AMISH, M. and ETTA-AGBOR, E. 2023. Genetic programming application in predicting fluid loss severity. *Results in engineering* [online], 20, article number 101464. Available from: https://doi.org/10.1016/j.rineng.2023.101464

Genetic programming application in predicting fluid loss severity.

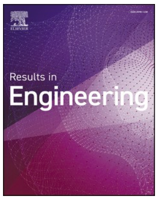
AMISH, M. and ETTA-AGBOR, E.

2023

Contents lists available at ScienceDirect

Results in Engineering

journal homepage: www.sciencedirect.com/journal/results-in-engineering



Genetic programming application in predicting fluid loss severity

Mohamed Amish [*], Eta Etta-Agbor

School of Engineering, Robert Gordon University, Garthdee Road, Aberdeen, AB10 7GJ, UK

ARTICLE INFO

Keywords:
Lost circulation
Machine learning
Multigene genetic algorithms
Drilling. non-productive time

ABSTRACT

Numerous wells worldwide encounter significant, costly, and time-consuming lost circulation issues during drilling or while deploying tubulars across naturally fractured or induced fractured formations. This can potentially lead to formation damage, wellbore instability, and even blowouts. Effectively addressing this problem and restoring fluid circulation becomes crucial to curbing non-productive time and overall operational expenses. Although numerous methods have been introduced, a universally accepted industry solution for predicting lost circulation remains absent due to the complex interplay of various factors influencing its severity. Anticipating the onset of circulation loss is imperative to mitigate its impacts, minimise costs, and reduce risks to personnel and the environment.

In this study, an innovative machine learning approach employing multigene genetic algorithms is utilised to analyse a dataset of 16,970 drilling datasets from 61 wells within the Marun oil field, located in Iran, where severe loss of circulation occurred. Geological characteristics, operational drilling parameters, and the properties of the drilling fluid were all considered. The dataset encompasses 19 parameters, of which seven are chosen as inputs for predicting lost circulation incidents. These inputs are then employed to construct a predictive model, employing an 85:15 training-to-test data ratio. To assess the model's performance, unseen datasets are utilised.

The novelty of this study lies in the proposed model's consideration of a concise set of relevant input parameters, particularly real-time surface drilling parameters that are easily accessible for every well. The model attains a remarkable level of prediction accuracy for fluid loss, as indicated by various performance indices. The results indicate a mean absolute error of 1.33, a root mean square error of 2.58, and a coefficient of determination of 0.968. The suggested prediction model is optimised not only for data reduction but also for universal prediction and compatibility with other existing platforms. Moreover, it aids drilling engineers in implementing suitable mitigation strategies and designing optimal values for key operational surface parameters, both prior to and during drilling operations.

1. Introduction

Drilling a well is a complex process fraught with numerous challenges, and among them is the issue of losing control over the flow of drilling fluid into the formation. The occurrence of lost circulation impacts around 20–25% of all drilled wells worldwide and even reaches up to 40% in North America [1,2]. The mitigation and prevention of lost circulation incur substantial costs, with estimates suggesting that the industry expends over 2 billion USD annually to tackle this concern [3]. Lost circulation is defined as the uncontrolled migration of wellbore drilling fluids into the formation, resulting in either partial or complete loss of drilling fluid [4]. A significant contributor to challenges faced during drilling operations is wellbore instability caused by lost circulation [5]. Dealing with this issue and reinstating fluid circulation demands considerable effort and time, thereby escalating non-productive time (NPT) and overall drilling costs [2]. Research by Ref. [6] uncovered that lost circulation accounted for 12% of NPT in the Gulf of Mexico region over a decade, with wellbore instabilities and kicks contributing to 18% of NPT. Furthermore, the intrusion of drilling fluid into the reservoir formation can inflict damage and curtail productivity [7]. In such scenarios, drilling costs can surge from USD 70 to USD 100 per foot, underscoring the imperative nature of effectively addressing lost circulation [6]. Fluid loss during drilling can be categorised by its severity, encompassing seepage loss, partial loss, and severe or total loss (Table 1), as well as the base fluid utilised in the drilling process, as outlined by Ref. [8]. These losses are most prone to occur in carbonate formations such as dolomite or limestone with characteristics like caverns, vugs, and fractures, as well as in formations

[*] Corresponding author.
E-mail address: m.amish@rgu.ac.uk (M. Amish).

https://doi.org/10.1016/j.rineng.2023.101464
Received 10 May 2023; Received in revised form 31 August 2023; Accepted 26 September 2023
Available online 2 October 2023

Table 1
Classification of fluid loss based on drilling fluid type (Adapted from [8]).

Fluid loss class	Water Based Muds (WBMs)		Oil Based Muds (OBMs)	
1. Seepage losses	<25 bbl/hr	4 m^3/h	<10 bbl/hr	1.6 m^3/h
2. Moderate losses	25 - 100 bbl/hr	4–16 m^3/h	10 - 30 bbl/hr	1.6–4.8 m^3/h
3. Severe losses	>100 bbl/hr	>16 m^3/h	>30 bbl/hr	>4.8 m^3/h
4. Total losses	no mud returns to the surface			

with induced fractures and high permeability. Particular concern arises in zones marked by a high incidence of severe, interconnected vugs, cavernous fractures, or total losses [9].

Numerous methods have been introduced, but the absence of an industry-wide solution is attributed to the intricate nature of drilling and the vast array of fluid losses that vary based on the formation being drilled. Despite significant emphasis on the development of loss circulation materials (LCM) to counter fluid losses, the efficacy of these materials is not always assured due to the uncertainties and unknown factors in subsurface conditions. The act of predicting and detecting fluid losses proves more efficacious than attempting to rectify the issue after its occurrence [10]. The prevalent practice within the industry to manage instances of lost circulation involves employing conventional methods, notably the addition of LCM (including fibrous, granular, and flaky materials) or deploying high-viscosity pills combined with LCM to handle seepage and partial losses. For higher-severity situations like severe or complete losses, alternative remedies are formulated and implemented. These can encompass the use of cement [11,12] and nanocomposite gels [13], primarily designed to seal existing fractures and thwart the occurrence of new fractures [14]. The urgency to minimise risks to rig personnel, the environment, and the financial burdens stemming from fluid loss is paramount.

1.1. *Lost circulation prediction using artificial intelligence and machine learning*

Artificial Intelligence (AI) and Machine Learning (ML) applications within drilling operations are now extensively employed in the oil and gas sector due to their adaptability in classification, selection, prediction, and optimisation tasks. These technologies have been harnessed to forecast lost circulation across various fields, modeling complex relationships and yielding time and cost savings [15]. Notably, the key distinctions between existing models pertain to the model type utilised, the chosen input parameters, and the precision of lost circulation prediction. The datasets used in prior studies can be categorised into three groups: those tied to drilling operations, formation characteristics, and drilling fluid properties. Instances of drilling operation data encompass variables like depth, drilling time, hole size, weight on the bit, pump rate, and circulation pressure. Formation properties include lithology, pore pressure, and fracture pressure, while drilling fluid is primarily represented by features such as viscosity, shear stress at shear rates of 600 and 300 rpm, gel strength, and solids content.

Although various studies have leveraged AI and ML to anticipate lost circulation, further research supported by technology is deemed essential to enhance these predictions, according to Ref. [1]. They proposed employing the Multi-Gene Genetic Programming (MGGP) approach, which stands as a data-driven methodology capable of eliminating errors and capturing non-linear interconnections between variables. While this technique has been applied across diverse disciplines and applications [16], it remains untapped in the domain of lost circulation prediction. MGGP holds the capacity to discern variables that exert significant influence on the dependent variable, along with their manner of impact. It proves to be a versatile tool for predictions and forecasts. However, before utilising these parameters for predictions, it is imperative to evaluate the strengths and directions of their effects on lost circulation [17–19]. The anticipation of lost circulation incidents is progressively gaining importance within drilling management. This capability empowers engineers to curtail fluid loss and implement fitting measures, thereby yielding improved economics and diminished Non-Productive Time (NPT) [20].

1.2. *Study objectives*

The aim of this study is to develop a state-of-the-art genetic programming model that exhibits high precision in forecasting early fluid loss occurrences. This model relies on a limited number of readily available input parameters from each well. Additionally, we have implemented several analysis and feature selection techniques to avert data overfitting, decrease computation time, and enhance prediction accuracy. The model's utility lies in its potential to avert incidents jeopardising well integrity, which can result in loss of life, environmental harm, and escalated operational expenses. In this context, we focus on the Marun oil field as a case study, renowned for its severe loss of circulation due to fractures.

