XLVI CILAMCE
VITÓRIA - ES - BRAZIL - 2025
XLVI IBERO-LATIN AMERICAN CONGRESS ON COMPUTATIONAL METHODS IN ENGINEERING

Automatic Detection and Classification of Relevant Events in Oil Wells Using Numerical Derivatives and Rule-Based Segmentation

Andressa Celestino Araujo da Silva[1], Igor de Melo Nery Oliveira[1], Lucas Gouveia Omena Lopes[1], Eduardo Toledo de Lima Junior[1], Thales Miranda de Almeida Vieira[2], Pedro Esteves Aranha[3]

[1]*Laboratório de Computação Científica e Visualização*
Universidade Federal de Alagoas
Maceió, AL, Brazil
andressa.silva@lccv.ufal.br, igornery@lccv.ufal.br, lucas.lopes@lccv.ufal.br, limajunior@lccv.ufal.br
[2]*Instituto de Computação*
Universidade Federal de Alagoas
Maceió, AL, Brazil
thalesv@gmail.com
[3]*Petrobras - Petróleo Brasileiro S.A.*
Rio de Janeiro, RJ, Brazil
pearanha@petrobras.com.br

Abstract. Anomaly detection in oil wells is challenging, especially in identifying critical events impacting operations. This study proposes an automated methodology to segment relevant events from multivariate time series, improving anomaly detection in Inflow Control Valves (ICVs) by isolating periods with significant signal variations. The approach uses a rule-based algorithm that processes sensor data via numerical differentiation, Gaussian smoothing, and peak detection. Decision rules based on signal stability and temporal proximity filters enable precise segmentation. Evaluation showed an 88% overlap (IoU) with manual annotations, confirming the method's effectiveness in identifying abrupt operational changes. A key contribution is reducing data complexity and bias by excluding low-variability segments, producing a concise dataset focused on meaningful variations. To validate this preprocessing, a Random Forest classifier trained on the segmented data achieved an F1-score and accuracy of 0.80, correctly identifying 8 of 10 anomalies. This performance surpassed the same model trained on the full dataset, which scored 0.60 on F1. These results validate the methodology's effectiveness and suggest its potential for broader use in detecting not only ICV anomalies but also other significant events reflected by notable variations in well data.

Keywords: data preprocessing, oil wells, anomalies, data segmentation, supervised classifiers.

1 Introduction

Anomaly detection in oil wells is a complex task, not only due to the need to identify critical events that impact operations, but also because of the challenging nature of the data involved. Although the oil industry increasingly uses pressure and temperature sensor data to monitor well integrity, traditional anomaly detection methods often struggle. This is due to the high dimensionality and complexity of these time series [1]. As Soriano-Vargas et al. [2] point out, such datasets typically contain hundreds, thousands, or even millions of instances and are influenced by factors such as system aging, equipment conditions, and operational variations, which make anomaly detection particularly challenging.

This growing emphasis on data selection reflects a central and intrinsic challenge in the process: the very nature of the data. As highlighted by Silva [3], production data often consist of long time series, and classifying such a large volume of observations becomes a computationally expensive task. In general, processing time is directly proportional to the series length, which makes eliminating unnecessary observations an essential step to reduce costs and optimize performance. In this context, Soriano-Vargas et al. [2] emphasize the need for strategies

to reduce the time spent identifying anomalies, thereby enabling improved decision-making processes.

The application of machine learning techniques for anomaly detection in oil wells has been growing, with a range of models being explored. These include Decision Trees (DT), valued for their interpretability [4], as well as more complex combinations such as Local Outlier Factor (LOF) and Convolutional Neural Network (CNN) [5], and even the application of transformer networks [6]. However, a crucial point is that the performance of these models is often linked to the quality and selection of the data. Studies have shown that dimensionality reduction [5] and precise data selection [6] significantly enhance both processing and accuracy, suggesting that large volumes of raw data may actually hinder accurate anomaly analysis. Meanwhile, Soriano-Vargas et al. [2] propose visual analysis approaches combined with machine learning to optimize the extraction of information from complex datasets.

