

## Original article

# A novel machine learning model for autonomous analysis and diagnosis of well integrity failures in artificial-lift production systems

Adel M. Salem<sup>1</sup>, Mostafa S. Yakoot<sup>2</sup>, Omar Mahmoud<sup>3</sup>✉\*

<sup>1</sup>Department of Petroleum Engineering, Faculty of Petroleum and Mining Engineering, Suez University, Suez 11252, Egypt

<sup>2</sup>District Engineering Division, Gulf of Suez Petroleum Co. (GUPCO), Cairo 11511, Egypt

<sup>3</sup>Department of Petroleum Engineering, Faculty of Engineering and Technology, Future University in Egypt (FUE), Cairo 11835, Egypt

### Keywords:

Machine learning  
well integrity  
risk assessment  
gas lift  
artificial lift systems  
oil and gas wells

### Cited as:

Salem, A. M., Yakoot, M. S., Mahmoud, O. A novel machine learning model for autonomous analysis and diagnosis of well integrity failures in artificial-lift production systems. *Advances in Geo-Energy Research*, 2022, 6(2): 123-142.

<https://doi.org/10.46690/ager.2022.02.05>

### Abstract:

The integrity failure in gas lift wells had been proven to be more severe than other artificial lift wells across the industry. Accurate risk assessment is an essential requirement for predicting well integrity failures. In this study, a machine learning model was established for automated and precise prediction of integrity failures in gas lift wells. The collected data contained 9,000 data arrays with 23 features. Data arrays were structured and fed into 11 different machine learning algorithms to build an automated systematic tool for calculating the imposed risk of any well. The study models included both single and ensemble supervised learning algorithms (e.g., random forest, support vector machine, decision tree, and scalable boosting techniques). Comparative analysis of the deployed models was performed to determine the best predictive model. Further, novel evaluation metrics for the confusion matrix of each model were introduced. The results showed that extreme gradient boosting and categorical boosting outperformed all the applied algorithms. They can predict well integrity failures with an accuracy of 100% using traditional or proposed metrics. Physical equations were also developed on the basis of feature importance extracted from the random forest algorithm. The developed model will help optimize company resources and dedicate personnel efforts to high-risk wells. As a result, progressive improvements in health, safety, and environment and business performance can be achieved.

## 1. Introduction

Well integrity (WI) is a key element of oil field strategy for safety and operating integrity (Anders, 2008; Kiran et al., 2017; Yakoot et al., 2020). The principle of lifecycle well integrity management system (WIMS) is important for ensuring continuous production stability, profitable operational activities, and a zero-leak target (Yakoot et al., 2021a). WIMS was established in many oil and gas (O&G) companies around the world, especially after the Macondo blowout in 2010. However, the level of system maturity differs among organizations. WI data management is a standalone system incorporated inside WIMS and has its pertinent maturity. Over time, not all companies have achieved the same level of

progress in terms of data maturity. This is mainly because the O&G business adopted artificial intelligence and the Internet of Things later than other industries. Machine learning (ML) has been used recently in the O&G industry to predict and forecast oil production, reservoir features, and artificial lift (AL) problems. Meanwhile, there is little published literature recording the adoption of such innovation to predict and analyze the WI status of oil/gas producers.

Successful implementation of WIMS implies a thorough knowledge of all risks that may be the source of uncontrolled flow to the atmosphere or surroundings. WI risk can be efficiently controlled and mitigated on the basis of robust risk analysis and classification (Elmousalami and Elaskary, 2020).

Generally, risk assessment (RA) related to WI operations is either qualitative or quantitative. In a qualitative RA, the industrial safety matrix is used to identify the probability and consequences of the predicted risks. Consensus on the integrity status of a well is considered a difficult process, and differs from one engineer to another in the absence of clear rules. Additionally, RA using a risk matrix is affected by the knowledge and subjective experience of the involved teams (Cox, 2008). In addition, many deficiencies of RA dependence on risk matrices were emphasized, such as;

- Poor resolution caused by range compression.
- Errors resulting from mismatching between qualitative and quantitative ratings.
- Ambiguous inputs and results because risk probability, severity, and rating require a subjective approach, which depends on assessor interpretation.

Further, RA was considered a tedious task by Dethlefs and Chastain (2012). They developed a qualitative RA model to classify failures of well barriers. In recent years, there has been an increasing amount of literature that focused more on quantitative RA using new methodologies. Loizzo et al. (2015) used an evidence-based RA approach to identify high-risk wells that require immediate intervention to reduce the risk level. Abimbola et al. (2016) applied the Bowtie model and Bayesian network to predict failure scenarios during casing and cementing wells. Brechan et al. (2018) developed a lifecycle model and performed RA using software. Zhao et al. (2019) used hierarchical Bayes analysis to assess WI failures. Adeyinka et al. (2020) assessed safety devices related to WI using a Swiss-cheese model. Recently, Yakoot et al. (2021a) have reviewed a large body of literature. They provided a summary of RA approaches over the last two decades. Finally, they identified the progress achieved in RA approaches related to WI from simple qualitative matrices to more complicated models.

### 1.1 Integrity of AL wells

At some stage in the life cycle of a well, AL is mandatory for most oil wells to lift fluids to the surface (Bates et al., 2004; Gupta et al., 2016). Among all AL methods, gas lift (GL) is considered one of the most important AL methods worldwide because of its successful history of operation (Rahmawati et al., 2020). GL is an AL technique that utilize high-pressure gas to improve well production and increase the ultimate oil recovery of a field (Elgibaly et al., 2021). Table 1 compares GL with pump-assisted lift methods from a WI perspective.

It can be inferred from Table 1 that GL is considered the highest risk AL method. Despite the long operational history of deploying the GL method, the documentation of WI management while operating such wells is limited in the literature. One reason for this is the underestimation of risk posed by GL operations for many decades. A typical mitigation for WI issues was to shut down the GL source. A good example illustrating the consequences of losing WI in GL wells is the Prudhoe Bay incident in 2002. As per the investigation report of Alaska Oil & Gas Conservation Commission (AOGCC, 2003), a GL well lost its integrity

while being brought on production. This led to a fire and explosion that badly injured the company operator. Recently, Miraglia (2020) has established a probabilistic model for GL wells. The numerical analysis of the model showed that the probability of completion failure, resulting from corrosion, is higher in GL wells than other AL wells. The high failure probability is attributed to the effect of corrosion on weakening the metal resistance to the GL pressure.

With that being said, it has therefore taken our interest to:

- Modernize and step-up WI management and RA from regular spreadsheets and non-user-friendly software into a system that utilizes ML.
- Develop a RA model for automated quantitative WI classification.
- Consider all barrier envelopes and leak paths in GL wells to produce a precise WI prediction model.

This study presents a ML model to perceive, identify, and avoid operational issues related to WI. In addition, the model captures the key features in GL wells, related to operational integrity and outflow system stability. The proposed model was validated with field data from a giant brownfield. ML was used to develop an automated model with the following features:

- It provides a unique method to convert the associated failure risk of each element in the well envelope into tangible numbers. These numbers sum up to show the total potential risk and hence the status of the well-barrier integrity system.
  - It can be used for any well stock with the same design parameters. As a result, the RA is quicker and easier to carry out, and avoids the time-consuming assessment for wells on an individual basis.
  - The layout can be simply adjusted to reflect the risk profile of any well type or field.
- Furthermore, the ML model has been recognized to:
- Effectively evaluate WI risks for thousands of wells.
  - Concentrate on the share of each integrity element in the whole envelope.
  - Successfully guide senior management for the most efficient dedication of existing resources to implement WI consistently.

This model will supersede a similar model introduced by Yakoot et al. (2021c) and also replace the common qualitative risk-evaluation process.

To the best of the authors' knowledge, this study presents several innovative contributions. First, the motivation behind the study was elucidated. The integrity of AL wells, generally, and GL, in particular, was discussed. Next, ML principles and types were overviewed. Then, building a ML model was described in detail. This part covers data gathering, data pre-processing, feature engineering, algorithm selection, model training, and model evaluation. Then, research results were investigated. The principle of risk mapping was introduced for the first time. Enumerative combinatorics were applied to produce limitless failure scenarios. ML model was analyzed and evaluated using traditional metrics. New evaluation metrics were introduced by applying domain knowledge experience. And finally, physical equations were developed for easier

**Table 1.** Comparison between GL and other AL methods in WI risks.

Major concern	GL	Pump-assisted lift
Annulus pressure	<ul style="list-style-type: none"> <li>Gas is applied with high pressures up to 3,000 psi.</li> <li>The damage of the gas injection line and/or wellhead stack causes gas venting at high pressure (Brodie, 2011).</li> </ul>	<ul style="list-style-type: none"> <li>Not Applicable.</li> </ul>
Annulus integrity and failure scenarios	<ul style="list-style-type: none"> <li>Annulus integrity is important.</li> <li>Failure of GL mandrel, valves, and completion means hydrocarbons with sustained pressure fill the annulus.</li> <li>In case of failure, gas will spread at the wellhead with high H<sub>2</sub>S concentrations.</li> <li>Check GL valves is a major WI compromise.</li> </ul>	<ul style="list-style-type: none"> <li>Not Applicable.</li> </ul>
System tripping	<ul style="list-style-type: none"> <li>Not Applicable.</li> </ul>	<ul style="list-style-type: none"> <li>Tripping the pump or cutting the power source causes immediate well dead.</li> </ul>
Production from annulus	<ul style="list-style-type: none"> <li>Not Applicable for the situation of injecting gas into the annulus.</li> </ul>	<ul style="list-style-type: none"> <li>Applicable.</li> <li>Most of the wells have hydrocarbons in the annulus.</li> </ul>
Gas-liquid ratio (GLR)	<ul style="list-style-type: none"> <li>GL is originally recommended for high GLR wells.</li> </ul>	<ul style="list-style-type: none"> <li>High GLR will harm the pump.</li> </ul>
Well production	<ul style="list-style-type: none"> <li>Inherent relationship between WI problems and production efficiency (Ismail and Trjanganung, 2014).</li> <li>Deterioration in surface and sub-surface equipment impair WI with holes and leaks.</li> <li>Corrosion causes difficulty in retrieving GL valves and optimizing performance.</li> <li>Shallow point of injection affects production.</li> <li>Inefficient liquid lifting in wells leads to instability in the system.</li> </ul>	<ul style="list-style-type: none"> <li>Not Applicable.</li> </ul>

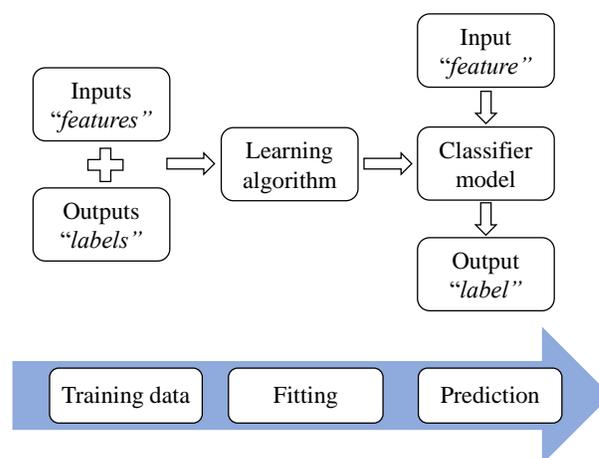
calculation.