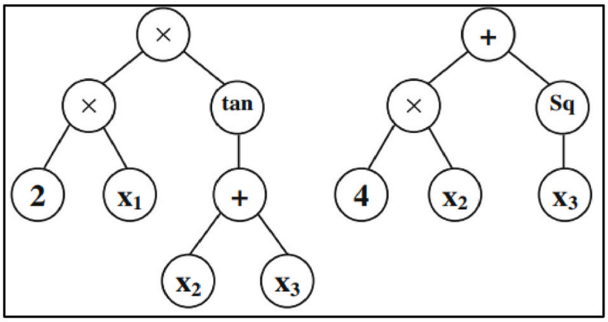
The structure of this paper encompasses five sections, outlined as follows: Section 1 introduces the concept of loss of circulation and outlines the prerequisites of a prediction model aimed at curbing its adverse effects. Section 2 provides an overview of the methodology employed. Moving forward, Section 3 delves into a comprehensive explanation of data description and preparation. Subsequently, Section 4 will lay out the model analysis and ensuing discussions. Lastly, in Section 5, the study concludes with some recommended insights.

2. Methodology overview

The study aims to employ Multigene Genetic Programming (MGGP) as the chosen machine learning algorithm for predictive purposes. MGGP has frequently demonstrated superior performance compared to other machine learning techniques like ANN, SVM, and several others in terms of predictability and model applicability [21]. This selection is grounded in a comprehensive literature review of past studies, encompassing reports, conference proceedings, and journal articles. These sources underscore the potential of MGGP, as seen in recent works like [22]; in providing mathematical models for intricate and non-linear parameters linked to loss circulation scenarios. The MGGP algorithm, devised by Ref. [21]; leverages a gene population to construct regression analysis models. Transformation-learn offers a range of machine learning algorithms, including Genetic Programming (GP), which serves as a foundation for implementing MGGP. As described by Ref. [23]; the MGGP algorithm follows the ensuing steps:

- Initiate by setting initial parameters such as function and terminal sets, generation count, population size, and maximum gene depth.
- Randomly generate the initial gene population.
- Utilise the least squares method to formulate models by amalgamating a set of genes.
- Assess model performance through the fitness function.
- Employ genetic operations to generate a new gene population.
- Evaluate model performance by comparing it with a benchmark and using a termination criterion. If unmet, return to step 5. If met, designate the evolved model with the best performance as the final solution [23].

Fig. 1 below illustrates a prototypical MGGP model. This model encompasses three input variables: x_1, x_2, and x_3. The individual genes d_0, d_1, d_2, …., d_m collaborate to form the overall model, which predicts the output variable ($y = d_0 + d_1(2x_1\tan(x_2+x_3) + d_2(4x_2 + x_3^2) + \ldots \ldots + d_m(\text{tree M})$) [21,24]. This construction can encompass simple mathematical functions like addition, multiplication, sine, or cosine, as well as more intricate mathematical expressions such as logarithmic or polynomial functions. Every prediction of the output variable y within multigene symbolic regression-based GP is derived from the weighted



$$y = d_0 + d_1(2x_1\tan(x_2+x_3)) + d_2(4x_2 + x_3^2) + \ldots\ldots + d_m(\text{tree } M))$$

Fig. 1. MGGP model.

output of each gene in the multigene individual, supplemented by a bias term. Each tree within the individual constitutes a function of one or more of the N input variables x_1, … x_n, where d_0 represents the bias (offset) term, d_1, …., d_m denotes gene weights, and M signifies the number of genes (i.e., trees) comprising the current individual. The weights (i.e., regression coefficients) for each multigene individual are determined automatically through a least squares procedure. A pseudolinear multigene model of predictor output y, with inputs x_1 to x_6; and calculates weights d_0, d_1, and d_2, automatically using least squares. In essence, a typical MGGP model is represented as a mathematical expression that incorporates multiple genes, merging into a complex function with the capability to accurately predict the target variable [23].

The GP algorithm entails considerable computational intensity, which makes it slow and resource intensive. However, the challenge was addressed by utilising Scikit-Learn. Scikit-learn stands as a Python library extensively employed in machine learning, offering a diverse array of supervised and unsupervised learning algorithms, along with tools for model evaluation, selection, and preprocessing. Its capabilities encompass regression, classification, clustering, model selection, and preprocessing [25]. Scikit-learn was harnessed to establish models and make predictions on pre-processed data, aiming to heighten accuracy and diminish computational burden. To utilise Scikit-learn's algorithmic functionality for MGGP, the following steps can be undertaken: Begin by defining the problem at hand, be it classification or regression. Proceed to prepare the data by segmenting it into input and output variables, followed by scaling the input variables. Subsequently, opt for the Genetic Programming algorithm and configure hyperparameters, including the mutation rate and population size. Afterward, train the model through the fit function and gauge its performance using metrics like accuracy or mean squared error with the predict function. If the model's performance falls short of expectations, refinement can be achieved by adjusting hyperparameters or the fitness function. Finally, once satisfied with the model's performance, deploy it to make predictions on fresh data. Scikit-learn furnishes potent algorithms and tools commonly utilised in machine learning for MGGP.

A comprehensive statistical depiction of the utilised datasets is formulated to capture the data's diversity. The statistical description includes parameters such as minimum, maximum, mean, range, mode, variation, kurtosis, skewness, and standard deviation. Data analysis aims to infuse significance into raw data, thereby eliciting meaningful insights. While this process can be demanding, its importance cannot be overstated. The statistical particulars of these parameters are presented in Table 2, which offers insight into their minimum and maximum values alongside corresponding units of measurement. Moreover, statistical measures like mean, standard deviation, kurtosis, variance, etc. are employed to delineate both input and output variables. This table furnishes a comprehensive overview of key data characteristics, rendering the previously intricate field report data in a lucid and intelligible format. The variance column gauges the tendency of a variable to deviate from its mean value (indicative of the average difference from the mean). The standard deviation, being the square root of the variance, provides a precise measurement of dispersion. The mean column represents the data's average, while the median column depicts its midpoint. Skewness values elucidate the distribution's imbalance; negative or positive skewness suggests an uneven distribution, leaning either left or right. The kurtosis values in the table denote the data distribution's flatness, indicating whether it possesses heavy or light tails. High kurtosis implies heavy tails or outliers, whereas low kurtosis points to lighter tails or their absence. With both positive and negative kurtosis values exhibited in the table, the data showcases a non-uniform distribution. Given machine learning algorithms' affinity for normally distributed input data [14], transformations might be necessary to rectify imbalances. In this study, a non-linear algorithm was leveraged to address uneven data distribution.

To heighten system accuracy, the data was normalized. This was undertaken to prevent biases stemming from variable magnitudes. Each variable was linearly scaled to the same range, which accelerated training speeds and slashed overall computational durations for each model. We applied the formula by Ref. [26] to normalise data within the range of −1 to 1. This was done by dividing the difference between the maximum and minimum values of each variable (xi) by their sum.

This formula is expressed mathematically in equation (1).

Formulas for normalisation of input and output data, RMSE, MAE, R^2

Data normalisation [26], expressed in equation (1)

$$x_i^n = 2 \times \frac{x_i - x_{min}}{x_{max} - x_{min}} - 1 \qquad 1$$

Table 2

Statistical summary of the data description used in predicting lost circulation.

PARAMETER	UNIT	MINIMUM	MAXIMUM	MEDIAN	MEAN	STD. DEV.	VARIANCE	KURTOSIS	SKEWNESS
Depth	ft	17	5662	2927	2818.8	927.42	860,111.66	0.223	-0.506
Pore pressure	psi	7.361	3398.4	1356.34	1643.5	814.10	662,759.24	-0.957	0.375
Fracture pressure	psi	11.56	4472.98	2607.39	2406.5	920.43	847,182.18	-0.583	-0.382
Mud pressure	psi	7.74	4922.54	1588.92	1854.6	892.34	796,262.17	-0.268	0.538
Hole size	inches	4.125	26	12.25	12.3	4.93	24.35	-0.392	0.316
ROP	ft/hr	10.5	88.86	6.91	9.69	18.43	330.32	20.638	2.541
WOB	kg	1000	70,000	20,000	20874	9418.59	88,709,809	1.915	1.045
Pump flow rate	m^3/hr	80	1000	530	548.41	277.44	76,974.91	-1.367	0.197
Pump pressure	psi	50	2950	2225	1969.6	838.63	703,306.51	-0.925	-0.666
Viscosity (MFVIS)	cp	27	100	44	47	12.12	146.89	-0.120	0.656
Solid % (RETSOLID)	%	0	61	18	22.85	16.82	282.94	-1.349	0.379
FAN600 (θ_{600})	lb/(100 ft^2	3	293	49	78.46	62.43	3897.07	-0.206	0.969
FAN300 (θ_{300})	lb/(100 ft^2	2	163	30	46.17	33.69	1134.94	-0.238	0.940
Gel Strength	lb/(100 ft^2	1	49	5	5.61	3.59	12.88	22.781	3.133
RPM	rpm	20	394	155	138.05	46.66	2177.67	-0.660	-0.420
MUDLOSSES	bbl/hr	0	999	25	97.74	160.99	25,919.34	7.779	2.615

*999 - corresponds to total loss (out of range of device measurement).

where x_i^n is the variable to be normalized.

x_i is the actual value of a particular variable.

x_{min} is the minimum value for each variable.

x_{max} is the maximum value for each variable.