This work proposes an automated approach to identify relevant events in oil wells, serving as a valuable pre-processing step. Our premise is that, throughout the lifecycle of a well, important events such as significant changes in pressure and temperature occur. We understand that anomalies represent a subset of these events, while other pressure and temperature (non-anomalous) changes are often the result of routine operations, such as valve adjustments. For quantitative analysis, we used a real dataset provided by the industry. This dataset covers long periods of normal well operation, but also includes specific periods annotated with anomalies caused by spurious closures of Inflow Control Valves (ICVs) in production and injection wells. The evaluation of the proposed approach was divided into two phases. Initially, we assessed the method's ability to correctly identify and delineate relevant events. Next, we evaluated the ability of a Random Forest (RF) classification model to distinguish between normal operations and the ICV anomaly present in the dataset. The RF model was trained in four different ways for evaluation. The classifier's performance was assessed using metrics such as accuracy and F1-score for the annotated anomalies and valve operations, ensuring a robust and comparative analysis of the approach.

2 Optimization of relevant event segmentation

The methodology employed in this study is an extension and application of the work developed by Silva et al. [7], which demonstrated effectiveness in optimizing the segmentation of relevant intervals in multivariate time series from oil wells. The main objective of this work is to focus on the practical application of this segmentation methodology, with advancements aimed at anomaly classification. The detailed flowchart of the methodological steps is presented in Figure 1, and the following sections describe its extensions and applications in depth.

2.1 Dataset description and preprocessing

This study employed real data provided by the industry, encompassing information from seven oil wells, including both production and injection wells, and capturing both normal operating conditions and occurrences of spurious closures of Inflow Control Valves (ICVs), which are components of the well completion system. The dataset consists of 10 instances of multivariate time series data, collected from sensors and valves installed across these different wells named from A to G.

2.2 Rule-based segmentation

The algorithm proposed by Silva et al. [7], focused on optimizing the segmentation of important intervals, consists of the following steps:

- Calculation of the time series derivative to highlight abrupt variations in each sensor's signal.
- Application of a Gaussian filter function [8], to smooth the derived time series and mitigate noise, using $\sigma = 6$ for the Gaussian kernel, based on preliminary sensitivity analyses conducted by the authors. Using the find_peaks function from the SciPy library [8], the maximum and minimum peaks are identified.
- By computing the standard deviation within a sliding window of $N_w = 40$ samples (equivalent to 20 minutes of data), periods of low variability are identified both before and after the peaks selected in the previous step. A threshold of 1×10^{-7} is used to classify a segment as having low variability. The indices that meet this condition, located both before and after the peaks, are stored as markers of the start and the end of intervals with abrupt changes.
- The final event segmentation is governed by a set of decision rules that validate the start of the variation before the corresponding minimum peak and the end of the variation after the maximum peak.

Additionally, a temporal proximity rule (T_{prox}=1500 samples) is applied to discard excessively brief or noisy segments, ensuring that only distinct and significant events are isolated.

Finally, a new dataset is generated from important segments throughout the well's lifecycle, while observations exhibiting constant behavior are discarded. Figure 1 illustrates the step-by-step rules of the segmentation methodology.

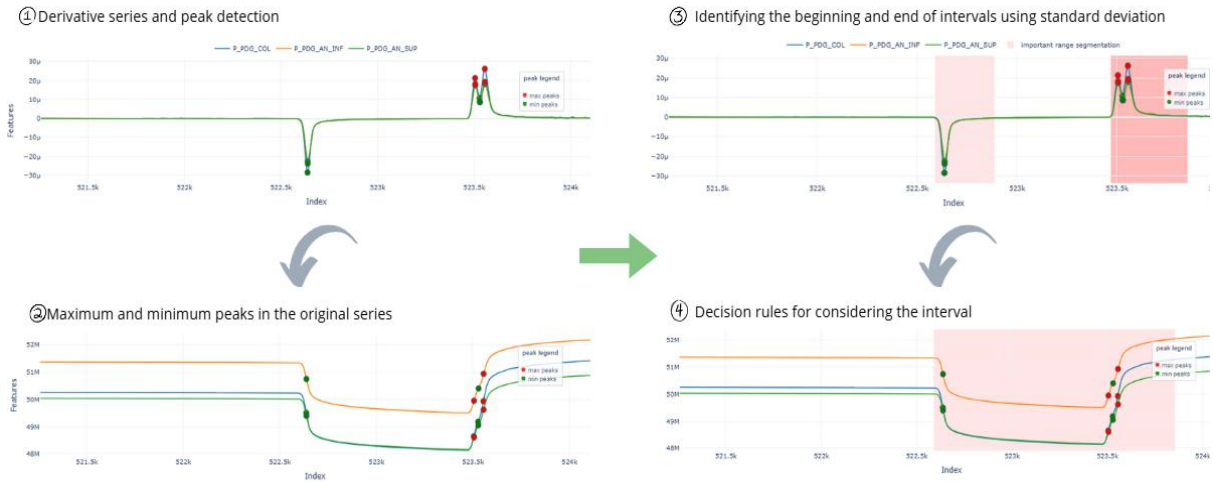


Figure 1: Segmentation methodology proposed by Silva et al. [7].

