Detection of Unexpected Events in Oil Wells Using Deep Learning with Local Outlier Factor and Autoencoders

Andressa Celestino Araújo da Silva[1], Pedro Esteves Aranha[2], Igor de Melo Nery Oliveira[1], Eduardo Toledo de Lima Junior[1], Thales Miranda de Almeida Vieira[1], Davi Leão Ramos[1]

[1]*Federal University of Alagoas, Av. Lourival Melo Mota, S/N, Tabuleiro do Martins, Maceió - AL, 57072-970*
andressa.silva@ctec.ufal.br, igornery@lccv.ufal.br, limajunior@lccv.ufal.br, thalesv@gmail.com, davi.ramos@lccv.ufal.br
[2]*Petrobras, R. Marquês de Herval, 58-78, Valongo, Santos – SP, 11010-310*
pearanha@petrobras.com.br

Abstract. In the oil and gas industry, investigating anomalies is crucial to prevent production losses, environmental accidents, and reduce maintenance costs. This study explores a density-based unsupervised machine learning model, the Local Outlier Factor (LOF), combined with autoencoders, to detect anomalies in subsea production/injection wells, through sensors. Two approaches are proposed: the first involves using well pressure and temperature sensor data, which are pre-processed with autoencoders to reduce dimensionality and capture key features, before being fed into the LOF model. The second approach applies the data directly to the LOF without using autoencoders and compares the results with previous studies on detecting important events. The experiments were conducted on real cases of wells with operational failures, focusing on different well anomalies defined in the 3W public database. After developing the models, the Autoencoder+LOF approach achieved the best performance, with an average F1 score of 0.8520 for the studied anomalies, compared to 0.7417 obtained by LOF itself. Additionally, there was a 62.22% reduction in computational time when using the Autoencoder. These results confirm that integrating autoencoders for data dimensionality reduction significantly improves anomaly detection overall, demonstrating the effectiveness of the proposed system in identifying unexpected events in time series.

Keywords: Anomaly Detection, Deep Learning, Local Outlier Factor, Autoencoders, Well monitoring.

1 Introduction

The processes in the oil and gas industry present a high level of complexity, requiring careful conduct and assertive monitoring to deal with possible failures (Sobrinho *et al.* [1]). In this case, the use of sensor-based data monitoring is a common practice, aiming to identify possible anomalies in structures and proactively prevent accidents. As emphasized by Vargas [2], the detection and classification of rare and undesirable events play a critical role in the oil industry. Vargas [2] also cites a study conducted by the Exploration and Production Operations Unit of Petrobras in Espírito Santo (UO-ES), which estimated a production loss of 1,514,000 barrels in 2016 due to anomalies in its offshore production wells operated by Natural Lift. Considering this estimate as an annual average, at $50 per barrel, the financial impact of these anomalies on UO-ES is $75.7 million per year.

In this context, the adoption of artificial intelligence and machine learning (ML) methodologies is being applied by various authors as a method for investigating critical intervals and monitoring well production, aiming to detect specific events and enhance the ability for proactive prevention and intervention. An example of this can be seen in Nascimento [3], where the author used one-class classifiers such as Support Vector Machine (SVM) and Isolation Forest to detect anomalies in wells and achieved recall scores higher than 0.98. Additionally, Fernandes Junior *et al.* [4] also compared one-class classifiers such as Isolation Forest, One-Class Support Vector

Machine (OCSVM), Local Outlier Factor (LOF), and Elliptic Envelope by applying them to a public database called 3W, with the best performance obtained by the LOF model with an F1 score of 88.2%, followed by Isolation Forest with 74.3%.

This article aims to apply and compare ML techniques for anomaly detection in oil production wells, using case studies presented in the 3W public database provided by Vargas [2], composed of multivariate time series of both real industry and simulated well sensor data. Two important event detection techniques were presented and compared: the first involved the direct application of data to the LOF model, and the second involved reducing data dimensionality using autoencoders before applying it to the LOF. Additionally, the performance of the two proposed approaches was evaluated in relation to the important events present in the database, as well as in comparison with related works from the literature.

