Advancing Anomaly Detection in Oil Production Wells with TranAD: A Deep Transformer Network Approach

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

[1]*Federal University of Alagoas, Av. Lourival Melo Mota, S/N, Tabuleiro do Martins, Maceió - AL, 57072-970*
igornery@lccv.ufal.br, thalesv@gmail.com, andressa.silva@ctec.ufal.br, davi.ramos@lccv.ufal.br, limajunior@lccv.ufal.br,

[2]*Petrobras, R. Marquês de Herval, 58-78, Valongo, Santos - SP, 11010-310*
pearanha@petrobras.com.br

Abstract. The oil and gas industry has been leveraging cutting-edge technologies, such as artificial intelligence and the Internet of Things, for well integrity analysis, aiming to enhance operational safety and reduce production losses. Timely detection of unexpected events through the well lifecycle, such as spurious closure of Downhole Safety Valves (DHSV) and rapid productivity loss events, is crucial. The integration of sensor-based monitoring and computational modeling provides vital insights for identifying and mitigating such anomalies, thereby bolstering industry's reliability and sustainability. However, the anomaly detection in oil and gas wells faces significant challenges due to the highly unbalanced nature of historical data, with few unexpected events, and the frequent valve changes that can disrupt pressure and temperature patterns, complicating unsupervised techniques. This paper utilizes TranAD, a deep transformer network-based multivariate time-series model that relies on attention-based sequence encoders to detect anomalies solely based on non-anomalous training data. It is assessed the effectiveness of TranAD in detecting anomalies using the 3W database, a public repository released by Petrobras containing real-world undesirable events in oil and gas wells. TranAD models trained on 3W are compared with benchmark techniques applied to this dataset, yielding promising results due to its data and time-efficient training strategy.

Keywords: anomaly detection, TranAD, oil wells.

1 Introduction

The oil and gas industry is undergoing a profound transformation, increasingly relying on cutting-edge technologies such as artificial intelligence, cloud computing, and the Internet of Things to enhance operational efficiency and safety. Within this evolving landscape, maintaining the integrity of oil production wells is crucial for ensuring operational safety, preserving the environment, and minimizing production losses. Detecting unexpected events—particularly anomalies like spurious closures of Downhole Safety Valves (DHSV) and rapid productivity loss events—requires timely intervention to mitigate risks and safeguard the industry's sustainability.

Sensor-based monitoring, combined with computational modeling, plays a vital role in providing insights necessary for identifying and addressing these anomalies. However, anomaly detection in oil production wells presents significant challenges. Historical data from wells tends to be highly unbalanced, with rare occurrences of unexpected events. Additionally, routine operations, such as valve changes, can substantially alter pressure and temperature behavior, complicating the detection process and potentially confounding unsupervised techniques.

To address these challenges, machine learning techniques have been increasingly applied in well production monitoring. These techniques aim to detect specific events that could signal potential problems, thereby improving preventive measures and enabling more proactive interventions. Various models have been explored in this context, each with its own set of advantages and limitations. For instance, Decision Trees (DT) have been recognized for their interpretability and performance, as highlighted by Alharbi et al. [1], while more complex classifiers like Random Forest (RF) and Adaptive Boosting (AdaBoost) have shown varying degrees of success in well production anomaly detection, as discussed by Turan and Jaschke [2]. The field of novelty detection in time series, often referred to as anomaly detection, remains a challenging but critical area of research due to its direct implications for real-time applications in the industry.

In this context, this paper evaluates TranAD (Tuli et al. [3]), a deep transformer network-based model designed for multivariate time-series anomaly detection. TranAD employs attention-based sequence encoders to detect anomalies using non-anomalous training data, which is particularly effective for highly unbalanced datasets with rare, unexpected events. Unlike traditional methods that may struggle with the non-linear and dynamic nature of sensor data, TranAD's architecture models complex interactions between multiple variables over time, effectively addressing subtle anomalies intertwined with normal operational fluctuations. This study uses the 3W dataset (Vargas et al. [4]) —Petrobras's first public repository of real-world undesirable events in oil wells—to assess TranAD's effectiveness, comparing it with established benchmark techniques to validate its efficacy and demonstrate its potential in advancing anomaly detection methodologies within the oil and gas industry.