2.3 Dataset Adaptation and Feature Engineering

To build upon the methodology proposed by Silva et al. [7], we developed a secondary dataset by engineering new variables from the original pressure sensor data. Specifically, absolute and relative percentage differences were computed for three key pressure variables: P_PDG_COL, P_PDG_AN_SUP, and P_PDG_AN_INF. Additionally, differences between these pressure variables were calculated to capture inter-sensor relationships.

Mathematically, the absolute difference between the final and initial values of a pressure is calculated as shown in Eq. 1:

$$\Delta P_{abs} = \left| P_{final} - P_{inicial} \right| \tag{1}$$

The relative percentage difference is calculated as Eq. 2:

$$\Delta P_{rel} = \frac{\left| P_{final} - P_{inicial} \right|}{P_{inicial}} \times 100 \tag{2}$$

The differences between the pressures were calculated as the following equations 3, 4 and 5:

$$P_{COL_INF} = P_PDG_COL - P_PDG_AN_INF \tag{3}$$

$$P_{COL_SUP} = P_PDG_COL - P_PDG_AN_SUP \tag{4}$$

$$P_{INF_SUP} = P_PDG_AN_INF - P_PDG_AN_SUP \tag{5}$$

Absolute differences allow measuring the magnitude of variations regardless of their direction, considering that the normal state of an oil well evolves over time. A variation once considered normal may, in the future, present a different magnitude and be interpreted as anomalous. Thus, these new variables enable the analysis to adapt to changes in the system's normal behavior.

In total, nine new variables were generated: six derived from the absolute and relative percentage differences of each pressure variable and three resulting from the differences between them. Our goal is, for this expanded feature set, to significantly enrich the dataset for training anomaly detection classifiers. In Figure 2, the methodology of the study is presented.

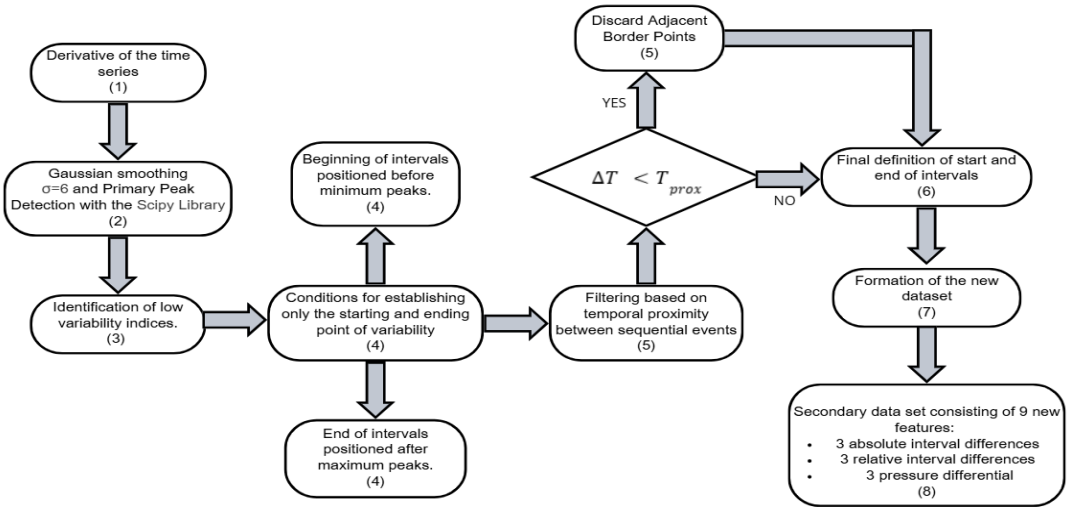


Figure 2: Flowchart containing the steps of the proposed methodology. Source: Adapted from Silva et al. [7]

3 Classification of relevant events

3.1 Proposed approach

For the practical application of the proposed methodology, both in this work and in previous studies [7], it is focused on detecting failures in Inflow Control Valves (ICVs). This choice was strategic due to the distinct way ICV failures appear in sensor data, typically as abrupt and pronounced changes in pressure, temperature, or flow after a period of stability. Such patterns strongly indicate potential ICV anomalies, guiding the algorithm's design to capture these behaviors efficiently.