2 On the Machine Learning techniques addressed

Local Outlier Factor (LOF) is a widely used technique in the context of anomaly detection, being defined as an unsupervised ML approach. Developed by Breunig *et al.* [5], LOF is designed to assess the degree of abnormality of data points in relation to their nearest neighbors, making it a powerful tool for anomaly detection.

According to Misra *et al.* [6], LOF aims to evaluate the degree of abnormality of points by comparing their local density with the density of their neighbors. Thus, points with significantly lower density than their neighbors are considered outliers, or anomalous points. It provides an anomaly score for each data point, enabling robust anomaly detection in complex and high-dimensional datasets.

In LOF, it is possible to adjust hyperparameters such as the number of neighbors to consider, the distance metric, and the contamination level in the dataset. Additionally, the parameter novelty can be configured to enable anomaly prediction in time series (Fernandes Junior *et al.* [4]). This aspect will be discussed in the development of the model in this paper.

The use of autoencoders (AE) to generate a reconstructed output that closely approximates the original data, but with reduced dimensions, has been explored by several authors in the literature, such as Chen *et al.* [7], Sobrinho *et al.* [2] and Aranha *et al.* [8]. This technique involves training the autoencoder to reconstruct data with reduced dimensions in the output, resembling the original data. This is possible because the nodes in the intermediate layers have a reduced number, which forces the model to learn weights that represent a condensed version of the input data (Fernandes Junior *et al.* [4]).

In addition to the use of autoencoders itself, it can be combined with other models. In the context of this study, we explore its integration with the Local Outlier Factor for anomaly detection, given that the combination of autoencoders with LSTM has proven effective in identifying complex temporal and spatial patterns in data, as demonstrated by previous studies such as Aranha *et al.* [9], Vargas [2], and Fernandes Junior *et al.* [4].

3 Materials and methods

The methodology applied for anomaly detection in oil well operations involves five distinct stages, starting with data collection and preprocessing, followed by feature extraction, and finally, classification and performance evaluation. The primary objective of this study is to compare two important event detection techniques and evaluate their performances: the first involves the direct application of data to the LOF model, and the second involves reducing data dimensionality using autoencoders before applying it to the LOF.

In this study, it has been used data from the 3W dataset, firstly published in Vargas [2]. In a recent update on July 25, 2024, various configurations were altered, resulting in the new version called 3W 2.0. This new version is divided into 2,228 instances of time series of oil well production. The instances are classified into operations under normal conditions and anomalies, with the latter now organized into nine categories. The anomalies are categorized according to their origin: historical real, simulated, or manually designed, covering 42 oil wells and 27 variables, with data from sensors and valves over time. Table 1 presents the quantities of instances that make up the 3W dataset after its update.

However, since the experiments in this study aim to evaluate the model's performance in real well cases, without any interference from simulated or manually designed data, as analyzed by Vargas [2], only historical real data were considered for training and testing the models, totaling 1,119 instances.

Table 1. 3W 2.0 dataset instances

Instance Type	Real	Simulated	Hand-Draw	TOTAL
0 - Normal Operation	594	0	0	594
1 - Abrupt Increase of BSW	4	114	10	128
2 - Spurious DHSV Closure	22	16	0	38
3 - Severe Intermittence	32	74	0	106
4 - Flow Instability	343	0	0	343
5 - Rapid Productivity Loss	11	439	0	450
6 - Rapid Restriction in CKP	6	215	0	221
7 - Scaling in CKP	36	0	10	46
8 - Hydrate in Production Line	14	81	0	95
9 - Hydrate in Service Line	57	150	0	207
TOTAL	1,119	1,089	20	2,228

After the data collection and separation phase, preprocessing was carried out, which included exploratory data analysis. In this stage, visualizations were generated to identify patterns and eliminate missing values, ensuring the uniformity and representativeness of the database. Based on this initial analysis, it was decided to consider only 4 of the 27 sensors present in the well database for model training and testing: P_PDG, P_TPT, T_TPT, P-MON-CKP. Additionally, the feature CLASS was used to determine the performance metrics of fit on test results. This selection was made to simplify the model and reduce its computational time by avoiding the consideration of features not influencing the detection of unexpected events, which often remained constant.

