Scientific Contributions Oil & Gas, Vol. 48. No. 4, December: 117 - 131

SCIENTIFIC CONTRIBUTIONS OIL AND GAS

Testing Center for Oil and Gas
LEMIGAS

Journal Homepage:http://www.journal.lemigas.esdm.go.id
ISSN: 2089-3361, e-ISSN: 2541-0520

An LSTM-Based Anomaly Detection on Subsea Oil-Producing Well

Dara Ayuda Maharsi, Syaloom Zefanya Tampi, and Ajeng Purna Putri Oktaviani

Universitas Pertamina
Teuku Nyak Arief, South Jakarta, Indonesia.

Corresponding author: Dara Ayuda Maharsi (dara.maharsi@universitaspertamina.ac.id)

Manuscript received: October 14th, 2025; Revised: October 30th, 2025
Approved: December 02th, 2025; Available online: December 12th, 2025; Published: December 12th, 2025.

ABSTRACT - The oil and gas industry faces substantial operational risks from anomalous events, necessitating effective abnormal event management (AEM) to mitigate production losses and safety hazards. This study presents a supervised anomaly classification approach using long hort-term memory (LSTM) networks on the 3W Dataset comprising over 2,000 real, simulated, and expert-drawn events from offshore wells. Focusing on real instances with sufficient normal-state duration, the dataset was refined and segmented using observation windows of 60, 120, and 180 seconds. The models were trained on four selected pressure and temperature features and evaluated using precision, recall, and F1-score. Comparative analysis with recurrent neural network (RNN) and gated recurrent unit (GRU) models shows that the LSTM model consistently performs best, achieving a peak F1-score of 92% at a 120-second window. Furthermore, event-level performance analysis highlights the LSTM model's strengths and limitations across different anomaly types. Compared to existing supervised and unsupervised methods on the 3W Dataset, the LSTM-based approach demonstrates competitive accuracy and robustness for real-time anomaly detection in offshore oil production systems.

Keywords: anomaly detection, time-series classification, long hort-term memory (LSTM), offshore wells, 3W dataset

How to cite this article:

Dara Ayuda Maharsi, Syaloom Zefanya Tampi, and Ajeng Purna Putri Oktaviani, 2025, An LSTM-Based Anomaly Detection on Subsea Oil-Producing Well, Scientific Contributions Oil and Gas, 48 (4) pp. 117-131. DOI.org/10.29017/scog.v48i4.1819.

INTRODUCTION

The oil and gas industry is one of the industries that have a high risk of operational failures, which are frequently caused by anomalous events. The ability to classify anomalous and normal events may help avoid production losses, environmental disasters, human casualties, and even financial impacts. The most effective way to reduce nonproductive in an operation is to prevent accidents (Antipova et al., 2019). In the oil and gas industry, abnormal event management (AEM) is the process of identifying, evaluating, and restoring operations to safe and regular functions (Vargas et al., 2019; Turan & Jaschke 2021). AEM automation has been the primary focus of several works created in conjunction with the advancement of machine learning. The idea of automation is to detect anomalous patterns in real-time data from operational sensors. However, because there is not a set of characteristics or guidelines that defines them, detecting anomalous patterns is a hard task (Fernandeset al., 2024; Alrifaey et al., 2021). Hence, the capability of interpreting patterns of historical data is the key of ensuring the desired outcome of complex technical systems.

In recent years, the application of time-series-based machine learning models has gained increasing attention in petroleum engineering research. For instance, Iskandar & Kurihara (2022) employed long short-term memory (LSTM) networks to forecast reservoir performance in carbon capture, utilisation, and storage (CCUS) operations, while Ulil et al. (2025) demonstrated the predictive capability of bidirectional gated recurrent unit (Bi-GRU) models for shear-wave velocity estimation from well-log sequences. Similarly, Kanyoma et al. (2023) implemented a time-series modelling approach to optimise odorant addition in natural gas distribution systems. Although these studies primarily focus on regression or forecasting rather than classification, they highlight the growing potential of deep learning architectures in handling sequential data within oil and gas operations.

In recent years, researchers have proposed a range of machine learning approaches for time series classification. Different architectures of deep neural networks, particularly those built on Long Short-Term Memory (LSTM), have been developed and shown to be effective (Lindemann et al., 2021). LSTM networks are a type of recurrent neural network (RNN) designed to handle the vanishing gradient problem, which often plagues traditional RNNs (Hochreiter & Schmidhuber 1997). LSTM can learn long-term correlations within a sequence, enabling it to capture patterns over specific time spans and accurately model complex multivariate time-series data (Malhotra et al., 2015). Therefore, LSTM networks are effective in learning long-term dependencies in sequential data, which is crucial for modeling time-series data typical in the oil and gas industry, where various variables like pressure, temperature, and flow rates need to be monitored simultaneously.

This paper will discuss the use of LSTM in time -series classification of anomalous events with the 3W Dataset. A number of classification methods have been developed along with the advancement of machine learning. Among these methods, time-series classification has emerged as an important learning approach. Time-series classification is mainly a supervised machine learning problem designed to label multivariate series of variable length (Batal et al., 2009; Tong et al., 2022). Time-series classification differs from non-time-series classification. Time-series classification will classify based on the relationship between consecutive data points for a period, while non-time-series classification will classify without depending on either previous data or the time variable. In other words, time-series classification has the ability to learn the trend for a period of time, analyse, and classify it. In the context of anomaly classification model, the model will analyse and recognize the normal behaviour of the sequence data, and the anomaly is a sequence that deviates from the normal behaviour of the data. For comparison, this paper will classify the data over three different periods: 60 seconds, 120 seconds, and 180 seconds, allowing the model to learn how the data are correlated across these intervals.

In the middle of 2017, the first realistic and public dataset of rare undesirable events in oil wells compiled by Petrobras known as 3W Dataset was released to the public with the primary objective of supporting study on creating new automated AEM that can quickly and effectively identify and categorize eight categories of undesired events in naturally flowing offshore wells operating under normal conditions. The dataset consists of over 2,000 instances, each represented a time series from a real or simulated offshore well (Vargas et al., 2019). When it comes to the 3W Dataset, there are already several automated AEM methods in use for detecting anomalous events (Turan & Jaschke 2021; Fernandes et al., 2024; Carvalho et al., 2021; Machado et al., 2024; Machado et al., 2022; Marins et al., 2021; Figueirêdo et al., 2021). However, there remains a gap in the application of LSTM-based method in classifying events of the dataset. This paper focuses on time-series classification of anomalies with LSTM on the 3W Dataset, guided by the following research questions: (1) How effective is the LSTM-based approach in classifying anomalies in sensor data of 3W Dataset and (2) How do the results compare to other deep learning algorithms, such as RNN and GRU?