2 Overview of the TranAD Model and the 3W Dataset

This paper evaluates the effectiveness of a deep transformer network, TranAD, using a public dataset. Both TranAD and the dataset are well-documented in their respective literature. This section provides an overview of TranAD's strengths and its application in various case studies. Additionally, it delves into the intricacies of the 3W dataset, highlighting the complexities and challenges inherent in real-world data, such as incompleteness, data imbalance, and feature variability.

2.1 TranAD overview

TranAD is a deep transformer network specifically designed for anomaly detection in multivariate time-series data developed by Tuli et al. [3], providing a robust solution to the complexities of modern industrial datasets. Leveraging the transformer architecture, TranAD captures long-range dependencies and contextual information within time-series data, addressing the limitations of traditional methods like ARIMA and LSTM-based models, which often struggle with high data volatility and computational inefficiency.

A standout feature of TranAD is its use of the Peak Over Threshold (POT) method for anomaly identification. The POT method dynamically selects threshold values by fitting the data distribution with a Generalized Pareto Distribution (GPD), focusing on the most extreme values in the data sequence. This approach is particularly effective in distinguishing between normal and anomalous behavior, even when anomalies are subtle. TranAD not only outputs one result per anomaly but also provides individual results for each feature. It then leverages these feature-level results to generate a comprehensive anomaly detection outcome, enhancing its precision in real-world applications where quick and accurate anomaly detection is crucial.

Another key strength of TranAD is its use of self-attention mechanisms, which allow the model to weigh the importance of different time steps, thereby improving its ability to detect context-dependent anomalies. TranAD's adversarial training process further amplifies reconstruction errors, making the model more sensitive to minor deviations that might indicate anomalies. This approach not only improves detection accuracy but also ensures efficient training, even when trained predominantly on non-anomalous data. This focus on non-anomalous training helps TranAD maintain high performance even in scenarios where labeled anomalous data is scarce.

Empirical studies have demonstrated TranAD's superiority over state-of-the-art models, with up to a 17% increase in F1-score and a 99% reduction in training times compared to traditional methods. For instance, on the Server Machine Dataset (SMD), TranAD outperformed other models, particularly in detecting anomalies that closely resemble normal behavior—a common challenge in industrial applications. This makes TranAD

particularly well-suited for complex, real-world datasets like the 3W dataset, where data incompleteness, imbalance, and high dimensionality pose significant challenges.

2.2 3W dataset overview

The 3W dataset is a unique and realistic public dataset designed to address the complexities of detecting undesirable events in offshore naturally flowing oil wells. It embodies several inherent challenges typical of real-world industrial data, such as data incompleteness, feature variability, and event rarity. These characteristics mirror the genuine conditions encountered in oil production monitoring, where data is often imperfect and unpredictable. This realism is intentional, as highlighted in the original paper, enabling the evaluation of various preprocessing techniques and their effectiveness in improving model performance across different tasks. By preserving the raw nature of the data—including missing values, frozen variables, and outliers—the 3W dataset provides a robust benchmark for developing and testing advanced anomaly detection methods that must operate effectively in realistic, noisy environments.

Specifically, one of the primary challenges of the 3W dataset is incompleteness. The dataset contains a significant number of missing values due to sensor malfunctions or communication issues in the hostile offshore environment. These missing data points can lead to sparsity, complicating the modeling process and requiring robust imputation or handling strategies to avoid biased results. Another complexity is the data imbalance. The 3W dataset includes instances of eight different types of undesirable events, but these anomalies are rare compared to the normal operating conditions. To mitigate this, the dataset includes not only real events but also simulated and hand-drawn instances to enrich the available data. However, the imbalance between normal and anomalous instances still presents a significant challenge for model training, often leading to models that are biased toward normal conditions.

