Hindawi
Applied Computational Intelligence and Soft Computing
Volume 2022, Article ID 1558381, 14 pages
https://doi.org/10.1155/2022/1558381

Research Article

Anomaly Detection Using Explainable Random Forest for the Prediction of Undesirable Events in Oil Wells

Nida Aslam,[1] Irfan Ullah Khan,[1] Aisha Alansari,[2] Marah Alrammah,[1] Atheer Alghwairy,[1] Rahaf Alqahtani,[1] Razan Alqahtani,[1] Maryam Almushikes,[1] and Mohammed AL Hashim[2]

[1]*Department of Computer Science, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia*
[2]*Computer Engineering Department, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia*

Correspondence should be addressed to Nida Aslam; naslam@iau.edu.sa

Received 25 May 2022; Revised 5 July 2022; Accepted 7 July 2022; Published 5 August 2022

Academic Editor: Manikandan Ramachandran

Copyright © 2022 Nida Aslam et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The worldwide demand for oil has been rising rapidly for many decades, being the first indicator of economic development. Oil is extracted from underneath reservoirs found below land or ocean using oil wells. An offshore oil well is an oil well type where a wellbore is drilled underneath the ocean bed to obtain oil to the surface that demands more stability than other oil wells. The sensors of oil wells generate massive amounts of multivariate time-series data for surveillance engineers to analyze manually and have continuous insight into drilling operations. The manual analysis of data is challenging and time-consuming. Additionally, it can lead to several faulty events that could increase costs and production losses since the engineers tend to focus on the analysis rather than detecting the faulty events. Recently, machine learning (ML) techniques have significantly solved enormous real-time data anomaly problems by decreasing the data engineers' interaction processes. Accordingly, this study aimed to utilize ML techniques to reduce the time spent manually to establish rules that detect abnormalities in oil wells, leading to rapid and more precise detection. Four ML algorithms were utilized, including random forest (RF), logistic regression (LR), k-nearest neighbor (K-NN), and decision tree (DT). The dataset used in this study suffers from the class imbalance issue; therefore, experiments were conducted using the original and sampled datasets. The empirical results demonstrated promising outcomes, where RF achieved the highest accuracy, recall, precision, F1-score, and AUC of 99.60%, 99.64%, 99.91%, 99.77%, and 1.00, respectively, using the sampled data, and 99.84%, 99.91%, 99.91%, 99.91%, and 1.00, respectively, using the original data. Besides, the study employed Explainable Artificial Intelligence (XAI) to enable surveillance engineers to interpret black box models to understand the causes of abnormalities. The proposed models can be used to successfully identify anomalous events in the oil wells.

1. Introduction

Oil is one of the most valuable energy sources and is considered the first indicator of economic development. Globally, oil demand has rapidly increased as it is used in various applications, such as heating buildings and generating electricity. According to OPEC, global oil demand will grow by 4.15 million barrels per day in 2022 [1]. Oil wells are deep narrow holes that combine multiple sensors, pneumatic, hydraulic, and mechanical systems to bring oil to the surface. An offshore oil well is an oil well type where a borehole is drilled under the seabed to bring oil to the surface, which requires more stability than other oil wells [2]. The sensors of oil wells produce tremendous amounts of multivariate time-series data. Presently, surveillance engineers manually analyze the generated data on a calendar basis, in which one engineer might be responsible for hundreds of wells. The exhaustive data analysis is causing engineers to put more effort and time into investigating the data rather than concentrating on more critical situations [3]. Consequently, a long lag between the occurrence of the problem and the detection allowed bypassing critical faulty, putting oil exploration and refining at risk for disruptions and heavy losses that can affect industries worldwide. Accordingly, detecting adverse events in oil and gas wells can help prevent production downtime, environmental accidents, and human casualties and reduce maintenance costs.

E. Oort et al. [4] developed a way to analyze the massive amount of data produced by the sensors on the wells more efficiently and accurately, called Automatic Rig-Activity Detection (ARAD). ARAD operates by detecting the drilling processes using the real-time data generated from the rig. However, ARAD is incapable of perceiving the abnormal events, as it classifies the anomaly activities as "unknown." However, these situations must be categorized as anomalous to aid engineers in understanding the performance of wells and examine the sources of these abnormalities for future prevention to diminish the possibility of severe injuries, loss of life, economic loss, or environmental pollution. Hence, employing recent technologies in this field is crucial for reducing its severe impact.

Machine learning (ML) gained wide attention by providing various robust tools showing promising results in multiple applications. Anomaly detection is one of the most prevalent problems solved by ML techniques. Moreover, the introduction of industrial 4.0 has completely transformed the whole process by integrating the Internet of Things (IoT), automated synchronized distributed data collection platforms like cloud computing, real-time data analysis, etc. The colossal amount of data generated using various sensors in several domains has been successfully analyzed and utilized for prediction in different fields using ML. Similarly, the ML techniques have also shown significant outcomes in the oil and gas industry [5].

Vargas et al. [6] provided a public dataset containing eight types of undesirable events of the offshore oil well that can be fed into an ML algorithm to automate the adverse event detection process. The faulty events include the abrupt increase of basic sediment and water, spurious closure of downhole safety value (DHSV), severe slugging, flow instability, rapid productivity loss, quick restriction in production choke (PCK), scaling in PCK, and hydrate in the production line. In this study, the abnormal events were combined with being considered the positive class, whereas the normal event was treated as the negative class. The dataset utilized suffers from class imbalance issues. Accordingly, two experiments were conducted for the prediction of rare adverse real-world events in oil wells. The first experiment trained four ML algorithms, including random forest (RF), logistic regression (LR), k-nearest neighbor (K-NN), and decision tree (DT), with the original data, whereas the second experiment trained the aforementioned algorithms using an upsampled data. The empirical results demonstrated promising outcomes, where RF achieved the highest accuracy, recall, precision, F1-score, and AUC of 99.60%, 99.64%, 99.91%, 99.77%, and 1.00, respectively, using the sampled data, and 99.84%, 99.91%, 99.91%, 99.91%, and 1 using original data.

Despite the robustness of ML algorithms in classification problems, it fails to provide informatics for nontechnical to deduce the models' classification behavior, limiting the possibility of deploying it in real-time applications. ML models are considered black box, in which their results are not explainable by nature. Consequently, Explainable Artificial Intelligence (XAI) gained wide popularity among researchers who started adopting it with ML models to demonstrate interpretability [7]. XAI provides insight into how a result was delivered to induce trustfulness by answering "wh" questions. Hence, it is believed that it can prevent life-threatening faults. This study used global surrogate, Shapley Additive Explanation (SHAP), and Local Interpretable Model-Agnostic Explanations (LIME) to interpret the black box models. The contribution of the study is as follows:

(i) Introduce a proactive model for detecting anomaly events in oil wells.

(ii) Employ XAI tools to assist surveillance engineers in understanding the causes of anomaly events. As per the authors' knowledge, no study has implemented XAI to identify anomaly events in oil wells.

(iii) Achieve better results than the benchmark studies that utilized the same dataset with lower complexity models.

The rest of the study is organized as follows: Section 1 provides a detailed review of related literature. Section 2 describes the materials and methods used, which includes a description of the dataset, the preprocessing techniques applied, and a technical description of the utilized ML algorithms. The experiment setup and results are explained in Section 2, whereas Section 2 represents the XAI. A brief conclusion with recommendations is provided in Section 3.

2. Review of Related Studies

In the industrial environment, increased demands for better functional safety, proficiency, supremacy, and energy efficiency necessitated classifying and detecting undesirable and unwanted oil wells to prevent the losses of expenses and collisions. Some recent studies on unpleasant events and locating oil reservoirs using ML have been discussed below in chronological order.

Al-Fadhli and Zaher [8] aimed to develop an automated monitoring and controlling system for oil refineries to replace the traditional techniques and enhance their performance. The model constantly assembles data from oil tanks and pipelines and performs early detection for any possible faults. LabVIEW and LabJack were used to implement the proposed system. The data were collected from sensors plugged in oil tanks and pipelines. The results proved the system's success in providing real-time and precise information about the oil tanks and pipelines.