Although this study analyzes two different LOF-based approaches, the data preprocessing procedures were standardized for both cases, so that the model's categorization could detect deviations in well behavior. This standardization ensures a fair comparison since both approaches used the same data treatment.

For both scenarios, 60% of the normal data from each instance was used for training, and the remaining normal data was used for validation. Anomalous periods were only used in the testing phase of the models. Additionally, a sliding window technique was employed solely during model validation, so that training was done with 60% of the normal data and validation occurred within temporal windows.

The data was normalized using the MinMaxScaler, a preprocessing technique that adjusts the values of each feature to a specific range, typically between 0 and 1. The MinMaxScaler was applied using the scikit-learn (sklearn) library (Pedregosa *et al.* [10]), with fit_transform on the training set and the transform method was applied to the test set, using the same scaling parameters defined from the training set. Additionally, a noise value of 0.05% of the data mean was applied to eliminate potential artifacts or small fluctuations that could be incorrectly interpreted as anomalies, thereby contributing to the robustness and reliability of the anomaly detection model.

3.1 Proposed approach

Two scenarios were studied for analysis and comparison with other anomaly detection models present in the literature.

- In Scenario 1, the 3W database was used after treatment and preprocessing, with the data being fed directly into the LOF detection model.
- In Scenario 2, the dataset was initially applied to the convolutional autoencoder model with the goal of generating new reconstructed input data that closely resembles the original data but with reduced dimensions, which were then applied to the LOF. This approach allows the autoencoder to significantly reduce computational time and simplify the model by effectively lowering the data dimensionality.

This approach aims to improve prediction results, as the combination of autoencoders with machine learning models has proven effective in identifying temporal patterns, as demonstrated by previously mentioned studies.

The proposed model in this work for Scenario 1 is the LOF, a simplified model that mainly relies on the optimization of hyperparameters to achieve the best possible fit. After preprocessing the data, the Grid Search function from the scikit-learn library was applied, which performs cross-validation to evaluate the performance of each hyperparameter combination and select the best configuration. After analyzing the most relevant hyperparameters, shown in Tab. 2, the values that achieved the best results were adopted.

Table 2. Parameters analyzed using the Grid Search function by Pedregosa *et al.* [10] (best results in bold).

Model	N_neighbors	Life_size	Metric distance
Local Outlier Factor	5; 15; **25;** 50; 100	5; **10;** 15; 20	**Euclidean;** Minkowski; Manhattan; Hamming

Thus, the hyperparameters adopted for the model were defined as follows: n_neighbors as 25, metric as euclidean for distance calculation, with a sliding window of 10 seconds.

Thus, after training the model (LOF) using the scikit-learn library, the testing phase was conducted. Each test instance was classified as 0 for normal instances and 1 for anomalies, based on the model's predictions. Subsequently, to verify the effectiveness of the model, its predictions were compared with the "CLASS" column present in the well database, which indicates its condition at each point in time.

In constructing the second detection model, a convolutional autoencoder was employed. This convolutional autoencoder consists of two encoders layers, with 32 filters in the first layer and 16 in the second, each with a kernel size of 7, "same" padding to maintain dimensionality, and strides of 2 to reduce data dimensionality, using Rectified Linear Unit (ReLU) activation to capture the main features. The model then includes two decoder layers that reconstruct the data, reversing the convolution process, with the same filter and kernel size characteristics. However, the final layer uses a sigmoid activation to ensure that the reconstructed outputs remain within the same range as the input data.

The model was trained with the Adam optimizer and the mean squared error (MSE) loss function for 50 epochs, with a batch size of 32, employing an early stopping mechanism to avoid overfitting. After training, the encoding part of the autoencoder was used to extract reduced representations of the data. These latent representations serve as input for the training of the Local Outlier Factor algorithm, which uses this compact and essential representation of the data to identify anomalies, reducing the computational complexity of the process. Thus, autoencoders are used for data preprocessing and dimensionality reduction of the input variables, facilitating the subsequent application in the LOF, which shares the same characteristics as in Scenario 1.

