction of this vector to be used in a classification and detection scheme.

The rest of this document is organized as follows. Section 2 presents the database, the mathematical theory, the proposed pipeline, and the feature selection algorithm. In Section 3, results are analyzed and discussed. Finally, conclusions and perspectives are presented in Section 4.

2 Material and Methods

2.1 Database

Signals were acquired from 12 patients at the Epilepsy Department of the Foundation for the Fight Against Pediatric Neurological Diseases (FLENI). An expert neurologist in epilepsy labeled 339 SW and 441 NW waveforms of the EEG signals, indicating the onset and duration of the epileptic waveform. A standard 10-20 EEG system with a sampling rate of 256 Hz was used with the following 22 channels: Fp1, Fp2, F7, F3, Fz, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T6, O1, O2, Oz, FT10, and FT9. Each waveform consists of a temporal sequence of amplitudes with morphological characteristics. SW waveforms are characterized by their regular and symmetrical morphology, combining spike peaks and smoother waves. This distinguishes them from NW waveforms, which have a less structured and more variable morphology. On average, the SW amplitudes are approximately 500, while the NW amplitudes are around 300. The total duration of the waveforms is approximately 100 to 264 seconds. Figure 1 shows representative examples of both types of waveforms. See [14] for more details of this database.

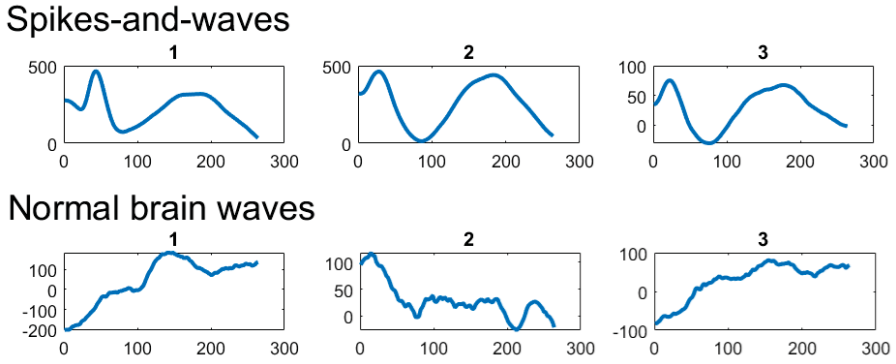

Fig. 1. Spike-and-waves and normal brain waves examples.

2.2 Pipeline methodology

The pipeline proposed in this study addresses the classification of spike-and-waves (SW) and normal brain waves (NW) using a comprehensive approach that combines data processing, polynomial regression techniques, and the Taylor series in a classifier scheme. The pipeline consists of two stages. The first stage is for decision-making, illustrated by the red dashed line in Figure 2. The second stage applies the optimal results from the first stage to feature selection and detect and classify spike-and-wave epileptiform patterns in EEG signals, illustrated by the blue dashed line in Figure 2.

The first stage begins with resizing the waveforms because all signals have different total durations in seconds. Thus, the maximum size of all waveforms was

calculated, and each SW and NW signal was resized to this maximum size using linear interpolation (Section 2.3). This process results in a single resolution size for all signals. Since the equation of the waveforms of interest is not known in advance, three polynomial regression models (Section 2.4) such as Fourier, Gaussian, and Sum-of-Sines were evaluated according to the metrics to identify which model best fits each resized waveform. The metrics used, such as Degrees of freedom for Error, Coefficient of determination R^2, Adjusted R^2, and the Root mean square error (Section 2.9), yielded that the Fourier regression was the best model. It produces an equation with coefficients representing each waveform. These coefficients were averaged to calculate a single overall coefficient to estimate the generalized mother waveform equation (Section 2.5). Finally, the Taylor series (Section 2.6) was utilized to approximate the equation of the generalized mother waveform at each point. The underlying idea is to assess how the series behaves relative to the original waveform and to determine whether it will be a good representation for subsequent analysis in the second stage of the pipeline. In addition, the truncation error (Section 2.7) is calculated at each point and cumulatively, providing a measure of the accuracy of the Taylor approximation for representing these waves.