Accordingly, the following subsections outline the foundational concepts for this study, including deep-learning models for time-series classification, the LSTM architecture, the 3W Dataset, and the metrics used to assess model performance.

Long short-term memory (LSTM)

LSTM, proposed by Hochreiter et al. (1997) to tackle the vanishing gradient problem that occurs with conventional RNNs and leads to the inability to learn long-term dependencies. This problem occurs when RNN weights tend to stop changing, causing the model to prioritize recent data and overlook historical patterns. Long-term dependencies, or relationships that repeat over an extended period, are therefore difficult to learn effectively. The purpose of LSTM construction is to regulate the entire neuronal information flow (Lindemann et al., 2021; Hochreiter & Schmidhuber 1997; Malhotra et al., 2015). For this purpose, a gating mechanism is introduced to control the process of adding and deleting information from the iteratively propagated cell state. Thus, the forgetting process can be controlled, and defined memory behavior is achieved to model both short-term as well as long-term dependencies. This mechanism works through several gates within the LSTM cell. A single cell consists of three main gates: input, output, and forget with individual activations defined as sigmoid functions. All three gates together form a feedback loop, preserving gradients during training. The main benefit for sequence learning is that LSTMs, to some extent, solve the vanishing gradient problem, i.e., long-term signals remain in the memory, whereas a simple feedforward architecture is prone to vanishing gradients (Škrlj et al., 2019).

One issue common to all neural network models is that they often overfit the data. One of the most common solutions is the introduction of dropout layers at each training step, a percentage of neurons are omitted from being trained. This is used for regularization. Large success of neural networks for classification is due to their capability of learning latent relationships in the data. Moreover, LSTM is now common in predicting time series data and various practical applications, including forecasting economic time series, resulting in the creation of neural network software models that offer precise predictions (Рапаков et al., 2023).

Time series classification (TSC) with deep learning

TSC can be defined as the task of training a classifier to predict the probability distribution over the class variable (Ismail Fawaz et al., 2019). A time series is a sequence of observations recorded at successive points in time (Batal et al., 2009). Univariate time series consists of the observations of a single variable. When multiple variables are observed simultaneously, the time series becomes multivariate, capturing the relationships and interactions among these variables over time (Batal et al., 2009; Ismail Fawaz et al., 2019). The dataset is a collection of pairs, where each pair comprises a time series and its corresponding one-hot label vector. The goal of TSC is to accurately classify new, unseen time series based on the learned mapping.

Time-series classification is more complex than traditional classification tasks because the sequence of the time series variable is connected to the input object (Tong et al., 2022). This inherent order makes the classification process more challenging. Current research in time series classification, depending on the availability of data labels, primarily focuses on supervised and semi-supervised learning methods. Supervised learning methods using labeled data generally yield better performance. However, in practical scenarios, there is a vast amount of unlabeled data. To tackle this issue, semi-supervised methods use both a limited amount of labeled data and a large quantity of unlabeled data. In the context of deep learning, deep neural networks (DNNs) have shown significant promise for TSC due to their ability to learn hierarchical representations of data. These networks consist of multiple layers, each containing neurons that apply non-linear transformations to the input received from the previous layer. This layered structure allows DNNs to capture complex patterns in the time series data. Training a DNN involves initializing the network's weights, performing forward passes to compute outputs, calculating prediction loss, and iteratively updating the weights through backpropagation to minimize the loss. During testing, the model is evaluated on unseen data to measure its generalization. Metrics like accuracy are commonly used to assess the model's effectiveness. One key advantage of DNNs is their probabilistic decision-making capability, which allows for confidence measurement in predictions.

Classification performance metrics

Performance of a model is commonly measured on a benchmark dataset which carries out a simple comparative analysis on different models by a set of metrics (Blagec et al., 2020). Common metrics in giving a complete view of a system’s performance are the three standard performance indicators: precision, recall, and F1-score. In classification, there are results to compare models’ predictions to actual data: true positive (TP), the model predicts positive class correctly; false positive (FP), when it incorrectly predicts the positive class; true negative (TN), the model predicts negative class correctly; false negative (FN), the model predicts negative class falsely, as shown in Table 1. These results are essential for calculating metrics. Precision is defined as the proportion of correctly predicted positive cases among all positive predictions, while recall is defined as the probability of true positive prediction to the actual positive samples. Finally, the F1 score is the harmonic mean of precision and recall, and it seeks a balance between the two (Blagec et al., 2020). Equation 1 shows the definition of precision, recall, and F1 score metrics.

In an anomaly classification task, the recall score is crucial to be prioritized in avoiding the misclassification of anomalous data which represents an unnoticed operational failure in the system. This paper focuses on reducing the number of false negatives or the number anomalies classified as normal. For this reason, high value of recall is an important metric to focus on.

Table 1. Confusion matrix

Confusion Matrix	Classified Normal	Classified Anomaly
Actual Normal	TP	FP
Actual Anomaly	FN	TN

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{F1} = \frac{2\ \times \text{Precision}\ \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

3W dataset

3W dataset is publicly available and can be used as a benchmark for machine learning development related to inherent difficulties of actual data. This dataset, as mentioned in the introduction section, has the objective of supporting the development of automated AEM with machine learning algorithms. It was collected from offshore naturally flowing wells and includes sensors installed on the well, subsea system, and platform. The sensors on subsea system, schemed in Figure 1.