Nwachukwu et al. [9] mentioned that the injector well location could predict the reservoir response using ML. The slightest change to the injector well location might significantly impact the reservoir response. As a result, they proposed well-to-well connectivity as a predictor variable to enhance the accuracy. The dataset was collected from numerical reservoir simulations, using different training sizes to test how it will impact the accuracy. The study employed the eXtreme Gradient Boosting (XGBoost) algorithm and evaluated its performance using the correlation coefficient (R^2) value with different training data sizes. The empirical results demonstrated that all the R^2 scores were higher than 0.85, indicating that the predictor variable improved the well-to-well connectivity model performance.

In another study, Vargas et al. [6] used a web application programming interface (API) simulator, which runs from a web client or as a hardware-in-the-loop (HIL) simulator from a control system environment with programmable logic controllers (PLCs). The authors aimed to test the oil well's data anywhere in the world without the need to install software, as the HIL functionality allows a workflow from early production for commercial pilots. In this study, the lack of a practical and scalable research environment for automated drilling systems slows advanced technological requirements and reduces the industry's ability to decrease costs and mitigate the carbon footprint. The authors suggested developing a drilling system that enhances the workflow to ensure decision-makers' competence in complex and high dynamic process operations.

Bronstad [10] implemented condition-based monitoring (CBD) system to detect and classify undesirable events in offshore oil wells using the 3W database originated by Petrobas. The proposed system achieved an overall accuracy of 90% in conducting three experiments, where two of them were associated with classification, and another was related to feature extraction.

On the other hand, Ghorbani et al. [11] predicted the oil flow using the flow affecting variables around orifice meters. The dataset was collected from the Cheshmeh Khosh oil field, including 1037 data records. They used adaptive neuro-fuzzy inference system, least squares support vector machine (SVM), multilayer perceptron (MLP), and genetic algorithm (GA). The results suggested that all the algorithms could predict the flow effectively, where MLP achieved the best results with a root mean square error (RMSE) of 8.70.

M. Marins et al. [12] proposed a system that utilized ML to detect and classify faulty events in oil and gas wells and lines as early as possible to avoid potential risks. The study used the 3W database developed by Petrobras, consisting of 2000 events. Moreover, the Bayesian approach and RF classifier were used to identify the faults and classify them into 9 classes by considering class 0 as the normal event and the remaining classes from 1 to 8 as different types of faults. In order to evaluate the proposed system's performance, the researchers conducted three experiments in various situations. The final result proved the proposed method successfully detected the faults with 94% accuracy.

In another study, Alsaihati et al. [13] proposed an intelligent system for predicting drilling torque profiles continuously to alert the crew in case of any operational accidents ahead of time. The authors collected the dataset by surface real-time transmitter sensors positioned at different locations within the rig site. Three ML models were trained, including RF, artificial neural network (ANN), and functional network. Results indicated that RF achieved the best results with an average absolute percentage (AAPE) of 1.46 and R^2 of 0.99 using the training set, whereas it achieved an AAPE of 3.98% and R^2 of 0.93 using the testing set. Additionally, the offered system could alert the crew 9 hours and 7 hours before incidents took place in Well-1 and Well-2, respectively.

Furthermore, Aljubran et al. [14] developed a DL-based model for early predicting fluid lost circulation incidents (LCIs) in drilling operations. The dataset utilized in their study was based on an analysis of historical drilling data derived from standard drilling rig equipment and apparatus. Three DL models were utilized, including RF, ANN, and long short-term memory (LSTM), with standard and window normalization. The results indicated that the CNN model with window normalization attained the best results with an accuracy of 92.55%, precision of 87.34%, recall of 73.40%, and F1-score of 79.77%. Additionally, the authors claimed that the proposed model is developed in such a way that it can be retrained with new sensor data and can be employed for the early prediction of other abnormal drilling events.

De Salvo Castro et al. [15] utilized the 3W Petrobras dataset and employed unsupervised techniques, namely control chart with three standard deviations and a fuzzy c-means algorithm to classify faulty events in oil wells. Additionally, the authors used the RF algorithm to evaluate the performance of the unsupervised models. The results indicated that the control chart outperformed fuzzy c-means in terms of sensibility with a minor difference. Furthermore, the results revealed that RF achieved a specificity and sensitivity of 100% and 99.91%, respectively, using the cohort chart, whereas it attained 99.98% and 94.01% using the fuzzy c-means.

More recently, Gurina et al. [16] proposed an algorithm to forecast drilling accidents using a large multivariate time-series mud telemetry dataset from Russia containing 6 drilling accidents. The authors utilized clustering and wavelet transform to convert the time-series data into a bag-of-features representation. The results demonstrated that 70% of the 6 drilling accidents could be classified using the proposed model with a false positive rate of 40%. The authors aim to reduce their false positive rate in the future.

Furthermore, Alharbi et al. [17] aimed to compare the performance of 6 ML algorithms in identifying anomalies in wells using two datasets. The ML algorithms included K-NN, RF, SVM, LR, DT, and rule fit classifier (RFC). The empirical results claimed that RFC achieved the highest results when trained using the first dataset with an F1-score of 0.92 and a complexity of 0.5. On the other hand, RF attained the best results when trained using the second dataset with an F1-score of 0.84 and a complexity of 0.4.

In another study, Alsaihati et al. [18] proposed an ML-based solution using the mechanical surface parameters for forecasting loss of circulation rate (LCR) during drilling based on mechanical surface parameters and active pit volume measurements. The authors collected the data from seven wells experiencing extreme or partial LCR. Three ML classifiers were utilized, including SVM, RF, and K-NN. Results indicated that K-NN attained the highest outcomes with an R^2 and RMSE of 0.90 and 0.17, respectively.

Cheng et al. [19] proposed a model to predict an oil depot's abnormal tank liquid level by analyzing its nonperiodic time-series data. Two datasets were utilized in this study to train two convolutional autoencoder algorithms, including recurrent neural network (RNN) and LSTM encoders. The empirical results demonstrated promising results with an accuracy of 98% and an F1-score of 82%.

According to the conducted literature, few research papers focused on classifying undesirable events in oil wells with considerable robust results using supervised ML algorithms. It is revealed that most of the reviewed studies utilized the Petrobras 3W dataset provided by Vargas et al. [6]. M. Marins et al. [12] achieved an overall accuracy of 94% in detecting faults, whereas Bronstad [13] achieved the highest accuracy of 90%. Conversely, De Salvo Castro et al. [15] employed unsupervised techniques to enhance the model's performance, where they achieved an overall specificity and sensitivity of 100% and 99.91%, respectively, using the cohort chart. Accordingly, this study aimed to explore the robustness of employing supervised ML algorithms to classify normal and undesirable events using the Petrobras 3W dataset. Additionally, it intended to eliminate the ambiguity associated with black box models by employing XAI techniques.

3. Material and Methods

This section contains the details related to the material and methods used in this study. The time-series dataset utilized in this study was collected using oil well sensors. The collected data were cleaned by carrying out several preprocessing steps. After applying the preprocessing steps, the Synthetic Minority Oversampling Technique (SMOTE) was applied to handle the data imbalance issue. Two experiments were conducted, that is using the original and upsampled datasets. Both the datasets were divided into training and testing using holdout (70–30). Four models were trained, namely RF, K-NN, LR, and DT. Next, the models were optimized using the GridSearchCV method. The models were evaluated in terms of accuracy, precision, recall, F1-score, AUC, and ROC. Finally, XAI was performed for the best-performing model to generate an explanation of the results for nontechnicals. Figure 1 indicates the methodology of the proposed study.

3.1. Description of the Dataset. This study utilized the 3W dataset developed by Petrobras [6], containing approximately 1984 events. The 3W database contains three different types of undesirable events, namely real, simulated, and hand-drawn, and 9 multivariate time series, namely normal and abrupt increase of basic sediment and water, spurious closure of downhole safety value (DHSV), severe slugging, flow instability, rapid productivity loss, quick restriction in production choke (PCK), scaling in PCK, and hydrate in the production line. Real events are the ones that happened in Petrobras in actual wells throughout the oil production. The utilization of simulated and hand-drawn events is essentially meant to reduce the imbalance of the dataset initially created by real events. The distribution of the samples per category is shown in Figure 2.

The dataset also includes time-series data with 8 tags obtained from 8 sensors, as shown in Table 1.