Mean square Error (MSE) expressed in equation (2)

$$\mathrm{MAE} = 1 \Big/ n \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{2}$$

where: n is the total number of samples in the dataset.

y_i is the i-th true value in the dataset.

$\hat{y}_i$ is the i-th predicted value in the dataset.

Σ is the summation operator.

The lower the value of the MSE, the better the model's prediction [24].

Root mean square error (RMSE) expressed in equation (3)

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (yi - \hat{y}_i)^2} \tag{3}$$

Coefficient of determination (R^2) expressed in equation (4)

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (f(x_i) - y_i)^2}{\sum_{i=1}^{n} f(x_i)^2 - \frac{\sum_{i=1}^{n} (y_i)^2}{n}} \tag{4}$$

where: n = number of observations in the dataset.

$f(x_i)$ = predicted value for the ith observation in the dataset

$\hat{y}_i$ = actual value for the ith observation in the dataset.

Σ = summation denotes the sum from i = 1.

The drilling data employed in this study was sourced from Ref. [20]. A substantial dataset was amassed from 61 drilled wells, extracted from daily drilling reports, with a focus on the most impactful parameters governing the severity of lost circulation. Geological characteristics, operational drilling parameters, and drilling fluid properties were all considered. Following data collection, a preprocessing stage was undertaken, involving the normalisation and scaling of variables to a consistent range. The aim was to discern any correlations between distinct input features within the dataset. Subsequently, the data was partitioned into training, testing, and evaluation sets.

A predictive model was constructed, adhering to an 85:15 training-to-test data ratio. An unseen dataset was then employed to assess the performance of the developed models, employing an array of performance metrics. Fig. 2 furnishes a visual representation of the methodology flowchart adopted for this study.

3. Data description and preparation

3.1. Marun field

The primary objective of this study is to enhance the prediction of fluid loss during drilling operations through the utilisation of data sourced from the Marun oilfield. Situated in the western south of Iran, the Marun oil field is of significant importance, ranking as the second largest within Iran and among the six largest onshore oilfields globally. It was initially discovered in 1963 and spans approximately 67 km in length and 7 km in width [27]. The Marun anticline aligns parallel to the Ahvaz and Aghajari structures, contributing to its complex geological characteristics. To manage its size, the oilfield has been segmented into eight sections, as illustrated in Fig. 3.

Within this field, there are two oil reservoirs (Asmari and Bangestan) and a gas reservoir (Khami). The Asmari reservoir incorporates a mixture of carbonate, shale, and sandstone lithology, while the Bangestan and Khami reservoirs primarily consist of carbonate and shale lithology. The Asmari formation, dating back to the Oligocene-early Miocene epoch, serves as the primary source of hydrocarbons within the Marun oilfield [29]. Due to tectonic activity, high-fracture zones are prevalent in this field [30], leading to considerable costs associated with mud loss control. It is within these zones that a substantial volume of data has been collected to facilitate the prediction of lost circulation. The Asmari formation, underlying the Gachsaran formation, is stratified into several sublayers (G1 to G7, cap rock), followed by the Mishan formation, as illustrated in Fig. 4.

A time breakdown analysis of over 200 development wells drilled within the Marun oil field is depicted in Fig. 5 [31]. This visual representation illustrates that approximately 10% of the rig's time was allocated to hole conditioning following instances of wellbore instability and lost circulation issues. Furthermore, about 3% of the rig's time was dedicated to fishing operations, a situation commonly arising from a stuck pipe. Another significant contributor to the drilling rig's schedule, accounting for 4% of drilling time, was equipment failure and the subsequent need for repairs. As a result, even a 1% reduction in nonproductive time carries substantial implications, particularly when millions of dollars are invested in well drilling operations. Consequently, the proposed MGPP model emerges as a potent tool capable of enhancing well integrity and curtailing nonproductive time associated with fluid loss and its related challenges.

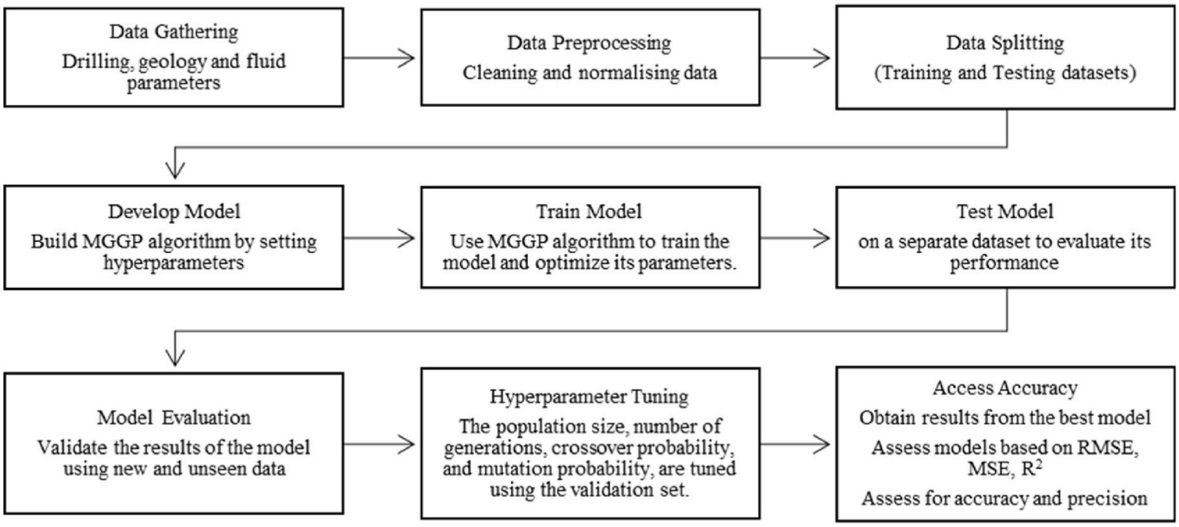

Data Gathering
Drilling, geology and fluid parameters
Data Preprocessing
Cleaning and normalising data
Data Splitting
(Training and Testing datasets)
Develop Model
Build MGGP algorithm by setting hyperparameters
Train Model
Use MGGP algorithm to train the model and optimize its parameters.
Test Model
on a separate dataset to evaluate its performance
Model Evaluation
Validate the results of the model using new and unseen data
Hyperparameter Tuning
The population size, number of generations, crossover probability, and mutation probability, are tuned using the validation set.
Access Accuracy
Obtain results from the best model
Assess models based on RMSE, MSE, R²
Assess for accuracy and precision

Fig. 2. Methodology flowchart.