After initial segmentation (Section 2.2), the detected important event segments are processed by classification models to determine their nature, distinguishing ICV failures from other operational events (e.g., production changes, system interventions). This two-phase approach separates general event detection from anomaly-specific classification, allowing the segmentation method to act as a universal event detector while specialized classifiers perform domain-specific categorization.

3.2 Classification model configuration

To evaluate the effectiveness of anomaly classification within the segmented intervals, we employed the Random Forest algorithm to train a classifier for ICV failures. This model was configured with n_estimators=100, meaning it builds 100 decision trees within the forest, and random_state=1200, to ensure the reproducibility of the results.

In addition, prior to training the classifier, class balancing was performed on the dataset using the random resampling technique, implemented through the resample function from the scikit-learn library [9]. This step is crucial because, typically, the number of anomalies in oil wells is significantly lower than the number of normal operating states. The imbalance between the volume of data in the minority class (anomalies) and the majority class (normal operation) can lead to biased classification models that tend to perform poorly in detecting anomalies, as they are predominantly trained with examples from the majority class.

3.3 Validation of the methodology

The proposed approach will be compared across four distinct cases using the training of the Random Forest (RF) classifier, aiming to demonstrate its advantages in terms of adaptability, accuracy, and robustness. The comparative analysis will include quantitative performance metrics such as F1-score, accuracy, Intersection over Union (IoU) and qualitative evaluation of the detected events, using the following datasets:

- **Case 1**: Using the complete dataset.

- **Case 2**: Using only intervals specifically annotated as valve changes.
- **Case 3**: Using the important intervals segmented by our preprocessing method.
- **Case 4**: Using the important intervals segmented by our method, but with the secondary features introduced in the methodology.

4 Results and discussions

4.1 Results of segmenting relevant events

This section presents the results obtained from applying the relevant event segmentation approach and discusses its effectiveness as a preprocessing step for anomaly classification in oil wells. As proposed in the previous work [7], the methodology initially focused on the analysis of pressure sensors, which proved to be the most significant and immediate indicators for detecting anomalies in ICVs.

The method was validated using a real dataset containing 10 cases of spurious ICV closures distributed across seven wells (A–G). Table 1 summarizes the performance, highlighting the accuracy in segmenting known anomalies and identifying normal operational events.

Table 1: Result of segmentation in the intervals of the studied anomaly.

Well	IoU Anomaly Segmentation	Segmented Instances in Normal Operations	Original Size	Filtered Size
A1	0.95	60	1,082,230	127,734
A2	0.92			
A3	0.98			
B	0.76	21	900,582	90,450
C1	0.90	28	615,271	89,886
C2	0.94			
D	0.98	17	181,419	53,400
E	0.84	26	293,375	39,266
F	0.73	26	411,824	104,385
G	0.81	24	403,003	184,364
Total	**0.88**	**202**	**3,887,704**	**689,485**

The results presented in Table 1 confirm the effectiveness of the proposed methodology. The IoU metric achieved an 88% overlap between manually annotated intervals and the important intervals identified by the method. This demonstrates success in the correct detection and segmentation of 10 cases of ICV anomalies, as well as the efficient identification of several relevant operational maneuvers during normal periods. The methodology proved to be an effective preprocessing tool, as it reduced the data volume by approximately 82% (from 3,887,704 to 689,485 observations). This reduction is achieved by eliminating periods of low variability, which are computationally expensive and have low informational value.

For a qualitative analysis, Figure 3 illustrates a comparison of the automatic segmentation proposed by the methodology (shaded area in red and gray bar) with the manual annotation of the important segment (yellow area), demonstrating the algorithm's effectiveness in accurately identifying and segmenting anomalous events into important intervals.