3.2 Evaluation indicators

Performance metrics are highlighted and used as a benchmark to assess the effectiveness of classification algorithms in anomaly detection. Therefore, a set of metrics was applied to compare and evaluate the unsupervised machine learning algorithm studied in this paper, based on other authors in the literature who assess classification models for anomaly detection. To do this, data from the confusion matrix were used to summarize the models' performance based on their numbers of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), which range from 0 to the total number of observations, with higher TP and TN values indicating better performance.

Thus, the performance metrics highlighted below were calculated from the confusion matrix:

- Accuracy (ACC) takes into account all normal and faulty samples and ranges from 0 to 1. ACC can be calculated using eq. (1) as follows:

$$ACC = \frac{TP+TN}{n_{total}}. \tag{1}$$

- Precision (PR) refers to the ratio of true positives to false negatives and ranges from 0 to 1. PR can be calculated using eq. (2) as follows:

$$PR = \frac{TP}{TP-FP}. \tag{2}$$

- Sensitivity (S), also known as the true positive rate or recall, assesses the ability to correctly detect positive instances in a dataset. It can be calculated using eq. (3) as follows:

$$S = \frac{TP}{TP+FN}. \tag{3}$$

- The F1 score is a harmonic mean between S and PR, and can be estimated using eq. (4) as follows:

$$F1 = \frac{2 \times P \times S}{P+S}. \tag{4}$$

Based on the performance evaluation metrics applied by Vargas [2], the performance assessment of the detection model in this study will focus on precision and sensitivity. Thus, the F-Measure was chosen as the

primary performance metric for evaluating the models.

4 Results and dissemination

The following results are the evaluations of the two models proposed in this study using the public database provided by Vargas [2]. The models were evaluated based on their respective metric values for all 9 cases of anomalies present in the database.

Additionally, the experiments were conducted following certain rules established in the benchmark for anomaly detection proposed by Vargas [2]:

- Only instances with normality periods longer than 20 minutes were used in the experiment. As a result, classes 3-Severe Intermittence and 4-Flow Instability were not tested because they only contain periods of anomalies.
- The number of training and testing rounds is equal to the number of instances, which, according to Tab. 2, resulted in approximately 150 cases of 7 anomalies.
- In each round, the performance metrics of the models, including F1 Score, accuracy, and precision, were computed, and the average of these metrics is presented for model comparison and comparison with previous works.

Thus, the two models studied and their respective results can be observed in Tab. 3 and Tab. 4, initially containing the values of F1 Score, accuracy, and precision, as well as the sum of False Positives (FP), False Negatives (FN), True Positives (TP), and True Negatives (TN) for each instance separated by anomalies. Subsequently, the average F1 Score for each instance of each anomaly is presented for comparisons with previous works presented in Tab. 5.

Table 3. Performance metrics per model studied and per anomaly.

Anomalies and Models		TN	TP	FP	FN	F1	ACC	PR	Time
1-Abrupt Increase of BSW	LOF	145648	51749	471	23050	0.8148	0.8935	0.9910	91 min
	AE+ LOF	146560	71755	152	2297	0.9842	0.9889	0.9979	25 min
2-Spurious DHSV Closure	LOF	81863	103430	95	6935	0.9676	0.9634	0.9991	35 min
	AE+ LOF	77482	104286	4704	3170	0.9850	0.9585	0.9568	15 min
5-Rapid Productivity Loss	LOF	32048	114931	36	35113	0.8674	0.8070	0.9997	107 min
	AE+ LOF	32084	113318	0	36242	0.8621	0.8005	1.0000	39 min
6-Rapid Restriction in CKP	LOF	30208	0	0	18331	0.0000	0.6223	0.0000	20 min
	AE+ LOF	30174	14772	2109	2829	0.8568	0.9010	0.8751	4 min
7-Scaling in CKP	LOF	94392	116755	285	189561	0.5516	0.5266	0.9976	80 min
	AE+ LOF	80303	242534	14374	63304	0.8620	0.8061	0.9441	33 min
8-Hydrate in Production Line	LOF	245426	703968	986	14368	0.9892	0.9841	0.9986	491 min
	AE+ LOF	234888	671259	11524	46586	0.9585	0.9397	0.9831	194 min
9 - Hydrate in Service Line	LOF	72538	108331	197	14071	0.9382	0.9269	0.9982	50 min
	AE+ LOF	67649	70573	5086	51560	0.7136	0.7093	0.9328	22 min
TOTAL	LOF	702123	1199164	2070	301429	0.8877	0.8623	0.9983	874 min
	AE+ LOF	669140	1288497	37949	205988	0.9135	0.8892	0.9714	330 min

The results of the two models analyzed in this study showed satisfactory performance in detecting important events for the seven classes present in the 3W database.