## 2. Building the ML model

As the Internet of Things is more widely adopted, O&G operators need to adapt to support the transformation. With technology advancements and a growing network of sensors, faster and higher frequency data gathering is possible. As a result, data volumes have increased significantly. Mishra and Datta-Gupta (2017) described big data by having velocity, volume, and variety. Big data is considered an integrated part of the O&G industry (Holdaway, 2014; Saputelli, 2016). It is currently normal practice to use ML algorithms to make predictions from complicated systems with several implicit variables (Wood, 2018). Rahmanifard and Plaksina (2019) highlighted the exceptional performance of ML and AI techniques. ML has been used extensively in O&G industry in analysis (Wood, 2022), prediction (Yavari et al., 2021), and evaluation (Mahdiani et al., 2020). Choubey and Karmakar (2021) reviewed AI and ML techniques and their contribution to the O&G sector. They provided a technical approach that enables gathering information from big data in the industry through an effective selection of ML and AI techniques.

ML is a branch of AI that focuses on inductive learning, in which predefined datasets can be trained to induce self-learning models. The created model can predict new or held-out data (Fig. 1). De Carvalho and Freitas (2009) listed classification problems as one of ML tasks, beside clustering, regression, and optimization.

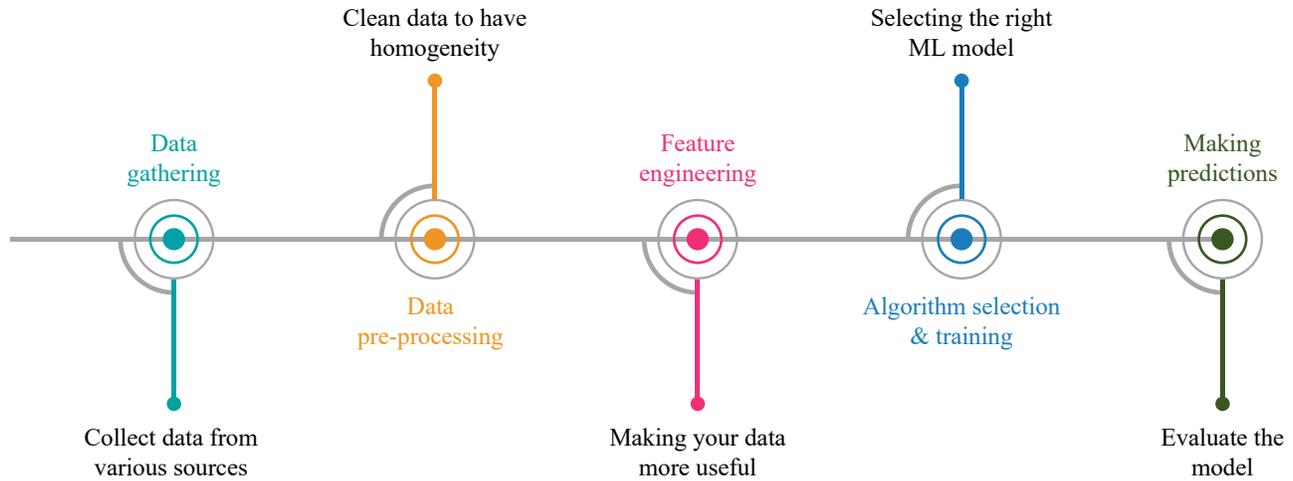
In ML and data mining fields, flat classification problems

**Fig. 1.** Supervised ML pipeline.

are divided into two-class (binary) or multi-class problems. Both problems are single-label classification, and both forecast only the classes at the structured leaf nodes (Silla and Freitas, 2011).

The objective of this paper is to analyze the GL wells and establish a ML model capable of simulating WI problems. Further, the model can be divided into three sub-models;

1. Predictive Model: to predict the accurate risk status of wells and classify their integrity level into five categories, rather than three broad-range categories.
2. Failure Model: to identify whether the well is considered in failure mode or not. In addition, the model can identify wells that require prompt mitigation and securing.



**Fig. 2.** Steps to solve a ML problem.

3. Suboptimal Model: to identify which wells are operating out of their integrity envelope and have some elements impaired.

The appropriate package was selected for scientific computing, performing different operations, data manipulation and analysis, and data visualization. Steps to solve the WI problem using ML are depicted in Fig. 2.

## 2.1 Data gathering

A typical production well configuration is shown in Fig. 3 and is considered during the classification process. All wells in the ML model have two sealed annuli. A-annulus is the production annulus between completion and production casing. B-annulus is the cemented annulus.

Almost 9,000 data arrays were collected from 800 GL wells for the last decade. Each data array represents a well event and comprises 23 predictors that are barrier elements related to the final status of WI. These elements (features) are divided into seven groups. The features and their relevant groups (envelopes) are listed in Table 2. The envelope of vertical barriers represents vertical Christmas tree (X-tree) valves and downhole safety valve (DHSV) that isolate the reservoir fluids. Upper master valve (UMV) and DHSV are safety-critical valves. They are connected to the tower emergency shut down (ESD) system. The horizontal barriers are mechanical valves downstream of the X-tree and on production/test headers. They provide isolation between well fluid and surface facilities. The gas line envelope represents the mechanical valves upstream of the gas injection choke, and they isolate high-pressure gas from production annulus. The wellhead envelope considers the integrity of both production and cemented annuli, in addition to wellhead valves. The structure envelope accounts for conductor movement, tower structure integrity, and risk of tower collision by ships. The flowing capability demonstrates the production rate of the well and its capability to produce naturally. This envelope greatly influences the WI risk severity. The health, safety, and environment (HSE) envelope justifies the impact of WI failures on people, environment, and production. The detailed

description of each envelope with its combined features is presented in Yakoot et al. (2021b).

## 2.2 Data pre-processing

Data gathering, processing, and analysis are defined as analytics (Bravo et al., 2014a). Data pre-processing is the longest phase to require effort to extract hidden knowledge in the data (Fernández et al., 2018). Pre-processing of data refers to the transformations applied to the data before feeding it to the algorithm. This exercise includes solving problems of missing values, outliers, biased data, etc (Bravo et al., 2014b).

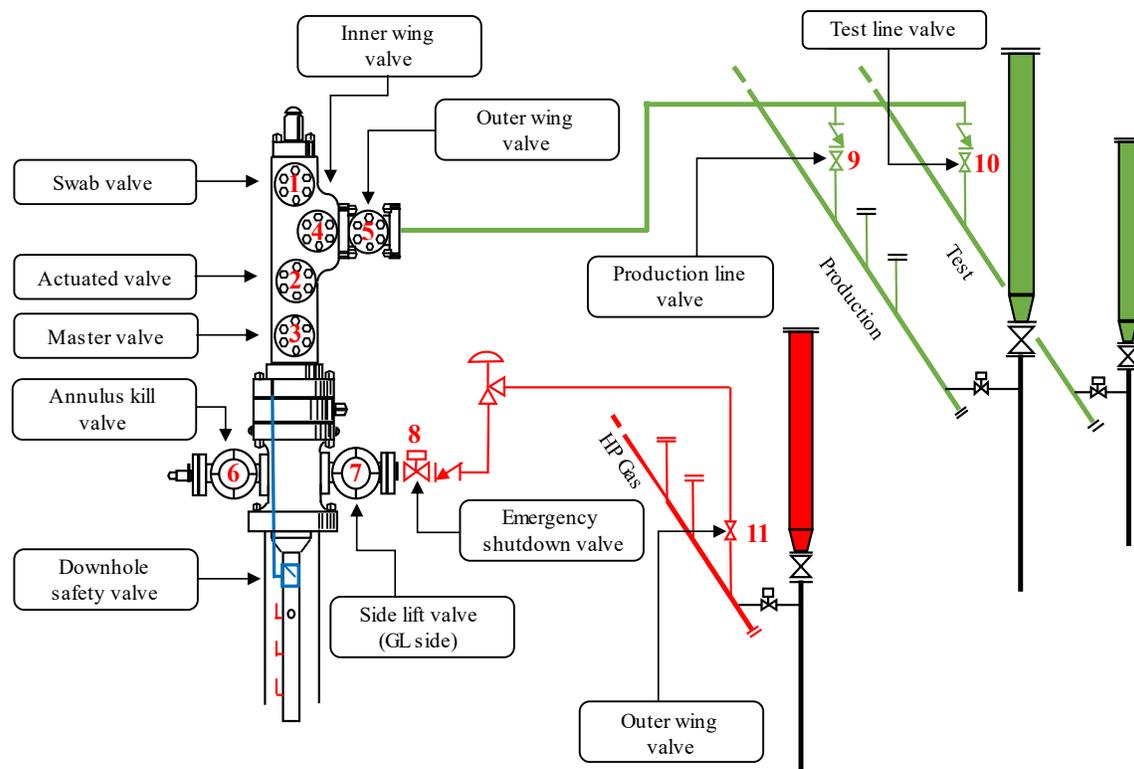
Data cleaning is an essential step that comes directly after data gathering to check any issues with the data and remove/fix data if required (Li et al., 2021). The relevance of independent variables to the dependent variables differs from one feature to another (Holdaway, 2014). The collected dataset couldn't be fed directly into the model, so data cleaning went through the following stages:

- All wells killed and/or secured by downhole plugs were excluded from the dataset to focus on the integrity status of operating wells.
- Execution date and status validity for all WI tests were removed because they don't contribute to the final risk of the well. Additionally, all statistical arrays were removed because they don't affect WI status.
- The preventive maintenance code and the due-date for the next test of different features were removed, because they are not relevant to WI status.
- All manufacturer data of the X-tree and DHSV were removed, because they don't not affect WI risk.
- Operation status, field name, complex platform, and satellite platform were removed, because they are not relevant to features influencing WI.