The sensors are pressure at permanent downhole gauge (P-PDG), temperature transducer (T-TPT), pressure transducer (P-TPT), while on production column there is down hole safety valve (DHSV), and on the platform there are chokes for production valve (PCK), flow, temperature, and pressure sensors. Table 1 describes eight variables of time series from sensors presented in 3W Dataset. This dataset presents three types of instances: real instances extracted from Petrobras’ actual wells during production, software-simulated instances

Table 2. Sensors in 3W Datasets

Variable	Description	Units
P-PDG	Fluid Pressure at PDG	Pa
P-TPT	Fluid Pressure at TPT	Pa
T-TPT	Fluid Temperature at TPT	°C
P-MON-PCK	Fluid Pressure upstream to PCK valve	Pa
T-JUS-PCK	Fluid Temperature downstream to PCK valve	°C
P-JUS-CKGL	Fluid Pressure downstream of gas lift	Pa
T-CKGL	Temperature downstream of gas lift choke	°C
QGL	Gas Lift flow rate	m^3/s

Table 3. Types of undesirable events in 3W Dataset

Class No.	Type of undesirable event	Real	Simulated	Hand-drawn	Total
0	Normal	597	-	-	597
1	Abrupt increase in BSW	5	114	10	129
2	Spurious closure of DHSV	22	16	-	38
3	Severe slugging	32	74	-	106
4	Flow instability	344	-	-	344
5	Rapid productivity loss	12	439	-	451
6	Quick restriction in PCK	6	215	-	221
7	Scaling in PCK	4	-	10	14
8	Hydrate in production line	3	81	-	84
TOTAL		1,025	939	20	1,984

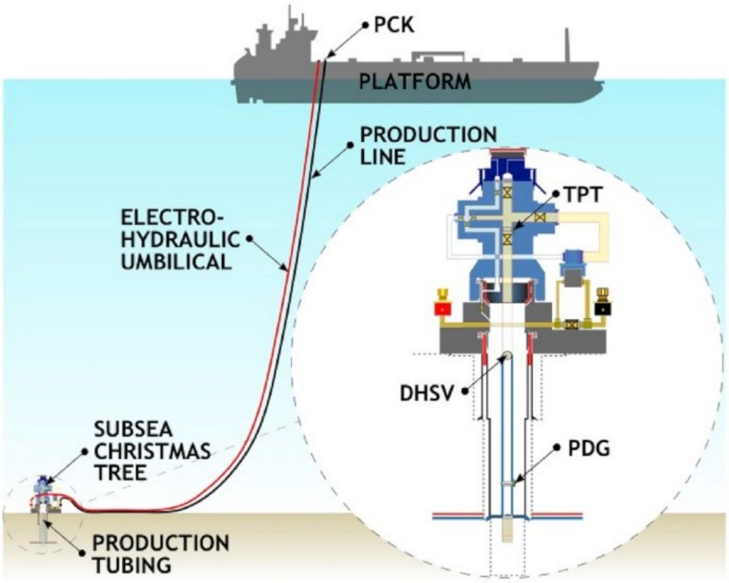

Figure 1. Schematic of the offshore production system in 3W Dataset, depicting the subsea Christmas tree, production tubing, electro-hydraulic umbilical, and production line connected to the platform. The figure also shows the locations of critical sensors for this study, including PCK, TPT, PDG, and DHSV. (Vargas et al., 2019)

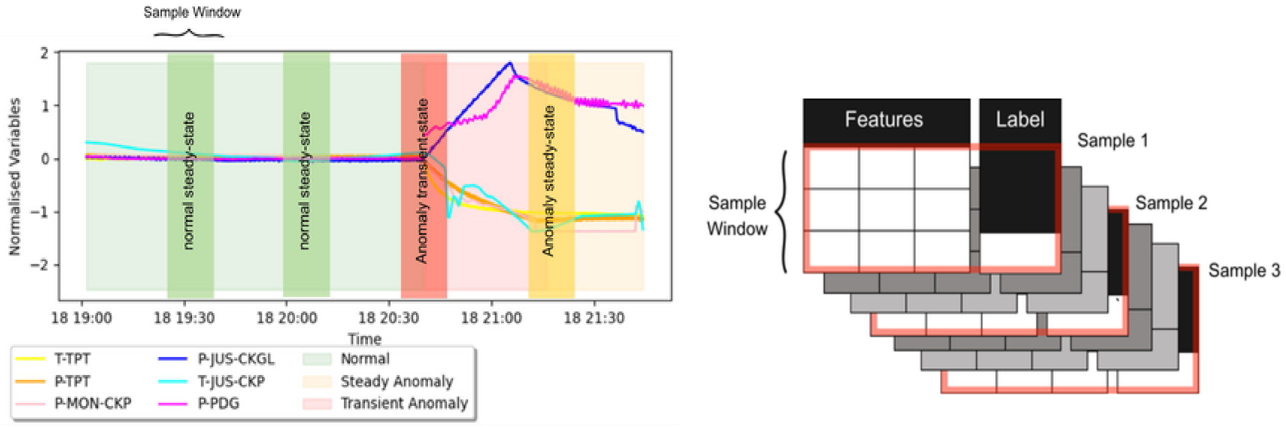

Figure 2. Illustration of an instance with the normal classed data (green) are labeled as normal, transient state and steady state (orange and red) are labeled as anomaly (left). The right panel shows the sampling process, in which a fixed-size sample window slides across the time series to create multiple samples, each comprising a window of

obtained from dynamic multiphase flow simulator OLGA, and hand-drawn instances digitized from experts' hand-drawn curves. The use of simulated and hand-drawn data is intended to decrease the imbalance of the dataset initially formed only by real instances (Vargas et al., 2019). There are 1,984 instances in this dataset with 1025 real instances obtained from 21 different wells, 939 simulated instances, and 20 hand-drawn instances. There are eight types of undesirable events categorized in this dataset, as shown in Table 3. They vary with the number of instances and event duration. Each data point has been labeled as three different states: normal, anomalous transient, and anomalous steady. In addition, the duration of each state differs greatly between each instance and event types. 3W Dataset faces several inherent difficulties: 31.17% variables were missing due to sensor or communication issues, 9.67% were frozen variables where a variable has a constant value due to sensor, system, or network issues. 0.01% of the observations were unlabeled due to tool limitations. Despite its inherent difficulties, 3W Dataset has been widely used for anomaly detection studies. Recent studies have applied machine learning techniques to classify and detect anomalies in oil well operations using 3W Dataset, demonstrating promising performance. A supervised decision tree classifier achieved an F1-score of 85% in identifying multiple classes of undesirable events (Turan & Jaschke 2021). In the unsupervised learning domain, a comparative study employing one-class classifiers, including Local Outlier Factor (LOF) and Autoencoder LSTM, was applied to multivariate time series data, where LOF outperformed with an F1-score of 85.9% compared to 62.7% from the Autoencoder LSTM (Fernandes et al., 2024). Furthermore, another unsupervised approach using LOF for anomaly detection in subsurface safety valves during offshore oil production reported an exceptionally high F1-score of 99.7%, highlighting the potential of novelty detection techniques in critical safety applications (Aranha et al., 2024).