3.2. Preprocessing. The 3W database of hand-drawn, simulated, and real instances contained multiple missing values, raising the need to apply preprocessing techniques to preserve the reliability of the final results. The target attribute in the dataset consisted of 9 integer values ranging from 0 to 8, each intended for a type of undesirable event. The dataset consists of 597 normal events and 1387 undesired events. The proposed study attempts to discriminate between the normal and undesired events without discriminating among the type of undesired oil well events. Various functions including "class_file_generator," "get_instances_with_undesirable_event," "load_instance," "load_downsample_instances," and "extract_samples_train" were applied in order to preprocess the data and remove missing values. Afterward, some statistical measures such as mean, median, variance, standard deviation, maximum, minimum, and root mean square were applied to the attributes. Initially, the dataset contained 41 attributes. After computing the statistical measures, duplicate records were removed, and features with a correlation above 0.8 with other available features were eliminated. Later, the StandardScaler preprocessing technique provided by the Sklearn learn library was performed. After applying the preprocessing techniques, the Synthetic Minority Oversampling Technique (SMOTE) with a random_state value of 42 was applied to balance the classes. The dataset finally contained 14 features and 5180 instances.

3.3. Description of the Classifiers. The study employed four different supervised ML algorithms to predict the occurrence of undesirable events in oil wells. The target class label was in binary format, indicating the existence of the abnormal event. The subsections below provide a theoretical background of the utilized algorithms.

3.4. Decision Tree. A decision tree (DT) is considered a supervised ML technique employed in both classification and regression problems. It is regarded as one of the simplest ML algorithms that can be easily interpreted and understood compared to different algorithms [20]. DT follows well-defined rules presented in a tree-like structure, including the root interior node, branches, and leaf node. The root node represents the attribute with the highest information gain, and the branches denote the attributes' values, whereas the leaf node represents the outcome. A DT is constructed by performing a greedy search to find the feature with the highest information gain. To calculate the information gain, the entropy is first computed. The following equation shows the formula for calculating the entropy of each feature, where P_x represents the positive samples and P_y represents the negative samples in S [21].

$$\text{Entropy}(s) = -\left(P_x \log_2 P_x + P_y \log_2 P_y\right). \quad (1)$$

The following equation shows the formula for calculating the information gain of each feature, where $V(A)$ represents all possible values for feature A and S_v is a subset of S, and the feature A has a value v [21].

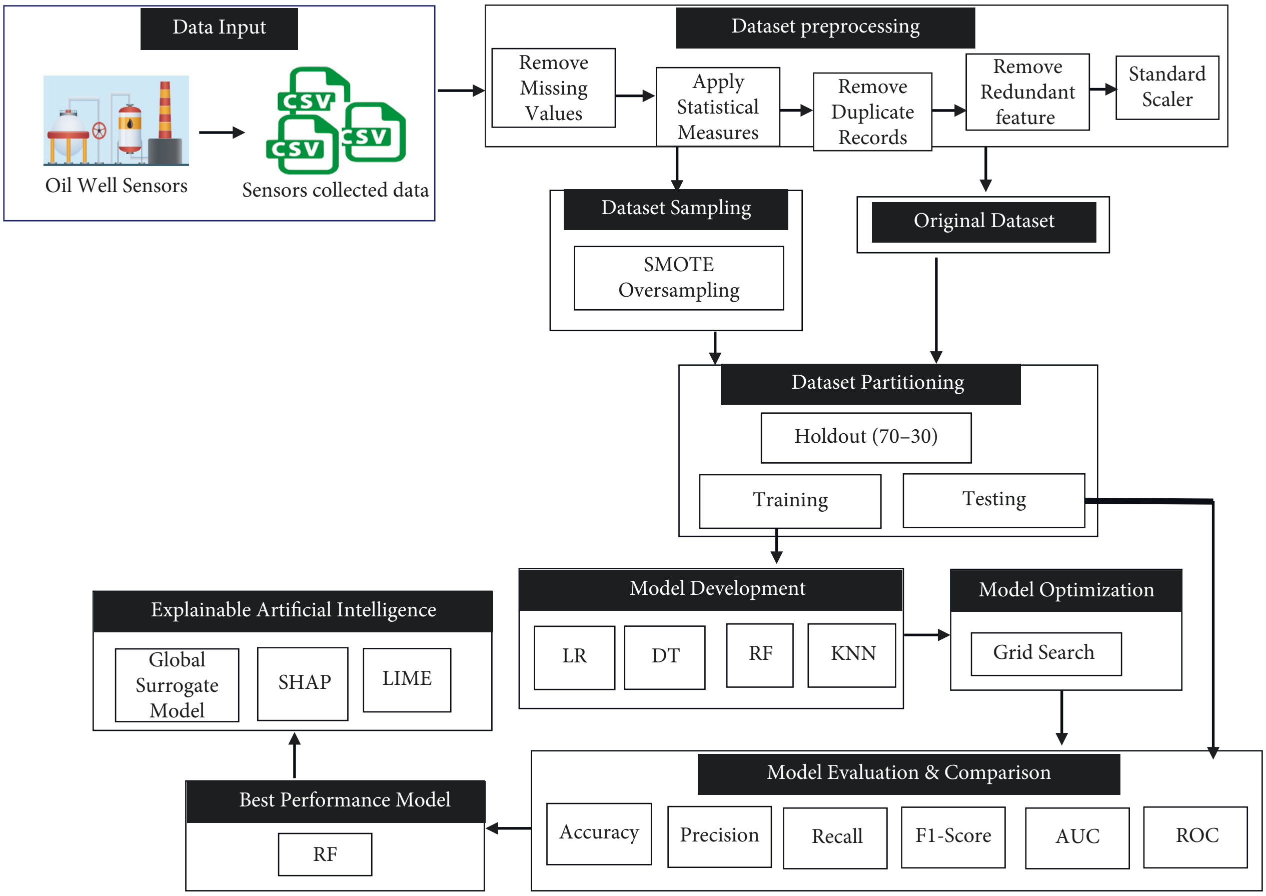


FIGURE 1: The proposed methodology for anomaly detection in oil wells.

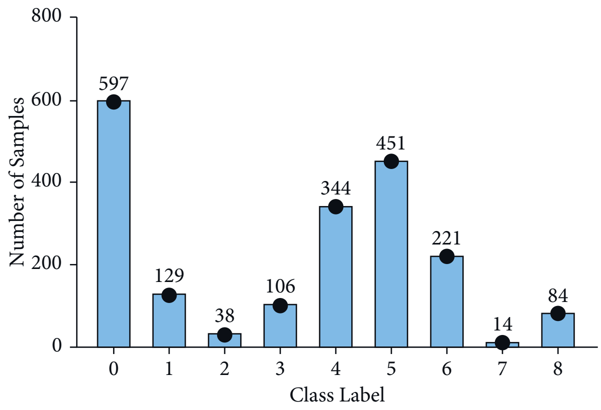


FIGURE 2: Distribution of samples in the 3W database in each event.

$$\text{Information Gain}\,(S, A) = \text{Entropy}\,(S) - \sum_{v\epsilon V(A)} \frac{|S_v|}{S}\text{Entropy}\,(S_v). \qquad (2)$$

The hyper-parameters selected for this algorithm for the original and SMOTE data are represented in Table 2.

3.5. Random Forest Classifier. Random forest (RF) is a procedural ML algorithm based on collective learning, which is flexible and easy to use. It combines numerous decision trees that result in a forest of trees, resulting in improved results. The advantages of using RF include handling a significant number of missing values by two approaches, either replacing them with the median or the mean by proximity weight. On the other hand, the disadvantage of using RF includes its low computational speed in generating the predictions due to the decision tree varieties [22]. The following equation represents the voting formula used by the RF classifier to generate the final result, where x

TABLE 1: The name, description, and measuring units of the tags in the 3W database.

Name	Description	Unit
P-PDG	Pressure at the permanent downhole gauge.	Pa
P-TPT	Pressure at the temperature and pressure transducer.	Pa
T-TPT	The temperature at the temperature and pressure transducer.	°C
P-MON-PCK	Pressure upstream of the production choke	Pa
T-JUS-PCK	Temperature downstream of production choke.	°C
P-JUS-CKGL	Pressure downstream of gas lift choke.	Pa
T-JUS-CKGL	Temperature downstream of gas lift choke.	°C
QGL	Gas lift flow rate.	m^3/s

Table 2: Optimal hyper-parameters of decision tree.