Part of the features in the dataset are continuous (oil/gas production rates, H<sub>2</sub>S concentration, and B-annulus pressure), and the rest are categorical. Imputation of categorical features was implemented on the dataset by replacing the missing values with the other compatible values using subject matter experience. All missing values in data of mechanical valves

**Table 2.** Data labeling for all features.

Barrier envelope	Feature	Integrity status								
		Hold	Pass	Leak	Not tested	Not available	Stuck close	Stuck open	Isolated	
Vertical barriers	DHSV	7014	330	349	511	252	75	852	0	
	Swab valve	7454	429	1011	407	52	0	26	4	
	Actuated valve	7367	322	915	430	68	0	277	4	
	Master valve	7467	261	1205	419	5	5	17	4	
Horizontal barriers	Inner wing valve	5759	256	1386	358	1474	2	40	108	
	Outer wing valve	5941	225	1433	503	842	20	262	157	
	Production line valve	4956	0	2795	889	134	0	175	434	
	Test line valve	3793	1	2187	943	1002	14	65	1378	
Gas line	ESD valve	4912	31	1747	1092	1368	0	45	188	
	Gas injection line valve	6546	0	644	830	564	10	290	499	
	Annulus kill valve	6705	179	930	1360	173	0	22	14	
	Side lift valve (GL side)	6370	62	324	2297	181	0	45	104	
Wellhead	Completion integrity	Yes (no communication between tubing and production annulus)			No (there is communication between tubing and production annulus)			Not tested		
		5205			1497			2231		
	B-annulus pressure	B-annulus pressure $\leq$ 100 psi			B-annulus pressure $>$ 100 psi			Not available		
		5030			762			3591		
Flowing capability	Natural flow capability	Natural-flow well				Gas-lift well				
		1815				7568				
	Rate of produced oil	$<$ 100 BOPD GL wells: 2097			100-1000 BOPD GL wells: 5235			$>$ 1000 BOPD GL wells: 236		
		Natural flow wells: 326 Total: 2423			Natural flow wells: 1267 Total: 6502			Natural flow wells: 222 Total: 458		
Rate of produced gas	$<$ 0.5 MMSCFD GL wells: 544			0.5-5 MMSCFD GL wells: 6231			$>$ 5 MMSCFD GL wells: 793			
	Natural flow wells: 181 Total: 725			Natural flow wells: 1443 Total: 7674			Natural flowwells: 191 Total: 984			
HSE	Tower manning	Complex (manned)				Satellite (unmanned)				
		1452				7931				
	H <sub>2</sub> S concentration	$\leq$ 100 ppm		$>$ 100 ppm		Not available		Not tested		
	4754		4297		287		45			
	Risk of pollution	Yes (close to resorts, populated areas, and/or reserves)				No (far from resorts, populated areas, and/or reserves)				
		1013				8370				
Structure	Conductor movement	Yes				No				
		384				8999				
	Risk of ship collision	Yes (close to shipping lane)				No (Far from shipping lane)				
		966				8417				
	Structure integrity	Yes (there is no structural problem)				No (there is a structural problem)				
		9368				15				
Total number of data arrays					9383					



**Fig. 3.** Typical production well configuration.

were treated the same way as non-integral valves. This may have increased the risk category of some wells. However, accurate prediction of WI status was still achieved. The same rule applied to B-annulus pressure. Wells without H<sub>2</sub>S reading are considered sweet wells with H<sub>2</sub>S <100 ppm. Wells without completion integrity tests were adjusted to integral completion. This process exhausted a long time, yet resulted in a more accurate model.

Data cleaning resulted in 23 independent variables believed to have the highest and most serious impact on WI status. They were then ready to be fed into the model for applying the ML algorithms. A description of pre-processed data is presented in Table 2.

### 2.3 Feature engineering

Feature engineering is the process of incorporating domain knowledge to identify attributes from raw data. These features can be used to boost the efficiency of ML algorithms (Bontempi, 2021). We adopted “one-hot encoding” in our research, because domain knowledge implies that each feature has a unique contribution to the final rank of WI status. “One-hot encoding” is one method of categorical encoding that converts categories into numbers. This method creates dummy variables for each data feature. Table 3 illustrates the weight scores for WI elements, which affect both probability and consequences (severity).

Essentially, the risk of WI failure is a product of probability and severity. Risk probability represents the likelihood of failure, while severity accounts for the impact of this failure on humans, environment, and assets. In our research, the

probability is related to barrier elements, specifically that their individual breakdown will increase the possibility of an entire WI failure. This includes mechanical isolation valves installed on the wellhead and X-tree, downhole integrity, and tower structure. Severity is related to elements that maximize and escalate the failure impact on HSE. This includes; capability of wells to produce naturally, their production rates, tower manning, pollution risk, and concentration of toxic gases.

Since many ML algorithms cannot run directly on categorical data and require all variables to be numeric, categorical data were converted into numbers, which ML models can better understand. Using categorical data directly in the ML model will generate unexpected results or poor performance, because the model assumes natural ordering of the data. It is important to weight feature on the basis of domain knowledge experience (Appendix A). As a result, ML models will perform better and the weight of features will be accurately reflected in the prediction of WI risk category.

### 2.4 Algorithm selection

The field of ML includes an enormous number of algorithms. Some of these algorithms are easy to use, while most of the algorithms are more complex and require more knowledge to understand (Khan et al., 2020).

Websites and internet blogs provide open-source material about ML algorithms. Additionally, the literature contains an abundance of informative textbooks and peer-reviewed papers (Raschka, 2015; Hastie et al., 2017; Fernández et al., 2018; Elmousalami and Elaskary, 2020; Shalaby et al., 2020; Bontempi, 2021; Choubey and Karmakar, 2021; Ragab et al.,

**Table 3.** Weight scores for model features.

Risk contribution	Variable	Weight score
Probability	1 Swab valve	0.02
	2 Actuated valve	0.20
	3 Master valve	0.02
	4 Inner wing valve	0.02
	5 Outer wing valve	0.04
	6 Annulus kill valve	0.10
	7 Side lift valve (GL side)	0.02
	8 ESD valve	0.10
	9 Production line valve	0.02
	10 Test line valve	0.02
	11 Gas injection line valve	0.02
	12 DHSV	0.10
	13 Completion integrity	0.12
	14 B-annulus pressure	0.04
	15 Conductor movement	0.06
	16 Risk of ship collision	0.02
	17 Structure integrity	0.06
	Safety factor	0.02
	Total	1
Consequences	18 Natural flow capability	0.30
	19 Tower manning (complex vs. satellite)	0.03
	20 Risk of pollution	0.15
	21 H <sub>2</sub> S concentration (ppm)	0.03
	22 Rate of produced oil (BOPD)	0.00 0.15 0.30
	23 Rate of produced gas (MMSCFD)	0.45
		0.04
		0.04
		0.04
		Total

2021; Tang et al., 2021; Salem et al., 2022).

Model performance may be the most important factor in selecting an algorithm for ML projects. Nevertheless, Vidiyala (2020) listed some other factors that guided us in selecting the best algorithm to address the WI problem. These factors can be summarized as:

- Interpretability: it is required to understand the model results. Decision trees, K-nearest neighbors (KNN), and

random forests are particular algorithms that generate easily interpretable models, while other models are less interpretable due to the high complexity of their construction.

- Model assumptions and data linearity: some algorithms, such as logistic regression and support vector machine (SVM), work properly with linearly separable data<sup>1</sup>, while ensemble models are a good choice for non-

<sup>1</sup>SVM works properly with linearly separable data (linear SVM). When a dataset has non-linearly separable data, we use Kernel functions (tricks) to transform data into another dimension and classify it easily.

linearly separable data. The WI problem has non-linearly separable data.

- Nature and size of data points and features: this factor is significant in algorithm selection. Data formats are either numerical or categorical. When categorical features are predominant, tree-based algorithms are more interpretable and hence considered a good choice. In the study, there are 19 categorical parameters.
- Accuracy: assumptions of each algorithm control the model accuracy in diverse scenarios. Maximizing accuracy is the primary target. Nevertheless, when the dataset contains highly imbalanced classes (i.e., one class has a bigger size compared with the others, similar to this study's research data), algorithms like logistic regression are not likely to perform well. SVM can be used since it can deal with class bias. Boosting techniques are originally used to improve the accuracy of any weak classifier.
- Rate of convergence: logistic regression and decision trees have a faster rate of convergence compared with SVM and random forests. This concern is mainly related to enormous datasets, in which learning time is typically hours.
- Learning time and memory requirements: some algorithms like KNN and logistic regression require less training time than others. Other algorithms tend to be more central processing unit and/or memory intensive. However, if computing power is sufficient to cover such requirements and conditions, this factor is negligible.

Considering the factors listed above, 11 ML algorithms<sup>2</sup> (classifiers) that apply to WI classification problem were selected:

- Logistic regression, which fits data to the logistic function to predict event probability.
- Naive Bayes (NB), which relies on the Bayes' theorem.
- Decision trees, in which a dataset is divided into multiple homogeneous sets.
- Random forests which consist of several decision trees.
- KNN that applies a voting process among K-nearest neighbors of a data point to classify it.
- SVM that attempts to draw an optimal hyperplane to differentiate the classes accurately.
- Stochastic gradient descent (SGD), in which a random selection of few samples is used to calculate the gradient for each iteration.
- Quadratic discriminant analysis (QDA), in which a specific covariance matrix is given for every class.
- Boosting techniques e.g., extreme gradient boosting (XGB), adaptive boosting (AdaBoost), and categorical boosting (CatBoost). In these boosting algorithms, a number of weak learners are combined to build a stronger model.

The selected models with their respective results are tabulated in the Supplements.