METHODOLOGY

This paper proposed a supervised machine learning method with LSTM approach for binary time-series classification, classifying normal and anomaly events in 3W Dataset. The study followed a structured workflow comprising of data preparation, sample generation using observation window and labeled state (i.e., normal or anomaly), data splitting, feature selection, and scaling, followed by model training and evaluation.

Data preparation

The 3W Dataset was a compilation of sensor data sourced from actual wells, simulated scenarios, and expert-constructed (hand-written) cases. It was organized into nine folders, each representing a different class of anomalous events, as outlined in Table 3. Each folder contained a number of CSV files, totaling 1,984 instances across the entire dataset. Each file represented one instance and included a sequence of three possible states: normal, anomalous transient, and anomalous steady states, as illustrates in Figure 2.

The data used in this study were sourced from real instances from wells, each containing a minimum of 20 minutes in the normal state. Consequently, event types 3 and 4 were excluded from this study, as they did not meet this criterion. Event type 8 was also excluded due to the limited number of real-well instances.

Sample generation and data splitting

To evaluate the model performance across varying temporal contexts, different observation window lengths (i.e., 60, 120, and 180 seconds) were tested. As illustrates in Figure 3, the *window* represented the duration of sensor readings considered by the LSTM algorithm when learning the time-series patterns. The *label* indicated the class assigned to the entire window. Each sample was associated with a single target label corresponding to the event type. As a result, a single instance could generate multiple samples: segments of normal-state data (green) were labeled as *normal*, while those overlapping with transient and steady anomaly states (orange and red) were labeled as *anomaly*.

Since both transient and steady anomalies were grouped under the *anomaly* label, a 7:3 ratio between transient and steady samples was maintained to ensure a representative and informative dataset. This ratio was intentionally

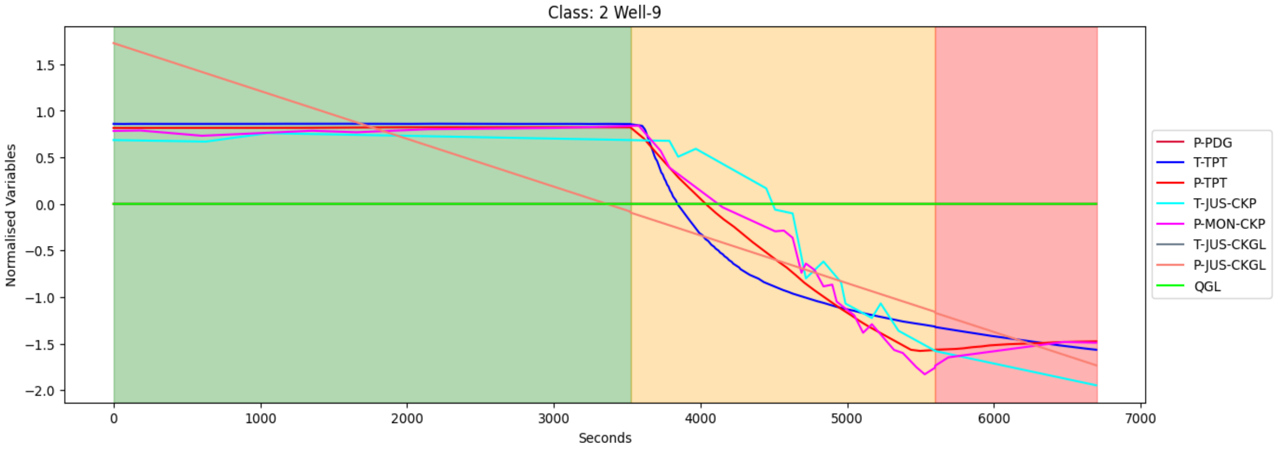

Figure 3. Example of a sample instance from a well, showing normalized sensor readings across normal (green), transient (orange), and steady anomaly (red) states for Event Type 2 (spurious closure of DHSV). The figure shows that several variables exhibit minimal variation across these states, reducing their predictive value. As discussed in the feature-selection section, only variables demonstrating meaningful temporal changes were retained as input features for the model.

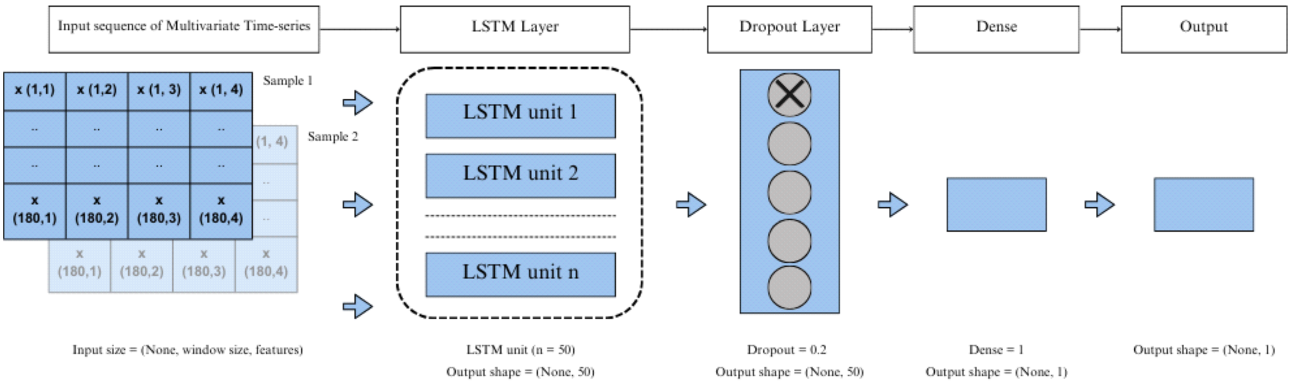

Figure 4. LSTM model architecture composed of an input sequence of multivariate time-series data, a single LSTM layer with 50 units, a dropout layer (rate = 0.2) for regularization, and a Dense output layer with sigmoid activation. The model processes sample windows of shape *(window size, features)* to produce a binary anomaly label.

chosen to emphasize early identification of anomalies, as transient states often preceded steady -state anomalies and provided critical cues for early intervention.

