Addressing Diverse Petroleum Industry Problems Using Machine Learning Techniques: Literary Methodology—Spotlight on Predicting Well Integrity Failures

Adel M. Salem,* Mostafa S. Yakoot,* and Omar Mahmoud*

ABSTRACT: Artificial intelligence (AI) and machine learning (ML) are transforming industries, where low-cost, big data can utilize computing power to optimize system performance. Oil and gas (O&G) fields are getting mature, where well integrity (WI) problems become more common and field operations are now more challenging. Hence, they are good candidates for transformation due to the low cost of data storage, highlighting the oil market decline, along with dynamic risk posed during operations. This paper is presenting a comprehensive compilation of different ML applications in diverse disciplines of the petroleum industry. The pool of AI and ML with respect to different areas of applications along with publication years has been categorized. The main focus of this study is classifying well integrity failures where the authors found that the potential of AI and ML in predicting well integrity failures has not been efficiently tapped, and there is an explicit gap in the literature. First, the applications of AI, ML, and data analytics in the O&G industry are discussed thoroughly, so this paper can be a comprehensive reference for readers and future researchers. Then data preprocessing is explained. This includes data gathering, cleaning, and feature engineering. Next, the different ML models are compared and discussed. Finally, model performance evaluation and best model selection are described. This study would be a concrete foundation in the design and construction of ML programs that can be deployed for WI risk management. The developed model can be simply used for any well stock, providing quick and easy assessment instead of subjective and tedious assessment. The layout can be simply adjusted to reflect the risk profile of any well type or any field.

INTRODUCTION

The oil and gas (O&G) sector is facing the challenge of significantly decreasing operating costs without compromising safety. Opportunely, artificial intelligence (AI) has been engaged widely to resolve challenges in the O&G industry. Technological improvements and the big data revolution can help provide the information needed for decisions to be made and reduce the time from identification to execution.

Diving into the literature, we compiled most of the case studies that implemented AI with success in multiple fields of petroleum engineering. However, going through the previous works, it has been detected that the application of AI in predicting well integrity (WI) failures has not been fully tackled. Not to mention, WI directly impacts health, safety, and the environment (HSE), loss of assets, governmental regulations, company reputation, society image, financial losses, employment, and production. Consequently, this work focuses on using multiple machine learning (ML) models to assess the risk of WI impairment and predict failures of barriers. The objectives of this paper are to:

- summarize and compile the applications of AI, ML, and data analytics in different disciplines of the O&G industry,
- elaborate the difference between ML and traditional programming,
- shed light on key elements and types of ML algorithms,
- explain steps taken to preprocess data before being fed into the model, which includes data gathering, cleaning, and feature engineering,
- compare and discuss the different ML models (advantages and disadvantages of each model are listed), and
- describe model efficiency evaluation, make comparison, and endorse selection of the best model based on confusion matrix technique.

This research paper presents a comprehensive review for AI and ML applications in the O&G industry over the past few
years. In addition, it gives a handy and valuable guide for both companies and researchers for the risk management of WI during operation of O&G wells. Figure 1 illustrates the sequence of steps taken to achieve the aforementioned objectives of the work.

- **PRINCIPLE OF WELL INTEGRITY MANAGEMENT**

WI issues may arise in any well, producing or not, particularly old wells that have been constructed as per the regulations and standards of their construction date. The standards of well construction have improved over time in terms of safety specifications. In addition, several wells that conform to current safety requirements may still have problems related to integrity. Although the necessity to enhance the integrity of wells is recognized, enhancement activities have been limited by past budget constraints. Therefore, Yakoot et al. concluded the significance of effectively managing the wide range of WI problems by adopting a concept of well integrity management system (WIMS). Maturity of WIMS can always be evaluated using different models and approaches. 

As WI is directly related to safety of well operations, it has become an increased focus by operators and is being published intensively in industry research. With that being said, WI should be modernized and stepped up from regular spreadsheets and non-user-friendly and full-of-bugs software to benefit from computational empowerment and self-educating machines to attain ultimate maturity in management and data maturity systems.

- **DIGITAL TRANSFORMATION IN THE O&G INDUSTRY**

As the world moves toward the Internet of things (IOT), O&G operators need to adapt to support the transformation. Big data is deemed the next phase of this transformation that enables industry leaders to make better decisions. The O&G industry is now rich with big data coming from real-time production parameters that can be measured every hour. According to Hajirahimova, “big data can be recognized by volumes, velocity, variety, value, and veracity.” Big data is said to be equivalent to return on investment. With technology advancements and a growing network of sensors, faster and higher frequency data gathering is enabled, and as a result, we have higher volumes of data.

AI techniques have been recognized recently as a prospective tool to overcome the uncertainties faced in most exploration and production (E&P) activities.Crockett and Baudoin listed, as shown in Table 1, the existing challenges in the industry and the way forward to benefit from IOT applications. Crockett explained how the IOT could map out the route to decision-making maturity, paving the way for automated handling of repetitive problems (i.e., faster decisions). The author concluded that adopting IOT results in monitoring by prediction, better visualization, big data connecting various databases, data and information can be sent to the operators using automated workflows, and finally, the ordinary way of working will transform it into expressive information for decision makers.

Analytics is an integrated process of collecting, processing, and analyzing big volumes of data to extract maximum value from the collected data. This can be achieved by identifying patterns that allow for automated detection of potential failures and help lessen the uncertainty in asset management. 

Data nowadays plays an extremely valuable role in E&P companies. In recent years, the quantity of available data has risen from kilobytes to terabytes, thereby being “big data” due to its incredibly broad scale, range, volume, and authenticity. The O&G industry are now utilizing ML methods to resolve and tackle problems actively.

Databanks of many O&G companies hold very encouraging potential for more applied decision-making practices. Moreover, there is a nonlinear growth in the flood of produced data from increasing data systems. Despite that, obtaining meaningful information from big-sized data remains challenging.

Implementation of certain algorithm(s) to explore hidden trends in data is known as data mining (DM). ML algorithms have been used for drill bit diagnosis, well-log analysis, surveillance of major drilling and completion activities, and reservoir simulation, encompassing seismic pattern identification, history matching, and reservoir characterization, prediction of permeability and porosity, pressure—volume—temperature (PVT) analysis, production optimization, and well performance evaluation. Figure 2 shows a better summarization of ML applications in the O&G industry.

ML algorithms have proved their capability to handle big data that can help to make quick and proper decisions. These algorithms are used for recognizing patterns and attaining valuable understanding based on the collected data. ML models have been effectively used in petroleum and reservoir

**Table 1. Problems and Challenges Facing Oil and Gas Operators**

| external pressures | challenges | response of O&G upstream operators |
|--------------------|------------|----------------------------------|
| reduce costs       | cost       | improve automation of data       |
| maximize production| increase   | leverage technology to give more insights |
| performance        | performance | leverage predictive analytics to reduce deferment and improve recovery efficiency |

Figure 1. General sequence of the research work.
engineering, such as history matching, production forecasting, and reservoir characterization.\footnote{4e,f}

\section*{OVERVIEW OF MACHINE LEARNING}
ML is a subset of AI around the idea that we should feed the machines with data and let them learn for themselves. If programming is automation, then ML is automating the process of automation. In traditional programming, data and programs are both run on the computing machine to compute the output, but in ML, both data and output are run on the computer together to develop a program. This program can be used in traditional programming.\footnote{5a} This concept is visualized in Figure 3.

\subsection*{Key Elements of Machine Learning}
Every ML algorithm has three stages during process of building the automated model:

- **Representation**: the algorithm used to represent the data pattern knowledge. Examples include regressors, classifiers, support vector machines (SVM), model ensembles and others.
- **Evaluation**: the way to evaluate the chosen algorithms’ performance. Examples include accuracy, precision, recall, mean squared error, cost function, etc.
- **Optimization**: the process of adjusting the model hyperparameters to improve performance.

\subsection*{Types of Learning Algorithms}
There are different styles in ML problems.\footnote{5b}

- **Supervised Learning**: These are algorithms in which, when given a sample of data and its desired outputs, the algorithm approximates a function that maps inputs to outputs. A model is developed through a training phase where prediction is made and then corrected when they are proved wrong. The training phase remains until the model reaches a desired accuracy level on the training data.

- **Unsupervised Learning**: These are algorithms which, when given a sample of data, do not contain the desired output. The algorithm learns the inherent pattern of the data without using explicitly provided labels. A model is prepared by inferring structures hidden in the input data.

- **Semi-supervised Learning**: The input data are a mixture of labeled and unlabeled examples that aim to label unlabeled data points using knowledge learned from the labeled data points. The model must learn the structures to organize the data as well as make predictions.

- **Reinforcement Learning**: It works in the following way: there is a presence of an agent and environment. The agent would be able to take some actions on the environment, based on which it would be rewarded or punished.

\section*{MACHINE LEARNING APPLICATIONS}

\subsection*{Petrophysics}
ML models have been successfully used in seismic data interpretation. Smart systems have been developed to handle complex and large-scale seismic data and processed to small-scale core data which have proven better results than those commonly used in industry practices.\footnote{6a–f} In addition, Vallabhaneni et al.\footnote{7a} introduced a ML-based model that integrates 3D seismic data with petrophysical properties. This ML model proved to be less costly, more precise, and faster than the orthodox geostatistical approach. The proposed approach helped to better understand reservoir uncertainties.