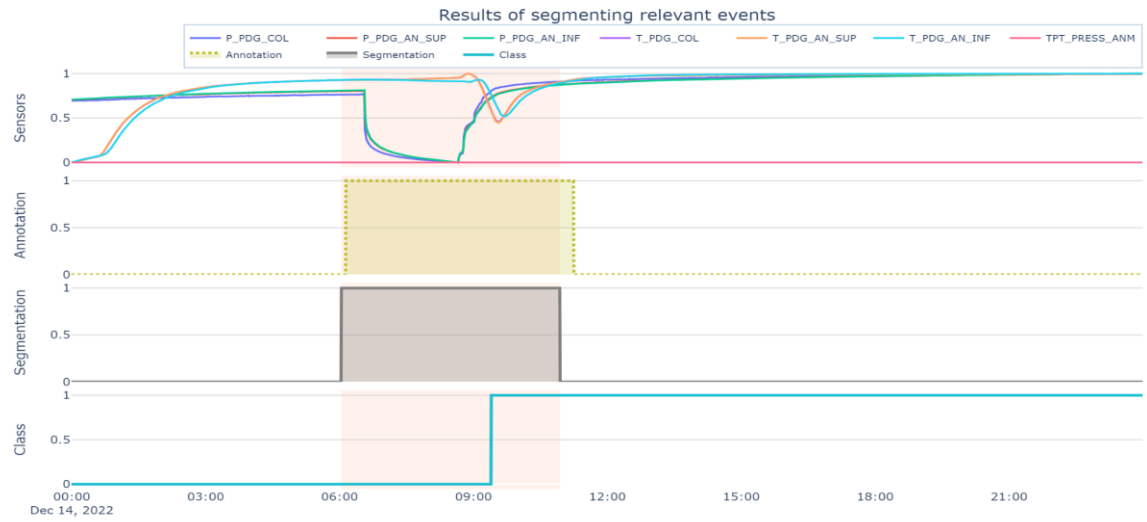


Figure 3: Application of the algorithm in one of the ICV spurious closure events.

4.2 Classification results using relevant events

To evaluate the effectiveness of our preprocessing method for training binary classifiers focused on anomaly classification, Table 2 presents the results of the binary RF classifier. This classifier was trained in four distinct scenarios, as described in the previous section, with the objective of distinguishing between anomalies (ICV failures) and normal well operation.

Table 2: Results of the classification of important segments.

Well	**Anomaly**								**Normal Operation**							
	Case 1		**Case 2**		**Case 3**		**Case 4**		**Case 1**		**Case 2**		**Case 3**		**Case 4**	
	ACC	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**
A1	1.00	1.00	0	0	1.00	1.00	1.00	1.00	0.84	0.91	1.00	1.00	0.41	0.58	1.00	1.00
A2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.76	0.87	1.00	1.00	0.75	0.86	1.00	1.00
A3	1.00	0.67	1.00	1.00	1.00	1.00	1.00	1.00	0.73	0.85	0.31	0.48	0.97	0.98	1.00	1.00
B	0	0	1.00	1.00	1.00	1.00	1.00	1.00	0.34	0.50	1.00	1.00	0.43	0.60	1.00	1.00
C1	0	0	0	0	0	0	1.00	1.00	0.10	0.19	1.00	1.00	1.00	1.00	1.00	1.00
C2	0	0	0	0	1.00	1.00	0	0	0.81	0.90	1.00	1.00	0.59	0.74	0.80	0.89
D	0	0	1.00	1.00	0	0	1.00	1.00	0.24	0.39	1.00	1.00	0.93	0.96	1.00	1.00
E	1.00	0.33	1.00	1.00	1.00	1.00	1.00	1.00	0.79	0.89	1.00	1.00	0.57	0.73	1.00	1.00
F	1.00	0.67	1.00	1.00	1.00	1.00	0	0	0.08	0.15	0.17	0.28	0.95	0.97	1.00	1.00
G	1.00	0.67	1.00	1.00	1.00	1.00	1.00	1.00	0.72	0.84	1.00	1.00	0.33	0.50	1.00	1.00
Total	**0.60**	**0.43**	**0.70**	**0.70**	**0.80**	**0.80**	**0.80**	**0.80**	**0.54**	**0.65**	**0.85**	**0.88**	**0.63**	**0.77**	**0.99**	**0.99**

From the analysis of Table 2, it is evident that Cases 3 and 4 demonstrate effectiveness in anomaly detection. For the anomaly data, both cases achieved an accuracy (ACC) and F1-score of 1.00 for most individual wells (A1, A2, A3, B, D, E, F, G), indicating robust performance in correctly identifying anomalies. This superior performance is especially clear when compared to Cases 1 and 2, which often show ACC and F1 values of 0 for anomaly detection in several wells. Overall, Cases 3 and 4 reached an ACC and F1 of 0.80 for anomalies, a significantly better result than the 0.60/0.43 of Case 1 and 0.70/0.70 of Case 2.