Table 4 summarizes the types of instances from which the samples were extracted, well identity, and the final counts of normal and anomalous samples. As shown in the table, samples were obtained from various wells to ensure diversity and generalizability across operating conditions. A total of 300 samples were used in this study, with a balanced distribution between normal and anomalous classes. Anomalous samples were drawn from a variety of event types to reflect realistic and heterogeneous failure scenarios. For model training and evaluation, the dataset was split into 60% for training and 40% for testing using stratified sampling to preserve the label distribution.

Feature selection and scaling

Figure 3 illustrates that not all variables in the dataset exhibited meaningful changes across different production states of the well. As a result, only a subset of variables was suitable for use as input features in time-series classification. Including variables with little to no variation could reduce model effectiveness, as they contributed to a high degree of sample homogeneity and offer limited predictive value.

An initial examination of the variables P-PDG and P-JUS-CKGL revealed unrealistic values. P-PDG exhibited a highly skewed distribution with an unusually large standard deviation, while P-JUS -CKGL contained negative pressure readings; an implausible condition that suggested data recording errors. T-CKGL contained only missing (NaN) values. QGL showed very limited variability, with a narrow range dominated by zero values. Based on availability across all event types and lower skewness, the features P-TPT, T-TPT, P-MON-CKP, and T-JUS-CKP were selected for modeling.

Following the feature selection, the data were standardized using the procedure defined in Equation 4. This standardization step adjusted each feature by removing its mean and scaling it to unit variance, ensuring equal contribution to the learning. The standardization was performed on the training set to avoid data leakage.

$$\text{Scaled Value} = \frac{\text{Input} - \text{Mean}}{\text{Standard deviation}} \tag{4}$$

Model architecture

The model used in this study was a single-layer LSTM network, constructed using the Keras Sequential API. Each LSTM layer consisted of 50 units, and was configured to return only the final output of each input sequence. A dropout layer with a rate of 0.2 was added for regularization, randomly dropping 20% of neurons during training to prevent overfitting. The final layer was a Dense layer with a sigmoid activation function, suitable for binary classification tasks, as it outputted a probability between 0 and 1.

The model was trained using the Binary Crossentropy loss function, which evaluated performance by comparing predicted probabilities with actual class labels. The Adam optimizer was employed due to its adaptive learning rate and efficiency, contributing to faster convergence and improved model performance. The input shape to the model followed the format (None, window size, number of features). For instance, with 255 samples, a window size of 180 seconds, and 4 features, the input shape became (None, 180, 4). Figure 4 illustrates the architecture of the LSTM model used in this study. This architecture was well-suited for capturing long-term dependencies in time-series data.

In addition to the LSTM model, this study also explored the performance of RNN and Gated GRU models using the same architecture and hyperparameters. This allows a direct comparison of their ability to capture temporal patterns in the data.

RESULT AND DISCUSSION

Table 5 summarises the classification metric scores for various window sizes and temporal nodes used in the model (i.e., LSTM, GRU, and RNN). Comparing window sizes across various models suggested that 120-second window was the most effective temporal resolution. As shown in

Table 5, the LSTM model performed best at 120-second window size with an F1-score of 0.92, followed by RNN and GRU models with F1-score of 0.91 and 0.90, respectively. This suggested that a 120-second window provided an optimal balance between capturing sufficient temporal context and avoiding overcomplication from longer sequences.

When comparing models at fixed window sizes, LSTM generally performed best at 120 seconds, while RNN slightly outperformed others at 60 seconds with faster training times. However, GRU matched LSTM's performance at 120 seconds but underperformed at other window lengths. With F1-score and recall dropping significantly at 60 and 180 seconds. These results indicated that model effectiveness varied with window size, and model selection should align with operational requirements such as latency. In anomaly detection tasks, the primary focus often depends on the application context. However, in most cases, recall is the most critical metric as the goal is minimizing false negatives. Missing an anomaly (false negative) could mean failing to detect events that have serious operational, safety, or financial consequences.

Both the LSTM and GRU models achieved the highest recall of 0.92 at a window size of 120 seconds, indicating that this window length captured enough temporal context for effective anomaly detection. The RNN model demonstrated the most consistent recall performance across different window sizes, with particularly strong results at 60 seconds (0.91 recall). Its efficiency and low training time made it a strong candidate for real-time anomaly detection, where quick response is essential. Meanwhile, the GRU model showed considerable variability in recall depending on window size, performing poorly at 60 and 180 seconds. This inconsistency may limit its reliability for anomaly detection unless tuned carefully. Without retraining or modifying the dataset, model performance can be further analyzed by evaluating

Table 5. Metric score on testing set

Model	Window Size	Precision	Recall	F1-Score
LSTM	60	0.88	0.90	0.89
	120	0.93	0.92	0.92
	180	0.83	0.86	0.84
RNN	60	0.91	0.91	0.90
	120	0.93	0.90	0.91
	180	0.87	0.80	0.83
GRU	60	0.81	0.68	0.73
	120	0.88	0.92	0.90
	180	0.80	0.72	0.76

Table 6. Metric score report across different type of anomalous event for LSTM model

Metric	Window size (s)	Event 1: abrupt increase of BSW	Event 2: spurious closure of DHSV	Event 5: rapid productivity loss	Event 6: quick restriction in PCK	Event 7: scaling in PCK
Precision	60	0.85	1.00	0.79	0.88	0.81
	120	1.00	0.98	0.87	0.91	0.93
	180	0.97	1.00	0.80	0.63	0.66
Recall	60	0.95	0.93	0.75	0.70	0.85
	120	0.78	1.00	0.80	0.50	0.83
	180	0.88	0.95	0.90	0.50	0.80
F1-score	60	0.89	0.96	0.76	0.78	0.83
	120	0.86	0.99	0.83	0.65	0.86
	180	0.92	0.97	0.84	0.56	0.72