In the second stage of the pipeline, a feature vector with two sets was built. The first set is based on the Taylor Series evaluation of degree 8 at each point. This indicates that the series is calculated in 1-second intervals, fully encompassing each SW and NW signal. Note that this data represents the results of the best decision-making model, the Fourier regression. The second set uses classical statistical properties, extracted directly from the original signals. The mean, median, standard deviation, kurtosis, and skewness improve the information available for the analysis.

At this point, the feature vector contains the Taylor Series approximation points and the statistical properties of each SW and NW signal (Section 2.10). This feature vector carries all the information needed to detect and classify spike-and-wave epileptiform patterns in EEG signals. The feature selection of this feature vector was performed using the proposed algorithm 1. This algorithm analyzes the points of the generalized mother equation that do not represent the classification model using the confidence interval. Subsequently, all features are normalized using the Min-Max Scaling technique from -1 to 1, ensuring that the data are in a uniform range and comparable (Section 2.8). To validate the effectiveness of the dataset, a 5-fold cross-validation is implemented, reserving 20% of the data for final testing. The theoretical framework used in the pipeline is introduced below.

2.3 Linear interpolation

Let $S_{max} = \max(\max(\mathrm{SW}), \max(\mathrm{NW}))$ be the maximum size of all waveforms. Let t_i and t_{i+1} be two successive points from each vector related to each waveform or class, $\mathrm{SW}(t)$ and $\mathrm{NW}(t)$. The goal is to find an intermediate point t between these two points. Then the linear 1D interpolation correspondent to each interval, $t_{i+1} - t_i$, and for $1 \leq \mathrm{SW}(t)|\mathrm{NW}(t) \leq S_{max}$ is given by:

$$t = (1 - t) * t_i + t * t_{i+1} = t_{i+1} + t(t_{i+1} - t_i) \tag{1}$$

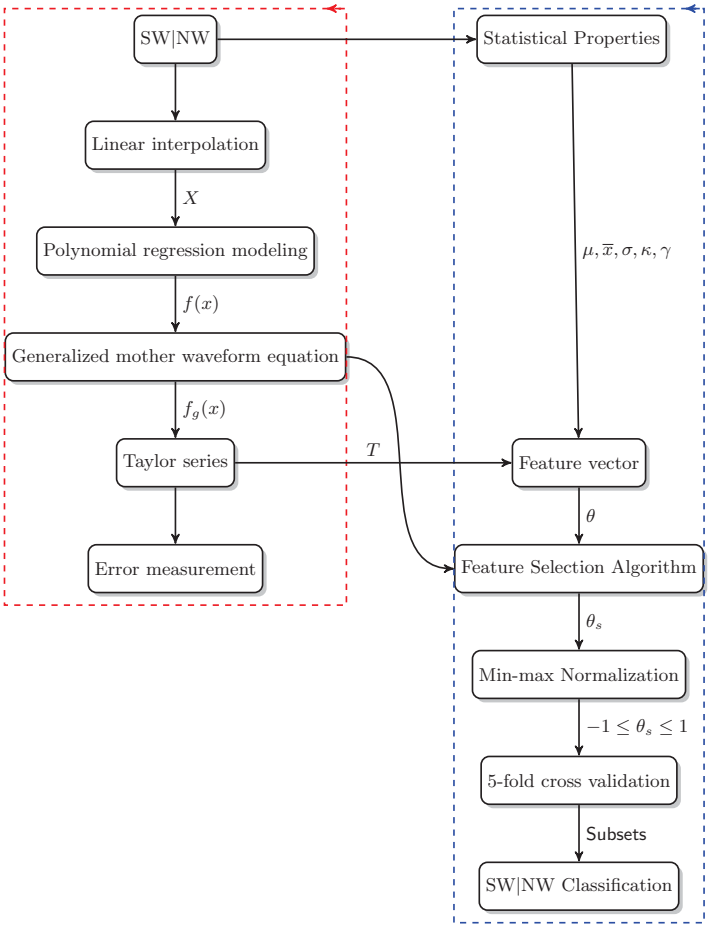

Fig. 2. Pipeline to detect and classify spike-and-wave epileptiform patterns in EEG signals. The decision-making stage includes all processes within the red dashed line. The feature selection and classification stage includes all processes within the red dashed blue.