## 2.5 Model training

Typically, when a dataset is separated into a training and testing sets, most of the data is used for training and a smaller portion of the data is used for testing. Separating data into training and testing sets is an important part of evaluating the model and determining its performance when it is used in real life (Olukoga and Feng, 2021). This action identifies the precision in the selected algorithm, depending on the result. A better way to check the accuracy of a model is to see its performance on data not used during training (Raschka, 2015). Our research dataset was divided into 70% training and 30% testing parts, as usually adopted (Yin et al., 2021).

Hyperparameters are defined as parameters that can control the learning process of ML models. Their values are set before starting the learning process, and subsequently tuned to produce the best predictions. The value of hyperparameters is significant in the learning process of ML algorithms. Optimizing hyperparameters enables the model to ideally solve the ML problem and minimize the loss (error) function. The traditional optimization approach was applied for hyperparameters, which is called "grid search". We defined (for each model) the hyperparameters that needed optimization for best performance. Grid search was used to train each algorithm with different combinations of hyperparameters and evaluated model performance by five-fold cross-validation on the training set. Then, the validation dataset was used to evaluate model performance after training. This process is called holdout cross-validation (Fig. 4). Finally, the function selected the best hyperparameters that fit the model on the training set. Table 4 summarizes the optimized hyperparameters for some algorithms. Other hyperparameters were set to the default values.

In this work, the data are obviously imbalanced. We have an unequal distribution of classes. Categories-3 and 4 (Cat. 3 and Cat. 4) have the highest observations, while categories-1

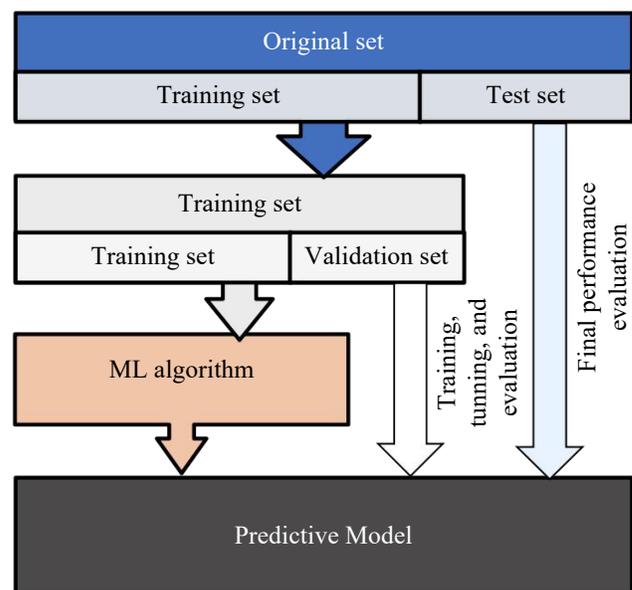


Fig. 4. Holdout cross-validation process (Raschka, 2015).

<sup>2</sup>Random forest, XGB, AdaBoost, and CatBoost are ensemble models that are decision-tree-based algorithms. Therefore, the number of algorithms is essentially fewer than 11.

**Table 4.** Hyperparameter tuning.

Algorithm	Hyperparameter	Hyperparameter tuning
Logistic regression	Regularization parameter (C)	0.1
SVM	Kernel (control the process of projecting input variables)	RBF (Radial basis function)
	Regularization parameter (C)	0.1
Random forest	Number of trees in the forest (n_estimators)	100
Stochastic gradient	Learning rate	Optimal
	Depth of the tree	6
	Learning rate	0.05
CatBoost	Number of trees (iterations)	100
	Learning rate	0.05
KNN	Number of neighbors	5
XGB	Learning rate	0.1
	Number of boosting stages to perform (n_estimators)	150
	Depth of each tree (max_depth)	6

**Table 5.** Number and percentage of samples in each class.

Well category	Number of samples	Percentage (%)
Cat. 1	101	1.1
Cat. 2	145	1.5
Cat. 3	3214	34.3
Cat. 4	3145	33.5
Cat. 5	2778	29.6

and 2 (Cat. 1 and Cat. 2) have the lowest observations. This is demonstrated in Table 5.

To avoid a high bias and enhance the performance of applied classifiers, imbalanced data were treated beforehand. The dataset was resampled using the synthetic minority over-sampling technique, in which synthesized samples were generated for minority classes. This technique was preferred over undersampling techniques to avoid removing observations that may have resulted in losing valuable information from the dataset (Mohammed et al., 2020).

## 2.6 Model evaluation

The evaluation of model performance is related to its ability to predict on hold-out test data (Hastie et al., 2017). A confusion matrix is simply a combination of the actual status and model predictions used to measure model performance (Kazak et al., 2021). It is a special table structure that enables for the evaluation of an algorithm's performance. All the predicted values by the ML model lie under one of four categories illustrated in Fig. 5. The ratio of true positive predictions compared to total positive predictions is called "Precision", while the ratio of true positive predictions compared to total actual positives is called "Recall". Accuracy refers to the ratio of accurate predictions compared to the total predictions. F1-score represents the harmonic average of precision and recall.

## 3. Results and discussion

Classification of WI failures is an essential step for safely

		Predicted	
		Predicted negative	Predicted positive
True (actual)	Actual negative	True negative (TN)	False positive (FP)
	Actual positive	False negative (FN)	True positive (TP)

**Fig. 5.** Confusion matrix.

and successfully operating O&G fields. Therefore, the categorization of WI anomalies and failures must be created for all elements of WI envelopes and converted into an appropriate risk level.

### 3.1 Principle of risk mapping

In addition to its original tasks, feature engineering enabled the authors to have a deeper view of the process of quantifying WI failures. Projecting levels of risk probability and severity on the company's 5X5 risk matrix distinguished between five different WI categories, instead of the previous three risk levels (high, medium, and low). Cat. 1 and Cat. 2, represented in red and amber zones (Fig. 6), are the highest risk and must be reported immediately upon identification. The risk must be reduced as soon as possible to as low as reasonably practicable level. Cat. 3 represented in the yellow zone are medium risk wells, for which the effectiveness of existing controls must be improved. Cat. 4 and Cat. 5 represented in the light and dark green zones are wells with the highest integrity.

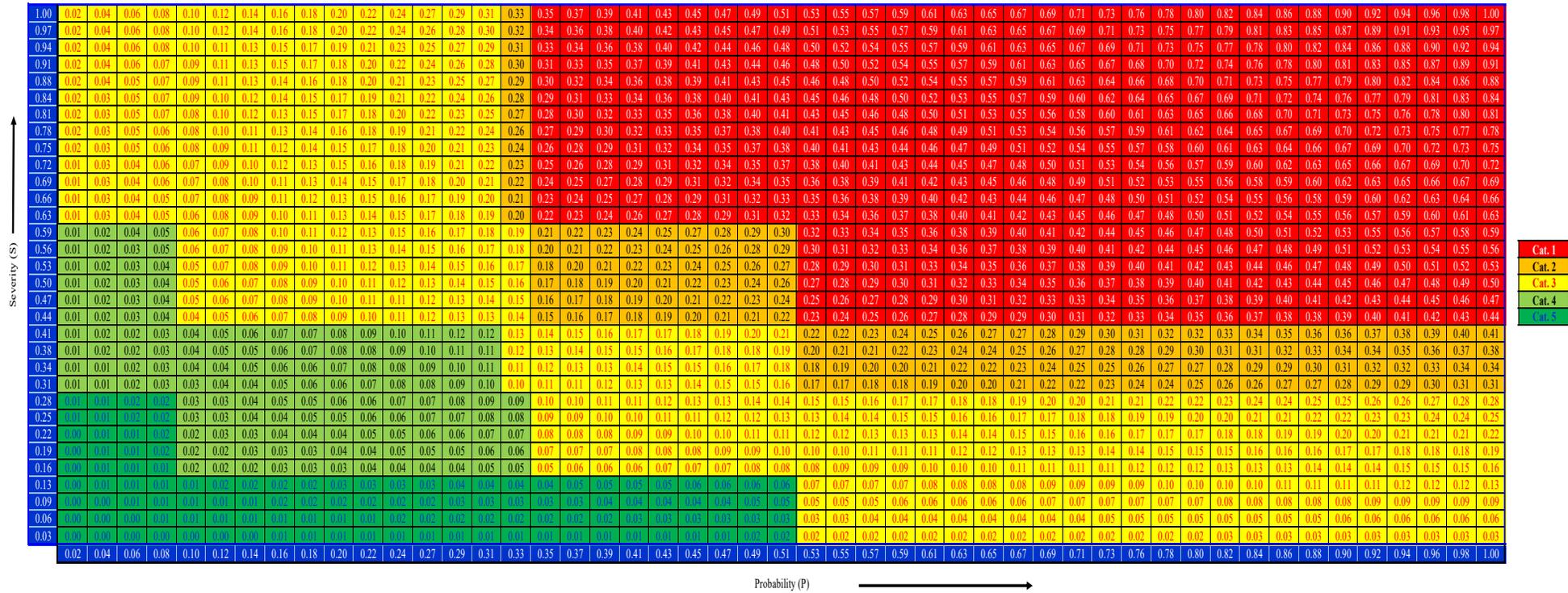
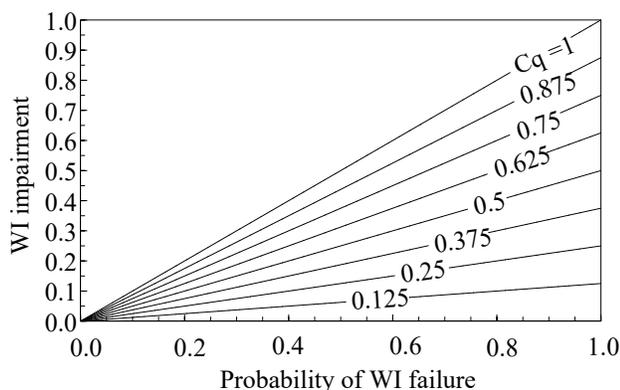


Fig. 6. Generated risk map.



**Fig. 7.** WI category at different severity layers.

The newly developed risk matrix represents a massive transformation in the approach of risk analysis for WI. However, it is still limited by the boundaries of categories, which makes the approach less effective in some critical cases. Accordingly, the risk map principle was introduced to eliminate the limits between categories and identify any WI impairment using the two main parameters of risk; risk probability and risk severity. The risk map which was developed in Fig. 6, empowered the risk analysis of integrity status and made the overall result more accurate.