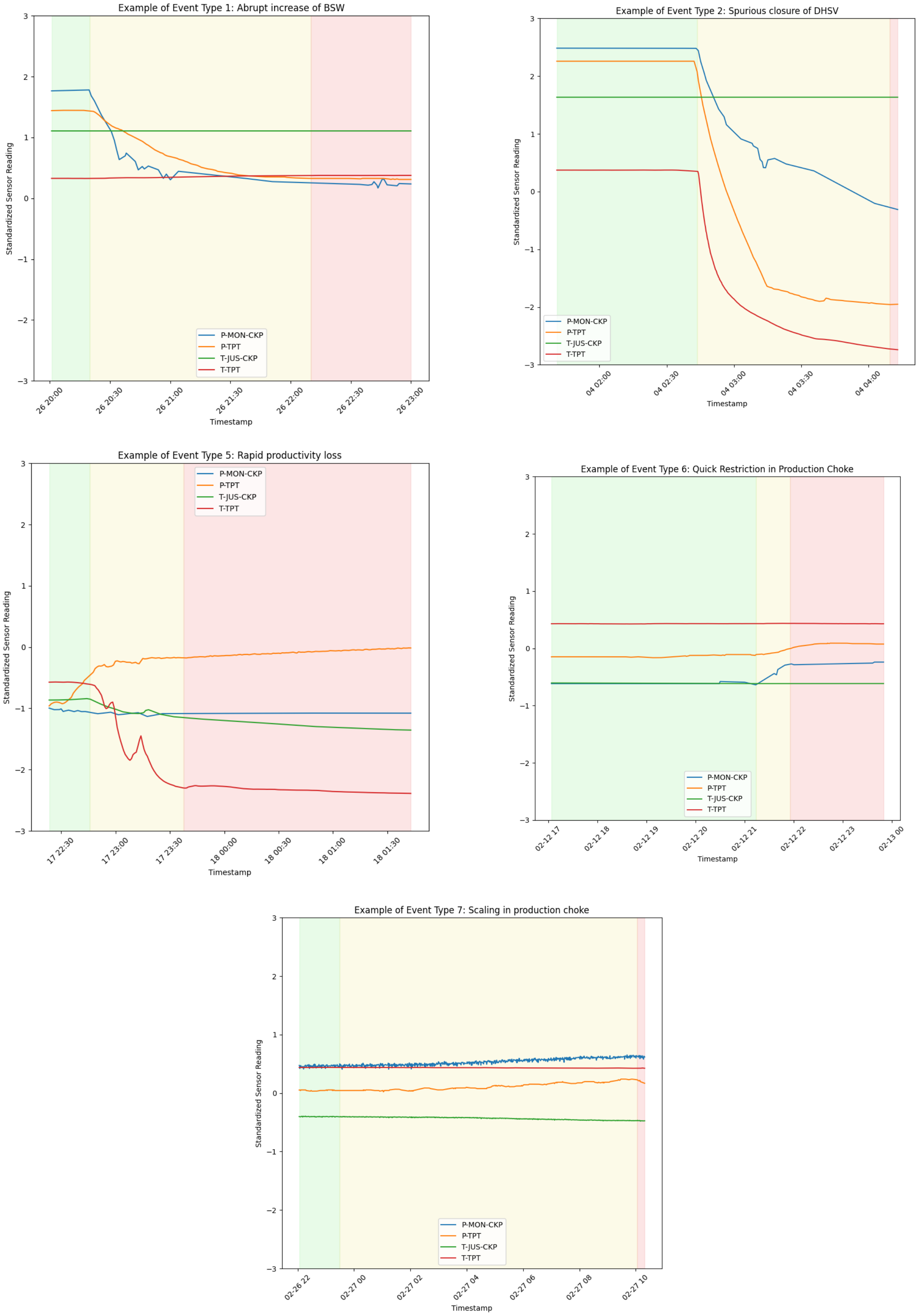

Figure 5. Standardized examples for several anomalous event types in the 3W Dataset across normal (green), transient (orange), and steady anomaly (red) states. The figure demonstrates how different anomalies manifest distinct temporal patterns, while some events (such as Event Type 7) exhibit only mild deviations that appear similar to normal state. These variations underscore the difficulty of detecting subtle anomalies in multivariate sensor data.

Table 7. Comparison with previous works

Reference	Algorithm	Task Type	Classification type	Dataset	Best Score (F1 or Accuracy)
Marins et al. (2020)	Random forest	Supervised	Binary	Real and Simulated	Accuracy: 97.1%
	Random forest	Supervised	Multiple binary	Real and Simulated	Accuracy: 96.2%-100%
	Random forest	Supervised	Multiclass (8 event types)	Real and Simulated	Accuracy: 94%
Turan and Jäschke (2021)	Decision Tree	Supervised	Multiclass (7 event types)	Real	F1: 85%
Fernandes Jr. et al. (2024)	LOF	Unsupervised	Binary	Real and Simulated	F1: 91.5% (simulated), 87.0% (real)
	Isolation Forest	Unsupervised	Binary	Real and Simulated	F1: 72.7%
	One-Class SVM	Unsupervised	Binary	Real and Simulated	F1: 47%
Aranha et al. (2024)	Local Outlier Factor (LOF)	Unsupervised	Binary	Real and Simulated	Accuracy: 99.9%
	Random Forest	Supervised	Binary	Real and Simulated	Accuracy: 87.1%
	Decision Tree	Supervised	Binary	Real and Simulated	Accuracy: 60%
This study	RNN	Supervised	Binary	Real	F1: 91%
	LSTM	Supervised	Binary	Real	F1: 92%
	GRU	Supervised	Binary	Real	F1: 90%

predictions across different types of anomalous events. This post hoc analysis provided insight into how well the model generalizes to the distinct temporal patterns and signal characteristics associated with each event type. This analysis also helped identify potential blind spots or biases in the model's behavior. Table 6 reports the metric score of LSTM model performance for different types of anomalous events. It should be noted that the following post hoc analysis was conducted solely on the LSTM model to explore its sensitivity to different anomaly types rather than to compare performance across RNN-based architectures.