In the field of logging interpretation, Al-Mudhafar\footnote{7b} applied probabilistic neural networks to model and predict lithofacies
distribution as a function of well logging interpretation data; then he used the generalized boosted regression model to build a nonlinear relationship between lithofacies, core permeability, and well logging data. Belozerov et al.\textsuperscript{7c} used ML to automatically interpret well logs. Dunham et al.\textsuperscript{7d} enhanced classification of well logs using semi-supervised learning algorithms. Shalaby et al.\textsuperscript{7e} provided robust evidence that ML models can identify the nonlinear correlation between the well log data and total organic carbon. Reinforced learning was applied by Bittar et al.\textsuperscript{7f} on depth matching of well logs, and they succeeded to achieve better definition of reservoir geometry.

Studying carbonate rocks, Tariq et al.\textsuperscript{8a} applied a functional network (FN) technique to estimate Poisson’s ratio. The developed model was further translated into a simple empirical equation. Later, Tariq et al.\textsuperscript{8b} applied FN optimized with particle swarm optimization (PSO) to predict water saturation in carbonate rocks. A simple equation was also presented and validated using three common water saturation models.

**Drilling Operations.** Considering the criticality of drilling operation, Shadravan et al.\textsuperscript{8c} used ML algorithms to design drilling fluids. They investigated and implemented a predictive data-driven technique, which deployed Gaussian Process Regression as a novel ML tool. The developed tool predicts composition of drilling fluids and suggests composition of spacer fluid or cement slurry by applying ML algorithms on the compiled experimental data. This model can predict rheological properties and allow engineers to keep the fluid rheological hierarchy to achieve better cement jobs and ensure WI. This model considerably reduces the testing expenses, optimizes the material consumption, and does not require involving complicated physics. Various applications of new intelligent tool are shown in Figure 4. Elkatatny et al.\textsuperscript{8d} predicted the rheological properties of drilling fluids using mathematical model from the artificial neural networks (ANNs). Bowie\textsuperscript{8e} optimized well construction design and completion processes for 262 wells. Further, Abdelgawad et al.\textsuperscript{8f} developed a new approach for predicting equivalent circulation density of drilling fluids applying AI models to surface drilling parameters such as drill pipe pressure, mud weight, and rate of penetration. They used 2376 data points that were collected during the drilling phase. The two used models were adaptive neuro-fuzzy inference system and ANN. Gasser et al.\textsuperscript{9a} presented an ANN model to predict the rheological properties of nanobased drilling fluids. Moreover, Gasser et al.\textsuperscript{9b} proposed new ANN model to forecast filtrate invasion of nanobased drilling fluids. For well drilling operations, Elmousalami and Elaskary\textsuperscript{9c} classified pipe stuck automatically using ML algorithms and were able to mitigate the different cases of pipe sticking. Deep reinforcement learning was applied by Yu et al.\textsuperscript{9d} to develop automated directional drilling model.

**Well Stimulation.** ML models were applied on a large scale to better identify and comprehend hydraulic fracturing. Alimkhanov and Samoylova\textsuperscript{10a} deployed 11 models on reservoirs with complex geology to identify criteria for hydraulic fracturing candidates. Further, Anderson et al.\textsuperscript{10b} classified hydraulic fractures by matching physics-based models with ML practices. Additionally, Nande\textsuperscript{10c} applied ANN to develop an innovative methodology to forecast fracturing closure pressure. Makhotin et al.\textsuperscript{10d} introduced a ML model for estimating the production gain after executing a hydraulic fracturing job at a Siberian oilfield. Desouky et al.\textsuperscript{10e} applied a ML classifier to predict the etching patterns that result from acid fracturing, based on treatment conditions and rock type. In addition, the authors used ANN to predict the fracture conductivity of three types of carbonate rocks. Recently, Tariq et al.\textsuperscript{10f} used ANN to predict the breakdown pressure of the unconventional reservoirs with 95% accuracy.

**Production Estimation and Forecasting.** Lately, thanks to advancement in computational capabilities and optimization algorithms, it has been noted that a major focus was put on the use of AI to improve the estimation of flow rates in real time by leveraging large amounts of readily available surface data.\textsuperscript{11a−c} In a like manner, Li et al.\textsuperscript{11d} detected production decline and set early warning using vector machine algorithm. Cao et al.\textsuperscript{11e} combined ANN with history matching from offset wells to estimate oil well production. Amr et al.\textsuperscript{12a} predicted monthly production from horizontal wells in unconventional reservoirs using ML models. Oloso et al.\textsuperscript{12b} used ML to develop ensemble for the prediction of oil. Further, Pankaj et al.\textsuperscript{12c} optimized completion design parameters, and forecasted production using the gradient boost algorithm. Furthermore, Tariq et al.\textsuperscript{12d} presented a FN model to predict flowing bottom hole pressure (FBHP) in vertical wells that produce naturally with multiphase flow. The developed model showed higher accuracy than mechanistic models and empirical correlations.

Noshi et al.\textsuperscript{13a} applied ML algorithms to predict cumulative production within the first 12 production months. The authors deployed gradient-boosted trees, support vector regression, and adaptive boosting (AdaBoost) to detect the most contributing parameters. They collected data from five wells located in the Volve field, North Sea. Six features were used to build the model. These factors are production period, average bean size, average wellhead flowing pressure, and bore oil, gas, and water volumes. The conclusion showed that the developed model showed better prediction compared to other analytical models. Recently, Alzahabi et al.\textsuperscript{13b} performed data analytics using response surface methodology to optimize production of a horizontal well as a function of five input variables: type of reservoir, fracturing fluid, proppants, cluster spacing, and stage length. Tariq et al.\textsuperscript{13c} presented new PSO-ANN optimized model to predict FBHP in vertical wells that produce naturally with multiphase flow. The developed model showed less prediction error than mechanistic models and empirical correlations. Also, Tariq et al.\textsuperscript{13d} developed the FN-PSO model to predict the black oil PVT properties, where the created model outperformed other ML techniques and empirical correlations.

**Gas Lift.** As an artificial lift method, gas lift (GL) is a cornerstone in production operations, researchers paid more
| major Study | author and year of publication | innovation |
|-------------|--------------------------------|------------|
| production estimation and forecasting | Gorjai et al.,11a Seidi and Sayahi11b | estimated flow rates in real time |
| | Li et al.11d | detected production decline |
| | Cao et al.11e | estimate oil well production |
| | Ghorbani et al.11b | estimated flow rates in real time |
| | Amr et al.12a | predicted monthly production from horizontal wells in unconventional reservoirs |
| | Oloso et al.12b | developed ensemble for the prediction of oil |
| | Pankaj et al.12c | optimized completion design parameters and forecasted production |
| | Tariq et al.12d,13c | predicted FBHP in vertical natural-flow wells that produce with multiphase flow |
| | Noshi et al.13a | predicted cumulative production within the first 12 production months |
| | Alzahabi et al.13b | optimized production of horizontal wells |
| | Tariq et al.13d | predicted the black oil PVT properties using FN-PSO model |
| | Nande10c | identified criteria for hydraulic fracturing candidates |
| | Makhotin et al.10d | classified hydraulic fractures |
| | Desouky et al.10e | forecasted fracturing closure pressure |
| | Tariq et al.10f | estimated the positive production deference after executing hydraulic fracturing job |
| | Khamelchi et al.14a | predicted the etching patterns, that result from acid fracturing, based on treatment conditions and rock type |
| | Khan et al.14b | used ANN to predict the breakdown pressure of the unconventional reservoirs |
| | Liu et al.14c | defined optimum tubing size, gas-lift injection depth, and gas-lift injection rate |
| | Pennel et al.14d | innovated a robust correlation to forecast production rates in gas-lift assisted wells |
| | Guo et al.14e | detected any possible failures or anomalies in rod pumps |
| | Gupta et al.14f | predicted failure of both tubing and sucker rod pump |
| | Sneed14g | predicted life span of ESP and forecast failure |
| | Shadra\v\v et al.14h | identified future ESP problems |
| | Elkatatny et al.14i | defined optimum tubing size, gas-lift injection depth, and gas-lift injection rate |
| | Abdelgawad et al.14j | developed an automated events detection tool for ESP operation, based on reinforcement learning scheme |
| | Elmousalami and Elskary9c | designed drilling fluids |
| | Gasser et al.14k | used ANN to predict the rheological properties of drilling fluids |
| | Yu et al.14l | optimized well construction design and completion processes |
| | El Ouahed et al.,6a Mohaghegh,6b Al-Thuwaini et al.,6c Piovesan and Kozman,6d Anifowose et al.,6e Shahkarami et al.6f | predicted equivalent circulation density of drilling fluids |
| | Al-Mudhafar7b | classified pipe stuck automatically |
| | Dunham et al.7d | applied ANN to predict the rheological properties and filtrate invasion in nanobased drilling fluids |
| | Vallabhaneni et al.7e | used deep reinforcement learning to develop automated directional drilling model |
| | Bittar et al.7f | handled complex and large-scale seismic data and processed to small-scale core data |
| | Shalaby et al.7g | predicted lithofacies distribution and built relationship between lithofacies, core permeability, and well logging data |
| | Tariq et al.14m | automatically interpreted well logs |
| | Kellogg et al.14n | applied FN technique to estimate Poisson’s ratio in carbonate rocks |
| | Arigbe et al.14p | enhanced classification of well logs using semi-supervised learning |
| | Elkatatny et al.14q | integrated 3D seismic data with petrophysical properties |
| | Ahmadi and Chen13d | applied reinforced learning to depth matching of well logs |
| | Gomaa et al.13e | identified the nonlinear correlation between the well log data and total organic carbon |
| | AlAjmi et al.16a | applied FN-PSO to predict water saturation in carbonate rocks |
| | Belozrov et al.7c | removed wellbore damage and restored permeability |
| | Tariq et al.14m | predicted oil—water relative permeability |
| | Tariq et al.14o | predicted the permeability of heterogeneous reservoirs |
| | Kellogg et al.14n | predicted permeability reduction (formation damage) |
| | Arigbe et al.14p | calculated the efficiency of a vertical sweep in oil reservoirs |
| | Elkatatny et al.14q | forecasted downhole integrity of casing strings in wells without corrosion logging data |
attention and successfully used ML in different artificial lift techniques. Khamehchi et al.\textsuperscript{14a} deployed genetic algorithms for defining optimum tubing size, gas-lift injection depth, and gas-lift injection rate. In their work, they represented reservoir performance utilizing a constant productivity index model. Khan et al.\textsuperscript{14b} used AI techniques to innovate a robust correlation to forecast production rates in gas-lift assisted wells. The AI techniques used in this research included artificial neuro-fuzzy inference systems, ANN, functional networks, and support vector machines. They collected test data from several GL wells and used ANN to develop an equation to forecast oil flow rate. Initially, they applied wide data analytics and then input data to the models that were compared to each other and to other empirical models. They could predict oil rates with accuracy exceeding 98%.