In terms of structure, the first release of the 3W dataset comprises 1,984 instances, each containing a complete contiguous time-series set of observations. These instances are derived from real, simulated, and hand-drawn sources, capturing data from 21 different wells and representing various operational conditions. The dataset includes 8 process variables (features), such as pressures and temperatures at different points in the production system. The instances encompass 8 different types of undesirable events, such as spurious closures of DHSV and severe slugging, in addition to non-anomalous data. This is visually illustrated in Fig 1, from the original paper, which presents a scatter map showing the temporal positioning of the gathered data from the 21 different wells, with instances color-coded to distinguish between the 8 types of anomalous events and the non-anomalous data. The files (instances) vary in size (number of observations) due to the different lengths of time-series data they contain, reflecting the real-world variability in monitoring periods and event durations. This variability, combined with the presence of feature variability—where key process variables exhibit diverse behaviors depending on the operational state of the well—adds another layer of complexity to the analysis. Moreover, some variables may be either frozen (showing constant values due to sensor failures) or exhibit outlier behaviors, further complicating the data processing.

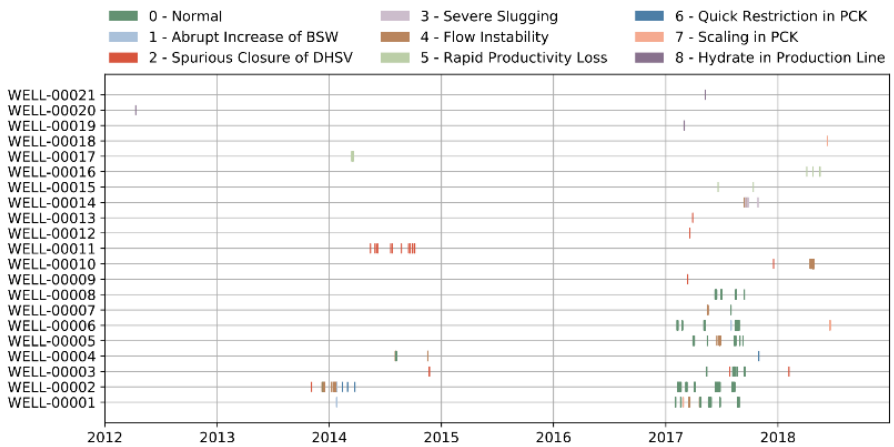

Figure 1. Scatter map of real instances of the 3W dataset, adapted from Vargas et al. [4]

Several studies have explored anomaly detection on the 3W dataset using various machine learning techniques. Vargas et al. [4] benchmarked models like Isolation Forest and One-Class SVM, with the former achieving an F1-score of 0.727. Fernandes Jr. et al. [5] further enhanced performance by testing classifiers such as Local Outlier Factor (LOF), which achieved an F1-score of 0.870 with feature extraction. Marins et al. [6] used Random Forests for classifying fault classes, obtaining a 97.1% accuracy, while Turan and Jaschke [2] applied a sliding window approach, where Decision Trees reached an F1-score of 85% in multi-class classification. These studies collectively demonstrate the effectiveness and challenges of various models in detecting anomalies within the 3W dataset.

3 Implementation details

In this section, we detail the specific steps and decisions made during the implementation of the TranAD model on the 3W dataset. Given the complexity of time-series data and the unique challenges presented by the 3W dataset, careful consideration was given to both data preprocessing and the analysis of the model's output. Our methodology is divided into two main parts: the preprocessing of the raw data to ensure its suitability for anomaly detection and the analysis of the results generated by the TranAD model.

3.1 Data Preprocessing of the 3W Dataset

As this paper focuses on benchmarking TranAD's ability to detect anomalies in the context of oil production wells, only real cases from the 3W dataset were used for this initial analysis. In total, there are 1,019 real instances used in this paper. The real instances in the non-anomalous dataset (event 0) are from wells 1 to 8, totaling 594 instances. In contrast, the anomalous instances (events 1 to 8) are distributed across these wells with 259 instances and an additional 166 instances in wells 9 to 18. Wells 19 to 21 and event 8 (hydrate in the production line) only contain simulated and hand-drawn instances, so they were omitted.