As shown in Table 6, across all window sizes, Event Type 2 showed excellent performance with recall ranging from 0.93 to 1.00 and F1-scores above 0.96. This demonstrated that the model is highly effective at detecting this event type, likely due to distinct signal patterns. These coherent, monotonic transitions align well with the LSTM's gated-memory structure, which excels at learning smooth, long-term dependencies within multivariate sequences (Lindemann et al., 2021; Iskandar & Kurihara, 2022). The duration of the transient-anomaly state varied between wells, ranging from 15 to 193 minutes. However, the signal patterns were identical, including decreasing pressure at the pressure transducer in the wellhead and the upstream of production choke, as well as decreasing temperature downstream of the production choke.

On the contrary, Event Type 6 was the most challenging to detect, as indicated by the recall scores that were consistently low across various window sizes (0.70, 0.50, and 0.50 for 60s, 120s, and 180s, respectively). Despite having distinct features of increasing pressure upstream of production choke and wellhead, the model failed to catch the event as anomalous. Upon investigation, the duration of the transient-anomaly state was short and could be as brief as 9 minutes. In addition, the temporal trend previously mentioned was much subtler compared to Event Type 2, and standardization resulted in the value being similar to normal events, as illustrates in Figure 5. Separate binary classification tasks that specialize in detecting this type of event might be useful for future development.

Table 7 presents a comparison of machine learning models applied to the 3W Dataset for anomaly detection and classification in oil well operations. Across the literature, unsupervised methods such as LOF have shown strong performance, with Aranha et al. (2024) reporting 99.9% accuracy, and Fernandes et al. (2024) achieving F1 scores of 91.5% (simulated) and 87.0% (real). Other unsupervised models, including Isolation Forest and One-Class SVM, performed notably lower. In supervised settings, Turan and Jäschke (2021) addressed multiclass classification and reported an F1 score of 85% using a Decision Tree. Marins et al. (2020) explored different settings for classification tasks using the Random Forest algorithm and reported 96.2% to 100% accuracy. In comparison, this study applied RNN-based models (RNN, LSTM, GRU) for binary classification on real data, with F1 scores ranging from 90% to 92%, which are broadly comparable to the top-performing models from prior studies, although the classification task differs in scope and type.

CONCLUSION

This study evaluated LSTM, RNN, and GRU models for classifying normal and anomalous data in the 3W Dataset using precision, recall, and F1-score across three window sizes: 60, 120, and 180. The results show that the LSTM model performed best overall, especially with a window size of 120, where it achieved the highest F1-score of 92%. LSTM was more consistent across different window sizes, likely due to its ability to capture long-term dependencies. RNN performed well at shorter window sizes but showed a drop in recall at larger ones. GRU generally underperformed but reached a 90% F1-score at a window size 120. Compared to previous studies using the same dataset, the LSTM-based approach demonstrates competitive results. These findings suggest that LSTM is a suitable model for time series anomaly detection in oil and gas operations. Building on the results of this study, future work could focus on expanding from binary to multiclass classification using the 3W Dataset, allowing models to identify specific types of faults rather than only detecting anomalies in general. Exploring semi-supervised approaches may also help leverage the numerous unlabeled or partially labeled instances in the 3W Dataset. Since LSTM showed strong performance, future studies could investigate attention-based or hybrid deep learning models to further improve accuracy and interpretability. Real-time implementation is another practical direction, where models process data streams instead of fixed windows.

ACKNOWLEDGEMENT

The authors would like to express their gratitude to the *Direktorat Penelitian dan Pengabdian Masyarakat* (DPPM) or Directorate of Research and Community Service of Universitas Pertamina for providing financial support for this research through its internal research funding program.

GLOSSARY OF TERMS

Symbol	Definition	Unit
3W Dataset	Public dataset by Petrobras containing real, simulated, and expert-drawn subsea well events.	–
Adam	Adaptive Moment Estimation; optimization algorithm that adjusts learning rates using gradient moments.	–
AEM	Abnormal Event Management; systematic process for identifying, evaluating, and restoring abnormal conditions in oil and gas operations.	–
Anomaly	Deviation in operational data indicating potential	–

BSW	Basic Sediment and Water; measure of impurities in produced crude oil.	%
CSV	Comma-Separated Values	–
DHSV	Downhole Safety Valve; subsurface valve that isolates reservoir flow during emergencies.	–
DNN	Deep Neural Network; multi-layered neural model for learning nonlinear relationships.	–
F1-score	Harmonic mean of precision and recall; evaluates classification performance.	–
FN	False Negative; actual positive case incorrectly classified as negative.	–
FP	False Positive; actual negative case incorrectly classified as positive.	–
GRU	Gated Recurrent Unit; simplified recurrent neural network similar to LSTM but computationally efficient.	–
JUS	“Downstream”; indicates sensor located after a valve or choke.	–
LSTM	Long Short-Term Memory; recurrent neural network architecture capable of learning long-term dependencies.	–
MON	“Monitoring point” or “Upstream”; indicates sensor	–
OLGA	Dynamic multiphase flow simulator for modeling pipeline and well performance.	–
PCK	Production Choke; valve regulating production rate and downstream pressure.	–
PDG	Permanent Downhole Gauge; sensor measuring pressure and temperature within the wellbore.	Pa, ° C
P-JUS-CKGL	Fluid Pressure downstream of gas-lift choke.	Pa
P-MON-PCK	Fluid Pressure upstream to Production Choke (PCK) valve.	Pa
P-PDG	Fluid Pressure at Permanent Downhole Gauge (PDG).	Pa
Precision	Ratio of true positives to predicted positives; measures accuracy of positive predictions.	–
P-TPT	Fluid Pressure at Temperature/Pressure Transducer (TPT).	Pa
QGL	Gas lift flow rate.	m^3/s
Recall	Ratio of true positives to actual positives; measures model sensitivity.	–
RNN	Recurrent Neural Network; neural model processing sequential data with internal state memory.	–
Supervised Learning	Machine learning using labeled input–output pairs for training.	–

T-CKGL	Fluid Temperature downstream of gas-lift choke.	°C
T-JUS-PCK	Fluid Temperature downstream to PCK valve.	°C
TN	True Negative; actual negative correctly identified as negative.	–
TP	True Positive; actual positive correctly identified as positive.	–
Transient State	Temporary unstable condition between steady operational states.	–
TSC	Time-Series Classification; task of categorizing sequential data based on temporal patterns.	–
T-TPT	Fluid Temperature at TPT.	°C

REFERENCES

Alrifaey, M., Lim, W. H., & Ang, C. K. (2021). A novel deep learning framework based RNN-SAE for fault detection of electrical gas generator. IEEE Access, 9. https://doi.org/10.1109/ACCESS.2021.3055427.