**Sucker Rod Pump.** Liu et al.\textsuperscript{14c} applied framework to remotely detect any possible failures or anomalies in rod pumps. Pennel et al.\textsuperscript{14d} reviewed problems of different artificial lifting methods and used ML algorithms to predict failure of both tubing and pump. Production data were collected for more than 1 year, from producing wells, and used to train ML models. Failure and performance data for sucker rod pumps were reviewed and identified. Random forest, neural network, and gradient-boosted tree models were used. The study proved high diagnostic model accuracy more than 99% and precision range between 50 and 60%. In terms of modeling long-term tubing failure, 31 days were found to be the accepted error compared to 100–600 days of a well life span. Suboptimal performance models showed accuracy more than 90%. Authors recommended applying the analytic workflow approach (that was presented in their paper to diagnose and predict both failure and suboptimal performance states in rod pumps) across different artificial lift types. They mentioned that the approach could be used to detect accurate injection rates, holes in tubing, hydrate plugging, identifying status of the GL valve, valve throttling, restrictions in flow lines, and more.

**Electrical Submersible Pump.** Guo et al.\textsuperscript{14e} applied ML models to predict life span of electrical submersible pumps (ESP) and forecast failure. Gupta et al.\textsuperscript{3a} used ESP performance-related statistics to develop a data-driven analytical solution using multivariate statistical techniques such as dimensionality reduction and pattern recognition. The authors combined engineering and mathematical principles in a real-time developed framework to identify future problems long before they arise, analyze potential sources, and recommend precautionary action. Their work helped to protect ESP operations and maximize time, prolong ESP life expectation, decrease intervention budgets, and enhance production. Analogous to Guo et al.,\textsuperscript{14e} Sneed\textsuperscript{4d} used ML algorithms to forecast ESP failures. Recently, Chong et al.\textsuperscript{14f} developed an automated events detection tool for ESP operation, based on reinforcement learning scheme.

**Wellbore Damage and Permeability.** In their research, Kellogg et al.\textsuperscript{15a} utilized a logistic regression approach for removing wellbore damage and restoring permeability. Aribge et al.\textsuperscript{15b} reviewed the existing relative permeability correlations and concluded that there are many limitations caused by their assumptions and complexity in a real-time update for different reservoirs. Therefore the authors used deep neural networks to predict oil-water relative permeability. Elkatatny et al.\textsuperscript{15c} applied ANN, adaptive neuro-fuzzy inference system (ANFIS), and SVM models to predict the permeability of heterogeneous reservoirs based on resistivity, neutron porosity, and bulk density logs. Ahmadi and Chen\textsuperscript{15d} employed ANN and a least-squares support vector machine to predict permeability reduction (formation damage) resulting from scale deposition throughout the water injection process. This helped them to have better and reliable comprehension of formation damage effects during water flooding without having costly laboratory measurements. Recently, Gomaa et al.\textsuperscript{15e} presented a new ANN model to calculate the efficiency of a vertical sweep in oil reservoirs.

**Well Integrity.** ML was remarkably used in managing and enhancing integrity of well barriers. AlAjmi et al.\textsuperscript{15f} used ANN to forecast downhole integrity of casing strings in wells without corrosion logging data. The authors collected more than 100 data sets to build the model. Yan\textsuperscript{15b} used the fault tree analysis method for many wellbore integrity factors to establish the damage analysis model of the wellbore. According to these risk factors, risk control measures are put forward, which provides reference for predicting if wellbore integrity to ensure the safety of O&G production runs smoothly. Khreibeh et al.\textsuperscript{15l} improved efficiency of WIMS in their study by the presentation of intelligent software analytics, integrating data from different sources. The WI data management process has decreased the duration spent on the search, quality assurance, and compilation of details for such well-integrity-based activities by between 68 and 87%. These efficiency gains were primarily accomplished by removing complicated excel sheets used for tracking, plotting, and monitoring annulus pressures and safety-sensitive equipment checks, which are essential details for risk evaluation and annulus pressure management.

In similar fashion, Noshi et al.\textsuperscript{15c} deployed nine ML models and identified the best algorithm performance to determine features affecting casing failure. Bilogan et al.\textsuperscript{15d} used real-time data for the enhancement of WI monitoring workflows. In their research, the authors presented the results of applying a pilot workflow for a giant onshore asset. They considered the advantage of real-time data visualization and implemented an online ML system to detect the anomalies and recognize the well

---

**Table 2. continued**

| major Study | author and year of publication | innovation |
|-------------|---------------------------------|------------|
| Yan\textsuperscript{16b} | | established the damage analysis model of the wellbore and predicted wellbore integrity |
| Khreibeh et al.\textsuperscript{17f} | | quality assurance and compilation of details for well activities |
| Noshi et al.\textsuperscript{17c} | | determined features affecting casing failure |
| Bilogan et al.\textsuperscript{17a} | | implemented an online machine learning system to detect the anomalies |
| Noshi et al.\textsuperscript{17a} | | identified the likely factors affecting casing failures |
| Elchev et al.\textsuperscript{17d} | | identified and classified well events for natural flowing wells |
| Ragab et al.,\textsuperscript{17a,b} Yakoot et al.\textsuperscript{17c,d} | | developed ML model to predict severity of WI failure in artificial lift systems |
events. Again, Noshi et al.\textsuperscript{16a} deployed ANN and boosted ensemble trees to assess 26 diverse features collected from drilling, fracturing, and geologic data. They identified the likely factors affecting casing failures using ML algorithms. Twenty wells with casing and tubing failure out of 80 onshore wells were studied. Eliechev et al.\textsuperscript{16d} suggested an approach that can identify and classify well events. The authors proposed an algorithm that allows retrospective analyses of data and precise identification of different events in the past or in real time. The algorithm was developed for natural flowing wells with ease of extending it to artificial lift wells. Recently, 11 ML algorithms were applied to data set of WI-related tests and failure events. This resulted in developing a novel model for automated analysis and of WI failures in artificial lift wells.\textsuperscript{17a–d}

As aforementioned, AI has been involved widely to tackle problems in the O&G industry, and the literature is rich with many case studies. Nevertheless, going through the previous works, it has been concluded that the application of AI in WI is still lacking more novelty. Some important advances achieved in the O&G industry by deploying AI and ML are presented in Table 2.

In summary, ML was lately utilized in the O&G industry to predict and forecast oil production, reservoir features, and artificial lift problems. Meanwhile, the literature is short of adopting such innovation to predict and analyze WI status of O&G producers. This research would be a milestone in the design and creation of the WI database management program through the combination of integrity and ML.

\section*{Building WI Machine Learning Model}

The appropriate package should be selected for scientific computing, performing different operations, data manipulation and analysis, and data visualization. In addition to having good data, you need to make sure that it is in a useful scale, format and even that meaningful features are included.\textsuperscript{18a} Further, the model can be subdivided into four submodels as follows:

1. \textbf{Predictive Model:} to predict the accurate risk status of wells and classify their integrity level into five categories rather than three broad-range categories as in qualitative risk classification. The categories are:
   - Category 1 (CAT-1), which is too risky.
   - Category 2 (CAT-2), which is still too risky but less risky than category 2.
   - Category 3 (CAT-3), which is medium risk but can be elevated to risky wells in case of additional barrier(s) failure occurs.
   - Category 4 (CAT-4), which is low risky but with some impaired barriers.
   - Category 5 (CAT-5), which is the lowest in risk among them.

2. \textbf{Failure Model:} to identify whether the well if considered in failure mode or not. In addition, the model can identify wells that necessitates prompt mitigation and securing.

3. \textbf{Suboptimal Model:} to identify which wells are operated out of the integrity envelope and has some elements impaired.

4. \textbf{Diagnosis Model:} to locate the root of cause of WI impairment and introduce brief action plan to restore WI.