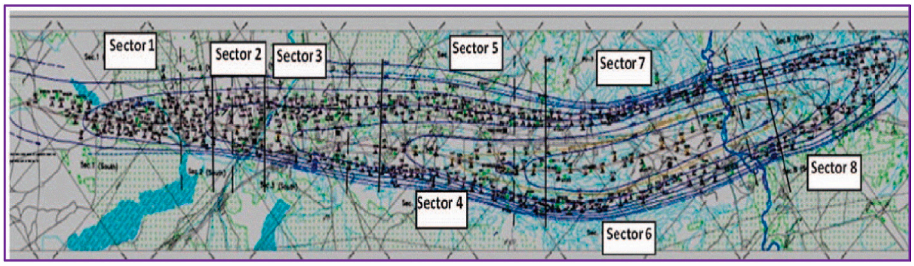
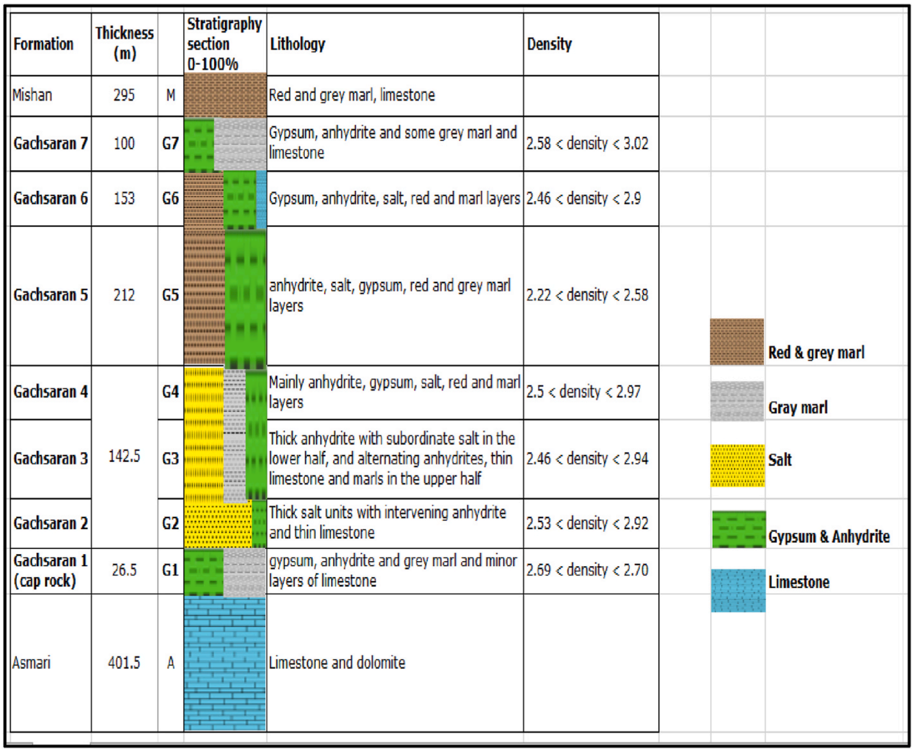


Fig. 3. Different sections of Marun oil field (Adapted from [28].



Formation	Thickness (m)		Stratigraphy section 0-100%	Lithology	Density
Mishan	295	M	🟫	Red and grey marl, limestone	
Gachsaran 7	100	G7	🟩⬜	Gypsum, anhydrite and some grey marl and limestone	2.58 < density < 3.02
Gachsaran 6	153	G6	🟫🟩🟦	Gypsum, anhydrite, salt, red and marl layers	2.46 < density < 2.9
Gachsaran 5	212	G5	🟫🟩	anhydrite, salt, gypsum, red and grey marl layers	2.22 < density < 2.58
Gachsaran 4	142.5	G4	🟨⬜🟩	Mainly anhydrite, gypsum, salt, red and marl layers	2.5 < density < 2.97
Gachsaran 3		G3	🟨⬜🟩	Thick anhydrite with subordinate salt in the lower half, and alternating anhydrites, thin limestone and marls in the upper half	2.46 < density < 2.94
Gachsaran 2		G2	🟨🟩	Thick salt units with intervening anhydrite and thin limestone	2.53 < density < 2.92
Gachsaran 1 (cap rock)	26.5	G1	🟩⬜	gypsum, anhydrite and grey marl and minor layers of limestone	2.69 < density < 2.70
Asmari	401.5	A	🟦	Limestone and dolomite	

🟫	Red & grey marl
⬜	Gray marl
🟨	Salt
🟩	Gypsum & Anhydrite
🟦	Limestone

Fig. 4. Sequence of Gachsaran lithology (Adapted from [28]).

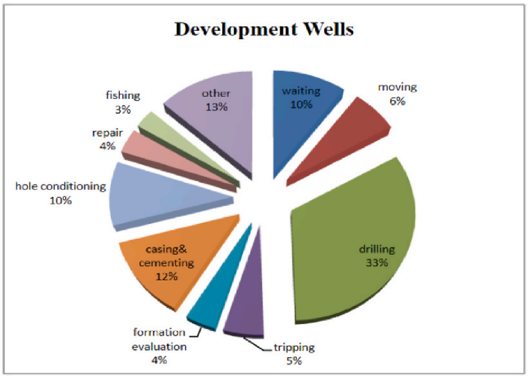


Fig. 5. Time break down diagram of drilled development wells in Marun oil field.

3.2. Data gathering

The proposed algorithm will be utilised for a classification task where the output data falls into categories. In this context, the model's objective is to learn and predict target classes based on the provided input data. Prior to data utilisation, the data must undergo preprocessing, a necessary step to prepare it for analysis. The pertinent and precise data for predicting fluid loss in the Marun field is drawn from several sources, including daily drilling reports (DDRs), end-of-well reports (EWRs), daily mud reports (DMRs), and mud logs. This amalgamated dataset, consisting of variables crucial for fluid loss prediction, is presented in Table 3. It offers a statistical summary of select essential features. The dataset comprises 20 variables, encompassing 19 inputs (11 related to drilling operations, 5 tied to drilling fluid, and 3 associated with formation parameters) as well as 1 output. The breakdown of these variables is as follows:

- Drilling Operation Parameters: Geographic coordinates (Northings and Eastings), drilling depth where fluid loss occurred (ft), pump flow rate (m^3/hr), mud circulating pressure (psi), hole size (inches), pump pressure (psi), drilling meterage (ft), drilling time (hr), weight on bit (WOB, 1000lb), and drill string rotation per minute (RPM).
- Drilling Fluid Properties: Marsh funnel viscosity (MFVIS, cp), solids content (Retort solids, %), Fan shear stress at rates of 300 rpm and 600 rpm (lb/100 ft^2), and gel strength (lb/100 ft^2).
- Formation Characteristics: Formation type, pore pressure (PP, psi), and fracture pressures (FP, psi).

Table 3
Statistical summary of the dataset for the proposed model validation.

	Hole size (in)	ROP (ft/hr)	WOB (1000lb)	Pump flow rate (gpm)	Pump pressure (psi)	Gel 10 min/Gel 10 s (100lb/ft2)	RPM	MUD LOSS (bbl/hr)
count	1794	1794	1794	1794	1794	1794	1794	1794
mean	12.312	3.126	21.551	567.870	2084.030	5.405	144.011	145.303
std	4.747	2.013	9.461	276.685	791.677	3.136	44.999	177.833
min	4.125	0.125	1	80	50	1	20	1
25%	8.375	1.714	15	300	1425	3	105	24
50%	12.250	2.771	20	543	2375	5	170	80
75%	17.5	4.031	28	860	2775	7	180	180
max	26.0	26.667	58	1000	2950	49	200	999

The output data pertains to the quantity of fluid loss (bbl/hr). The dataset employed for this research was derived from the study conducted by Ref. [32]. The data hails from daily drilling operation reports from 61 wells drilled in the Marun field, as depicted in Fig. 6.

To facilitate the application of machine learning methodologies to textual or symbolic data, such as the “type of formation” variable, a conversion into a numeric form is imperative. Several techniques exist to achieve this, including class numbering, unary encoding, and binary encoding [33]. For this study, the approach of class numbering has been adopted to translate the “type of formation” into a numeric representation, as illustrated in Table 4. Maintaining data quality sourced from the field remains a significant challenge, largely due to the inherent uncertainty in data measurements collected during drilling operations, often arising from human error or equipment malfunctions [33]. Consequently, data from these sources underwent thorough analysis and validation to identify and eliminate incorrect entries, commonly referred to as “outliers”. Such outliers can exert a considerable impact on the efficacy of machine learning, both during the training and prediction stages. To counter this concern, a comprehensive review of the data was undertaken, leading to the exclusion of data points with unusual values, including well trajectory, leak-off Test (LOT), and wellbore temperature. Following the data cleansing process, out of the initial 19,578 data points, a total of 16,970 data points from the 61 wells were identified as possessing valid data and were thus utilised for the development of the model. Conversely, 2608 data points were discarded.

Table 4
Codes generated for the lithology in the Marun’s field subsurface.

Formation type	Code
Aghajary	**1**
Mishan	2
Gachsaran 7	3
Gachsaran 6	4
Gachsaran 5	5
Gachsaran 4	6
Gachsaran 3	7
Gachsaran 2	8
Gachsaran 1 (Cap rock)	9
Asmari	10
Pabdeh	11
Gurpi	12
Ilam	13
Sarvak	14
Fars	15

Table 5
Hyperparameter settings for the MGGP algorithm.