Hyper-parameter	Value
max_features	Auto
ccp_alpha	0.001
Criterion	Entropy
max_depth	9
Splitter	Best
Random state	3

is the data point, C_i is the result of i^{th} sample and x, and $\hat{y}_f$ is the computed result [23].

$$\hat{y}_f = \text{mode}\{C_1(x), C_2(x), \ldots C_n(x)\}. \tag{3}$$

The hyper-parameters selected for this algorithm for the original and SMOTE data are represented in Table 3.

3.6. *K-Nearest Neighbor.*

K-nearest neighbor (K-NN) is one of the simplest statistical-based nonparametric ML algorithms used in classification and regression problems. K-NN is considered a lazy algorithm, as it does not train on the supplied data but stores the training samples and classifies the new instances based on a chosen distance measure. Before applying the distance measure, the K value, which is referred to as the number of nearest neighbors, is specified. The unknown label is then classified based on the most often appearing class around the assigned K value [24]. The advantage of using K-NN includes its simplicity and ease of use. However, it is limited to its low computational speed with large datasets since it needs to calculate the distance between the unknown point and all the known points. The hyper-parameters selected for this algorithm are represented in Table 4.

3.7. *Logistic Regression Classifier.*

Logistic regression (LR), regardless of its name, is a classification model as opposed to a regression model. LR is a statistical ML technique that employs a logistic function to foresee the probability of occurrence of a binary event. The logistic function is a sigmoid equation that accepts all real values in the 0 and 1 range [25]. Additionally, it can deal with quite a few variables, either categorical or numerical. The sigmoid function is as follows:

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \tag{4}$$

Table 3: Optimal hyper-parameters of random forest.

Hyper-parameter	Value
Criterion	Entropy
max_depth	8
max_features	Auto
n_estimators	30
Random state	3

Table 4: Optimal hyper-parameters of k-nearest neighbor.

Data	Hyper-parameter	Value
Original data	Metric	Minkowski
	n_neighbors	1
SMOTE data	Metric	Manhattan
	n_neighbors	1

Compared to other supervised classification techniques, such as ensemble classifiers or kernel SVM, LR is comparatively fast. However, it suffers to some extent in its accuracy, as it is much too simplistic for dealing with complex relationships between variables. The hyper-parameters selected for this algorithm for the original and SMOTE data are represented in Table 5.

3.8. *Performance Measure.*

The final models were evaluated using five evaluation metrics, namely accuracy, precision, recall, F1-score, area under the curve (AUC), and receiver operating characteristic (ROC). Initially, confusion matrices evaluate the models in terms of true positive (TP), false positive (FP), true negative (TN), and false negative (FN), where:

(i) TP presents the number of correctly classified records as the presence of any undesirable event.

(ii) FP presents the number of incorrectly classified records as the presence of any undesirable event.

(iii) TN presents the number of correctly classified records as normal.

(iv) FN presents the number of incorrectly classified records as normal.

Accuracy represents the number of correctly classified undesired and normal events to the number of oil well events in the dataset. It is expressed as follows:

$$\text{Accuracy} = \frac{\text{Correctly classified normal and undesirable well events}}{\text{Total number of well events in the dataset}}. \tag{5}$$

Precision is the percentage of the correctly classified undesired well events to the sum of classified undesired well events. It is represented as follows:

$$\text{Precision} = \frac{\text{Correctly classified undesirable well events}}{\text{Total number of classified undesirable well events}}. \tag{6}$$

Recall is the percentage of correctly classified undesired well events to the number of undesired oil well events in the dataset. It is represented as follows:

$$\text{Recall} = \frac{\text{Correctly classified undesirable well events}}{\text{Total number of undesirable well events in the dataset}}. \tag{7}$$

Table 5: Optimal hyper-parameters of logistic regression.

Hyper-parameter	Value
Penalty	12
C	60
Solver	Liblinear
Random state	3

Similarly, F1-score represents the weighted average of the precision and the recall and is represented mathematically as follows:

$$F1 - \text{Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}. \quad (8)$$

3.9. Experimental Setup and Results. The proposed models for predicting the presence of undesirable events in oil wells were implemented using Python programming language (ver. 3.10.0). Several libraries were used, including Sklearn, Numpy, Pandas, and Matplotlib. Sklearn library was mainly used to split the data with a random_state of 52 into a stratified ratio of 70 : 30 to train and test four ML algorithms, including decision tree, random forest, k-nearest neighbor, and logistic regression. Additionally, GridSearchCV was utilized to obtain the optimal hyper-parameters for each model using the original and upsampled training data with stratified 10-folds cross-validation and random_state value of 15. The final models were evaluated using five evaluation metrics, namely accuracy, precision, recall, F1-score, and area under the curve (AUC), and also using confusion matrices.

This section represents the results attained using the trained models. A comparison between the results obtained while training the models with the original and upsampled datasets is discussed in Table 6 in terms of five performance measures, namely accuracy, precision, recall, F1-score, and AUC.

The table indicates a comparative performance between the models trained using the original and SMOTE datasets. Overall, the precision rates were better in the SMOTE data than in the original dataset, whereas the recall rates were better using the original dataset. It is revealed that the LR and RF classifiers performed better when trained using the original data. Conversely, DT and K-NN attained higher accuracy scores when trained using the SMOTE dataset.

To be more specific, the table suggests that RF outperformed all classifiers using the SMOTE dataset in terms of accuracy, recall, F1-score, and AUC, achieving an accuracy, recall, and F1-score of 99.60%, 99.64%, and 99.77%, respectively, followed by K-NN, which attained accuracy, recall, and F1-score of 99.36% and 99.37%, and 99.64%, respectively. LR achieved the lowest accuracy among the other classifiers at 95.74% using the SMOTE data. Despite the unsatisfactory performance of LR in terms of accuracy, it attained the highest precision of 100%. Additionally, DT achieved the same precision rate of 100%. However, with the original data, RF achieved the highest outcome with all the measures, attaining accuracy, precision, recall, F1-score, and AUC of 99.84%, 99.91%, 99.91%, 99.91%, and 1, respectively. Similar AUC was achieved for both the original and SMOTE data for all the measures for RF.

To visualize the attained results more precisely, the confusion matrices for all the four classifiers using the SMOTE data are represented in Figure 3, and the confusion matrices for the models trained using the original dataset are shown in Figure 4. Overall, the best results were achieved using the original dataset.

The confusion matrices reveal that the difference in FPs between the developed models is insignificant, whereas the difference in FNs is considerable. Moreover, the matrices conclude that the best algorithm for predicting undesirable events is the RF model since it attained the lowest FNs of 1, followed by the DT model, which missed the classification of 5 instances as undesirable events using the original data. On the other hand, using the SMOTE data, it is observed that RF is the best model with 4 FNs, followed by K-NN with 7 FNs. Besides, RF and LR are the best models in terms of predicting the normal events correctly, as they have the lowest FP rates using the original data, whereas LR and DT are the best using the SMOTE data. Since it is more crucial to predict undesirable events correctly than normal events for their serious negative impact, it is more important to consider low FN values. Accordingly, it is concluded that RF achieved the best results among all other classifiers using the SMOTE and original dataset.