In this context, two methodologies were applied: In the first, a local model was trained using the non-anomalous data from each well individually and tested on the corresponding anomalous data from the same well. In the second approach, a global model was trained using all the non-anomalous data and then tested across all wells.

Feature selection was based on four key sensors: P-PDG, P-TPT, T-TPT, and P-MON-PCK, chosen according to the cost-benefit analysis provided by Vargas et al [4]. Approximately 2% of faulty data were removed, and some instances with extensive NaN values were excluded. However, instances with frozen data, such as P-PDG in wells 1, 2, 4, and 5, and P-MON-PCK in well 8, were retained to ensure there was sufficient training data.

In neural network models, feature scaling is essential to ensure that no single feature disproportionately influences the model's learning process. Without scaling, features with larger ranges can dominate, leading to biased predictions. Neural networks are particularly sensitive to the scale of input data, making it critical to normalize features so they contribute equally to the model's training. As highlighted by Goodfellow et al. [7] and Bishop [8], proper scaling enhances model performance and mitigates bias. To address this, Min-Max scaling was applied, with scaling done by time segments within wells for non-anomalous data and on a per-instance basis for anomalous data due to their complexity and variability.

3.2 Model Training Optimization

In the training phase, it was observed that 21 epochs provided an optimal balance between underfitting and overfitting for this type of data. This choice was made to ensure that the model learned effectively without overfitting to the specificities of the training set, which could diminish its generalizability.

3.3 Performance Evaluation Metrics

For evaluating the model's performance, some key metrics were used, while the F1-score was the mandatory metric. A brief explanation of those metrics is presented below, but a more extensive context can be found on Bishop [8] and Goodfellow et al. [7].

- Confusion Matrix: This matrix summarizes the model's performance by displaying the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). These values range from 0 to the total number of observations, with higher TP and TN values indicating better performance.

- Precision: Measures the accuracy of positive predictions and ranges from 0 to 1, where a higher value indicates fewer false positives. Precision is calculated as:

$$precision = \frac{TP}{TP+FP}. \tag{1}$$

- Recall: Recall, also known as True Positive Rate, measures the ability of the model to correctly identify positive instances. Like precision, its value ranges from 0 to 1, with higher values indicating better detection of true positives. Recall is calculated as:

$$recall = TPR = \frac{TP}{TP+FN}. \tag{2}$$

- Specificity: Specificity, also known as True Negative Rate, is the ratio of correctly predicted negative observations to all actual negatives. It's value ranges from 0 to 1, with higher values indicating better detection of negative values, that is the correct labelling of non-positives. Specificity is calculated as:

$$specifity = TNR = \frac{TN}{TN+FP}. \tag{3}$$

- F1-score: This metric is the harmonic mean of precision and recall, providing a single metric that balances both. It is calculated as:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{4}$$

  and ranges from 0 to 1. The F1-score is particularly important in this application, given the imbalanced nature of the data, where anomalies are much rarer than normal events. It effectively combines precision and recall into a single metric that accounts for both false positives and false negatives, ensuring that the model detects anomalies accurately while minimizing false alarms. This balance is crucial for maintaining the reliability and relevance of the anomaly detection system in an operational context.

- Balanced Accuracy (ACCb): Balanced accuracy is the average of the True Positive Rate (recall) and the True Negative Rate (specifity). It is calculated as follows:

$$ACCb = \frac{1}{2} \times (TPR + TNR) = \frac{1}{2} \times \left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{5}$$

4 Results and analysis

This section breaks down how the TranAD model performed in detecting anomalies in oil production wells. It starts by examining the results for individual wells, analyzing how the model handled different situations. Following that, the analysis broadens to assess the model's overall performance across all wells, highlighting both its strengths and limitations.