Parameter	Range	Settings
Function set	-, +, x,/, √, In, sin, square, cos, exp, tanh	
Population size	100–150	150
Generation count	50–100	100
Parent count	50–100	100
Mutation rate/percent	0.01	0
Cross over rate/percent	0.05	0.05
Reproduction rate/percent	0.03	0.03
Selection mode	Random	Random
Gene count	120–170	170
Gene length	5	5

Normalisation of both input and output data constitutes a pivotal stride in augmenting the precision of models [26]. introduced a formula for normalising data between −1 and 1, accomplished by dividing the disparity between the maximum and minimum values of each variable (xi) by their cumulative sum. This formula is mathematically represented as equation (1). Within the Python programming environment, the MGGP algorithm was designated from the Scikit-Learn library, and a spectrum of hyperparameters was established as detailed in Table 5. Various combinations of values were experimented with to identify the optimal configuration for constructing the model. Subsequent to data normalisation for heightened accuracy, within the pool of 16,970 data points, 85% (14,426 data points) were randomly selected to establish the model and functioned as the training dataset. This subset facilitated the training of the algorithm. Concurrently, 15% (2546 data points) were earmarked for testing the model, enabling an assessment of its performance. Additionally, a novel dataset of 1794 data points was introduced to validate the MGGP model, as illustrated in Fig. 7. A comprehensive breakdown of the data distribution and outcomes is presented in Table 6.

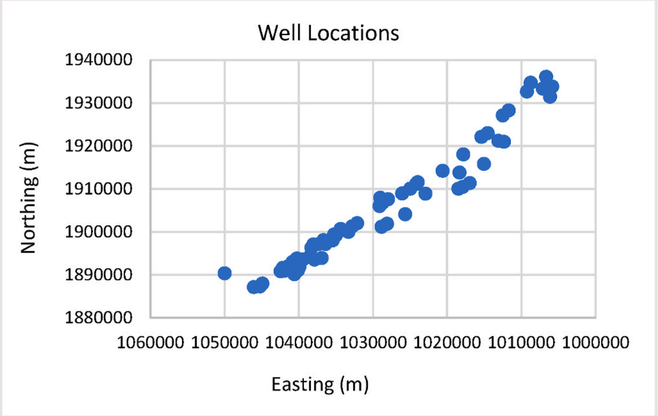


Fig. 6. The distribution of the 61 Wells in Marun field.

3.3. Data analysis and visualisation

Data analysis and visualisation play a pivotal role in comprehending the interplay between input features, such as pump pressure, RPM, and WOB and the resulting output, namely mud loss. Fig. 8 illustrates the pressure distribution (formation pore pressure and mud pressure) in relation to depth for the various formations existing within the field. A shared pattern among these formations is the elevation of pressure with increasing depth. However, at specific depths, certain abnormal pressure values have been observed, manifesting as deviations from the typical trend line associated with each formation. These atypical pressure values can be attributed to a range of factors, including tectonic

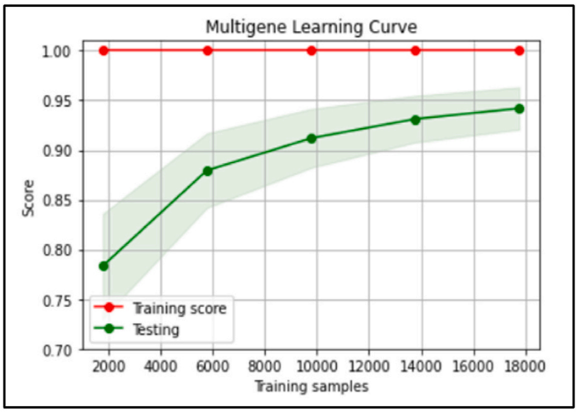


Fig. 7. Train – Test performance on sample data.

Table 6
Percentage of data splitting of 16970 data points and results.

Training (%)	16970	Testing (%)	16970	R^2
50	8485	50	8485	0.901
60	10182	40	6788	0.908
75	12728	25	4242	0.943
85	14426	15	2546	0.968

activity in the field, fluid migration, or geological anomalies.

The drilling progress (rate of penetration. ROP), represented by the depth gained during drilling, along with the corresponding WOB required to penetrate each formation, has been graphically depicted in Fig. 9 for this study. Through a comprehensive analysis of this plot, drilling engineers are empowered to optimise drilling performance. This optimisation can be achieved by refining bit design, bit hydraulics, WOB, and other pertinent drilling parameters such as pump flow rate and pump circulation pressure. The adjustments can be tailored to the distinct attributes of each formation. This strategic approach holds the potential to enhance drilling meterage and elevate the average rate of penetration across the drilling process. Consequently, it becomes feasible to accomplish drilling the entire hole section within a single bit run, thereby yielding substantial reductions in well drilling duration. Furthermore, it mitigates the need for frequent tripping operations, averting the advent of surge pressures that might precipitate induced fractures and mud losses.

Fig. 10 illustrates the historical progression of circulation and pump pressure derived from the daily drilling reports (DDRs), delineated separately for each formation. Within the plot, the pump flow rate signifies the volume of drilling fluid introduced into the wellbore, while the pump pressure reflects the pressure exerted by the drilling fluid within the wellbore. Notably, elevated solids content, escalated pump flow rates (associated with heightened annular pressure loss), or excessively swift tripping operations (resulting in surge pressures) can all contribute to elevated wellbore pressures exerted against the formation. These factors possess the capacity to elevate the mud's equivalent circulating density and instigate fractures, culminating in the occurrence of lost circulation events. This depiction holds valuable insights for drilling engineers, who can leverage this information to diligently monitor the performance of the drilling fluid system. Subsequently, they can fine-tune a spectrum of parameters, encompassing both drilling fluid properties and drilling parameters, to orchestrate efficient drilling operations. The meticulous optimisation strives to prevent the emergence of induced fractured formations, which in turn forestalls fluid loss occurrences.

3.4. Input parameters selection from available data

Building a comprehensive database for AI models is a time-consuming and demanding process. One of the primary challenges is determining the impact of various input parameters, as drilling operations often yield a plethora of them. However, employing all these parameters as input data can result in an unwieldy network that diminishes learning efficiency and speed. It is imperative to identify the optimal set of relevant and valid variables to address the issue of lost circulation. The success of the predictive model hinges on the synergy between algorithm performance and computational prowess [15]. Scikit-learn, a Python library for machine learning, presents an array of supervised and unsupervised learning algorithms, along with tools for model selection, evaluation, and preprocessing. Designed to be user-friendly and efficient, Scikit-learn offers a unified interface that simplifies the transition between diverse algorithms and models, thereby enhancing accuracy while curbing computational time [25]. Though Scikit-learn accommodates various data formats, it particularly excels at handling numpy arrays or Pandas' data frames. In predicting solutions for lost circulation, Pandas—a Python data manipulation and analysis library—will be enlisted. Pandas encompasses a rich suite of functions and methods for data cleaning, manipulation, and visualisation, rendering it an encompassing tool for every phase of the data analysis workflow. It caters to an array of data manipulation operations, such as filtering, selecting,

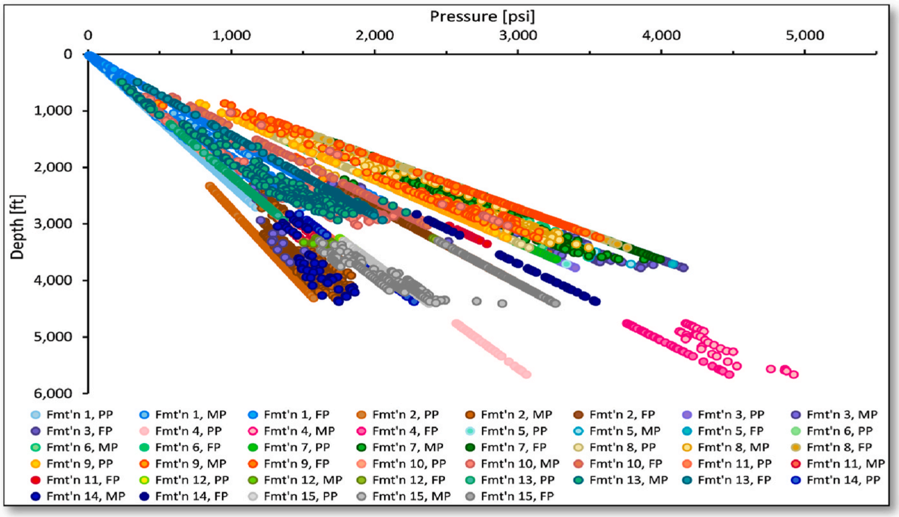


Fig. 8. Pressure distribution with depth drilled.