2.4 Polynomial regression modeling:

Since the mathematical waveform of spike-and-wave is not known in advance, a mathematical model was created to describe its characteristics. Gaussian, Fourier, and Sum-of-sines polynomial regression models were evaluated to fit the EEG signals. These models can capture the specific and regular morphology of the waveforms, which is important for accurate pattern recognition. Each one is introduced below.

Gaussian Regression: A statistical method that uses the Gaussian function to describe a relationship between waveforms, approximating the function to fit

the peaks. The Gaussian model is expressed as:

$$f(x) = \sum_{i=1}^{n} a_i \exp\left[-\left(\frac{x - b_i}{c_i}\right)^2\right] \tag{2}$$

where a_i is the peak height or amplitude, b_i is the peak's center position or location, c_i is the peak width, and n is the number of peaks to fit.

Fourier Regression: A statistical method that uses the Fourier series to describe a relationship between waveforms as a sum of sine and cosine functions. The trigonometric Fourier series is given by:

$$f(x) = a_0 + \sum_{i=1}^{N} a_i \cos(i\omega x) + b_i \sin(i\omega x) \tag{3}$$

where a_0 is the intercept, a constant term associated with the $i = 0$ cosine term, a_i and b_i are the Fourier coefficients, n is the number of terms, and ω is the fundamental frequency.

Sum-of-sines regression: A statistical method that fits a weighted sum-of-sines functions to data. The mathematical expression for this model is:

$$f(x) = \sum_{i=1}^{n} a_i \sin(b_i x + c_i) \tag{4}$$

where a_i, b_i, and c_i are adjustable parameters that control the amplitude, frequency, and phase of each sin component respectively, and n is the number of terms. Note that this model includes the phase constant, and does not include the intercept term. This is the main difference from the Fourier Regression method.

2.5 Generalized mother waveform equation

Let $f(x)$ be the polynomial regression model that best fits each X EEG waveform. Let $\mathcal{C} \in R^{f \times c}$ the coefficients' matrix from each $f(x)$, where f is each equation and c each coefficient. The mean of all coefficients was calculated to yield a generalized representative alignment, which captures the studied waveforms' main periodic and morphological characteristics.

2.6 Taylor approximation

The Taylor series expansion provides a way to represent a function as an infinite sum of terms. Its derivatives are calculated at a specific point to approximate complex functions to feasible polynomials [15]. The idea is to approximate the generalized alignment waveform of the set of SW and NW to a polynomial function. In this case, the best polynomial regression model that fits the EEG patterns, see Section 2.4. Thus, it is necessary to define the function to be expanded, the variable, the initial point, and the number of terms in the Taylor series. :

$$f(x_{i+1}) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_i)}{n!}(x_{i+1} - x_i)^n \tag{5}$$

where $f^{(n)}(x_i)$ is the n-derivative of f evaluated at x_i, and x_i is the point around which the function is expanded.

2.7 Error Measurement

It is used to assess the accuracy of an approximation or model compared to the true value. It helps determine how close or far an estimated result is from the true value, allowing for improved methods and informed decisions. Different metrics are used for this purpose.

True Error (E_t) : It is the difference between the exact value of a function or series and its finite approximation. It is the error generated by putting a finite number of decimals in an approximation.

$$E_t = \text{true value} - \text{approximate value} \tag{6}$$

Percent Relative Error (ϵ_t) : It indicates how significant the difference is between the prediction and the actual value.

$$\epsilon_t = \frac{E_t}{\text{true value}} 100\% \tag{7}$$

Normalized Percent Error (ϵ_a) : It measures errors when the actual approximation value is unknown.

$$\epsilon_a = \frac{\text{present approximation} - \text{previous approximation}}{\text{present approximation}} 100\% \tag{8}$$

Truncation error E_ξ : It arises from using the Taylor series approximation instead of an exact mathematical expression method. The complete expansion of the Taylor series Eq. (5) is defined as

$$f(x_{i+1}) = f(x_i) + f'(x_i)h + \frac{f''(x_i)}{2!}h^2 + \cdots + \frac{f^{(n)}(x_i)}{n!}h^n + R_n \tag{9}$$

$$R_n = \frac{f^{(n+1)}(\xi)}{(n+1)!}h^{n+1} \tag{10}$$

where the subindex n of R indicates the residue of the n order approximation, ξ is the truncation error, a value of x that is located somewhere between x_i and x_{i+1}. For a comprehensive mathematical treatment of truncation errors, see [15].