4.1 Well-specific model training results

Due to the absence of real instances for event 8 and for event 5 across wells 1 to 8, only events 1, 2, 3, 4, 6, and 7 were included in the well-specific model training. After processing the data, a total of 259 instances were tested, the compiled results are presented in Tab. 1. The last row of totals shows the count of instances across events, with TN, TP, FP, and FN also aggregated. The overall F1-score and ACCb are weighted means based on the instance count. By selecting the best F1-score feature-wise for each test, the average F1-score across all tests was 0.983, with a balanced accuracy of 0.985. Notably, across every event, the feature 'P-MON-CKP' consistently produced the highest F1-score.

Table 1. Compiled results of well-specific model training

Event	Instances	TN	TP	FP	FN	F1	ACCb	Feature
1	5	168,818	84,764	26,558	31,996	0.692	0.720	P-MON-CKP
2	4	3,756	40,771	0	0	1.000	1.000	P-MON-CKP
3	1	0	17,975	0	1	1.000	1.000	P-MON-CKP
4	240	0	1,718,835	0	113	1.000	1.000	P-MON-CKP
6	6	28,554	20,258	7,765	1	0.877	0.907	P-MON-CKP
7	3	107,771	39,282	0	231,135	0.333	0.333	P-MON-CKP
TOTAL	259	308,899	1,921,885	34,323	263,246	0.983	0.985	P-MON-CKP

Despite the generally strong performance metrics, event 1 (Abrupt Increase of BSW) performed slightly worse than the other events, while event 7 (Scaling in PCK) suggests that TranAD may not be well-suited for detecting this type of anomaly. It's important to note that, except for event 4 (Flow Instability), which had 240 test instances, the number of instances tested was relatively small, contributing to the concise performance metrics.

4.2 Single training results

A single model training was conducted using all non-anomalous data. A total of 574 instances were utilized for training, excluding 20 instances from well 6 during preprocessing. For testing, 411 instances were used, 259 of which were previously tested in the well-specific model, and 152 were new test instances from wells 9, 10, 14, 15, 16, 17, and 18. Some wells did not contain real data, and 14 anomalous instances were excluded during preprocessing. The compiled results are presented in Tab. 2, constructed similarly to Tab. 1. The F1-score and ACCb metrics showed a decline, with respective values of 0.797 and 0.798, reflecting a 19% drop compared to the first methodology.

Table 2. Compiled results of single training broader model

Event	Instances	TN	TP	FP	FN	F1	ACCb	Feature
1	5	195,376	51,052	0	65,708	0.400	0.400	P-MON-CKP
2	8	0	61,780	13,184	2	0.870	0.904	T-TPT
3	32	0	568,405	0	30	1.000	1.000	P-MON-CKP
4	344	0	1,984,536	0	475,899	0.808	0.808	P-MON-CKP
5	11	95,297	282,534	32,633	140,424	0.686	0.716	P-MON-CKP
6	6	36,319	9,352	0	10,907	0.167	0.167	P-MON-CKP
7	5	152,039	0	0	313,937	0.000	0.000	-
TOTAL	411	479,031	2,957,659	45,817	1,006,907	0.797	0.798	P-MON-CKP

The worst performance was observed in event 7 (Scaling in PCK), where no anomalies were identified, resulting in a zero cumulative sum of positive values (both TP and FP). Event 6 (Quick Restriction in PCK) also performed poorly, with both metrics at 0.167, and event 1 (Abrupt Increase of BSW) had metrics at 0.400. The feature 'P-MON-CKP' remained the most effective for anomaly detection in most events. The exception was event 2 (Spurious Closure of DHSV), where 'T-TPT' emerged as the best feature, showing strong metrics.

4.3 Comparative Analysis with Related Work

To better understand the effectiveness of the TranAD model applied in this study, it is essential to compare its performance with previous research conducted on the 3W dataset. Various studies have explored different machine learning techniques to detect and classify anomalies within oil production wells, each focusing on specific types of events and models.