Antipova, K., Klyuchnikov, N., Zaytsev, A., Gurina, E., Romanenkova, E., & Koroteev, D. (2019). Data-driven model for the drilling accidents prediction. Proceedings - SPE Annual Technical Conference and Exhibition. https://doi.org/10.2118/195888-MS.

Aranha, P. E., Policarpo, N. A., & Sampaio, M. A. (2024). Unsupervised machine learning model for predicting anomalies in subsurface safety valves and application in offshore wells during oil production. Journal of Petroleum Exploration and Production Technology, 14(2). https://doi.org/10.1007/s13202-023-01720-4.

Batal, I., Sacchi, L., Bellazzi, R., & Hauskrecht, M. (2009). Multivariate time series classification with temporal abstractions. In Proceedings of the 22nd International Florida Artificial Intelligence Research Society Conference (FLAIRS-22).

Blagec, K., Dorffner, G., Moradi, M., & Samwald, M. (2020, August). A critical analysis of metrics used for measuring progress in artificial intelligence.

Carvalho, B. G., Vaz Vargas, R. E., Salgado, R. M., Munaro, C. J., & Varejao, F. M. (2021). Flow instability detection in offshore oil wells with multivariate time series machine learning classifiers. In IEEE International Symposium on Industrial Electronics (ISIE 2021). https://doi.org/10.1109/ISIE45552.2021.9576310.

Fernandes, W., Komati, K. S., & Assis de Souza Gazolli, K. (2024). Anomaly detection in oil-producing wells: A comparative study of one-class classifiers in a multivariate time series dataset. Journal of Petroleum Exploration and Production Technology, 14(1). https://doi.org/10.1007/s13202-023-01710-6.

Figueirêdo, I. S., Vargas, R. E. V., Munaro, C. J., Varejao, F. M., & Salgado, R. M. (2021). Unsupervised machine learning applied to multivariate time series data of a rotating machine from an oil and gas platform. In IMCIC 2021 - 12th International Multi-Conference on Complexity, Informatics and Cybernetics - Proceedings.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. A. (2019). Deep learning for time series classification: A review. Data Mining and Knowledge Discovery, 33(4), 917–963. https://doi.org/10.1007/s10618-019-00619-1.

Iskandar, U. P., & Kurihara, M. (2022). Long short-term memory (LSTM) networks for forecasting reservoir performances in Carbon capture, utilisation, and storage (CCUS) operations. Scientific Contributions Oil and Gas, 45(1), 35–51. https://doi.org/10.29017/SCOG.45.1.943.

Kanyoma, I. R., Venriza, O., & Kushariyadi, K. (2023). Optimalisasi penambahan odorant pada gas menggunakan metode time series di PT. XYZ. Lembaran Publikasi Minyak dan Gas Bumi, 57(2), 43–53. https://doi.org/10.29017/LPMGB.57.2.1584

Lindemann, B., Maschler, B., Sahlab, N., & Weyrich, M. (2021). A survey on anomaly detection for technical systems using LSTM networks. Computers in Industry, 129, 103498. https://doi.org/10.1016/j.compind.2021.103498.

Machado, A. P. F., Munaro, C. J., Ciarelli, P. M., & Vargas, R. E. V. (2024). Time series clustering to improve one-class classifier performance. Expert Systems with Applications, 243, 122895. https://doi.org/10.1016/j.eswa.2023.122895.

Machado, A. P. F., Vargas, R. E. V., Ciarelli, P. M., & Munaro, C. J. (2022). Improving performance of one-class classifiers applied to anomaly detection in oil wells. Journal of Petroleum Science and Engineering, 218, 110983. https://doi.org/10.1016/j.petrol.2022.110983.

Malhotra, P., Vig, L., Shroff, G., & Agarwal, P. (2015). Long short term memory networks for anomaly detection in time series. In 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2015) - Proceedings.

Marins, M. A., Vargas, R. E. V., Salgado, R. M., Munaro, C. J., & Varejao, F. M. (2021). Fault detection and classification in oil wells and production/service lines using random forest. Journal of Petroleum Science and Engineering, 197, 107879. https://doi.org/10.1016/j.petrol.2020.107879.

Рапаков, Г. Г., Горбунов, В. А., Дианов, С. В., & Елизарова, Л. В. (2023). Research of the LSTM neural network approach in time series modeling. Cherepovets State University Bulletin, 3 (114). https://doi.org/10.23859/1994-0637-2023-3-114-4.

Škrlj, B., Kralj, J., Lavrač, N., & Pollak, S. (2019). Towards robust text classification with semantics-aware recurrent neural architecture. Machine Learning and Knowledge Extraction, 1(2), 646–666. https://doi.org/10.3390/make1020034.

Tong, Y., Zhang, D., Guo, C., Yuan, Y., He, Y., & Li, X. (2022). Technology investigation on time series classification and prediction. PeerJ Computer Science, 8, e982. https://doi.org/10.7717/peerj-cs.982.

Turan, E. M., & Jaschke, J. (2021). Classification of undesirable events in oil well operation. In Proceedings of the 2021 23rd International Conference on Process Control (PC 2021). https://doi.org/10.1109/PC52310.2021.9447527.

Ulil, M. R., Winardhi, S., & Dinanto, E. (2025). Machine learning-based prediction of shear wave velocity: Performance evaluation of Bi-GRU, ANN, and the Greenberg-Castagna empirical method. Scientific Contributions Oil and Gas, 48 (3). https://https://doi.org/10.29017/scog.v48i3.1797.

Vargas, R. E. V., de Souza, C. J. M., Varejão, F. M., de Almeida, L. R., de Oliveira, L. T., de Azevedo, J. A., & dos Santos, L. R. (2019). A realistic and public dataset with rare undesirable real events in oil wells. Journal of Petroleum Science and Engineering, 181. https:// doi.org/10.1016/j.petrol.2019.106223.