2.8 Min-max normalization

It is a method for scaling data to a fixed range of values from minimum to maximum. It is beneficial to prevent data analysis from being influenced by the variation in time.

$$x' = \frac{2 \times (x - \min(x))}{\max(x) - \min(x)} - 1 \tag{11}$$

where x is the original value of the wave, $\min(x)$ is the minimum value of the wave in the data set, $\max(x)$ is the maximum value of the wave in the data set, and x' is the normalized wave value, scaled in the range -1 to 1.

2.9 Metrics performance

The following metrics were used to evaluate the fit quality of the regression models used in this study:

Sum of Squares Error (SSE): It measures the discrepancy between the observed values and the values predicted by the regression model. It is calculated by summing the squares of the differences between the actual values y_i and the predicted values $\hat{y}_i$ [16]. A lower SSE indicates a better fit of the model to the data. It is given by:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{12}$$

Degrees of Freedom for Error (DFEs): It represents the number of independent observations in a model minus the number of estimated parameters, including the intercept. DFEs are primarily used in assessing the statistical significance of regression coefficients because they influence the error variance estimation. It is given by:

$$DFE = n - p - 1 \tag{13}$$

where n is the total number of observations and p is the number of predictors in the model [17].

Coefficient of Determination $\mathbf{R^2}$: It measures the proportion of the variance in the dependent variable explained by the regression model. R^2 ranges between 0 and 1, where a value of 1 indicates that the model perfectly explains the variability observed in the data. It is given by

$$R^2 = 1 - \frac{SSE}{SST} \tag{14}$$

where SST is the Total Sum of Squares, representing the total variability in the data, although R^2 provides a general measure of model fit [18].

Adjusted $\mathbf{R^2}$: It is a modified version of the R^2 that takes into account the number of predictors in the model. Unlike R^2, adjusted R^2 penalizes adding additional predictors that do not significantly improve the model. It is calculated as:

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1} \tag{15}$$

where n is the number of observations, and p is the number of predictors [19].

Root Mean Square Error (RMSE): RMSE measures the average magnitude of the error in the model's predictions. It is the square root of the average of the squared errors and is expressed in the same units as the dependent variable, which makes it easier to interpret [20]:

$$RMSE = \sqrt{\frac{SSE}{n}} \tag{16}$$

2.10 Feature vector

Let $\mu, \overline{x}, \sigma, \kappa, \gamma$ be the mean, median, standard deviation, kurtosis, and skewness, the statistical properties respectively of each SW and NW waveform. Let $T' \in R^{p \times s}$ be the vector of p points and s seconds corresponding to the Taylor Series approximation points in each sec. The final feature vector is defined as

$$\theta = [T', \mu, \overline{x}, \sigma, \kappa, \gamma] \tag{17}$$

2.11 Feature Selection

In this paper, a feature selection algorithm was designed to reduce the number of variables in the dataset, see Algorithm 1. Remember that the dataset contains the Taylor approximation points and the statistical properties. For each second, a Taylor approximation point was computed. Thus, each point represents a feature. This algorithm focuses on reducing these points because not all points are equally representative of the overall signal. Therefore, those points that do not provide significant information are identified and eliminated. For this purpose, the algorithm compares the waveforms between the generalized equation and the polynomial regression model of each SW and NW pattern, identifying the points inside and outside the confidence interval. The points within the confidence interval are considered redundant because they do not provide new or significant information about the signal. These points are marked for elimination. Therefore, the points of interest are those outside the confidence interval, see Figure 3. This process allows a dimensional reduction of the dataset without losing the most relevant aspects of the signal.