As summarized in Tab. 3, Vargas et al. [4] benchmarked traditional models like Isolation Forest across all eight types of anomaly events, achieving an F1-score of 0.727. Fernandes Jr. et al. [5] improved detection performance using LOF, with an F1-score of 0.870, while their Autoencoder model performed less effectively. Marins et al. [6] employed Random Forests across all events, reporting from 97.1% to 99.0% balanced accuracy, demonstrating the robustness of ensemble methods. Similarly, Turan and Jaschke [2] used Decision Trees in a multiclass classification, achieving an F1-score of 0.921.

Table 3. Comparison with Related Work

Author	Proposed methodology	Mean (and STD) of F1-score
Vargas et al. [4]	Isolation forest	0.727 (0.1822)
Fernandes Jr. et al. [5]	Local outlier factor	0.870 (0.14)
Turan and Jaschke [2]	Decision tree multiclass classifier	0.921 (0.1027)
Marins et al. [6]	One-class classifier	0.971*
	Multiple binary classifiers	0.990* (0.0070)
	Single multiclass classifier	0.918* (0.0817)
This paper	TranAD (well-specific)	**0.983 (0.1142)**
	TranAD (single broader model)	0.797 (0.3978)

**only ACCb was provided from author Marins et al. [6]*

In comparison, the TranAD model in this study showed superior performance in well-specific training, with an F1-score of 0.983, but encountered challenges in broader generalization during single training across wells, where the F1-score dropped to 0.797. These results highlight the potential of TranAD in targeted applications, while also indicating areas for further improvement, particularly in handling diverse well conditions.

5 Conclusions

The TranAD model demonstrated high efficacy in anomaly detection when applied to well-specific models, achieving a mean F1-score of 0.983, surpassing benchmarks established by related works on the 3W dataset. However, its performance declined notably in broader model training, with an F1-score dropping to 0.797. This contrast underscores the model's sensitivity to well-specific conditions and suggests that TranAD is particularly suited for targeted anomaly detection rather than generalized applications. A key observation is the consistent performance of the 'P-MON-CKP' feature across various events, highlighting its potential as a reliable indicator in oil well anomaly detection.

Despite these promising results, the methodology's reliance on the best F1-score feature-wise selection indicates a potential bias due to data incompleteness, warranting further investigation. Future work should focus on addressing these limitations by exploring alternative transformer-based models and evaluating their performance across a broader range of scenarios and datasets, potentially incorporating advanced feature selection techniques to enhance robustness and generalization.

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] Alharbi et al., “Explainable and Interpretable Anomaly Detection Models for Production Data”. *SPE Journal*, vol. 27, pp. 349–363, 2022.
[2] E. M. Turan and J. Jaschke, "Classification of undesirable events in oil well operation." In: *2021 23rd International Conference on Process Control (PC)*, Strbske Pleso, Slovakia, pp. 157-162, 2021.
[3] Tuli et al., “TranAD: Deep Transformer Networks for Anomaly Detection in Multivariate Time Series Data”. *Proceedings of VLDB*, vol. 15, no. 6, pp. 1201-1214, 2022.
[4] Vargas et al., “A realistic and public dataset with rare undesirable real events in oil wells”. *Journal of Petroleum Science and Engineering*, vol. 181, 106223, 2019.
[5] Fernandes et al., “Anomaly detection in oil-producing wells: a comparative study of one-class classifiers in a multivariate time series dataset”. *J. Petrol. Explor. Prod. Technol.*, vol. 14, pp. 343–363, 2024.
[6] Marins et al., “Fault detection and classification in oil wells and production/service lines using random forest”. *Journal of Petroleum Science and Engineering*, vol. 197, 107879, 2021.
[7] Goodfellow et al., *Deep learning (Adaptive Computation and Machine Learning series)*. MIT Press, 2016.
[8] C. M. Bishop, *Pattern recognition and machine learning (information science and statistics).* Springer New York, 2007.