\textbf{Steps of Building the Model. Data Gathering.} Different data sets should be collected from many wells and/or fields. Each data array should represent a well event and comprise different features that are related to the final status of WI. The status of each feature can be reported in specified groups across the entire arrays.

\textbf{Data Pre-Processing.} Data gathering, processing, and analysis is defined as analytics.\textsuperscript{18b} Data preprocessing is a long and burdensome process to get unseen knowledge.\textsuperscript{18c} Data analytics extracts the significant information from data that supports real-time recognition of failures, early prevention of them, forecasting, and risk quantification of assets. Preprocessing of data refers to the transformations applied to the data before feeding it to the algorithm (i.e., cleaning data to have homogeneity). This includes solving problems of missing values, outliers, biased data, etc.\textsuperscript{18d} Data preparation may be one of the most difficult steps in any ML project. The reason is that each data set is different and highly specific to the project.\textsuperscript{18e} Raw data typically cannot be used directly in ML models because of the following reasons:

- ML algorithms require data to be numbers.
- Some ML algorithms impose requirements on the data. Errors and statistical noise existing in the gathered data should be corrected.
- Complex nonlinear relationships may find difficulty in obtaining data.

Data cleaning is the following data gathering stage and aims at checking errors/mistakes with the data and to remove/fix data if needed.\textsuperscript{18a} Feature relevancy to the output is different from feature to another one, so to make the created model more effective, it was mandatory to scrutinize all data gathered to evaluate their significance and remove features with low relevancy to the output.\textsuperscript{18a}

\textbf{Missing Values.} Real data often have missing values. Data may include missing values for many reasons such as nonrecorded observations and/or data corruption. Handling missing data is important as many ML algorithms do not support data with missing values. Missing values are a common occurrence, and there should exist a strategy for treating them. A missing value can signify a number of different things in research data such as lack of data, irrelevance of data, non-occurrence of the event itself, or human error during data entry. DM methods may ignore the missing values, exclude any relevant record, use variable mean instead, or conclude them from existing values. Approaches of missing values replacement are as follows:

- Ignore the records with missing values.
- Replace missing values with a universal constant (e.g., "?").
- Fill in missing values manually based on domain knowledge.
- Replace missing data with the mean value (if variable is numerical) or the most recurring value (if variable is categorical).
- Use modeling techniques.

Categorical imputation was implemented on the data set by replacing the missing values with the other compatible values based on subject matter experience.

\textbf{Detecting Outliers.} Outlier is a data point that differs significantly from other observations. An outlier may result in challenging problems during statistical analysis. Outliers can occur by chance in any distribution. Outlier points may be an indication for incomplete data or flawed procedures. However, in big data samples, a smaller number of outliers is to be expected and not because of faulty condition.
Feature Engineering. Feature engineering is the process of incorporating domain knowledge to identify attributes from raw data. Efficiency of ML algorithms can be enhanced using the step of feature engineering. Relevancy of features to the output varies from one feature to another, so to make the proposed model more accurate, it is mandatory to investigate all the data set features and assess their importance and focus on features with high relevancy to the output. One of the most common feature engineering techniques that is extensively used is categorical encoding and feature scaling.

Categorical Encoding. Any structured data set contains various columns with a combination of numerical and categorical features. Machines recognize numbers, not text. Each text category needs to be encoded into numbers (Figure 5) for processing them using model algorithms.

The process of categorical encoding converts categories into numbers. The two methods most commonly adopted are label encoding and one-hot encoding. Label encoding includes assigning a distinctive integer to each text label considering alphabetical ordering. In this type of encoding, there is a high probability that the model captures misleading relationship between features. Here is the concept of one-hot encoding which creates dummy variables for each data feature.

WI classification was first implemented qualitatively using WI elements and resulting in three risk categories, namely, high, medium, and low category wells. Each WI element was identified as being either integral or nonintegral. Further, a quantitative approach was developed to replace the status of each WI element with a weight score representing the contribution of each element into the risk status of WI.

Scaling. Feature scaling is an important methodology applied to normalize the range of data features. It is also defined as data normalization and normally applied during the step of data preprocessing. Normalization is vital for some ML algorithms when there is a wide range of raw data. It enables using a common range for the numeric features without losing information. It is also paramount for efficient data modeling in some algorithms because combining values of features with different scale could cause unseen problems. Data scaling avoids such problems and creates normalized values that maintain the same ratios and distribution of the original values. The simplest method for normalization and feature scaling is “min–max scaling” or called “min–max normalization”. The general formula for a min–max of [0,1] is given as the following equation (eq 1):

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$  \hspace{1cm} (1)

where $x$ = the original value and $x'$ = the normalized value.

Standardization is another scaling method that centers the values around their mean with one standard deviation. This technique does not restrict values to a specific range. The standardization formula is given as the following equation (eq 2):

$$x' = \frac{x - \mu}{\sigma}$$  \hspace{1cm} (2)

where $x$ = the original value, $x'$ = the standardized value, $\mu$ = the mean of values, $\sigma$ = the standard deviation.

Preference for using a particular scaling technique depends on the problem nature and applied ML algorithm. Sometimes, we can normalize and standardize the data and then compare the best achieved results.

Algorithm Selection and Training. The field of ML includes enormous number of algorithms, some of which are easy to use, while some others necessitate more difficult understanding of the algorithms. Predictive modeling is the method of developing a model using historical data to make a prediction on new data where we do not have the answer. It can be described as a mathematical problem of approximating a mapping function ($f$) from input variables ($x$) to output elements.
variables \( (y) \). Below is a list of the ML algorithms that were duly applied in this research with their advantages and disadvantages.

**Logistic Regression.** It is a popular classification algorithm that is applied for categorical variables. The idea of logistic regression is to find a relationship between features and probability of a particular outcome.\(^{19d}\) Its technique is similar to linear regression and why it is named “logistic regression”. The term “logistic” is taken from the logit function that is used in this method of classification.\(^{19e}\)

Logit function used in logistic regression can be expressed as eq 3:

\[
\log \left( \frac{p(x)}{1 - p(x)} \right) = \beta_0 + \beta_1 x
\]

where the left-hand side is called the logit (log-odds) function, and \( p(x)/(1 - p(x)) \) is called the odds ratio (ratio of success probability to failure probability). If we take an inverse of the above function, we get eq 4:

\[
p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}
\]

Decision boundary assists to discriminate probabilities into negative and positive classes. Decision boundary might be linear or nonlinear, as shown in Figure 6.

**Decision Trees.** Decision tree algorithm is a supervised learning algorithm. Unlike other supervised learning algorithms, they are also known as CART (classification and regression trees), as they can be used for both. It builds various models in a tree-structured form. It splits a data set into smaller parts; meanwhile, an associated decision tree is developed.\(^{9c}\) Decision trees classify the examples by sorting them down the tree from the root to some leaf/terminal node, with the leaf/terminal node providing the classification of the example.\(^{19f}\) Figure 7 shows an example of a simple decision tree. Decision tree splits nodes on the available features to generate the most homogeneous subnodes.
Random Forest. The prediction of a single decision tree might be wrong, as it may have a variance. To solve this, random forests take predictions from hundreds or thousands of trees and average them together. Decision trees are the basic building blocks of a random forest which is a supervised ML model that map data to outputs during the model-building phase. In order to arrive at an estimate, each decision tree in the random forest asks a series of questions, with each question narrowing the possible values until it is confident enough to make a prediction. A random forest example is illustrated in Figure 8.

Support Vector Machines (SVM). Separating two data groups can be achieved by straight line, which is considered to be a one-dimension, flat plane as two dimensions or a hyperplane with \( n \) dimensions. SVM can be used for both classification and regression. SVM performs classification by finding the hyperplane that maximizes the margin between the two classes (Figure 9). If we have space that has \( n \) dimensions, subspace with \( n - 1 \) dimensions is then called the hyperplane. For a two-dimensional space, its hyperplane will be one dimension, which is just a line. A hyperplane with two dimensions is used for space with three dimensions.

SVM can be applied in linearly separable classes where margins are the perpendicular distances between the line and those dots closest to the line. If predicted classes are classified properly (linear case), SVM tries to identify the hyperplane that can maximize the margin, yet data sets are probably never linearly separable, so the condition of 100% correctly classified by a hyperplane will never be met. Figure 10 shows linear versus nonlinear SVM.

K-Nearest Neighbors. The K-nearest neighbors (KNN) is considered to be supervised ML algorithms that are applied for regression or classification and models. The algorithm works on the principle that the value of a data point is determined by data points around it. The KNN algorithm uses a majority voting principle to classify data points. For example, setting \( K = 4 \), the classes of 4 closest points are then checked. Prediction is done according to the majority class. Data points are determined to be close by measuring the distance between them. There are many methods to measure the distance; the most common one is calculating the Euclidean distance (Figure 11), which is the ordinary straight-line distance between two points.

Stochastic Gradient Descent (SGD). It is considered as a stochastic approximation of gradient descent optimization. It significantly replaces the gradient calculated from the whole data set by another estimate, including gradient calculated from a random selection of data subset. SGD can be effectively applied to various ML problems.
Adaptive Boosting (AdaBoost). It is a meta-estimator that begins by fitting a classifier on the original data set and then fits additional copies of the classifier on the same data set but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases. The core principle of AdaBoost is to fit a sequence of weak learners on repeatedly modified versions of the data. The predictions from all of them are then combined through a weighted majority vote (or sum) to produce the final prediction.