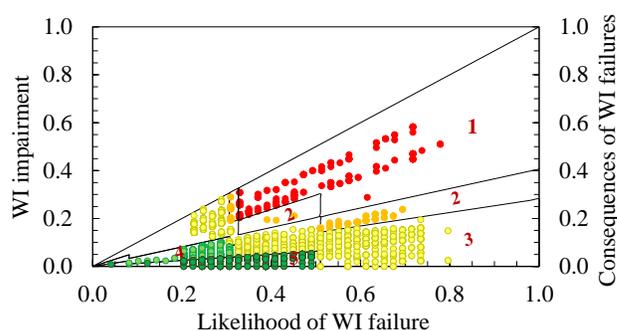
The new map converted risk probability and severity to numerical continuous values, and as a result, risk calculation became more specific. As risk is a function of probability ( $P$ ) and severity ( $S$ ), the volume of risk can be expressed as double integration of their product (Volume of Risk =  $\int \int P S dP dS$ ). A 3D-map can be visually presented for each category. To demarcate the WI failures more precisely, the level of WI impairment was identified for each layer of failure severity (Fig. 7) and dataset observations were projected on the developed WI monograph (Fig. 8). Supplementary to these results, Yakoot et al. (2022) delineated boundaries for each category and presented a 3D-map for the wells of Cat. 5 based on the “volume of risk” principle.

### 3.2 Application of enumerative combinatorics

There are 17 probability-related features, as predefined in Table 3. For more simplicity, different integrity scenarios for each feature have been converted into binary scenarios (i.e., integral or non-integral). In the same manner, four severity-related features have been converted into binary scenarios. O&G rates were converted into four scenarios, in which the rate may be zero, less than 100, between 100 and 1,000, or more than 1,000 BOPD for oil production. For gas production, the rate was split into zero, less than 0.5, between 0.5 and 5, or more than 5 MMSCFD. Applying simple mathematical combinations to all failure scenarios of barrier elements and variables, a massive number of WI risk scenarios and 1,568 unique risk categories can be concluded.

### 3.3 Application of enumerative combinatorics

The WI classification model is hypercritical because it has a direct impact on field operations. Classifying the risk level of WI has resulted in a 5-class classification model. The data-



**Fig. 8.** WI monograph.

set is composed of nearly 10,000 WI tests. After some pre-processing, analysis resulted in 9,383 data points distributed amongst the different classes.

This model is a multi-class model, and evaluation metrics are directly derived from the following four groups:

- 1- True Positive (TP): Labels are truly positive and predicted as positive.
- 2- False Positive (FP): Labels are truly negative and predicted as positive.
- 3- True Negative (TN): Labels are truly negative and predicted as negative.
- 4- False Negative (FN): Labels are truly positive and predicted as negative.

Any multi-class taxonomy is actually a group of many binary classifications. For example, considering the wells of Cat. 1 as a true-positive prediction means that those wells from the data set are correctly predicted as Cat. 1. Any well from the Cat. 1 that was predicted as a different category means false negative prediction. All other categories from the dataset are considered false positives if they are predicted as Cat. 1 and true negative if they are predicted as other categories. This applies to all other categories, as presented in Fig. 9.

The most common evaluation metrics for ML classification models and their formulas are shown in the Supplements (Sokolova and Lapalme, 2009).

From the ML model perspective, it's required to increase accuracy and minimize errors in all metrics used for model evaluation. However, from the WI domain knowledge perspective (in which, the integrity of wells is classified, i.e. how safely can the field be operated), model evaluation stems from the number of non-integral wells that can be predicted correctly. The model error is not as important as if the non-integral well is incorrectly identified as integral. For this type of classification models, the confusion matrix assists to consider the implications from all perspectives. We calculated evaluation metrics by running a test split of 30% of the total dataset as previously mentioned. The results are displayed in the Supplements for all 11 models.

Reviewing the details of each classification model shows that metrics in many classifiers are excellent with high accuracy. This is detailed in Table 6. From Table 6 and the evaluation metrics shown in the Supplements, it is obvious that almost all metrics are excellent across the five categories, with minimal errors in CatBoost and XGB.

In a WI problem, prediction errors have serious implica-

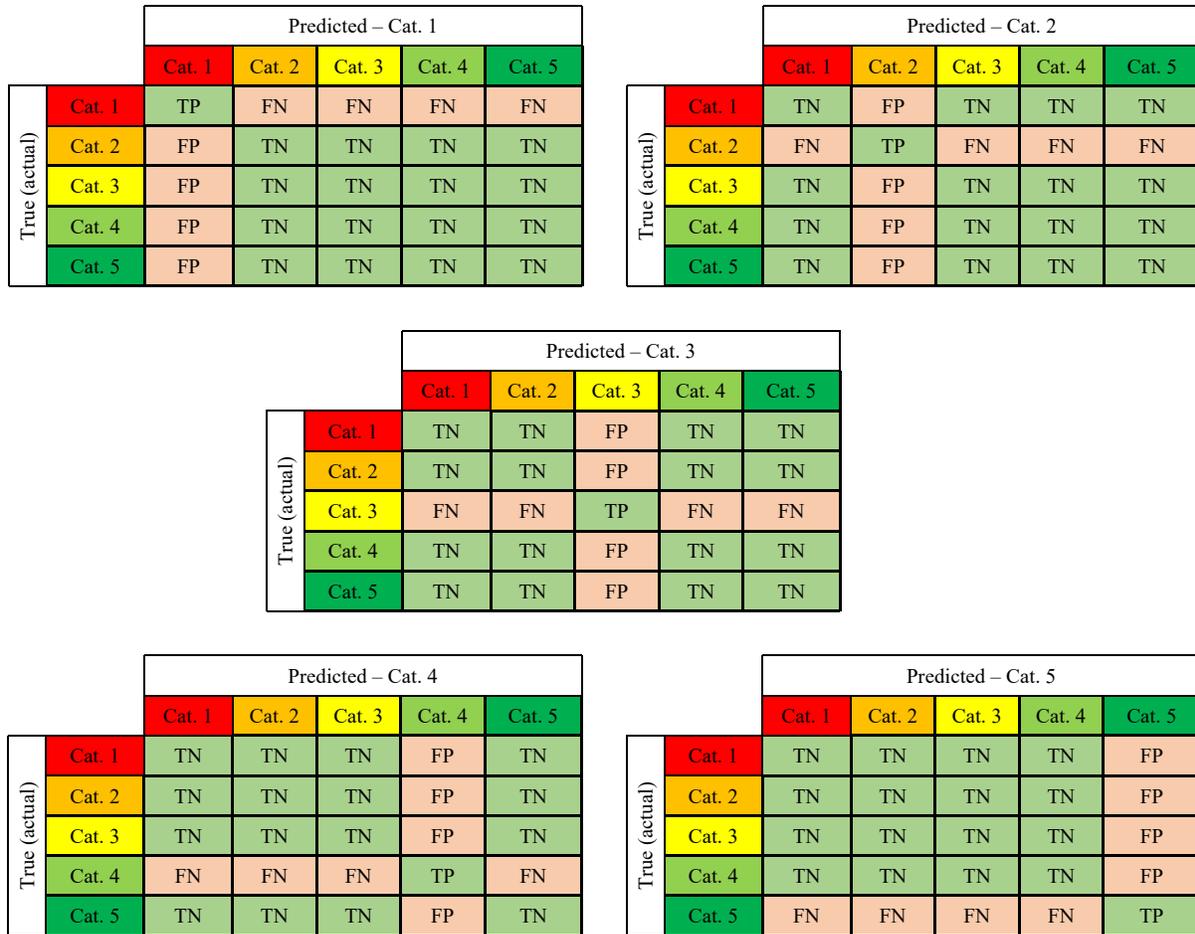


Fig. 9. Confusion matrix for five well categories.

Table 6. Best classifier for each evaluation metric.

Well Category	Precision	Recall	F1-Score	Accuracy
Cat. 1	SGD, CatBoost, XGB	XGB	XGB	Random forest, CatBoost, XGB
Cat. 2	XGB	CatBoost, XGB	XGB	Decision tree, XGB
Cat. 3	CatBoost	CatBoost, XGB	Decision tree, CatBoost, XGB	CatBoost, XGB
Cat. 4	Random forest, CatBoost, XGB	QDA	CatBoost, XGB	Decision tree, CatBoost, XGB
Cat. 5	Decision tree, CatBoost, XGB	NB, XGB, Decision tree, Random forest, CatBoost	Decision tree, Random forest, CatBoost, XGB	Decision tree, CatBoost, XGB
Total (macro-average)	XGB	XGB	XGB	CatBoost, XGB
Total (micro-average)	Decision tree, CatBoost, XGB	Decision tree, CatBoost, XGB	Decision tree, CatBoost, XGB	CatBoost, XGB

**Table 7.** Modified accuracy calculation per class.

Class	WI model accuracy	Modified accuracy
Cat. 1	XGB	1.00
Cat. 2	CatBoost XGB	0.96
Cat. 3	CatBoost XGB	0.99
Cat. 4	QDA AdaBoost	1.00
Cat. 5	Random forest CatBoost XGB	1.00

**Table 8.** Modified accuracy for all models.

Model	Average modified accuracy
CatBoost	0.991
XGB	0.991
Decision tree	0.985
Random forest	0.983
KNN	0.966
SVM	0.940
Logistic regression	0.924
SGD	0.808
AdaBoost	0.388
NB	0.364
QDA	0.348

tions. The regular metrics don't consider domain knowledge behind the classification problem. Taking the importance of model error into account, we have introduced a new metric evaluation to ensure that the ML model prediction isn't affecting operation safety and WI. Model accuracy was calculated by summing up all true positive predictions only and dividing by the total number of instances. New WI model accuracy calculations are tabulated in Table 7 for each class, and the average value for each model is listed in Table 8.