Algorithm 1 Feature selection algorithm

Input: $f(x)$ (regression), z (Confidence interval), n (Regression duration),
Output: $cols$ (Non-representative points of the regression)

$cols \leftarrow []$
$i \leftarrow 0$
$z \leftarrow 0.2$ // Choice according to the criterion
while $i < n$ **do**
    $int\_sup \leftarrow f(i) + z \cdot \sigma$ //σ is the std
    $int\_inf \leftarrow f(i) - z \cdot \sigma$
    for $j \leftarrow i + 1 : n$ **do**
        if $f(j) \leq int\_sup$ and $f(j) \geq int\_inf$ **then**
            $cols \leftarrow j$
        else
            $break$
        end if
    end for
    $i \leftarrow j$
end while
return $cols$

3 Results

This section presents the results from the proposed pipeline for detecting and classifying spike-and-wave epileptiform patterns in EEG signals. Three polynomial regression models in the decision-making pipeline, namely Fourier, Gaussian, and sum-of-sines, were evaluated with their metrics, to select the best model that fits the EEG signal waveform. Table 1 shows the performance metrics for the three polynomial regression models evaluated. Fourier regression stands out for its superiority in all the assessed metrics. A higher R^2 and lower RMSE values suggest a better fitting capacity and accuracy in representing SW waveforms. Gaussian regression also shows acceptable performance, with relatively high value for R^2 and adjusted R^2. In contrast, the Sum-of-sines regression presented significantly inferior performance, with a negative R^2 value and a high RMSE value, suggesting a lack of ability to capture the characteristics of SW waveforms. The Fourier regression was fitted for each EEG waveform from the

Table 1. Comparison of the three Polynomial regression models evaluated

Model	**SSE**	**DFE**	**R^2**	**Adjusted R^2**	**RMSE**
Gauss	755346.3235	155.8053	0.7367	0.7089	53.6163
Fourier	320986.8081	152.7788	0.9093	0.9020	26.5285
Sum-of-sines	825650.9445	146.7876	0.2540	-0.0936	64.9567

dataset. Remember that the dataset contains 339 SW and 441 NW signals, see Section 2.1. This process yields 18 coefficients for each waveform. These are averaged to generate a final generalized mother waveform equation, see equation (18) with a morphology similar to that observed in the original signals, see signal with blue color in Figure 3.

$$\begin{aligned} f_{\mathrm{SWS}}(x) = {} & 38.2296 + 14.6907\cos(0.0312\cdot x) - 134.9640\sin(0.0312\cdot x) \\ & - 95.5140\cos(0.0624\cdot x) - 4.6392\sin(0.0624\cdot x) \\ & - 32.4604\cos(0.0936\cdot x) + 55.7023\sin(0.0936\cdot x) \\ & + 30.1549\cos(0.1248\cdot x) + 4.1776\sin(0.1248\cdot x) \\ & + 17.1263\cos(0.156\cdot x) - 22.0911\sin(0.156\cdot x) \\ & + 5.1722\cos(0.1872\cdot x) - 9.7515\sin(0.1872\cdot x) \\ & + 6.1877\cos(0.2184\cdot x) - 1.8771\sin(0.2184\cdot x) \\ & + 4.3840\cos(0.2496\cdot x) + 0.1376\sin(0.2496\cdot x) \end{aligned} \tag{18}$$

Note that different Taylor series degrees generate good coefficients for a classifier and detection scheme with low computational complexity and small errors. The Taylor series approximation errors were calculated at each point of the generalized Fourier function with the average of the coefficients. Remember that these values are part of the input from feature vector Eq. 17 for the second stage of the pipeline. Table 2 shows the cumulative errors of all signal points in each function degree. Note that, in both Figures, as the degree of the Taylor series approximation grows until n=8, the error decreases until it becomes imperceptible.

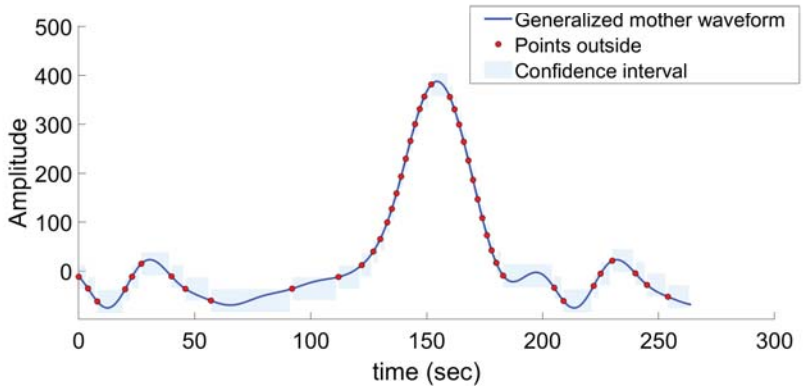

Fig. 3. Graphical representation of the generalized mother waveform (blue color) from the Fourier regression, with its confidence interval of $z = 0.2$ (light blue rectangles) and its points outside the confidence intervals from the Taylor series approximation (red circles).