Naive Bayes. It is a classification technique based on Bayes’ theorem with an assumption of independence among the independent variables. In simple terms, the naive Bayes’ classifier assumes that the presence of a feature in a class is unrelated to the presence of any other feature. Bayes’ theorem provides a way of calculating posterior probability \( P(cl|x) \) from the class prior probability \( P(c) \), predictor prior probability \( P(x) \), and the likelihood \( P(x|c) \). This is explained in eq 5:

\[
P(cl|x) = \frac{(P(cl|x) \times P(c))}{P(x)}
\]

where \( P(cl|x) \) is the posterior probability of class \( (c, \text{target}) \) given predictor \( (x, \text{attributes}) \) and \( P(c) \) is the prior probability of class. \( P(cl|x) \) is the likelihood in which the probability of \( P(x) \) is the prior probability of predictor.

Quadratic Discriminant Analysis (QDA). A classifier with a quadratic decision boundary, generated by fitting class conditional densities to the data and using Bayes’ rule. The model fits a Gaussian density to each class. This classifier is powerful because it has closed-form solutions that can be easily computed, is inherently multiclass, has proven to work well in practice, and has no hyper parameters to tune.

Artificial Neural Networks. It is composed of artificial neurons which are inspired from biological neurons. Each neuron has inputs and produces a single output that can be sent to other multiple neurons. ANN can be applied in both classification and regression problems. Essentially, an ANN comprises layers that consist of interconnected nodes that contain activation function. ANN may contain and input layer (represents the raw data), a hidden layer (one or more hidden layers based on weights on the connections between input and hidden units), and an output layer (depends on the weights between hidden and output units).

Table 3 shows the advantages and disadvantages of all of the previously discussed algorithms.

**Table 3. Advantages and Disadvantages of Different Machine Learning Algorithms**

| ML algorithm | advantages | disadvantages |
|--------------|------------|--------------|
| logistic regression | • it is used extensively because of its efficiency, highly interpretability, and no scaling or tuning required<br>• logistic regression has higher efficiency when irrelevant and correlated features removed | • other algorithms will outperform logistic regression in case of similar or correlated variables |
| decision trees | • easier data preparation during preprocessing phase<br>• decision tree does not prerequisite data normalization or scaling<br>• missing values in the data also does not affect the process of building decision tree to any considerable extent | • instability of model can result from any minor change in data<br>• long time of model training |
| random forest | • it can handle large number of features (dimensions)<br>• it can calculate the contribution of each feature<br>• less training time | • difficult interpretation of some random forest models<br>• for very large data sets, the size of the trees can take up a lot of memory |
| SVM | • higher accuracy in case of having classes with clear separation margin<br>• SVM is shows more efficiency in case of high dimensional spaces | • presence of noise in data badly affects SVM performance<br>• model speed decrease with increasing number of data points increases<br>• not memory efficient<br>• sensitive to outliers |
| KNN | • simple and easy to interpret<br>• it can be directly implemented in nonlinear cases because it does not include any assumption | • requiring a number of hyper parameters and being sensitive to feature scaling<br>• sensitive to noisy data and outliers |
| SGD | • efficiency and ease of implementation | • it is almost impossible that we get a set of predictors which are completely independent |
| AdaBoost | • simple to implement<br>• few parameters | • performance severely decline as number of predictor variables approaches sample size |
| naive Bayes | • better performance compared to other models in case of independent predictors<br>• it performs well in case of categorical input variables compared to numerical variables | • needs a large and diverse training data for real-life applications<br>• in many cases, referred to as “black box” as it offers little insights |
| QDA | • more efficient compared to other models like logistic regression in case of having more than two non-ordinal classes<br>• more stable than other models in case of well-separated classes | • it is almost impossible that we get a set of predictors which are completely independent |
| ANN | • good performance with linear and nonlinear data<br>• capable to learn from the analyzed data and reprogramming not required | • needs a large and diverse training data for real-life applications<br>• in many cases, referred to as “black box” as it offers little insights |

Adaptive Boosting (AdaBoost). It is a meta-estimator that begins by fitting a classifier on the original data set and then fits additional copies of the classifier on the same data set but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases. The core principle of AdaBoost is to fit a sequence of weak learners on repeatedly modified versions of the data. The predictions from all of them are then combined through a weighted majority vote (or sum) to produce the final prediction.

Artificial Neural Networks. It is composed of artificial neurons which are inspired from biological neurons. Each neuron has inputs and produces a single output that can be sent to other multiple neurons. ANN can be applied in both classification and regression problems. Essentially, an ANN comprises layers that consist of interconnected nodes that contain activation function. ANN may contain and input layer (represents the raw data), a hidden layer (one or more hidden layers based on weights on the connections between input and hidden units), and an output layer (depends on the weights between hidden and output units).

Table 3 shows the advantages and disadvantages of all of the previously discussed algorithms.

**Table 3. Advantages and Disadvantages of Different Machine Learning Algorithms**

| ML algorithm | advantages | disadvantages |
|--------------|------------|--------------|
| logistic regression | • it is used extensively because of its efficiency, highly interpretability, and no scaling or tuning required<br>• logistic regression has higher efficiency when irrelevant and correlated features removed | • other algorithms will outperform logistic regression in case of similar or correlated variables |
| decision trees | • easier data preparation during preprocessing phase<br>• decision tree does not prerequisite data normalization or scaling<br>• missing values in the data also does not affect the process of building decision tree to any considerable extent | • instability of model can result from any minor change in data<br>• long time of model training |
| random forest | • it can handle large number of features (dimensions)<br>• it can calculate the contribution of each feature<br>• less training time | • difficult interpretation of some random forest models<br>• for very large data sets, the size of the trees can take up a lot of memory |
| SVM | • higher accuracy in case of having classes with clear separation margin<br>• SVM is shows more efficiency in case of high dimensional spaces | • presence of noise in data badly affects SVM performance<br>• model speed decrease with increasing number of data points increases<br>• not memory efficient<br>• sensitive to outliers |
| KNN | • simple and easy to interpret<br>• it can be directly implemented in nonlinear cases because it does not include any assumption | • requiring a number of hyper parameters and being sensitive to feature scaling<br>• sensitive to noisy data and outliers |
| SGD | • efficiency and ease of implementation | • it is almost impossible that we get a set of predictors which are completely independent |
| AdaBoost | • simple to implement<br>• few parameters | • performance severely decline as number of predictor variables approaches sample size |
| naive Bayes | • better performance compared to other models in case of independent predictors<br>• it performs well in case of categorical input variables compared to numerical variables | • needs a large and diverse training data for real-life applications<br>• in many cases, referred to as “black box” as it offers little insights |
| QDA | • more efficient compared to other models like logistic regression in case of having more than two non-ordinal classes<br>• more stable than other models in case of well-separated classes | • it is almost impossible that we get a set of predictors which are completely independent |
| ANN | • good performance with linear and nonlinear data<br>• capable to learn from the analyzed data and reprogramming not required | • needs a large and diverse training data for real-life applications<br>• in many cases, referred to as “black box” as it offers little insights |
data sets). For better generalization, cross-validation is required during the training phase to ensure all portions of data set are exploited. The nature of the ML algorithm and model complexity also play an obvious role in model generalization.

Achieving a good fit on the ML model is eminently essential. This includes a trade-off between bias and variance (regression) or precision and recall (classification). A confusion matrix is simply a table that is frequently employed to outline the performance of a classification model. It tabulates the combinations of actual and predicted status. Moreover, it allows the visualization of the performance of an algorithm. All of the predicted values by the ML model lie under one of four categories illustrated in Figure 12.

![Confusion matrix](https://doi.org/10.1021/acsomega.1c05658)

**Figure 12.** Confusion matrix.

Concisely, ML is completely different from traditional programing, and it includes multiple algorithms that can be applied efficiently in regression and classification problems. ML models are broadly categorized into four categories comprising supervised, unsupervised, semi-supervised, and reinforcement learning. Building the ML model has definitive steps that must be followed accurately to obtain a reliable model with high accuracy. Selecting the ML model that matches data collected is a key stage in building any model.

### SUMMARY AND CONCLUSIONS

ML algorithms are booming as a firm tool to develop a powerful model and explore hidden correlations between parameters in O&G wells. Introducing ML models has allowed an effective estimation of the risk level for WI. Management of WI is a fundamental feature of the oil field philosophy for safety and operational integrity. The success of any model starts from gathering data representing different groups of wells and multiple services in different environments. Including more features that are believed to contribute to the WI envelope means stronger model prediction and higher accuracy. Data preprocessing and feature engineering must be duly applied. Based on discussion and data analysis, the following conclusions can be drawn:

1. This paper has presented a comprehensive compilation of different ML applications in diverse disciplines of petroleum industry. We categorized the pool of AI and ML with respect to different areas of applications along with publication years.

2. In this work, we divided applications into nine groups, i.e., petrophysics, wellbore damage, drilling, production, fracturing, ESP, GL, sucker rod pumps, and well integrity. The main focus of this study was well integrity failures where the authors found an obvious gap in the literature regarding failure predictions.

3. ML is a robust tool whenever sufficient field data are available for training a model. Furthermore, sufficient engineering judgment is essential for both data analysis and preprocessing.

4. AI and ML models are modern and thriving approaches to develop models for ranking integrity of critical wells and give them top priority. This would help to utilize company resources efficiently and dedicate personnel efforts to the highly risked wells. As a result, progressive improvements on business, safety, and environmental performance should be achieved.