Regarding normal operation cases, Cases 3 and 4 also show very strong performance, particularly Case 4, which achieved total ACC and F1 values of 0.99. Case 3 follows closely with an ACC of 0.63 and F1 of 0.77. Although Case 2 also delivered good results in normal operation (ACC of 0.85, and F1 of 0.88), the combined performance in both anomaly detection and normal operation highlights Cases 3 and 4 as the most promising for the overall goal of the study. This suggests that the proposed segmentation and the use of secondary features are effective in optimizing the detection of anomalous events.

5 Conclusions

The results of this study demonstrate the efficiency of the proposed algorithm. The methodology for segmenting important intervals achieved an 88% overlap (IoU) with manual annotations. This method not only identified 202 important events during normal operation periods but also reduced the data volume by approximately 82%. To validate this preprocessing stage, a binary Random Forest classifier was trained and correctly classified 8 out of the 10 annotated anomalies. The best performance was observed in Cases 3 and 4, with Case 4 showing a significant reduction in false positives through the use of secondary features.

These findings confirm that focusing the analysis on more informative intervals is a highly effective strategy for ICV anomaly detection, as it minimizes bias introduced by long periods of stability. The methodology serves as a crucial preprocessing step, producing a reduced and relevant dataset that captures the main operational variations of the well. By isolating the segments of interest, the subsequent classification process operates on more targeted data, resulting in superior accuracy and precision.

Acknowledgements. The authors would like to thank PETROBRAS for the financial and technical support.

Authorship statement. The authors hereby confirm that they are the sole liable persons responsible for the authorship of this work, and that all material that has been herein included as part of the present paper is either the property (and authorship) of the authors, or has the permission of the owners to be included here.

References

[1] Machado, A.P.F., Munaro, C.J., Ciarelli, P.M. (2025). Enhancing one-Class classifiers performance in multivariate time series through dynamic clustering: A case study on hydraulic system fault detection, Expert Systems with Applications, vol. 286, 2025. https://doi.org/10.1016/j.eswa.2025.128088.
[2] Soriano-Vargas, A. et al. A visual analytics approach to anomaly detection in hydrocarbon reservoir time series data. Journal of Petroleum Science and Engineering, v. 206, 2021, 108988. https://doi.org/10.1016/j.petrol.2021.108988.
[3] Silva, A. M. (2016). Classificação de Séries Temporais baseada em Análise de Recorrência e Extração de Características. 2016. Universidade Federal de Mato Grosso do Sul, Master's dissertation.
[4] Alharbi et al. (2022). Explainable and Interpretable Anomaly Detection Models for Production Data. SPE Journal, vol. 27, pp. 349–363, 2022.
[5] Silva, A. C. A., Aranha, P. E., Oliveira, I. M. N., Lima Junior, E. T., Ramos, D. L., Vieira, T. M. A. (2024). Detection of unexpected events in oil wells using deep learning with Autoencoders and Local Outlier Factor. In: XLV Ibero-Latin American Congress on Computational Methods in Engineering, 2024. https://doi.org/10.55592/cilamce.v6i06.8231.
[6] Oliveira, I. M. N., Silva, A. C. A., Aranha, P. E., Vieira, T. M. A., Ramos, D. L., Lima Junior, E. T. (2024). Advancing Anomaly Detection in Oil Production Wells with TranAD: A Deep Transformer Network Approach. In: XLV Ibero-Latin American Congress on Computational Methods in Engineering, 2024. https://doi.org/10.55592/cilamce.v6i06.8224.
[7] Silva, A. C. A., Oliveira, I. M. N., Lima Junior, E. T., Vieira, T. M. A., Lopes, L. G. O., Aranha, P. E. (2025). Automatic Detection of Relevant Events in Oil Wells Using Numerical Derivatives and Rule-based Segmentation. In: X Encontro] Nacional de Cosntrução de Poços de Petróleo, 2025.
[8] Virtanen P. *et al.* SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, v. 17, p. 261–272, 2020. DOI: 10.1038/s41592-019-0686-2.
[9] Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.