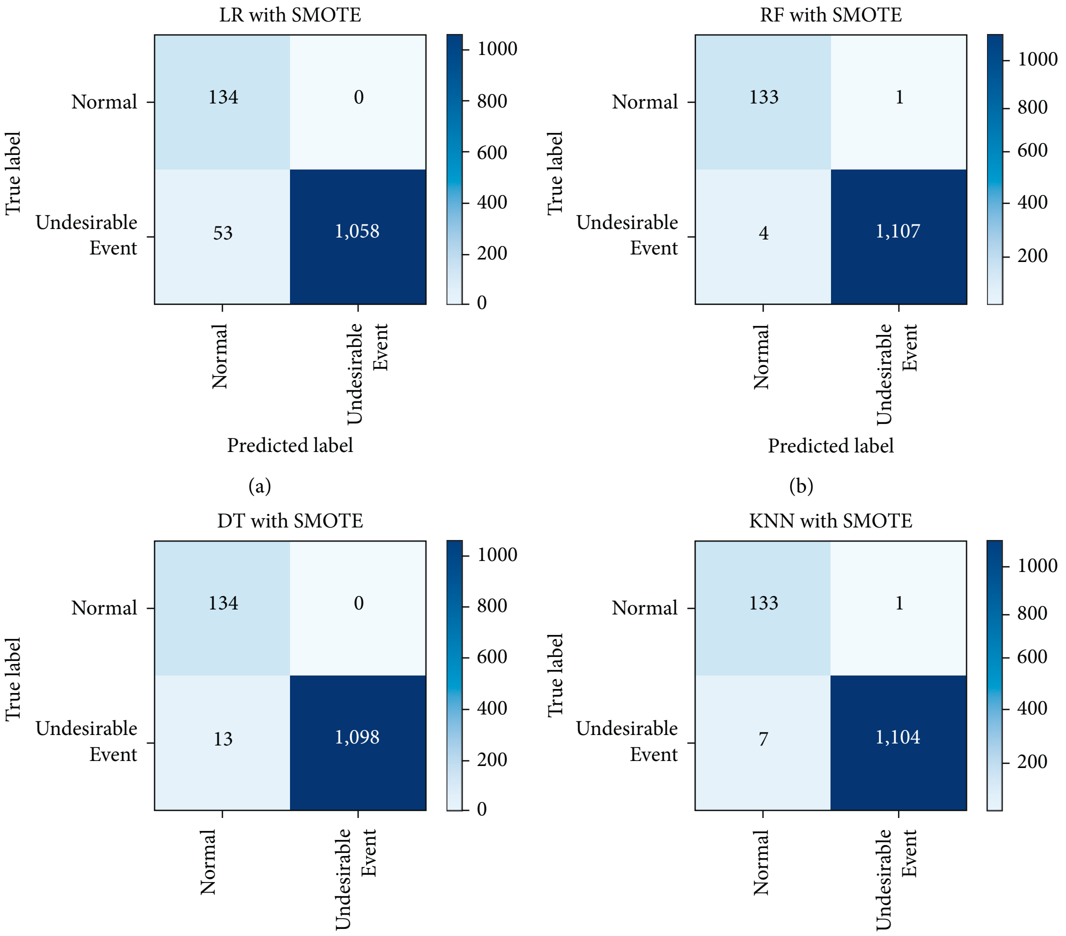
To further evaluate the effectiveness of the models and support the conclusion of considering RF the best among all, the area under the receiver operating characteristics (AUROC) or ROC curves were constructed to assess the discrimination ability of the classifiers with varying thresholds using original and SMOTE data. Figure 5 represents the ROC curve of all algorithms using the SMOTE data, whereas Figure 6 illustrates the ROC curve of all algorithms using the original data.

Figure 5 shows that all the models provide perfect classification, where RF achieved the highest AUC of 1.000, followed by K-NN attaining an AUC of 0.993. Although DT achieved higher accuracy than LR, it is revealed that LR achieved a higher AUC of 0.992 than DT, which achieved an AUC of 0.976. Besides, Figure 6 points out that DT performed significantly better when trained using the original data than the SMOTE data, where it achieved an AUC of 0.991. However, the discrimination ability of LR and K-NN was lower using the original data, where they attained an AUC of 0.990 and 0.982, respectively. It is also proved that RF attained the best results with an AUC of 1.000.

The extreme weather fluctuations in offshore oil wells zones may result in severe disruption in the installed components, increasing the possibility of failure. Additionally, the case of water and natural gas forming a crystalline compound can halt the production for several days. To date, the oil and gas industry lacks accurate measurement instruments to automatically detect the occurrence of undesirable events, resulting in severe injuries, loss of life, economic loss, or environmental pollution [6]. Compared to the benchmark studies, our proposed model achieved higher accuracy than M. Marins et al. [12], who attained an overall accuracy of 94% in detecting faults, and Bronstad [13], who reached the

Table 6: The final results obtained with the optimal hyper-parameters.

Dataset	Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
Original dataset	LR	97.99	97.92	99.82	98.86	0.99
	RF	99.84	99.91	99.91	99.91	1
	DT	99.39	99.55	99.73	99.64	0.99
	K-NN	99.20	99.46	99.64	99.55	0.99
SMOTE dataset	LR	95.74	100	95.23	97.56	0.992
	RF	99.60	99.91	99.64	99.77	1
	DT	98.96	100	98.83	99.41	0.998
	K-NN	99.36	99.10	99.37	99.64	0.993



(a)

(b)

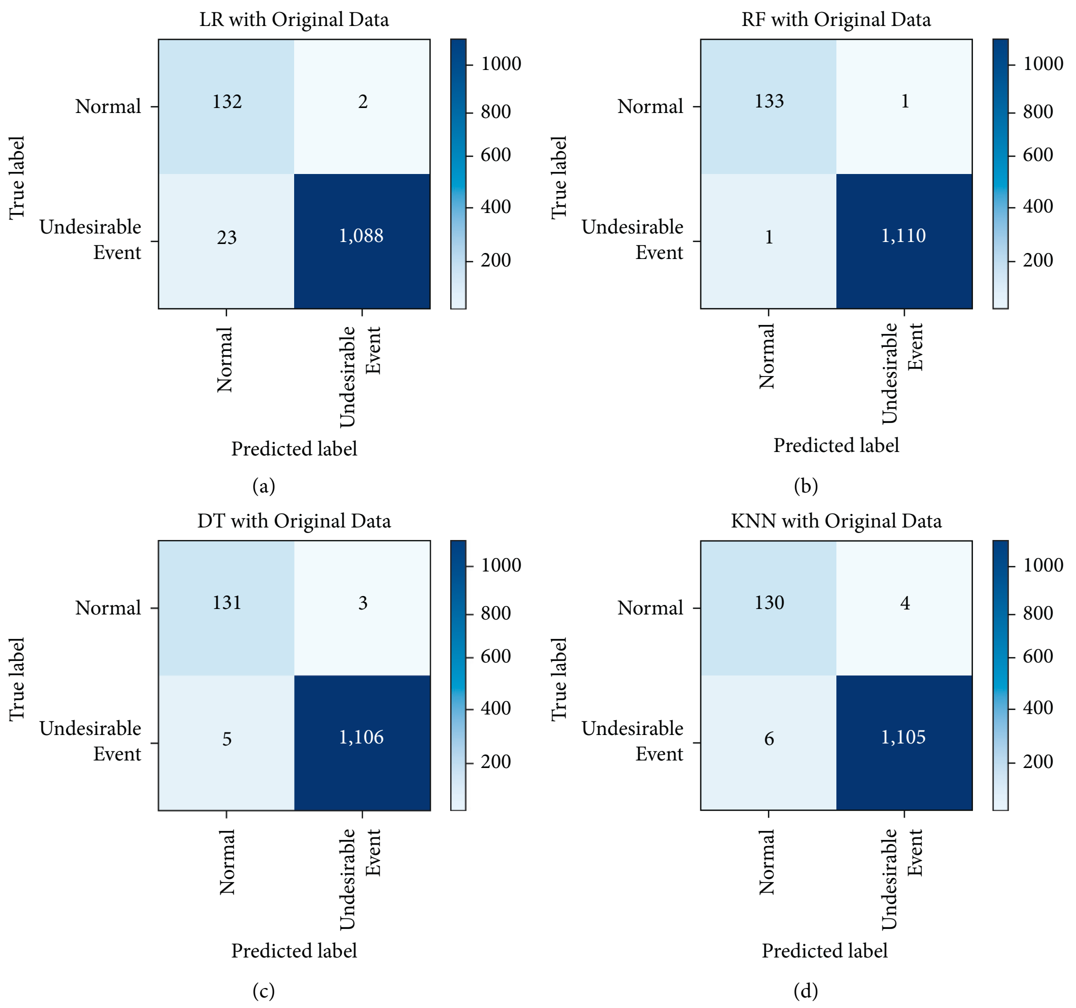
(c)

(d)

Figure 3: Confusion matrix with the SMOTE data: (a) LR, (b) RF, (c) DT, and (d) K-NN.

highest accuracy of 90%. However, the proposed model achieved an accuracy lower than De Salvo Castro et al. [15], who employed unsupervised techniques to enhance the model's performance, achieving an overall specificity and sensitivity of 100% and 99.91%, respectively, using the cohort chart. Although De Salvo Castro et al. [15] reached higher accuracy, unsupervised algorithms are known to be more complex than supervised learning models. Moreover, they are considered to be less reliable and trustworthy. In this study, the proposed RF model achieved promising results that can significantly reduce the possibilities of the eight previously mentioned abnormal events in oil wells. Therefore, deploying the proposed model can play a vital role in detecting and reducing the likelihood of an undesirable occurrence in oil wells.