5. The ML model is helpful to the organization in terms of essential classification and data quality for the analysis of hundreds of WI issues, enabling careful evaluation, quick follow-up, and data collection to recommend effective action or repair as applicable.

6. Effectiveness of the model confirmed by ensuring its continuity and capability to be used across all similar assets. In other words, the model must be functional and can replace all previous processes and systems completely.

7. The efficient model must act as an automated systematic tool for calculating an imposed risk category of any well. It should be competently used in all operating wells.

In summary, ML applications have been compiled to be a valuable reference for the reader. The application of ML can help to improve the strategy of operating brownfield assets, optimize well-intervention expenses, obtain more efficient management of risks throughout the field, boost business performance, and maximize production. Moreover, the methodology for developing a ML tool has been introduced to manage WI in mature fields. This tool can (a) predict well risk level, (b) provide a unique methodology to convert associated failure risk of each element in the well envelope into a tangible value, (c) show the total potential risk, and (d) identify the status of overall well-barrier integrity.

### RECOMMENDATIONS

This study has covered almost all of the ML applications in the industry during the past few years. Further, the inferred conclusions provide the following important implications and insights for future research:

1. E&P companies should accelerate digital transformation in all components in the industry (e.g., people, organization, structure, and processes). In other words, they should accelerate their digitalization maturity through strong business cases rather than having multiple isolated digital initiatives.

2. O&G operators should pay more attention to the value of data in the age where data are considered as new oil. The importance of data is often misunderstood, which does not make it influential as it should be or even recorded.

3. Data science should be applied to artificial lift techniques where it can effectively reduce cost, increase productivity,
improve competence, and ultimately increase return on investment.

4. WI should be transformed from legacy tools to IOT and AI in all related aspects. Being a critical discipline involved in all well life cycles makes it more urgent to O&G operators to modernize and step-up from regular spreadsheets and non-user-friendly software to a self-learning machine. This will empower the achievement of ultimate system maturity.

It is obvious in the era of AI that innovating a WI prediction tool based on ML is helpful to organizations in terms of having essential classification and reliable analysis of hundreds of WI issues, enabling careful evaluation, quick follow-up, and data collection to recommend effective action or repair as applicable. In addition, all teams in the company are now having the same language regarding WI challenges. Challenges and recommendations for future research are also discussed and thoroughly presented.

AUTHOR INFORMATION

Corresponding Authors
Omar Mahmoud — Future University in Egypt (FUE), New Cairo 11835, Egypt; orcid.org/0000-0002-9223-6349; Email: omar.mahmoud@aggienetwork.com
Adel M. Salem — Suez University, Suez 43511, Egypt; Email: adel.salem@suezuni.edu.eg
Mostafa S. Yakoot — Gulf of Suez Petroleum Co. (GUPCO), Cairo 11511, Egypt; Email: mostafayakoot@yahoo.com

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.1c05658

Notes
The authors declare no competing financial interest.

Biographies
Adel M. Salem is currently the Head of the Petroleum Engineering Department at Suez University. He obtained a Ph.D. (2008) in Petroleum Engineering from Leoben University, Austria. He also earned a B.Sc. and M.Sc. from Suez Canal University, Egypt, all in Petroleum Engineering. Previously, he was assistant professor at American University in Cairo (AUC) for 4 years, and Future University in Egypt (FUE) for 2 years. Dr. Adel has published more than 80 papers on EOR, characterization of formation damage, nanotechnology of EOR, smart drilling fluids, oil shale, well testing, hydraulic fracturing, radial drilling, and well performance using nodal analysis. Moreover, he gained international experience on NMR and production simulation at Bologna University, Italy, and at Leoben University on simulation. Before joining the university in 1997, He worked as a field Petroleum Production Engineer at Qarun Petroleum Company, Western Desert, Egypt.

Mostafa S. Yakoot has completed his B.Sc. and M.Sc. from Suez University, Egypt, and is currently enrolled as a Ph.D. student, all in petroleum engineering. He is currently a senior petroleum engineer in GUPCO, Egypt, responsible for well integrity management and well intervention. In addition, he is an accredited IWCF/IADC instructor in well intervention pressure control. He did more than 1000 teaching hours in well intervention and well integrity to candidates in Egypt, Africa, Middle East, Asia, and Europe. Mostafa held many roles in GUPCO, including production technologist, well-site leader, and well integrity engineer.

Omar Mahmoud is currently working as assistant professor at the Department of Petroleum Engineering, Faculty of Engineering and Technology, Future University in Egypt (FUE). Previously, he was working as a petroleum reservoir engineer with the North American Unconventional Resources (NAUR) Team at Apache Corporation’s San Antonio Region, USA. Before joining Apache Corporation, he worked as a graduate research assistant/Ph.D. candidate at the Harold Vance Department of Petroleum Engineering, Texas A&M University, USA. Before that, he worked for 8 years as an assistant lecturer and teaching/research assistant at the Faculty of Petroleum and Mining Engineering, Suez Canal University, Suez, Egypt. He also served as a petroleum engineering intern at Apache Corporation (USA) during the summer of 2017. Mahmoud has demonstrated history of working in higher education and research industry in addition to having good experience of working with petroleum companies. His research interests include drilling fluids, rheology, filter cake characterization, nanoparticle applications, oilfield chemistry, formation damage, well stimulation, unconventional resources, reservoir simulation, artificial intelligence, machine learning, and pressure transient/production data analysis. Mahmoud has authored/co-authored 40+ technical articles and conference proceedings and served as technical reviewer/editor for SPE Journal, SPE Production & Operations, SPE Drilling & Completions, ASME Journal of Energy Resources Technology, Journal of Petroleum Exploration & Production Technology, Energies, and the Canadian Journal of Chemical Engineering. He also is serving as associate editor for the Petroleum & Petrochemical Engineering Journal (PPEJ) Medwin Publishers. He received a Ph.D. degree from Texas A&M University and B.Sc. and M.Sc. degrees from Suez University, Egypt, all in petroleum engineering.

ACKNOWLEDGMENTS

The authors would like to express appreciation for the support of the Gulf of Suez Petroleum Company (GUPCO) and permission to publish this work. The authors would also like to acknowledge Prof. Ahmed A. Elgibaly, who passed away on January 15, 2021. The second author would like to extend his appreciation for the faculty members in the department of petroleum engineering, Suez University, Egypt, for their continuous encouragement and support.

ACRONYMS

AAPE = Average Absolute Percentage Error
AdaBoost = Adaptive Boosting
AI = Artificial Intelligence
AME = Mean Absolute Error
ANFIS = Adaptive Neuro-Fuzzy Inference System
ANN = Artificial Neural Network
CAT = Category
DM = Data Mining
E&P = Exploration and Production
EOR = Enhanced Oil Recovery
ESP = Electric Submersible Pump
FBHP = Flowing Bottom Hole Pressure
FN = Functional Network
GL = Gas Lift
HSE = Health, Safety, and Environment
IOT = Internet Of Things
KNN = K-Nearest Neighbors algorithm
MIT = Multifinger Imaging Tool
ML = Machine Learning
O&G = Oil and Gas
PSO = Particle Swarm Optimization
PVT = Pressure Volume Temperature
QC = Quality Control
QDA = Quadratic Discriminant Analysis
R² = Coefficient of determination
RMSE = Root Mean Squared Error
SGD = Stochastic Gradient Descent
SVM = Support Vector Machine
WI = Well Integrity
WIMS = Well Integrity Management System

REFERENCES

(1) (a) Yakoot, M. S.; Elgibaly, A. A.; Ragab, A. M. S.; Mahmoud, O. Well Integrity Management in Mature Fields: A State-of-the-Art Review on the System Structure and Maturity. *J. Pet. Explor Prod. Technol.* 2021, *11*, 1833−1853. (b) Yakoot, M. S.; Elgibaly, A. A.; Ragab, A. M. S.; Mahmoud, O. A Comprehensive Review and Analysis of Maturity Model for Well Integrity in Brownfield. *IADC Drilling Middle East 2020 Conference & Exhibition* (Virtual), 2020; https://www.iadc.org/wp-content/uploads/2020/12/ME20-Mostafa-Yakoot.pdf (accessed 2022-01-06). (c) Perrons, R. K.; Jensen, J. W. Data as an Asset: What the Oil and Gas Sector Can Learn from Other Industries About "Big Data." *Energy Policy.* 2015, 81, 117−121. (d) Mishra, S.; Datta-Gupta, A. Basic Concepts. In *Applied Statistical Modeling and Data Analytics: A Practical Guide for the Petroleum Geoscientists*, 1st ed.; Elsevier: Amsterdam, The Netherlands, 2017; pp 1−13. (e) Hajirahimova, M. S. Opportunities and Challenges Big Data in Oil and Gas Industry. *National Supercomputer Forum*; Russian Federation: Pereslavl-Zalessky, 2015. (f) Kheirabadi, H. A.; Shamsi, S. A.; Vasillo, E.; Farag, F.; Lunt, G.; Bhardwaj, J. Optimizing Well Integrity in Intelligent Oil Fields. *SPE/IADC Middle East Drilling Technology Conference and Exhibition*; Abu Dhabi, UAE, January 29−31, 2018; DOI: 10.2118/189389-MS.