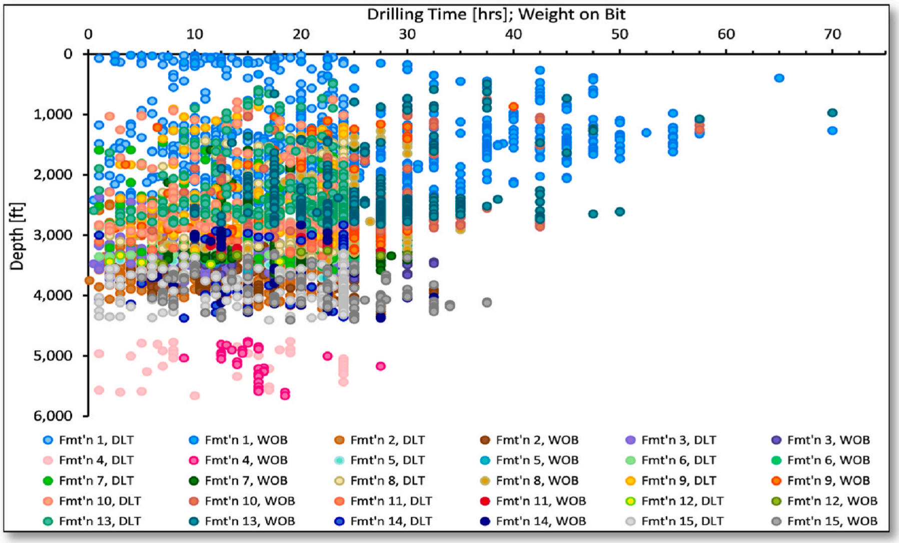


Fig. 9. Depth gained and their corresponding drilling time with the weight on bit.

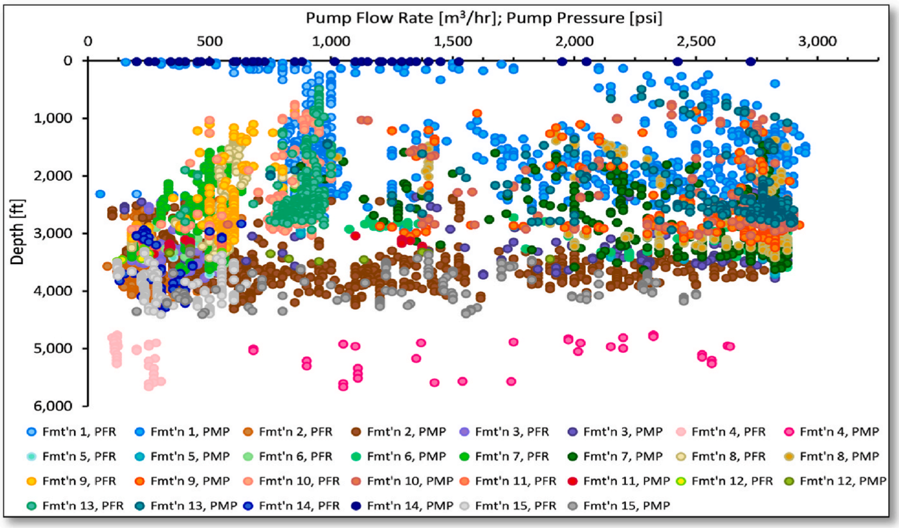


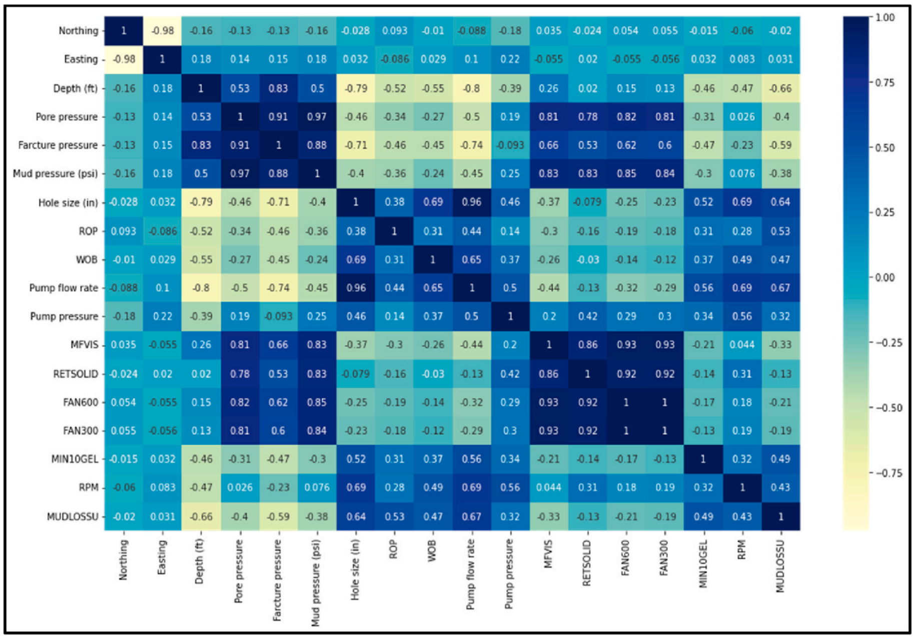
Fig. 10. Pump flow rate and pressure distribution with different formation depths.

grouping, and aggregating. Additionally, Pandas offers capabilities for managing missing data, categorical data, as well as encoding and decoding data. Its prowess also extends to computing diverse statistical measures and establishing correlations among input features. Pandas will play a pivotal role in the preprocessing and preparation of data before feeding it into Scikit-Learn models. Scikit-learn will be leveraged for constructing models and generating predictions on the pre-processed data [25]. From the 19 input variables found in the field data, a selection process led to the exclusion of 12 variables. Using the correlation matrix approach, variables not directly related were filtered out, ultimately retaining the seven most pertinent variables. These variables—hole size, RPM, WOB, pump flow rate, pump pressure, mud gel strength, and drill string RPM—are considered the input parameters for predicting lost circulation. The correlation matrix technique, illustrated in Figs. 11 and 12 for feature selection, provides insights into the relationship between different variables within a dataset. It aids in determining the optimal number of relevant variables for minimising the objective function. The correlation matrix provides valuable information about the impact of specific input variables on the severity of fluid loss. It demonstrates how one variable's behavior changes with fluctuations in another, quantified by correlation coefficients spanning −1 to +1. Values approaching absolute 1 indicate a potent relationship, with positive values implying direct proportionality and negative values signifying inverse proportionality. Fig. 11 presents the correlation of all parameters in the drilling report and field data, while Fig. 12 narrows down a subset of selected parameters for modeling. The process of variable subset selection is pivotal, as too few variables can lead to significant model biases, while an excessive number can compromise predictive capabilities and heighten variance in predictions. The correlation matrix critically informs multivariate analysis by exposing relationships and potential multicollinearity among variables [34]. This facilitates the examination of interdependencies among various data parameters, underscoring the correlation matrix's invaluable role in multivariate investigations.

4. Results and discussions

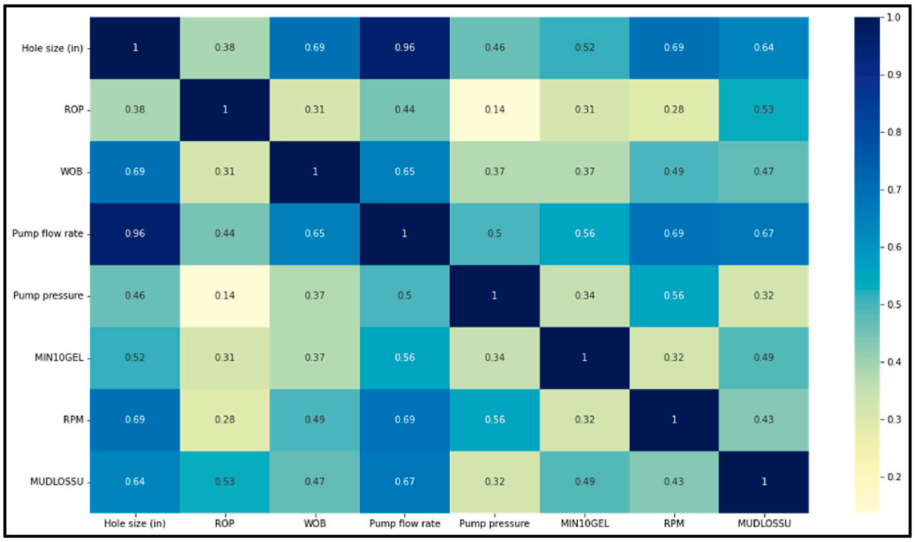
4.1. Features ranking

"Feature selection" is the process of choosing a specific subset of available features (variables) that best contributes to satisfactory