Table 2. Cumulative errors. E_t: True error. E_ξ: Truncation error. E_t:Percent relative error. ϵ_a: Normalized percent error

Taylor degree	E_t	E_ξ	E_t	ϵ_a
0	-60.7772	-77.4907	-7.7491e+05	1.5266e+04
1	1.1755	-2.8812	1.0822e+03	9.0007e+03
2	0.2302	0.5151	66.0135	2.4176e+04
3	-6.2901e-04	-0.0066	0.1901	1.5160e+04
4	-5.2751e-04	-0.0011	0.1765	1.5225e+04
5	-1.6621e-06	2.3260e-05	0.0076	1.5235e+04
6	5.9759e-07	1.1875e-06	1.6747e-04	1.5235e+04
7	4.0414e-09	-2.8451e-08	2.9340e-06	1.5235e+04
8	-4.4729e-10	-7.9035e-10	1.2316e-07	1.5235e+04

For the second stage of the pipeline, the feature vector contains the Taylor series approximation points T of degree 8 at each point, with the statistical properties $\mu, \overline{x}, \sigma, \kappa, \gamma$ of the SW and NW signals. The proposed feature selection algorithm 1 achieved a dimensionality reduction of 23%. For illustration, Figure 3 shows the confidence interval (light blue rectangles), the points T outside the confidence interval (red color) from the generalized Fourier function (blue color). Remember that for the feature selection algorithm θ_s, the points of interest are those outside the confidence interval. θ_s is the feature selection from the feature vector θ, this vector was normalized to be tested in a machine-learning scheme.

Three classical machine models, such as Decision Trees, SVM with Gaussian kernel, and 10-nearest neighbors, were tested with θ_s. Table 3 shows the variations of the classification models depending on the degree of Taylor approximation in terms of accuracy. All models perform well, but the Gaussian SVM excels compared to the other models as the Taylor degree increases.

Table 3. Classification models comparison in terms of accuracy.

Taylor degree	Decision Tree	SVM	10-NN
0	0.8970	0.9310	0.8140
1	0.9420	0.9620	0.8330
2	0.8910	0.9290	0.8190
3	0.8900	0.9280	0.8200
4	0.8900	0.9300	0.8200
5	0.8890	0.9290	0.8210
6	0.8870	0.9280	0.8200
7	0.8860	0.9280	0.8210
8	0.8860	0.9300	0.8200

4 Conclusions

This work proposed an original two-stage pipeline to classify spike-and-wave epileptiform patterns in EEG signals. The first stage is for decision-making and the second is for feature selection and classification. At the decision-making stage, polynomial regression models of Fourier, Gaussian, and sums-of-sines were analyzed to find the best model that fits all the EEG waveform patterns, such as SW and NW. The best model was Fourier regression based on error metrics. From this model, a generalized waveform equation was computed, averaging all its coefficients for all waveform patterns. This generalized equation was evaluated through the truncation error of the Taylor series. In the second stage, a feature selection algorithm was designed. The algorithm computes the confidence interval for the generalized equation and the Taylor coefficients given by the polynomial regression model of each SW and NW pattern. The points inside and outside the confidence interval are detected and compared. Only the points outside the confidence interval were considered to yield a dimensional reduction of this data. Finally, the algorithm output coupled with the statistical properties of the SW and NW waveforms builds a vector to be used in a classification and detection scheme. The Fourier regression achieved an accuracy of 96.2% using the SVM classifier with a Gaussian kernel, allowing the detection of spike-and-wave patterns.

In addition to its excellent performance, the proposed pipeline has a low computational cost. The proposed pipeline's main limitation is that it does not explicitly consider physiological and non-physiological artifacts. Future work will focus on evaluating the proposed pipeline more extensively and studying robust feature extraction methods using highly imbalanced data.

Bibliography

[1] World health organization. epilepsy. https://www.who.int/news-room/fact-sheets/detail/epilepsy, 2025. Accessed: 2025-02-02.

[2] Bruce J. Fisch. *Epilepsy and Intensive Care Monitoring: Principles and Practice*. Demos Medical, 2009.