(2) (a) Okwu, M.; Nwachukwu, A. N. A Review of Fuzzy Logic Applications in Petroleum Exploration, Production and Distribution Operations. *J. Pet. Explor Prod. Technol.* 2019, *9*, 1555−1568. (b) Irons-Mclean, R.; Greengrass, J. The Internet of Things IoT and Security: A Practical Strategy to Secure an OT and IT Integrated Process Control Domain. *SPE Intelligent Energy International Conference and Exhibition*; Aberdeen, Scotland, U.K., 2016; DOI: 10.2118/181002-MS. (c) Crockett, B. J. The Internet of Things in Upstream Oil and Gas – How Can Decisions Be Made in Real-Time and Safely Manage Risk? *SPE Intelligent Energy International Conference and Exhibition*; Aberdeen, Scotland, U.K., 2016; DOI: 10.2118/181052-MS. (d) Baudoin, C. R. Deploying the Industrial Internet in Oil & Gas: Challenges and Opportunities. *SPE Intelligent Energy International Conference and Exhibition*; Aberdeen, Scotland, U.K., 2016; DOI: 10.2118/181107-MS.

(3) (a) Gupta, S.; Saputelli, L.; Nikolaou, M. Applying Big Data Analytics to Detect, Diagnose, and Prevent Impending Failures in Electric Submersible Pumps. *SPE Annual Technical Conference and Exhibition*; Dubai, UAE, 2016; DOI: 10.2118/181510-MS. (b) Brule, M. R. Big Data in Exploration and Production: Real-Time Adaptive Analytics and Data-Flow Architecture. *SPE Digital Energy Conference*; The Woodlands, Texas, USA, 2013; DOI: 10.2118/165721-MS. (c) Gross, K. C.; Vaidyanathan, K.; Valiollahzadeh, S. Advanced Pattern Recognition for Optimal Bandwidth and Power Utilization for Wireless Intelligent Motes for IoT Applications. *17th International Conference on Wireless Networks*; Las Vegas, Nevada, USA, 2016. (d) Andrianova, A.; Simonov, M.; Perets, D.; Marguit, A.; Serebryakova, D.; Bogdanov, Y.; Budenny, S.; Volkov, N.; Tsanda, A.; Buckharest, A. Application of Machine Learning for Oilfield Data Quality Improvement. *SPE Russian Petroleum Technology Conference*; Moscow, Russia, 2018; DOI: 10.2118/191601-1RPTC-MS. (e) Elichev, V.; Muñoz, E. Case Study of Digital Oilfield Implementation in the Southern North Sea. *SPE Russian Petroleum Technology Conference*; Moscow, Russian Federation, 2018; DOI: 10.2118/191596-1RPTC-MS. (f) Holdaway, K. R. Fundamentals of Soft Computing. In *Harness Oil and Gas Big Data with Analytics: Optimize Exploration and Production with Data-Driven Models*, 1st ed.; Wiley and SAS Business Series: Cary, North Carolina, USA, 2014; pp 1−31. (g) Saputelli, L. Technology Focus: Petroleum Data Analytics. *J. Pet. Technol.* 2016, 68 (10), 66.

(4) (a) Noshi, C. I. A Brief Survey of Text Mining Applications for the Oil and Gas Industry. *International Petroleum Technology Conference 2019*, DOI: 10.2523/19382-MS. (b) Noshi, C. I.; Assem, A. I.; Schubert, J. J. The Role of Big Data Analytics in Exploration and Production: A Review of Benefits and Applications. *SPE International Heavy Oil Conference and Exhibition*; Kuwait City, Kuwait, 2018; DOI: 10.2118/193776-MS. (c) Noshi, C. I.; Schubert, J. J. The Role of Machine Learning in Drilling Operations; A Review. *SPE/AAPG Eastern Regional Meeting*; Pittsburgh, PA, USA, 2018; DOI: 10.2118/191823-18ERM-MS. (d) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature 2015*, *521*, 436−444. (e) Alvarado, V.; Ranson, A.; Hernandez, K.; Manrique, E.; Matheus, J.; Liscano, T.; Prosperi, N. Selection of EOR/IOR Opportunities Based on Machine Learning. *European Petroleum Conference; Aberdeen, Scotland*, U.K., 2002; DOI: 10.2118/78332-MS. (f) Tarrahi, M.; Jafarpour, B.; Ghassami, A. Integration of Microseismic Monitoring Data into Coupled Flow and Geomechanical Models with Ensemble Kalman Filter. *Water Resour. Res.* 2015, *51* (7), 5177−5197.

(5) (a) Brownlee, J. Basic Concepts in Machine Learning. *machinelearningmastery.com/a-tour-of-machine-learning-algorithms/ (accessed 2021-02-17). (b) Brownlee, J. A Tour of Machine Learning Algorithms. *machinelearningmastery.com, August 12, 2019; https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/ (accessed 2012-02-17).

(6) (a) El Ouahed, A. K.; Tiab, D.; Mazouzi, A.; Jokhio, S. A. Application of Artificial Intelligence to Characterize Naturally Fractured Reservoirs. *SPE International Improved Oil Recovery Conference in Asia Pacific*; Kuala Lumpur, Malaysia, 2003; DOI: 10.2118/84870-MS. (b) Mohaghegh, S. D. Recent Developments in Application of Artificial Intelligence in Petroleum Engineering. *J. Pet. Technol.* 2005, *57*, 86−91. (c) Al-Thuwaini, J.; Zangl, G.; Phelps, R. E. Innovative Approach to Assist History Matching Using Artificial Intelligence. *Intelligent Energy Conference and Exhibition; Amsterdam, The Netherlands*, 2006; DOI: 10.2118/99882-MS. (d) Piovesan, C. M.; Kozman, J. Cross-Industry Innovations in Artificial Intelligence. *SPE Digital Energy Conference and Exhibition*; The Woodlands, Texas, USA, 2011; DOI: 10.2118/143873-MS. (e) Anifowose, F. A.; Abdulraheem, A.; Al-Shuhail, A.; Schmitt, D. P. Improved Permeability Prediction From Seismic and Log Data using Artificial Intelligence Techniques. *SPE Middle East Oil and Gas Show and Conference*; Manama, Bahrain, 2013; DOI: 10.2118/164465-MS. (f) Shahkarami, A.; Mohaghegh, S. D.; Gholami, V.; Haghhighat, S. A. Artificial Intelligence (AI) Assisted History Matching. *SPE Western North American and Rocky Mountain Joint Meeting*; Denver, Colorado, USA, 2014; DOI: 10.2118/169507-MS.

(7) (a) Vallabhaneni, S.; Saraf, R.; Priyadarshy, S. Machine-Learning-Based Petrophysical Property Modeling. *SPE Europec featured at 81st EAGE Conference and Exhibition*; London, England, U.K., 2019; DOI: 10.2118/195436-MS. (b) Al-Mudhafar, W. J. Integrating Well Log Interpretations for Lithofacies Classification and Permeability Modeling Through Advanced Machine Learning Algorithms. *J. Pet. Explor Prod. Technol.* 2017, *7*, 1023−1033. (c) Belozzerov, B.; Bukhanov, N.; Egorov, D.; Zakirov, A.; Osmonalieva, O.; Golitsyna, M.; Reshytko, A.; Semenikhin, A.; Shindin, E.; Lipets, V. Automatic Well Log Analysis Across Priobskoe Field Using Machine Learning Methods. *SPE Russian Petroleum Technology Conference*; Moscow, Russia, 2018; DOI: 10.2118/191604-1RPTC-MS. (d) Dunham, M. W.; Malcolm, A.; Welford, J. K. Improved well log classification using semi-supervised algorithms. *SEG International Exposition and Annual Meeting*; San Antonio, Texas, USA, 2019; (e) Shalaby, M. R.; Malik, O. A.; Lai, D.; Jumat, N.; Islam, M. A. Thermal Maturity and TOC Prediction Using Machine Learning Techniques: Case Study From the Cretaceous−Paleocene source rock, Taranki Basin, New Zealand. *J. Pet. Explor Prod. Technol.* 2020, *10*, 2175−2193. (f) Bittar, M.; Wang, S.; Chen, J.; Wu, X. Reinforced learning technique for multi-well logs depth matching
yield better reservoir delineation. SEG International Exposition and Annual Meeting (Virtual), 2020;

(8) (a) Tariq, Z.; Abdulraheem, A.; Mahmoud, M.; Ahmed, A. A Rigorous Data-Driven Approach to Predict Poisson’s Ratio of Carbonate Rocks Using a Functional Network. Petrophysics. 2018, 59, 761–777. (b) Tariq, Z.; Mahmoud, M.; Abdulraheem, A. Real-time prognosis of flowing bottom-hole pressure in a vertical well for a multiphase flow using computational intelligence techniques. J. Pet. Explor Prod. Technol. 2020, 10, 1411–1428. (c) Shadravan, A.; Tarrahi, M.; Amani, M. Intelligent Tool to Design Fracturing, Drilling, Spacing and Cement Slurry Fluids Using Machine Learning Algorithms. SPE Kuwait Oil & Gas Show and Conference; Mishref, Kuwait, 2015; DOI: 10.2118/175338-MS. (d) Elkattany, S.; Tariq, Z.; Mahmoud, M.; Real time prediction of drilling fluid rheological properties using Artificial Neural Networks visible mathematical model (white box). J. Pet. Sci. Eng. 2016, 146, 1202–1210. (e) Bowie, B. Machine Learning Applied to Optimize Duvernay Well Performance. SPE Canada Unconventional Resources Conference; Calgary, Alberta, Canada, 2018; DOI: 10.2118/189823-MS.