	Northing	Easting	Depth (ft)	Pore pressure	Farcture pressure	Mud pressure (psi)	Hole size (in)	ROP	WOB	Pump flow rate	Pump pressure	MFVIS	RETSOLID	FAN600	FAN300	MIN10GEL	RPM	MUDLOSSU
Northing	1	-0.98	-0.16	-0.13	-0.13	-0.16	-0.028	0.093	-0.01	-0.088	-0.18	0.035	-0.024	0.054	0.055	-0.015	-0.06	-0.02
Easting	-0.98	1	0.18	0.14	0.15	0.18	0.032	-0.086	0.029	0.1	0.22	-0.055	0.02	-0.055	-0.056	0.032	0.083	0.031
Depth (ft)	-0.16	0.18	1	0.53	0.83	0.5	-0.79	-0.52	-0.55	-0.8	-0.39	0.26	0.02	0.15	0.13	-0.46	-0.47	-0.66
Pore pressure	-0.13	0.14	0.53	1	0.91	0.97	-0.46	-0.34	-0.27	-0.5	0.19	0.81	0.78	0.82	0.81	-0.31	0.026	-0.4
Farcture pressure	-0.13	0.15	0.83	0.91	1	0.88	-0.71	-0.46	-0.45	-0.74	-0.093	0.66	0.53	0.62	0.6	-0.47	-0.23	-0.59
Mud pressure (psi)	-0.16	0.18	0.5	0.97	0.88	1	-0.4	-0.36	-0.24	-0.45	0.25	0.83	0.83	0.85	0.84	-0.3	0.076	-0.38
Hole size (in)	-0.028	0.032	-0.79	-0.46	-0.71	-0.4	1	0.38	0.69	0.96	0.46	-0.37	-0.079	-0.25	-0.23	0.52	0.69	0.64
ROP	0.093	-0.086	-0.52	-0.34	-0.46	-0.36	0.38	1	0.31	0.44	0.14	-0.3	-0.16	-0.19	-0.18	0.31	0.28	0.53
WOB	-0.01	0.029	-0.55	-0.27	-0.45	-0.24	0.69	0.31	1	0.65	0.37	-0.26	-0.03	-0.14	-0.12	0.37	0.49	0.47
Pump flow rate	-0.088	0.1	-0.8	-0.5	-0.74	-0.45	0.96	0.44	0.65	1	0.5	-0.44	-0.13	-0.32	-0.29	0.56	0.69	0.67
Pump pressure	-0.18	0.22	-0.39	0.19	-0.093	0.25	0.46	0.14	0.37	0.5	1	0.2	0.42	0.29	0.3	0.34	0.56	0.32
MFVIS	0.035	-0.055	0.26	0.81	0.66	0.83	-0.37	-0.3	-0.26	-0.44	0.2	1	0.86	0.93	0.93	-0.21	0.044	-0.33
RETSOLID	-0.024	0.02	0.02	0.78	0.53	0.83	-0.079	-0.16	-0.03	-0.13	0.42	0.86	1	0.92	0.92	-0.14	0.31	-0.13
FAN600	0.054	-0.055	0.15	0.82	0.62	0.85	-0.25	-0.19	-0.14	-0.32	0.29	0.93	0.92	1	1	-0.17	0.18	-0.21
FAN300	0.055	-0.056	0.13	0.81	0.6	0.84	-0.23	-0.18	-0.12	-0.29	0.3	0.93	0.92	1	1	-0.13	0.19	-0.19
MIN10GEL	-0.015	0.032	-0.46	-0.31	-0.47	-0.3	0.52	0.31	0.37	0.56	0.34	-0.21	-0.14	-0.17	-0.13	1	0.32	0.49
RPM	-0.06	0.083	-0.47	0.026	-0.23	0.076	0.69	0.28	0.49	0.69	0.56	0.044	0.31	0.18	0.19	0.32	1	0.43
MUDLOSSU	-0.02	0.031	-0.66	-0.4	-0.59	-0.38	0.64	0.53	0.47	0.67	0.32	-0.33	-0.13	-0.21	-0.19	0.49	0.43	1

Fig. 11. Correlation matrix for the general field data (parameters).



	Hole size (in)	ROP	WOB	Pump flow rate	Pump pressure	MIN10GEL	RPM	MUDLOSSU
Hole size (in)	1	0.38	0.69	0.96	0.46	0.52	0.69	0.64
ROP	0.38	1	0.31	0.44	0.14	0.31	0.28	0.53
WOB	0.69	0.31	1	0.65	0.37	0.37	0.49	0.47
Pump flow rate	0.96	0.44	0.65	1	0.5	0.56	0.69	0.67
Pump pressure	0.46	0.14	0.37	0.5	1	0.34	0.56	0.32
MIN10GEL	0.52	0.31	0.37	0.56	0.34	1	0.32	0.49
RPM	0.69	0.28	0.49	0.69	0.56	0.32	1	0.43
MUDLOSSU	0.64	0.53	0.47	0.67	0.32	0.49	0.43	1

Fig. 12. Correlation matrix for the selected parameters for modeling.

prediction performance with respect to an objective function [35]. In various machine learning applications, the number of selected features has ranged from less than ten to over forty variables in certain cases. Eliminating insignificant and/or duplicated variables not only enhances prediction accuracy but also streamlines computational efficiency. The presence of these variables often hampers the efficiency of many machine-learning applications. The goals of feature selection encompass preventing data overfitting, streamlining analyses, decreasing computation time, bolstering accuracy and resolution, and crafting more streamlined models with heightened efficiency. The hierarchy of significance for variables is determined by the chosen predictive model. As such, the selection of a reliable model for feature selection is of paramount importance. The methodology employed here for feature selection leverages data analysis, visualisation, field experience, and the correlation matrix approach, as depicted in Figs. 11 and 12. This approach underscores the relationship between different variables within a dataset, shedding light on the optimal number of pertinent variables needed to minimise the objective function. It delivers insightful information on how specific input variables impact the severity of fluid loss. Among these variables, seven exhibit the highest correlation coefficients, ranging from 0.32 to 0.67, with fluid loss severity. Based on the compiled dataset, these seven input variables exert a significant influence on the severity of lost circulation in the Marun field. Intriguingly, introducing more than seven features does not impact prediction accuracy but instead impairs model performance. Through feature selection, variables like Northing, Easting, depth, formation type, formation pore pressure, formation fracture pressure, marsh funnel viscosity, retort solids content, and Fan300 and Fan600 from the Marun field dataset were deemed irrelevant to lost circulation. Their inclusion would likely hinder the prediction accuracy achieved by the considered technique. In another context, these variables act as dependent variables, offering no supplemental insight into fluid loss. Consequently, they contribute noise to the predictor, particularly when juxtaposed with numerous other input variables. It's worth noting that these variables were previously considered alongside the selected variables, as discussed in the data visualisation section. As feature ranking results are specific to the Marun field, distinct feature selections (i.e., priority input variable combinations) could be identified for other oil fields and geological scenarios. Previous research by Ref. [20] examined how feature selection for Marun field data relates to the combination of inputs rather than individual input parameters in minimising the objective function. Their sensitivity analysis yielded contrasting results to those derived from the proposed feature ranking method. For instance, pore pressure, fracture pressure, and certain mud-related properties were identified as the most influential parameters for predicting lost circulation. The authors' prior work in the Marun field utilised various techniques to identify the most significant selected features, working with a dataset containing around 19 parameters for predicting lost circulation events, as detailed in Table 7. The outcomes indicated that prediction accuracy increased as the number of selected features grew, although the effect tapered off once the number exceeded ten [36]. Accordingly, this study selected the following seven features, in order, as input parameters: pump flow rate, hole size, ROP, gel strength, WOB, RPM, and pump pressure (Fig. 13).