[3] J. Gotman. Automatic recognition of epileptic seizures in the EEG. *Electroencephalography and Clinical Neurophysiology*, 54(5):530–540, 1982. https://doi.org/10.1016/0013-4694(82)90038-4.

[4] Jonathan J. Halford. Computerized epileptiform transient detection in the scalp electroencephalogram: Obstacles to progress and the example of computerized ECG interpretation. *Clinical Neurophysiology*, 120(11):1909–1915, 2009. https://doi.org/10.1016/j.clinph.2009.08.007.

[5] U. Rajendra Acharya, S. Vinitha Sree, G. Swapna, Roshan Joy Martis, and Jasjit S. Suri. Automated EEG analysis of epilepsy: A review. *Knowledge-Based Systems*, 45:147–165, 2013. https://doi.org/10.1016/j.knosys.2013.02.014.

[6] Ali H. Abdulwahhab, Alaa Hussein Abdulaal, Assad H. Thary Al-Ghrairi, Ali Abdulwahhab Mohammed, and Morteza Valizadeh. Detection of epileptic seizure using EEG signals analysis based on deep learning techniques. *Chaos, Solitons & Fractals*, 181:114700, 2024. https://doi.org/10.1016/j.chaos.2024.114700.

[7] Risto J. Ilmoniemi and Jukka Sarvas. *Brain Signals: Physics and Mathematics of MEG and EEG*. The MIT Press, 2019.

[8] A. Ananthi, M.S.P. Subathra, S. Thomas George, Geno Peter, Albert Alexander Stonier, and N.J. Sairamya. A fusion wavelet-based binary pattern approach for enhanced electroencephalogram signal classification. *Computers and Electrical Engineering*, 123:110019, 2025. https://doi.org/10.1016/j.compeleceng.2024.110019.

[9] Ali Shahidi Zandi, Manouchehr Javidan, Guy A. Dumont, and Reza Tafreshi. Automated real-time epileptic seizure detection in scalp EEG recordings using an algorithm based on wavelet packet transform. *IEEE Transactions on Biomedical Engineering*, 57(7):1639–1651, 2010. https://doi.org/10.1109/TBME.2010.2046417.

[10] Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of Neural Engineering*, 16 (5):051001, 2019. https://doi.org/10.1088/1741-2552/ab260c.

[11] Aradia Fu and Fred A Lado. Seizure detection, prediction, and forecasting. *Journal of Clinical Neurophysiology*, 3(41):207–213, 2024. https://doi.org/10.1097/WNP.0000000000001045.

[12] Brandon M. Brown, Aidan M. H. Boyne, Adel M. Hassan, Anthony K. Allam, R. James Cotton, and Zulfi Haneef. Computer vision for automated

seizure detection and classification: A systematic review. *Epilepsia*, 65(5): 1176–1202, 2024. https://doi.org/10.1111/epi.17926.

[13] N. Rehab, Y. Siwar, and Z. Mourad. Machine learning for epilepsy: A comprehensive exploration of novel EEG and MRI techniques for seizure diagnosis. *Journal of Medical and Biological Engineering*, 3(44):317–336, 2024. https://doi.org/10.1007/s40846-024-00874-8.

[14] Antonio Quintero-Rincón, Valeria Muro, Carlos D'Giano, Jorge Prendes, and Hadj Batatia. Statistical model-based classification to detect patient-specific spike-and-wave in EEG signals. *Computers*, 9(4):1–14, 2020. https://doi.org/10.3390/computers9040085.

[15] Steven C. Chapra. *Applied Numerical Methods with MATLAB for Engineers and Scientists*. McGraw Hill, 2023.

[16] John Neter, Michael H. Kutner, Christopher J. Nachtsheim, and William Wasserman. *Applied Linear Statistical Models*. McGraw-Hill, Boston, MA, 4th edition, 1996.

[17] Samprit Chatterjee and Ali S. Hadi. *Regression Analysis by Example*. Wiley, Hoboken, NJ, 5th edition, 2012.

[18] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer, New York, NY, 2013.

[19] Norman Draper and Harry Smith. *Applied Regression Analysis*. Wiley, New York, NY, 3rd edition, 1998.

[20] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge, UK, 2007.