It can be inferred from the results presented in Tables 7 and 8 that CatBoost and XGB are the best classifiers for our dataset, showing the highest modified accuracy among all other models and also across all classes. CatBoost and XGB have been developed from the original gradient boost algorithm (Tang et al., 2021). The wells of Cat. 2 are relatively less accurate in terms of the model prediction. However, they still have excellent accuracy. Having 4 misclassifications out of 100 predictions is considered excellent, but in WI perspective, these 4 misclassified wells may be predicted as low-risk categories (Cat. 3, 4, or 5), which means non-integral well was diagnosed as an integral well. With that being mentioned,

		Predicted	
		Non-integral	Integral
True (actual)	Non-integral	TN	FP
	Integral	FN	TP

**Fig. 10.** Modified confusion matrix for WI classification.

		Predicted				
		Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5
True (actual)	Cat. 1	TN	TN	FP	FP	FP
	Cat. 2	TN	TN	FP	FP	FP
	Cat. 3	FN	FN	TP	TP	TP
	Cat. 4	FN	FN	TP	TP	TP
	Cat. 5	FN	FN	TP	TP	TP

**Fig. 11.** Rearrangement of the confusion matrix for WI classification.

the authors introduced an innovative approach to interpret the accuracy of the ML model.

The developed evaluation metric depends on converting the confusion matrix into a new one on the basis of WI domain knowledge, as shown in Fig. 10. In this matrix, we refer to Cat. 3, Cat. 4, and Cat. 5 as integral wells (true positive values) while the wells of Cat. 1 and Cat. 2 are non-integral (true negative values). It can be viewed as having meta-classes of five categories out of two major taxonomies, in which wells are either integral or non-integral.

Predicting the well of Cat. 2 as Cat. 1 is less risky than predicting the same well as Cat. 3, 4 or 5. In the same manner, prediction the well of Cat. 3 as Cat. 4 or 5 means more model efficiency than predicting this well as Cat. 1 or 2. Accordingly, the confusion matrix for multi-class problems was rearranged in a specific order to assert optimum model performance without having faulty predictions. A new confusion matrix for the ML model is presented in Fig. 11.

Using this matrix will lead to the correct prediction of the WI status of any well. The integrity category can be tuned using high-performance models, depending on regular ML evaluation metrics. For that reason, well integrity index-1 ( $WII - 1$ ), well integrity index-2 ( $WII - 2$ ), and well integrity

**Table 9.** Comparison of model accuracy.

	Average model accuracy	Modified model accuracy	<i>WII</i> – 1	<i>WII</i> – 2
Summary	“Uses confusion matrix in Fig. 9”	“Uses regular multi-class confusion matrix”	“Uses modified confusion matrix”	“Similar to <i>WII</i> -1 but with less risk prediction”
Decision tree	0.99	0.985	1.00	1.00
CatBoost tree	1.00	0.991	1.00	1.00
XGB	1.00	0.991	1.00	1.00
NB	0.75	0.364	0.83	1.00
SVM	0.98	0.940	1.00	1.00
Model Random forest	0.99	0.983	1.00	1.00
KNN	0.99	0.966	1.00	1.00
AdaBoost	0.76	0.388	0.86	0.99
Logistic regression	0.97	0.924	0.99	0.99
SGD	0.92	0.808	0.98	0.98
QDA	0.74	0.348	0.97	0.97

false positives (*WII* – 3) are introduced as evaluation metrics in this work. All calculations are completed based on Eqs. (1), (2), and (3). Models are sorted in order of the least classification error. *WII* – 1 represents the summation of accurate predictions of different well categories. *WII* – 2 represents the summation of safe predictions, including correct predictions, plus the prediction of integral wells as high-risk wells. This misclassification causes lost time and wasted team effort to confirm the integrity of predicted wells, when in reality, there is no hidden risk at all. *WII* – 3 represents the riskiest scenario, in which non-integral wells are predicted as integral ones. It’s required to keep *WII* – 3 as low as possible.

$$WII - 1 = \sum_{i=1}^k \left( \frac{TP_i + TN_i}{TP_i + TN_i + FN_i + FP_i} \right) \quad (1)$$

$$WII - 2 = \sum_{i=1}^k \left( \frac{TP_i + TN_i + FN_i}{TP_i + TN_i + FN_i + FP_i} \right) \quad (2)$$

$$WII - 3 = \sum_{i=1}^k \left( \frac{FP_i}{TP_i + TN_i + FN_i + FP_i} \right) \quad (3)$$

Results are illustrated in Table 9, demonstrating the comparison between different evaluation metrics considered to determine model performance and classification efficiency. Average model accuracy is macro-average, which gives an equal weight for each well category. The modified model accuracy calculated only the summation of true positive predictions across all classes to project WI knowledge on the ML model. *WII* – 1 and *WII* – 2 are two indices introduced by the authors to tune the models and maximize the benefit from the ML model in such a critical classification problem. It can be observed that some models show the same average accuracy across all metrics (e.g., decision tree, CatBoost, XGB, and random forest). Some models were found to be a robust classifier when new WI indices were introduced, such as NB, AdaBoost, and QDA. Classifiers such as SVM and KNN,

logistic regression, and SGD showed a slight improvement with new indices. In conclusion, CatBoost and XGB, followed by decision tree and random forest, are the best classifiers from all perspectives, either proper prediction of integral wells or accurate well category allocation.

Random forest was used to identify the relative importance (feature importance) of independent variables (Fig. 12). It can be seen from the results that the failure of WI in high-rate wells (either oil or gas production or both) can cause disastrous effects and prolonged impact, especially in naturally flowing wells. Automatic closure valves (UMV, DHSV, and ESD) were flagged as having high value owing to their safety-critical importance in securing the well remotely in case of a hydrocarbon leak into the surroundings. On the other side, structure integrity, risk of ship collision, and conductor movement were the least important features. This resulted from fewer non-integral wells associated with these features in the research dataset. Therefore, the model assumed their low importance. The results of feature importance have proven to match the actual reliability data of the field under study. In other fields, the importance of features may vary respectively. For example, some fields may have all their towers manned, or most of them in a shipping lane exposed to frequent risk of collision or located in a region with severe weather condition. In this situation, feature importance will change to reflect the impact of each failure event.

This extraction of feature importance from the random forest model assisted the authors to develop handy equations for calculating failure of WI (product of probability and severity) as shown in Eqs. (4) and (5);

$$P = 0.45 \cdot ESD + 0.24 \cdot ANN + 0.28 \cdot XT + 0.03 \cdot ST \quad (4)$$

where:

*P* = probability of failure, and it depends on specific variables listed and numbered in Table 3.

*ESD* = emergency shut down valve (automatic closure

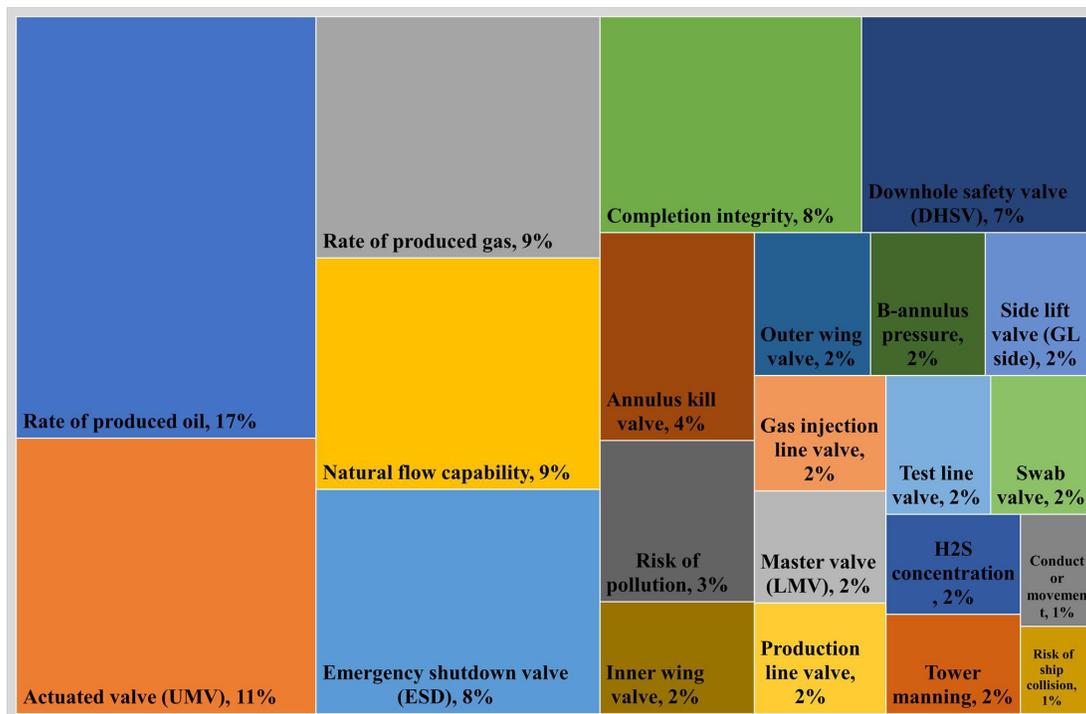


Fig. 12. Feature importance of WI elements.

valves), referring to variables 2, 8, and 12.

$ANN$  = True Negative (TN): Labels are truly negative and predicted as negative.

$XT$  = X-tree/wellhead valves, referring to variables 1, 3, 4, 5, 7, 9, 10, and 11.

$ST$  = well/tower structure factors, referring to variables 15, 16, and 17.

$$S = 0.21 \cdot ENF + 0.62 \cdot Q + 0.17 \cdot HSE \quad (5)$$

where:

$S$  = severity of failure, and it depends on specific variables listed and numbered in Table 3.

$NF$  = well natural flowing capability, referring to variable 18.

$Q$  = well production rate, referring to variables 22 and 23.

$HSE$  = actors related to HSE, referring to variables 19, 20, and 21.

The ML model has demonstrated success in classifying well-risk categories for similar wells, or an entire well stock. The assessment results can classify wells into 5 categories (Cat. 1, 2, 3, 4, and 5). Fig. 13 summarizes the major contributions of this paper, in which the legacy matrix used for classifying WI failures was transformed into a more quantitative tool. After that, a risk map was developed to calculate the percentage of failure through an innovative monograph. Finally, an automated ML model was developed to predict WI with precise results. In addition, this research provides three functional equations that can be used as precise evaluation metrics for similar ML models related to WI. Physical equations were developed from the ML model, and it can be used by applying well data with no AI or ML knowledge or software expertise.