3.10. Explainable Artificial Intelligence. Machine learning (ML) is considered the next Internet, in which intensive work has been done to introduce its applications in several



(a)

(b)

(c)

(d)

Figure 4: Confusion matrix with the original data: (a) LR, (b) RF, (c) DT, and (d) K-NN.

domains. However, ML algorithms remain a mystery, as they do not reveal information about their internal workings. Therefore, it is crucial to introduce transparency to those models to enable users to understand the reasons and decisions made. There has been a widespread argument that intelligent systems must explain their results. As a result, XAI became a trending topic utilized with ML and deep learning (DL) models. It provides techniques that give insight into how a result was delivered for users to trust, whether local or global. This study employs three XAI techniques, including global surrogate model using DT, Shapley Additive Explanation (SHAP), and Local Interpretable-Agnostic Explanation (LIME).

Global surrogate model is a model-agnostic method, applicable to all algorithms, trained to approximate a black box model's predictions. The technique interprets black box models without taking the models' interior logic into account. In this study, a DT classifier was trained using the predictions produced by the proposed RF model and the original attributes to provide interpretability. Figure 7 illustrates the global surrogate model using DT for the proposed RF model, in which it is concluded that the most influential attribute for the RF classifier to predict the target class is the "T-TPT_standard_deviation."

Below are the rules generated from the global surrogate using DT for RF.

(i) if (T-TPT__sDev>0.0) and (T-JUS-CKP__ sDev>0.0) and (T-JUS-CKP__ sDe >0.0).

then response: 0.0 | based on 3,261 samples.

(ii) if (T-TPT__ sDev≤0.0) and (T-JUS-CKP__ max≤0.576) and (T-JUS-CKP__ sDev≤0.0) and (P-MON-CKP__ sDev≤0.0).

then response: 1609.0 | based on 1,629 samples.

(iii) if (T-TPT__ sDev≤0.0) and (T-JUS-CKP__ max>0.576) and (T-TPT__ sDev>0.0).

then response: 0.0 | based on 298 samples.

(iv) if (T-TPT__ sDev≤0.0) and (T-JUS-CKP__ max≤0.576) and (T-JUS-CKP__ sDev>0.0) and (P-PDG__ sDev>0.0).

then response: 225.0 | based on 252 samples.

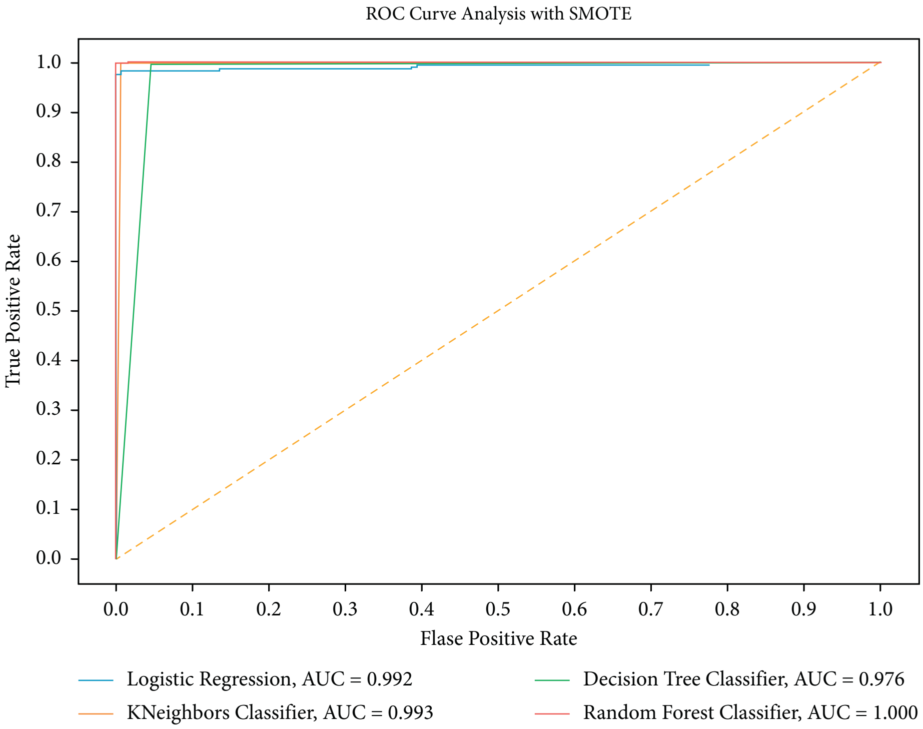


Figure 5: ROC curve of models with the SMOTE data.

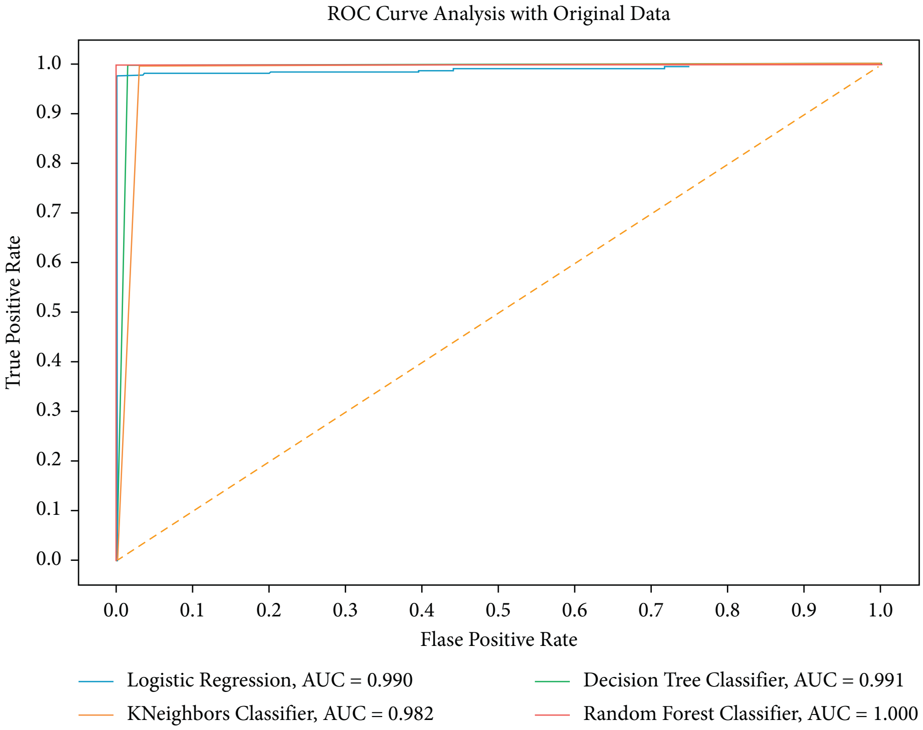


Figure 6: ROC curve of models with the original data.