(f) Abdelgawad, K. Z.; Elzenary, M.; Elkattany, S.; Mahmoud, M.; Abdulraheem, A.; Patil, S. New Approach to Evaluate the Equivalent Circulating Density (ECD) Using Artificial Intelligence Techniques. J. Pet. Explor Prod. Technol. 2019, 9, 1569–1578.

(g) Gasser, M.; Mahmoud, O.; Ibrahim, F.; Abadir, M. Using Artificial Intelligence Techniques in Modeling and Predicting the Rheological Properties of Nano-Based Drilling Fluids. Proceedings of the ASME 40th International Virtual Conference on Offshore Mechanics and Arctic Engineering (OMAE) (Virtual), 2021; DOI: 10.1115/OMAE2021-63749. (b) Gasser, M.; Naguib, A.; Mahmoud, O. New Approach to Predict the Filtrate Invasion of Nanoparticle-Based Drilling Mud Using Artificial Neural Network. Proceedings of the ASME 41st International Conference on Offshore Mechanics and Arctic Engineering (OMAE); Hamburg, Germany, 2022. (c) Elmousalami, H. H.; Elskary, M. Drilling Stuck Pipe Classification and Mitigation in the Gulf of Suez Oil Fields Using Artificial Intelligence. J. Pet. Explor Prod. Technol. 2020, 10, 2055–2068. (d) Yu, Y.; Chen, W.; Liu, Q.; Chau, M.; Vesselinov, V.; Meehan, R. Training an Automated Directional Drilling Agent with Deep Reinforcement Learning in a Simulated Environment. SPE/IADC International Drilling Conference and Exhibition (Virtual), 2021; DOI: 10.2118/204105-MS.

(h) Alimkhanaev, R.; Samoylova, I. Application of Data Mining Tools for Analysis and Prediction of Hydraulic Fracturing Efficiency for the BV8 Reservoir of the Povk Oil Field. SPE Russian Oil and Gas Exploration & Production Technical Conference and Exhibition; Moscow, Russia, 2014; DOI: 10.2118/171332-MS. (b) Anderson, R. N.; Xie, B.; Wu, L.; Kressner, A. A.; Frantz, J. H.; Ockree, M. A.; Brown, K. G.; Carragher, P.; McLane, M. A. Using Machine Learning to Identify the Highest Wet Gas Potential Mix of Hydraulic Fracture Classes and Technology Improvements in the Marcellus Shale. Unconventional Resources Technology Conference; San Antonio, Texas, USA, 2016; DOI: 10.15530/urtec-2016-2430481.

(i) Nande, S. Application of Machine Learning for Closure Pressure Determination. SPE Annual Technical Conference and Exhibition; Dallas, Texas, USA, 2018; DOI: 10.2118/194042-STU.

(d) Makhtoon, I.; Koroteev, D.; Bumaev, E. Gradient Boosting to Boost the Efficiency of Hydraulic Fracturing. J. Pet. Explor Prod. Technol. 2019, 9, 1919–1925.

(e) Desouky, M.; Tariq, Z.; Aljawad, M. S.; Alhouri, H.; Mahmoud, M.; AlShehri, D. Data-Driven Acid Fracture Conductivity Correlations Honoring Different Mineralogy and Etching Patterns. ACS Omega. 2020, 5 (27), 16919–16931. (f) Tariq, Z.; Aljawad, M. S.; Murtaza, M.; Mahmoud, M.; Al-Shehri, D.; Abdulraheem, A. A Data-Driven Approach to Predict the Breakdown Pressure of the Tight and Unconventional Formation. SPE Annual Technical Conference and Exhibition; Dubai, UAE, 2021; DOI: 10.2118/206136-MS.

(g) Gholhoseyari Gorigaei, R.; Songolzadeh, R.; Torkaman, M.; Safari, M.; Zargar, G. A Novel PSO-LSSVM Model for Predicting Liquid Rate of Two Phase Flow Through Wellhead Chokes. J. Nat. Gas Sci. Eng. 2015, 24, 228–237. (b) Seidi, S.; Sayahi, T. A New Correlation for Prediction of Sub-Critical Two-Phase Flow Pressure Drop Through Large-Sized Wellhead Chokes. J. Nat. Gas Sci. Eng. 2015, 26, 264–278. (c) Ghorbani, H.; Moghadasi, J.; Wood, D. A. Prediction of Gas Flow Rates from Gas Condensate Reservoirs Through Wellhead Chokes Using a Firefly Optimization Algorithm. J. Nat. Gas Sci. Eng. 2017, 45, 256–271. (d) Li, J.; Jichen, Z.; Hanqiao, J.; Wenchao, F.; Hongxia, L.; Wu, X. A New Early Warning Model with Vector Machine for Abnormal Production Prediction. SPE Annual Caspian Technical Conference and Exhibition; Astana, Kazakhstan, 2014; DOI: 10.2118/172307-MS.

(e) Cao, Q.; Banerjee, R.; Gupta, S.; Li, J.; Zhou, W.; Jayachandra, B. Data Driven Production Forecasting Using Machine Learning. SPE Argentina Exploration and Production of Unconventional Resources Symposium; Buenos Aires, Argentina, 2016; DOI: 10.2118/189848-MS.

(f) Amr, S.; El Ashhab, H.; El-Saban, M.; Schietinger, P.; Caile, C.; Kaheel, A.; Rodriguez, L. A Large-Scale Study for a Multi-Basin Machine Learning Model Predicting Horizontal Well Production. SPE Annual Technical Conference and Exhibition; Dallas, Texas, USA, 2018; DOI: 10.2118/191538-MS.

(g) Oloso, M. A.; Hassan, M. G.; Bader-El-Den, M. B.; Buick, J. M. Ensemble SVM for Characterisation of Crude Oil Viscosity. J. Pet. Explor Prod. Technol. 2018, 8, 531–546. (c) Pankaj, P.; Geetan, S.; MacDonald, R.; Shukla, P.; Sharma, A.; Menasria, S.; Xue, H.; Judd, T. Application of Data Science and Machine Learning for Well Completion Optimization. Offshore Technology Conference; Houston, Texas, USA, 2018; DOI: 10.4043/28632-MS.

(h) Tariq, Z.; Mahmoud, M.; Abdulraheem, A.; Al-Shehri, D.; Khan, M. R.; Janjua, A. N. An Intelligent Solution to Forecast Pressure Drop in a Vertical Well Having Multiphase Flow Using Functional Network Technique. PAPG/SPE Pakistan Section Annual Technical Conference and Exhibition; Islamabad, Pakistan, 2018; DOI: 10.2118/195656-MS.

(i) Noshi, C. I.; Eissa, M. R.; Abdalla, R. M. An Intelligent Data Driven Approach for Production Prediction. Offshore Technology Conference; Houston, Texas, 2019; DOI: 10.4043/29243-MS.

(j) Alzahabi, A.; Trindade, A. A.; Kamel, A.; Harouaka, A. Optimizing Initial Oil Production of Horizontal Wolfcamp Wells Utilizing Data Analytics. J. Petrol. Explor. Prod. Technol. 2020, 10, 2357–2371.

(k) Tariq, Z.; Mahmoud, M.; Abdulraheem, A. An intelligent data-driven model for Dean–Stark water saturation prediction in carbonate rocks. Neural Comput. Appl. 2020, 32, 11919–11935. (d) Tariq, Z.; Mahmoud, M.; Abdulraheem, A. Machine Learning-Based Improved Pressure—Volume—Temperature Correlations for Black Oil Reservoirs. J. Energy Resour. Technol. 2021, 143, 113003.

(l) Hsuing, M.; Hsiau, J.; Putcha, V. B. Detecting Failures and Optimizing Performance in Artificial Lift Using Machine Learning Models. SPE Western Regional Meeting; Garden Grove, California, USA, 2015; DOI: 10.2118/174015-MS.

(m) Pennel, M.; Huang, J.; Putcha, V. B. Detecting Failures and Optimizing Performance in Artificial Lift Using Machine Learning Models. SPE Western Regional Meeting; Garden Grove, California, USA, 2018; DOI: 10.2118/190090-MS.

(n) Guo, D.; Raghavendra, C. S.; Yao, K.-T.; Harding, M.; Anvar, A.; Patel, A. Data Driven Approach to Failure Prediction for Electrical Submersible Pump Systems. SPE Western Regional Meeting; Garden Grove, California, USA, 2015; DOI: 10.2118/174062-MS.

(o) Snee, J. Predicting ESP Lifespan with Machine Learning. SPE/AAPG/SEG Unconventional Resources Technology Conference; Austin, Texas, USA, 2017; DOI: 10.15530/urtec-2017-2669988.

(p) Chong, J.; Kelly, D.; Agrawal, S.; Nguyen, N.; Monzon, M. Reinforcement Learning Control Scheme for Electrical Submersible Pumps. SPE Gulf Coast Section Electric Submersible Pumps Symposium (Virtual); The Woodlands, Texas, USA, 2021; DOI: 10.2118/204508-MS.

(q) Kellogg, R. P.; Chessum, W.; Kwong, R. Machine Learning Application for Wellbore Damage Removal in the Wilmington Field. SPE Western Regional Meeting; Garden Grove, California, USA, 2018; DOI: 10.2118/190307-MS.

(r) Arigbe, O. D.; Oyeneyin, M. B.; Arana,