## 4. Conclusions

In this work, ML algorithms were used to develop a resilient and reliable model to forecast WI status and well risk category in GL wells. The ML algorithms used in this research include logistic regression, SVM, KNN, NB, decision trees, random forest, XGB, AdaBoost, CatBoost, QDA, and SGD. All developed models were compared with conventional risk ranking methods. In addition, a combined model was developed that uses a voting process among all models and automatically selects the highest accurate result. The developed model can forecast WI status with precise accuracy that is highly functional. Three new metric equations were developed to suit our project. Moreover, physical equations were developed to predict the risk ranking of any well. Based on discussion and data analysis, the following conclusions can be drawn:

- 1) The developed model can effectively challenge problems in O&G wells, alert WI events promptly, decrease the risk of WI deterioration, reduce failures, and prolong the lifetime of O&G wells.
- 2) The categorization process generates the WI status, relying on the status of barrier components and their conditions. The well can be classified into one of five categories on the basis of criticality. Cat. 4 and Cat. 5 are indicating very good to excellent WI status, Cat. 3 means WI status in the watch list, which is the main feeder of high-risk wells. Meanwhile, Cat. 1 and Cat. 2 are considered problematic wells with many integrity issues, and they have priority to perform barrier restoration and repair jobs.

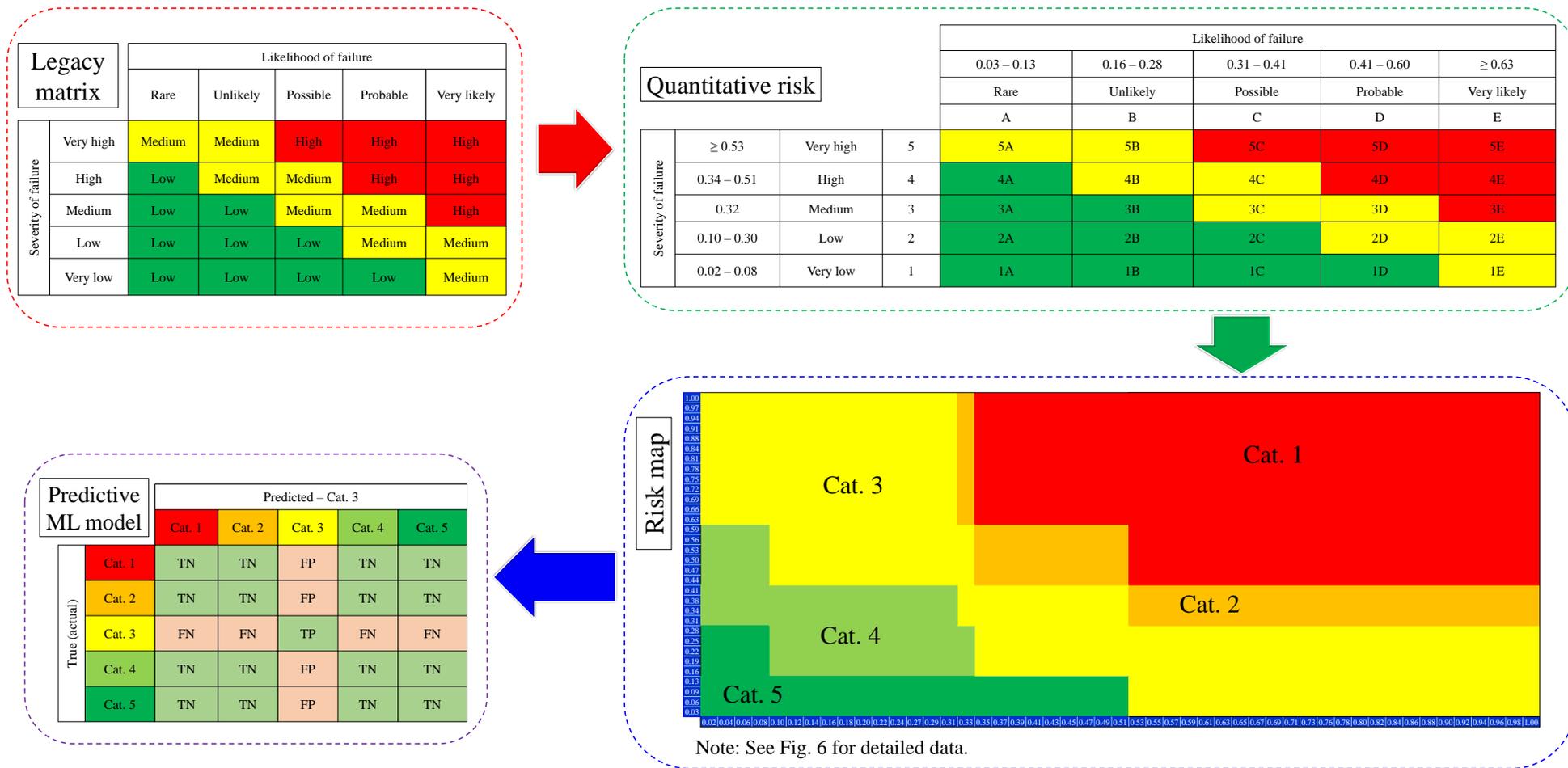


Fig. 13. Summary of research contributions.

- 3) Risk mapping has proven its ability to identify WI's failure and the exact percentage of impairment.
- 4) The developed metrics assisted in properly evaluating ML models by applying the WI knowledge domain.
- 5) Among all the computational models developed, the ones produced using XGB, CatBoost, and decision tree were the most optimum. They have the best performance, as confirmed by traditional metrics and also the newly-developed metrics. They can accurately identify integral wells and also predict well category precisely. SGD and QDA showed the lowest performance. This may be due to having all features as categorical variables, and the random forest algorithm was able to better predict the risk level of WI better than any other algorithm.
- 6) Two unique equations were developed for both probability and severity of failure, considering WI elements contributing to both risk factors and the weight of contribution of each element as per the random forest algorithm.

## Acronyms

3D = 3-dimensional  
 AdaBoost = Adaptive boosting  
 AL = Artificial lift  
 BOPD = Barrel oil per day  
 Cat. = Category  
 CatBoost = Categorical boosting  
 DHSV = Downhole safety valve  
 ESD = Emergency shut down  
 GL = Gas lift  
 GLR = Gas-liquid ratio  
 H<sub>2</sub>S = Hydrogen sulfide  
 HSE = Health, safety, and environment  
 KNN = K-nearest neighbor  
 LMV = Lower master valve  
 ML = Machine learning  
 MMSCFD = Million standard cubic feet per day  
 NA = Not available  
 NB = Naive Bayes  
 NF = Natural flow  
 NT = Not tested  
 O&G = Oil and gas  
 ppm = Part per million  
 QDA = Quadratic discriminant analysis  
 RA = Risk assessment  
 SCFH = Standard cubic feet per hour  
 SGD = Stochastic gradient descent  
 ST. CL = Stuck closed  
 ST. OP = Stuck open  
 SSSV = Subsurface safety valve  
 SVM = Support vector machine  
 UMV = Upper master valve  
 WI = Well integrity  
 WIMS = Well integrity management system  
 XGB = Extreme gradient boosting  
 Gas-liquid ratio  
 X-tree = Christmas tree

## Acknowledgement

The authors would like to express appreciation to the Gulf of Suez Petroleum Company (GUPCO) and the faculty members in the department of petroleum engineering, Suez University, Egypt, for their continuous encouragement and support. The authors would also like to acknowledge Prof. Ahmed A. Elgibaly who passed away on January 15, 2021. May Allah forgive his sins, have mercy on him, and admit him Jannatul Firdauss. Society of Petroleum Engineers (SPE) is acknowledged for authorization to republish the SPE-206129-MS paper in *Advances in Geo-Energy Research* (order license ID 1183775-1, 2022).

## Supplementary file

<https://doi.org/10.46690/ager.2022.02.05>

## Conflict of interest

The authors declare no competing interest.

**Open Access** This article is distributed under the terms and conditions of the Creative Commons Attribution (CC BY-NC-ND) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## References

- Abimbola, M., Khan, F., Khakzad, N. Risk-based safety analysis of well integrity operations. *Safety Science*, 2016, 84: 149-160.
- Adeyinka, A., Tsakporhore, A., Arije, O., et al. Rethinking well integrity for sustainability: Adopting a risk-based approach. Paper SPE 203672 Presented at the SPE Nigeria Annual International Conference and Exhibition, Virtual, 11-13 August, 2020.
- Anders, J. L., Rossberg, R. S., Dube, A. T., et al. Well integrity operations at Prudhoe Bay, Alaska. *SPE Production & Operations*, 2008, 23(2): 280-286.
- AOGCC. Investigation of Explosion and Fire at Prudhoe Bay Well A-22 North Slope, Alaska. Alaska, USA, Alaska Oil & Gas Conservation Commission Staff Report, 2003.
- Bates, R., Cosad, C., Fielder, L., et al. Taking the pulse of producing wells-ESP surveillance. *Oilfield Review*, 2004, 16(2): 16-25.
- Bontempi, G. *Statistical foundations of machine learning*. Brussels, Universite Libre de Bruxelles, 2021.
- Bravo, C., Rodriguez, J., Saputelli, L., et al. Applying analytics to production workflows: Transforming integrated operations into intelligent operations. Paper SPE 167823 Presented at the SPE Intelligent Energy Conference and Exhibition, Utrecht, Netherlands, 1-3 April, 2014a.
- Bravo, C., Saputelli, L., Rivas, F., et al. State of the art of artificial intelligence and predictive analytics in the E&P industry: A technology survey. *SPE Journal*, 2014b, 19(4): 547-563.
- Brechan, B., Dale, S. I., Sangesland, S. Well integrity risk assessment-software model for the future. Paper OTC 28481 Presented at the Offshore Technology Conference Asia, Kuala Lumpur, Malaysia, 20-23 March, 2018.
- Brodie, A. Gas-lift valve design addresses long-term well