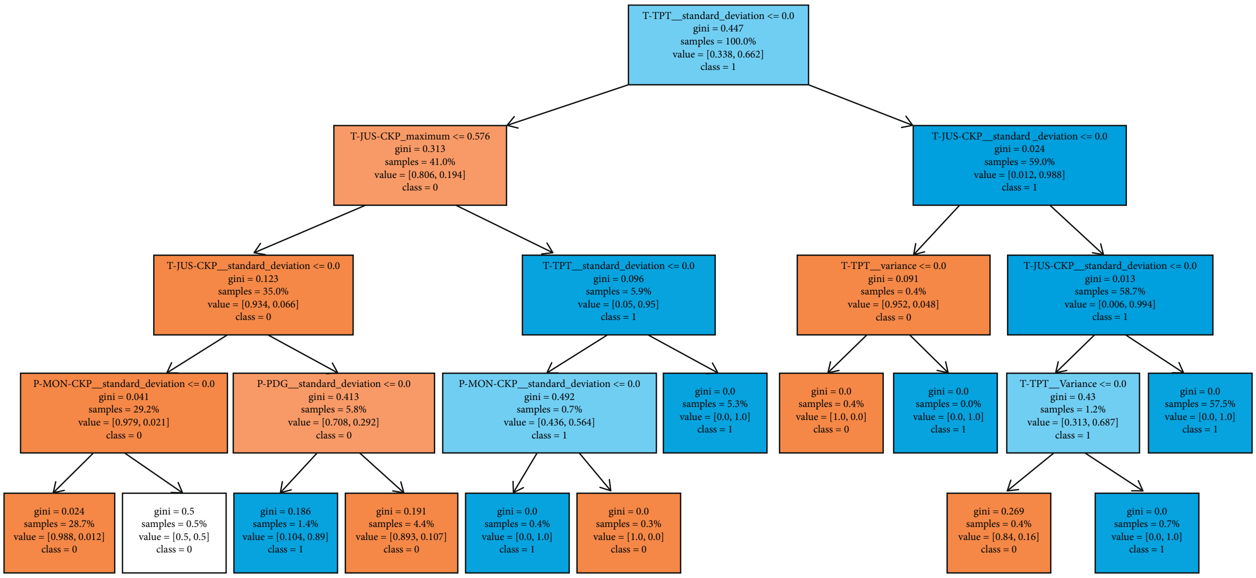


Figure 7: Global surrogate model using DT for RF.

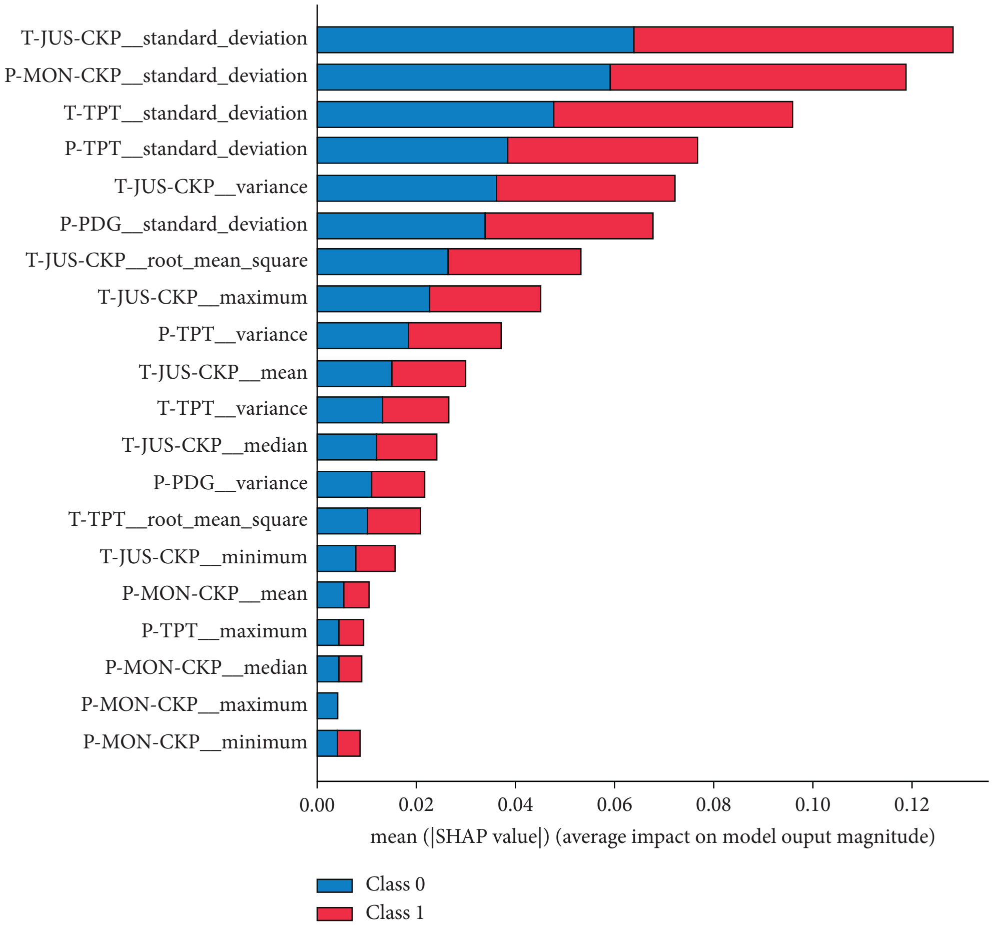


Figure 8: Global explanation using SHAP.

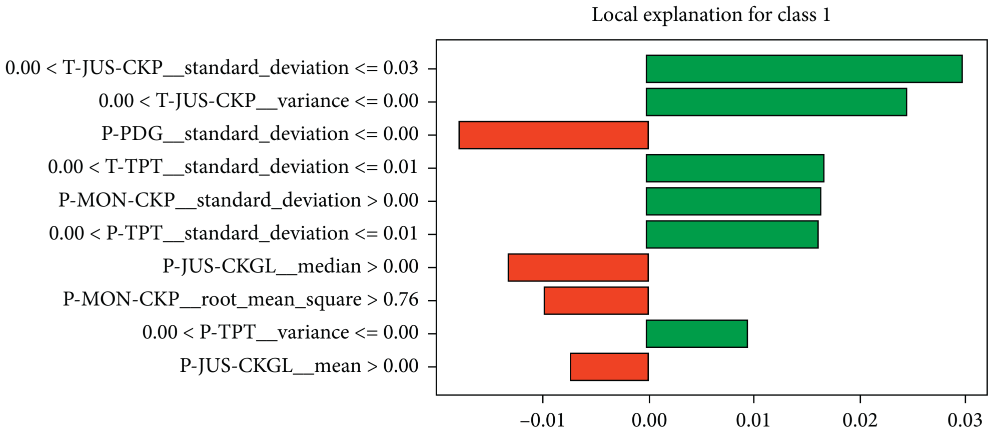


Figure 9: Local explanation using LIME.

(v) if (T-TPT__sDev≤0.0) and (T-JUS-CKP__ max≤0.576) and (T-JUS-CKP__ sDev>0.0) and (P-PDG__ sDev≤0.0).

then response: 8.0 | based on 77 samples.

(vi) if (T-TPT__sDev>0.0) and (T-JUS-CKP__ sDev>0.0) and (T-JUS-CKP__sDev≤0.0) and (T-TPT__var>0.0).

then response: 0.0 | based on 42 samples.

(vii) if (T-TPT__sDev≤0.0) and (T-JUS-CKP__ max≤0.576) and (T-JUS-CKP__sDev≤0.0) and (P-MON-CKP__sDev>0.0).

then response: 15.0 | based on 30 samples.

(viii) if (T-TPT__sDev>0.0) and (T-JUS-CKP__ sDev>0.0) and (T-JUS-CKP__sDev≤0.0) and (T-TPT__var≤0.0).

then response: 21.0 | based on 25 samples.

(ix) if (T-TPT__sDev≤0.0) and (T-JUS-CKP__ max>0.576) and (T-TPT__sDev≤0.0) and (P-MON-CKP__sDev≤0.0).

then response: 0.0 | based on 22 samples.

(x) if (T-TPT__sDev>0.0) and (T-JUS-CKP__ sDev≤0.0) and (T-TPT__var≤0.0).

then response: 20.0 | based on 20 samples.

(xi) if (T-TPT__sDev≤0.0) and (T-JUS-CKP__ max>0.576) and (T-TPT__sDev≤0.0) and (P-MON-CKP__sDev>0.0).

then response: 17.0 | based on 17 samples.

(xii) if (T-TPT__sDev>0.0) and (T-JUS-CKP__ sDev≤0.0) and (T-TPT__var>0.0).

then response: 0.0 | based on 1 sample.

Shapley Additive Explanation (SHAP) is one of the XAI techniques extended from the optimal Shapley value game theory. It calculates the contribution of each feature affecting the outcomes for each instance. Consequently, it informs users of the impact of each feature on the outcome. Figure 8 illustrates the Shapley values for classifying the oil well events. It is indicated that the most influential features are the "T-JUS-CKP_standard_deviation," "P-MON-CKP_standard_-deviation," and "T-TPT_standard_deviation" as they have the highest Shapley values. On the other hand, the least significant features for predicting the target class are the "T-TPT_Maximum," "P-MON-CKP_median," "P-MON-CKP_maximum," and "P-MON-CKP_minimum." Additionally, it is observed that all features contribute equally to both classes.