The LOF model showed consistent results. In previous studies, LOF achieved an F1 score of 0.870 for the best instance in Fernandes Junior *et al.* [4] and 0.9969 for class 2 cases, as studied by Aranha *et al.* [9]. However, in this study, the LOF results were notable for the following anomalies: Anomaly 2 (Spurious DHSV Closure): F1 score of 0.9676, Anomaly 8 (Hydrate in Production Line): F1 score of 0.9892, Anomaly 9 (Hydrate in Service

Line): F1 score of 0.9382, Anomaly 1 (Abrupt Increase of BSW): F1 score of 0.8148 e Anomaly 5 (Rapid Productivity Loss): F1 score of 0.8674. These results highlight the effectiveness of LOF in detecting important events. However, it showed moderate performance for Anomaly 7 (Scaling in CKP) with an F1 score of 0.5516 and completely failed to detect anomalies in class 6 (Rapid Restriction in CKP), where it could not identify any abnormalities in the data.

In contrast, the Convolutional Autoencoder + LOF model outperformed the LOF model in four of the seven studied anomalies, demonstrating robust and consistent performance. Its highest F1 score was observed for Anomaly 1 (Abrupt Increase of BSW), with a value of 0.9842. On the other hand, the LOF model achieved the best results in Anomalies 5 (Rapid Productivity Loss), 8 (Hydrate in Production Line), and 9 (Hydrate in Service Line), proving effective in detecting these specific classes of anomalies.

However, another important metric in the event detection process is the computational time required for model processing. In this aspect, the Convolutional Autoencoder + LOF model significantly stood out. The total processing time for the autoencoder was 330 minutes, compared to 874 minutes for the LOF model, representing an approximate reduction of 62.22% in computational time.

Additionally, the Convolutional Autoencoder + LOF model also showed a superior global F1 score for the seven anomalies, achieving 0.9135 compared to 0.8877 obtained by the LOF model. These results indicate that the Convolutional Autoencoder + LOF not only offers superior performance in terms of precision for various anomalies but is also more efficient in terms of computational time, making it a more effective choice for anomaly detection in complex databases

Table 4. Means and standard deviation of the metrics considered, by algorithm.

Anomalies and Models		Mean F1	STD
1-Abrupt Increase of BSW	AE + LOF	0.9845	0.0078
	LOF	0.8887	0.1569
2-Spurious DHSV Closure	AE + LOF	0.9572	0.0901
	LOF	0.9479	0.2069
5-Rapid Productivity Loss	AE+ LOF	0.8468	0.0766
	LOF	0.7397	0.7828
6-Rapid Restriction in CKP	AE+ LOF	0.6883	0.2116
	LOF	0.0000	0.0000
7-Scaling in CKP	AE+ LOF	0.7968	0.1250
	LOF	0.6980	0.3944
8-Hydrate in Production Line	AE+ LOF	0.9538	0.0344
	LOF	0.9865	0.0078
9 - Hydrate in Service Line	AE+ LOF	0.6578	0.1221
	LOF	0.9312	0.0661
TOTAL	AE+ LOF	0.8520	0.0954
	LOF	0.7417	0.2307

Table 5. Comparison with related work.