- integrity needs. *Offshore*, 2011, 71(2): 98-99.
- Choubey, S., Karmakar, G. P. Artificial intelligence techniques and their application in oil and gas industry. *Artificial Intelligence Review*, 2021, 54(5): 3665-3683.
- Cox, L. A. J. What's wrong with risk matrices? *Risk Analysis*, 2008, 28(2): 497-512.
- De Carvalho, A. C. P., Freitas, A. A. A tutorial on multi-label classification techniques, in *Foundations of Computational Intelligence*, edited by A. Abraham., A. E. Hassanien, and V. Snášel, Berlin Heidelberg, Germany, pp. 177-195, 2009.
- Dethlefs, J., Chastain, B. Assessing well-integrity risk: A qualitative model. *SPE Drilling & Completion*, 2021, 27(2): 294-302.
- Elgibaly, A. A., Ghareeb, M., Kamel, S., et al. Prediction of gas-lift performance using neural network analysis. *AIMS Energy*, 2021, 9(2): 355-378.
- Elmousalami, H. H., Elaskary, M. Drilling stuck pipe classification and mitigation in the Gulf of Suez oil fields using artificial intelligence. *Journal of Petroleum Exploration and Production Technology*, 2020, 10(5): 2055-2068.
- Fernández, A., García, S., Galar, M., et al. Introduction to KDD and Data science, in *Learning From Imbalanced Data Sets*, edited by A. Fernández, S. García, M. Galar, et al., Springer, Berlin, pp. 1-16, 2018.
- Gupta, S., Saputelli, L., Nikolaou, M. Applying big data analytics to detect, diagnose, and prevent impending failures in electric submersible pumps. Paper SPE 181510 Presented at the SPE Annual Technical Conference and Exhibition, Dubai, UAE, 26-28 September, 2016.
- Hastie, T., Tibshirani, R., Friedman, J. Model assessment and selection, in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, edited by T. Hastie, R. Tibshirani and J. Friedman, Springer, New York, pp. 219-259, 2017.
- Holdaway, K. R. Fundamentals of soft computing, in *Harness Oil and Gas Big Data with Analytics: Optimize Exploration and Production with Data-Driven Models*, edited by K. R. Holdaway, Wiley and SAS Business Series, North Carolina, pp. 1-31, 2014.
- Ismail, W. R., Trjanganung, K. Mature field gas lift optimisation: Challenges & strategies, case study of D-field, Malaysia. Paper IPTC 17896 Presented at the International Petroleum Technology Conference, Kuala Lumpur, Malaysia, 10-12 December, 2014.
- Kazak, A., Simono, K., Kulikov, V. Machine-learning-assisted segmentation of focused ion beam-scanning electron microscopy images with artifacts for improved void-space characterization of tight reservoir rocks. *SPE Journal*, 2021, 26(4): 1739-1758.
- Khan, M. R., Tariq, Z., Abdulraheem, A. Application of artificial intelligence to estimate oil flow rate in gas-lift wells. *Natural Resources Research*, 2020, 29(1): 4017-4029.
- Kiran, R., Teodoriu, C., Dadmohammadi, Y., et al. Identification and evaluation of well integrity and causes of failure of well integrity barriers (A review). *Journal of Natural Gas Science and Engineering*, 2017, 45: 511-526.
- Li, B., Billiter, T. C., Tokar, T. Rescaling method for improved machine-learning decline curve analysis for unconventional reservoirs. *SPE Journal*, 2021, 26(4): 1759-1772.
- Loizzo, M., Bois, A., Etcheverry, P., et al. An evidence-based approach to well-integrity risk management. *SPE Economics & Management*, 2015, 7(3): 100-111.
- Mahdiani, M. R., Khamsehchi, E., Hajirezaie, S., et al. Modeling viscosity of crude oil using k-nearest neighbor algorithm. *Advances in Geo-Energy Research*, 2020, 4(4): 435-447.
- Miraglia, S. A data-driven probabilistic model for well integrity management: Case study and model calibration for the Danish sector of North Sea. *Journal of Structural Integrity and Maintenance*, 2020, 5(2): 142-153.
- Mishra, S., Datta-Gupta, A. *Applied Statistical Modeling and Data Analytics: A Practical Guide for the Petroleum Geosciences*. Amsterdam, Netherlands, Elsevier, 2017.
- Mohammed, R., Rawashdeh, J., Abdullah, M. Machine learning with oversampling and undersampling techniques: Overview study and experimental results. Paper ICICS Presented at the 11<sup>th</sup> International Conference on Information and Communication Systems, Irbid, Jordan, 7-9 April, 2020.
- Olukoga, T. A., Feng, Y. Practical machine-learning applications in well-drilling operations. *SPE Drilling & Completion*, 2021, 36(4): 849-867.
- Ragab, A. M. S., Yakoot, M. S., Mahmoud, O. Application of machine learning algorithms for managing well integrity in gas lift wells. Paper SPE 205736 Presented at the SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition, Virtual, 12-14 October, 2021.
- Rahmanifard, H., Plaksina, T. Application of artificial intelligence techniques in the petroleum industry: A review. *Artificial Intelligence Review*, 2019, 52(4): 2295-2318.
- Rahmawati, S. D., Chandra, S., Aziz, P. A., et al. Integrated application of flow pattern map for long-term gas lift optimization: A case study of Well T in Indonesia. *Journal of Petroleum Exploration and Production Technology*, 2020, 10(4): 1635-1641.
- Raschka, S. Building good training sets-data preprocessing, in *Python Machine Learning: Unlock Deeper Insights into Machine Learning with This Vital Guide to Cutting-Edge Predictive Analytics*, edited by S. Raschka, Safari Books Online, Packt Publishing, Birmingham, United Kingdom, pp. 99-127, 2015.
- Salem, A. M., Yakoot, M. S., Mahmoud, O. Addressing diverse petroleum industry problems using machine learning techniques: Literary methodology-spotlight on predicting well integrity failures. *ACS Omega*, 2022, 7: 2504-2519.
- Saputelli, L. Technology focus: Petroleum data analytics. *Journal of Petroleum Technology*, 2016, 68(10): 66-66.
- Shalaby, M. R., Malik, O. A., Lai, D., et al. Thermal maturity and TOC prediction using machine learning techniques: Case study from the Cretaceous-Paleocene source rock, Taranaki Basin, New Zealand. *Journal of Petroleum Exploration and Production Technology*, 2020, 10(6): 2175-2193.

- Silla, C. N., Freitas, A. A. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 2011, 22(1): 31-72.
- Sokolova, M., Lapalme, G. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 2009, 45(4): 427-437.
- Tang, J., Fan, B., Xiao, L., et al. A new ensemble machine-learning framework for searching sweet spots in shale reservoirs. *SPE Journal*, 2021, 26(1): 482-497.
- Vidyalala, R. [How to select the right machine learning algorithm. Towards Data Science, 2020.](#)
- Wood, D. A. A transparent Open-Box learning network provides insight to complex systems and a performance benchmark for more-opaque machine learning algorithms. *Advances in Geo-Energy Research*, 2018, 2(2): 148-162.
- Wood, D. A. Gamma-ray log derivative and volatility attributes assist facies characterization in clastic sedimentary sequences for formulaic and machine learning analysis. *Advances in Geo-Energy Research*, 2022, 6(1): 69-85.
- Yakoot, M. S., Elgibaly, A. A., Ragab, A. M. S., et al. A comprehensive review and analysis of maturity model for well integrity in brownfield. Paper Presented at the IADC Drilling Middle East Conference & Exhibition, Virtual, 14-15 December, 2020.
- Yakoot, M. S., Elgibaly, A. A., Ragab, A. M. S., et al. Well integrity management in mature fields: A state-of-the-art review on the system structure and maturity. *Journal of Petroleum Exploration and Production Technology*, 2021a, 11(4): 1833-1853.
- Yakoot, M. S., Ragab, A. M. S., Mahmoud, O. Machine learning application for gas lift performance and well integrity. Paper SPE 205134 Presented at the SPE Europec Featured at 82<sup>nd</sup> EAGE Conference and Exhibition, Amsterdam, The Netherlands, 18-21 October, 2021b.
- Yakoot, M. S., Ragab, A. M. S., Mahmoud, O. Multi-class taxonomy of well integrity anomalies applying inductive learning algorithms: Analytical approach for artificial-lift wells. Paper SPE 206129 Presented at the SPE Annual Technical Conference and Exhibition, Dubai, United Arab Emirates, 21-23 September, 2021c.
- Yakoot, M. S., Ragab, A. M. S., Mahmoud, O. Developing analytical model for calculating the wellbore-integrity comprehensive risk in artificial-lift system. Paper OMAE 2022-80222 to be Presented at the ASME 41<sup>st</sup> International Conference on Ocean, Offshore and Arctic Engineering, Hamburg, Germany, 6-9 June, 2022.
- Yavari, H., Khosravanian, R., Wood, D. A., et al. Application of mathematical and machine learning models to predict differential pressure of autonomous downhole inflow control devices. *Advances in Geo-Energy Research*, 2021, 5(4): 386-406.
- Yin, Q., Yang, J., Tyagi, M., et al. Machine learning for deepwater drilling: Gas-kick-alarm Classification using pilot-scale rig data with combined surface-riser-downhole monitoring. *SPE Journal*, 2021, 26(4): 1773-1799.
- Zhao, L., Yan, Y., Wang, P., et al. A risk analysis model for underground gas storage well integrity failure. *Journal of Loss Prevention in the Process Industries*, 2019, 62: 103951.

## **Appendix A. Basis of assigning the weight score in Table 3**

X-tree, subsurface safety valve (SSSV), and surface valves provide barriers to isolate the reservoir energy source. As more valves leak, the higher the probability of a loss of WI. Automatic closure of the UMV, SSSV, and the gas injection line valve is linked to the tower ESD. These valves are flagged with high value to reflect their safety critical importance. Other factors increase the probability of failure (e.g., risk of ship collision, tower structure integrity, and well conductor integrity). The consequence of a loss of WI is strongly related to whether it can produce naturally or not, the flow rate, and risks of environmental pollution. The following thresholds have been considered intolerable risks:

- Wells with natural flow capability and no vertical barriers, or the downhole completion has tubing/annulus communication,
- Wells with no barriers at production or gas injection sides,
- Wells with very high buildup rate in bled B-annulus, and
- Surfaces leak of well fluid or high-pressure gas.

The probability of failures for well barrier elements was weighted as well as the severity of events on the basis of field reliability data. Accordingly, failure scenarios have been drawn against different levels of probability and severity to assign a weight score for each barrier element. The defined score depends on the impact that barrier could have on the loss of fluid containment. A total WI status score is calculated from the number of anomalies.