Local Interpretable Model-Agnostic Explanations (LIME) interprets a model locally to observe its behavior using a single record selected randomly from the test set. In this study, a sample from the positive target class was selected and interpreted in Figure 9. The green color denotes the attributes contributing to undesirable oil well events, while red symbolizes the features contributing to the normal event. It is noted that 6 features contribute to the positive class, whereas the other 4 contribute to the negative class. Moreover, the "T-JUS-CKP_standard_deviation" feature is the most noteworthy feature for predicting the undesirable well event class, whereas the P-PDF_standard_deviation is the most significant feature for predicting the normal event.

4. Conclusion

Severe complications were reported in recent years in offshore oil well zones, resulting in several destruction and production losses. Machine learning (ML) techniques have produced promising results in several domains, especially anomaly detection. Accordingly, the main objective was to develop an ML model to predict the rare undesirable events in oil wells using a realistic and public dataset. Four ML classifiers, namely logistic regression (LR), decision tree (DT), random forest (RF), and k-nearest neighbor (K-NN), were utilized. Furthermore, two experiments were conducted to build the ML models, where the models were trained using the original dataset in the first experiment and upsampled data in the second experiment. The results proved the proposed ML models' effectiveness in classifying undesirable and normal oil well events. Random forest attained the best results with accuracy, recall, precision, F1-score, and AUC of 99.84%, 99.91%, 99.91%, 99.91%, and 1 using original data, while for the upsampled data, it attained 99.60%, 99.64%, 99.91%, 99.77%, and 1.00, respectively. Accordingly, it is concluded that similar results were achieved with original and SMOTE data using RF. Besides, Explainable Artificial Intelligence (XAI) was employed to explain the outcomes produced by the proposed models for users to understand the reasons behind anomaly events. It is observed from the global surrogate model using DT that the most influential attribute for the RF model is "T-TPT_standard_deviation." On the other hand, it is indicated from the Shapley Additive Explanation (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) models that "T-JUS-CKP_standard_deviation" is the most significant feature for classifying the oil well events. Despite the considerable results achieved in this study, the proposed solution is limited to the undesirable events available in the utilized dataset. Moreover, the proposed model is incapable of specifying the type of undesirable event that is likely to happen. Accordingly, this study opens the doors for researchers to extend the experiment by conducting further investigations to develop reliable models that can classify the eight types of undesirable events. Feature analysis could be applied to identify the key features contributing to each undesirable event to enable engineers to take proactive actions and focus on other critical tasks. Additionally, experiments need to be performed on more than one dataset and also a dataset with a huge size.

Data Availability

The study was performed using 3W oil well dataset and can be accessed from the web link, https://www.kaggle.com/datasets/afrniomelo/3w-dataset.

Conflicts of Interest

"The authors declare that there are no conflicts of interest regarding the publication of this article."

References

[1] A. Lawler, *OPEC Sees Upside to 2022 Oil Demand Forecast On Strong Pandemic Recovery*, Arab News, 2022, https://www.arabnews.com/node/2022246/amp.

[2] S. Chakrabarti, J. Halkyard, and C. Capanoglu, "Historical Development of Offshore Structures," *Handbook of Offshore Engineering*, pp. 1–38, 2005.

[3] M. H. Hasan, A. A. Malik, and M. Jasamai, "A Review on Anomaly Detection Methods for Optimizing Oil Well Surveillance," *International Journal of Computer Science and Information Security*, vol. 17, no. 11, 2017.

[4] E. van oort, E. Taylor, G. Thonhauser, and E. Maidla, "Real-time rig-activity detection helps identify and minimize invisible lost time," 2022, https://www.worldoil.com/magazine/2008/april-2008/special-focus/real-time-rig-activity-detection-helps-identify-and-minimize-invisible-lost-time.

[5] A. Sircar, K. Yadav, K. Rayavarapu, N. Bist, and H. Oza, "Application of machine learning and artificial intelligence in oil and gas industry," *Petroleum Research*, vol. 6, no. 4, pp. 379–391, Dec. 2021.

[6] R. E. V. Vargas, C. J. Munaro, P. M. Ciarelli et al., "A realistic and public dataset with rare undesirable real events in oil wells," *Journal of Petroleum Science and Engineering*, vol. 181, Article ID 106223, 2019.

[7] F. K. Dosilovic, M. Brcic, and N. Hlupic, "Explainable artificial intelligence: A survey," in *Proceedings of the 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 210–215, Opatija, Croatia, Jun. 2018.

[8] M. Al-Fadhli and A. Zaher, "A smart SCADA system for oil refineries," in *Proceedings of the 2018 International Conference on Computing Sciences and Engineering (ICCSE)*, pp. 1–6, Kuwait, Jun. 2018.

[9] A. Nwachukwu, H. Jeong, M. Pyrcz, and L. W. Lake, "Fast evaluation of well placements in heterogeneous reservoir models using machine learning," *Journal of Petroleum Science and Engineering*, vol. 163, pp. 463–475, 2018.

[10] CC. Brønstad, *"Data-driven detection and identification of undesirable events in subsea oil wells," Master dissertation*, University of South-Eastern Norway, Norway, 2020.

[11] H. Ghorbani, D. A. Wood, A. Choubineh et al., "Prediction of oil flow rate through an orifice flow meter: Artificial intelligence alternatives compared," *Petroleum*, vol. 6, no. 4, pp. 404–414, Dec. 2020.

[12] M. A. Marins, B. D. Barros, I. H. Santos et al., "Fault detection and classification in oil wells and production/service lines using random forest," *Journal of Petroleum Science and Engineering*, vol. 197, Article ID 107879, 2021.

[13] A. Alsaihati, S. Elkatatny, A. A. Mahmoud, and A. Abdulraheem, "Use of machine learning and data analytics to detect downhole abnormalities while drilling horizontal wells, with real case study," *Journal of Energy Resources Technology*, vol. 143, no. 4, 2021.

[14] M. Aljubran, J. Ramasamy, M. Albassam, and A. Magana-Mora, "Deep learning and time-series analysis for the early detection of lost circulation incidents during drilling operations," *IEEE Access*, vol. 9, pp. 76833–76846, 2021.

[15] A. O. De Salvo Castro, M. De Jesus Rocha Santos, F. R Leta, C. B. C. Lima, and G. B. A Lima, "Unsupervised methods to classify real data from offshore wells," *American Journal of Operations Research*, vol. 11, no. 05, pp. 227–241, 2021.

[16] E. Gurina, N. Klyuchnikov, K. Antipova, and D. Koroteev, "Forecasting the abnormal events at well drilling with machine learning," *Applied Intelligence*, vol. 52, no. 9, pp. 9980–9995, 2022.

[17] B. Alharbi, Z. Liang, J. M. Aljindan, A. K. Agnia, and X. Zhang, "Explainable and interpretable anomaly detection models for production data," *SPE Journal*, vol. 27, no. 01, pp. 349–363, 2022.

[18] A. Alsaihati, M. Abughaban, S. Elkatatny, and D. A. Shehri, "Application of machine learning methods in modeling the loss of circulation rate while drilling operation," *ACS Omega*, vol. 7, no. 24, p. 20696, 2022.

[19] L. Cheng, G Dong, L Ren et al., "Anomaly detection and analysis based on deep autoencoder for the storage tank liquid level of oil depot," *Journal of Physics Conference Series*, vol. 2224, no. 1, p. 012013, 2022.

[20] H. Haji Ali Afzali and J. Karnon, "Specification and implementation of decision analytic model structures for economic evaluation of health care technologies," *Encyclopedia of Health Economics*, pp. 340–347, 2014.

[21] B. Charbuty and A. Abdulazeez, “Classification based on decision tree algorithm for machine learning,” *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, 2021.

[22] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[23] D. Sarkar and V. Natarajan, *Ensemble Machine Learning Cookbook: Over 35 Practical Recipes to Explore Ensemble Machine Learning Techniques Using Python*, p. 327, Packt Publishing, Birmingham. UK, 2019.

[24] P. Cunningham and S. J. Delany, “k-nearest neighbour classifiers: 2nd edition (with python examples),” *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–25, 2020.

[25] M. P. LaValley, “Logistic regression,” *Circulation*, vol. 117, no. 18, pp. 2395–2399, 2008.