Author	Analyzed event	Model	Best F1 score (STD)
Vargas [2]	All anomalies with mean and standard deviation of metrics	Isolation Forest	0.727 (0.182)
Fernandes Junior *et al*. [4]	All anomalies with mean and standard deviation of metrics	LOF	0.870 (0.14)
		Autoencoder	0.590 (0.14)
This Paper	All anomalies with mean and standard deviation of metrics	LOF	0.9865 (0.0078)
		Autoencoder+LOF	0.9845 (0.0078)

According to Tab. 4 and Tab. 5, both the Local Outlier Factor (LOF) and the combined Convolutional Autoencoder + LOF models proposed in this study outperformed the models from previous works. The LOF achieved an F1 Score of 0.9865, while the Convolutional Autoencoder + LOF reached an F1 Score of 0.9845. These results are notably superior to those of the Isolation Forest by Vargas [2] and the autoencoder by Fernandes Junior *et al*. [4], indicating the effectiveness of the proposed models for anomaly detection in the analyzed classes.

5 Conclusions

This study conducts a quantitative comparative analysis of anomaly detection techniques in offshore oil production wells, using a public database provided by Vargas [2]. Two distinct scenarios for detecting significant events were investigated: the first using the single-class classifier, Local Outlier Factor, and the second combining Convolutional autoencoder with Local Outlier Factor.

The experiments were conducted by applying these algorithms to the dataset, using seven different classes of anomalies for training and validating the models. The main evaluation metric was the F1 score, used to assess the effectiveness of the models in detecting anomalies. Statistical test results indicated that the Convolutional Autoencoder + LOF combination performed better compared to the model in scenario 1, which used only LOF. Specifically, the Autoencoder + LOF model achieved an average F1 score of 0.8520 across the seven anomaly classes and an F1 score of 0.9155 considering the total amount of data tested.

These results suggest that the Convolutional Autoencoder + LOF model is more suitable in terms of precision for various anomalies and more efficient in terms of computational time, with a reduction of approximately 62.22% compared to scenario 1. Although scenario 1 presented slightly lower results, both models can be considered robust and efficient tools for the analysis and detection of anomalies in complex databases.

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] E.S.P. Sobrinho; F. L. Oliveira; J.L.R. Anjos; C. Gonçalves; M.V.D. Ferreira; L.G.O. Lopes; W.W.M. Lira; J.P.N. Araújo; T.B. Silva; L.P. Gouveia. Uma ferramenta para detectar anomalias de produção utilizando aprendizagem profunda e árvore de decisão. Rio Oil & Gas Expo and Conference, 2020.
[2] R.E.V. Vargas. Base de dados e benchmarks para prognóstico de anomalias em sistemas de elevação de petróleo. Doctoral thesis, Universidade Federal do Espírito Santo, 2019.
[3] R.S.F. Nascimento. Detecção de anomalias em poços de produção de petróleo offshore com a utilização de autoencoders e técnicas de reconhecimento de padrões. Masters dissertation, Universidade Federal de Lavras, 2021.
[4] W. Fernandes Junior; K. S. Komati; K. A. S. Gazolli. Anomaly detection in oil-producing wells: a comparative study of one-class classifiers in a multivariate time series dataset. J Petrol Explor Prod Technol 14, p. 343–363, 2024.
[5] M. M. Breunig et al. Lof: identifying density-based local outliers. In: Acm Sigmod International Conference on Management Of Data. Proceedings… [S.l.: s.n.]. p. 93–104, 2000.
[6] S. Misra; H. Li; J. He. Machine Learning for Subsurface Characterization. [S.l.]: Elsevier, 2020.
[7] J. Chen et al. Outlier detection with autoencoder ensembles. In: Siam International Conference on Data Mining. Proceedings… [S.l.]: SIAM. p. 90–98, 2017.
[8] P. E. Aranha; L. G. O. L. Lopes; E. Paranhos Sobrinho; I. de M. N. Oliveira; J. P. de Araújo; B. B. dos Santos; et al. A system to detect oilwell anomalies using deep learning and decision diagram dual approach. SPE Journal, 29 (3), p. 1540-1553, 2024.
[9] P. E. Aranha; N. A. Policarpo; M.A. Sampaio. Unsupervised machine learning model for predicting anomalies in subsurface safety valves and application in offshore wells during oil production. J Petrol Explor Prod Technol 14, p. 567–581, 2024.
[10] F. Pedregosa et al. Scikit-learn: machine learning in python. Journal of Machine Learning Research, v. 12, p. 2825–2830, 2011.