WOB, pump flow rate, pump circulation pressure, and drill string rotation per minute (RPM) were included due to their status as driller-controlled variables at the rig site. Along with gel strength, these variables exert a significant influence on other parameters (e.g., ROP) that impact lost circulation in naturally fractured and induced fractured formations. Mechanical surface drilling parameters play a crucial role in drilling and are readily available for each well. These parameters serve as sensitive indicators for detecting lost circulation and display noticeable changes in their values following such occurrences. Increased durability leads to enhanced meterage drilled and an improved average ROP over the course of the bit run. An excessive pump flow rate significantly affects annular hydraulics by elevating the equivalent circulating density and subsequently increasing fluid loss rates. If fluid losses transpire, pump pressure (including drill string, annular, and drill bit pressure drops) and pump flow rate will promptly decrease. Elevated annular pressure drops contribute to a higher equivalent circulating density, which in turn can fracture formations. Moreover, WOB and RPM are pivotal features, potentially linked to detecting smaller fractures or surrounding larger fractures or caves, resulting in abrupt WOB changes. Similarly, ROP might display sudden increases when encountering small fractures.

4.2. *Prediction of onset lost circulation using the multigene genetic programming*

For precise prediction of the onset of loss of circulation, this study harnessed the identified significant parameters to construct a multigene genetic model. The resultant output of this network model is the prediction of the onset of loss of circulation. The algorithm was developed using the Python programming language, employing modified versions of existing packages. This research emphasises the practical application of the multigene genetic algorithm, omitting the presentation of the mathematical formulas underpinning its operation [39]. To fine-tune its performance, the control variables were adjusted, and the optimal configuration was ascertained through a sequence of trial-and-error assessments. Fig. 14 displays the test results of the model's predictions, with the horizontal and vertical axes representing actual and predicted values, respectively.

$y = 0.9279x + 11.764.$

“y" = dependent variable or the variable that is being predicted.

“x" = independent variable or the variable that is used to make the prediction.

Units of predicted and actual values = bbl/hr.

The developed multigene genetic model underwent a comparison with actual measured data to determine its accuracy in representing the targeted physical phenomenon. The comparison between the model's

Table 7
Reviewed Marun field publications using machine learning methods for prediction of fluid loss.

Authors	Prediction method	No. of input variables	No. of data points	Input parameters
[37]	ANN.	18	589	Depth, well trajectory, drilling time, length of the open hole section, formation top, bit size, Average pump flow rate, average pump discharge pressure, Mud weight, Solid percentage, FAN 300, FAN 600, mud filtrate, mud volume lost, porosity, rock type, permeability, minimum horizontal stress profile.
[31]	ANN	15	32 wells	Depth from ground, depth from sea level, daily drilling time, formation top, well northing, easting, hole size, average pump flow rate, average pump pressure, mud weight, solid percentage, FAN 300, FAN 600, mud filtrate, mud volume lost.
[32]	DT; ANFIS; ANN; GA-MLP.	19	1900	Drilling length, North, east, hole size, WOB, flow rate, pump pressure, viscosity, FAN 300, FAN 600, gel10 m, drilling time, depth, solid percentage, bit rotational speed RPM, drilling meterage, pore pressure, mud weight, fracture pressure.
[38]	Data Mining; ANFIS.	18	42,948	Drilling meterage, drilling time, mud velocity, hole size, WOB, flow rate, pump pressure, viscosity, FAN 300, FAN 600, GS10 min, solids content, RPM, pore pressure, mud pressure, fracture pressure, formation type, loss severity.
[20]	MLP, MLP-GA, MLP-PSO, MLP-COA, LSSVM, LSSVM-GA, LSSVM-PSO, LSSVM-COA	10	2820	North, east, formation type, hole size, pore pressure, fracture pressure, pump pressure, FAN600/FAN300, gel10 m/gel10s, RPM.
[36]	LSSVM; CNN; COA-MELM; PSO-MELM; GA-MELM; COA-LSSVM; PSO-LSSVM; GA-LSSVM.	9	2783	Pump pressure, mud weight, fracture pressure, pore pressure, depth, gel10 m/gel10s, FAN600/FAN300, flow rate, and formation type.

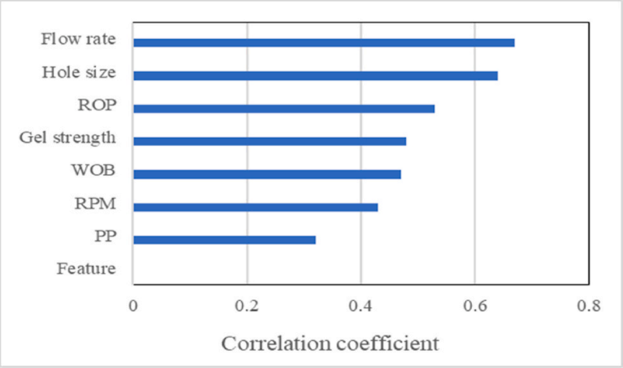


Fig. 13. Feature importance of the input variables.

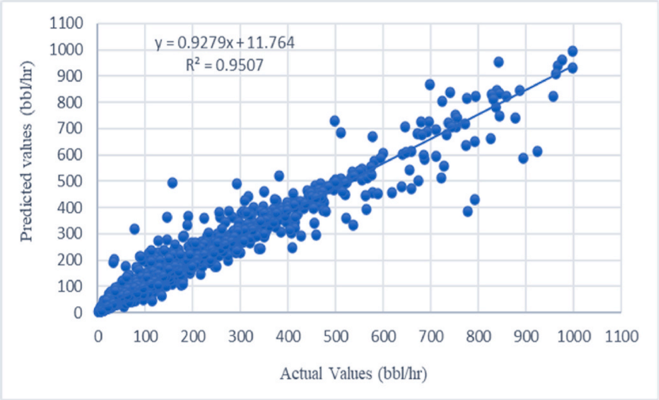


Fig. 14. The Test Result prediction of the Multigene Genetic Programming.

output and the field data yielded an evaluation of the model's performance and robustness. The regression coefficient (R^2) of the model was found to be highly acceptable (0.951), signifying that the model adequately predicts the onset of lost circulation. To further evaluate the model's accuracy, statistical performance indices such as the root mean square error (RMSE) and mean absolute error (MAE) were employed. We utilised the combination of metrics recommended by Ref. [40]. The model demonstrated low values for MAE (7.57) and RMSE (8.46) due to data quality issue, indicating its satisfactory performance in predicting the onset of lost circulation.

4.3. Out-of-sample validation of multigene genetic network prediction

Ensuring the reliability of the generated model is of utmost importance, and validation plays a critical role in achieving this by quantifying the model's performance on unseen data. Model validation holds significant significance in the realm of machine learning, as it assesses whether the model can generalise its learning to new, previously unseen data [32]. Given that different models can yield varying degrees of accuracy, model validation is typically conducted using a distinct dataset that the model has not been exposed to during training. The statistical metrics employed in validating the Multigene Genetic model showcased minimal values for MAE (1.33) and RMSE (2.58) due to data quality issue, alongside an excellent R^2 value of 0.968. These findings demonstrate a strong alignment between the model's predictions and actual data, thus establishing the reliability and utility of the proposed model for predicting drilling fluid loss (Fig. 15). This model has the potential to significantly assist drilling engineers in accurately predicting loss circulation onset at various depths, both prior to and during drilling operations.

4.4. Model optimisation

Preventing fluid loss during drilling operations through the anticipation of its onset and prudent planning represents an effective approach to pre-emptively addressing the issue. Table 8 provides a comparative analysis of the outcomes generated by the proposed Multigene Genetic model against existing literature on models predicting the onset of fluid loss. It is noteworthy that the authors of this study employed the same dataset from the Marun oilfield as the basis for their predictions, albeit using distinct techniques and models. It's important to acknowledge that, while machine learning models hinge on the data at their disposal, there might be some variance in the data used for model training despite a substantial overlap. Thus, the primary focus of this comparison is on elucidating how input features are selected and employed to elucidate the target variable, rather than undermining the findings of fellow researchers. The Multigene Genetic model presented in this study displayed a notably higher coefficient of determination (R^2), which signifies the extent of agreement between the predicted values and the field dataset. Moreover, it was evident that this model provides the advantage of interpretability while maintaining a high degree of accuracy. The fundamental principle underpinning the evaluation of accuracy revolves around employing specific metrics to juxtapose the original target with the predicted outcomes. Notably, the proposed model hinges on a selection of pertinent input parameters, namely the real-time surface drilling parameters that are readily accessible for every well.

5. Conclusions

This study effectively demonstrated the utility of employing AI algorithms, particularly the Multigene Genetic Model, to predict and identify instances of lost circulation in drilling operations. The study's outcomes underscored the reliability and efficiency of the developed model as a dependable solution for forecasting fluid loss. This capability holds the promise of substantial cost reduction, prevention of mud loss incidents, and time savings in drilling operations. The research establishes a replicable and comprehensive workflow for predicting fluid loss, which could find application in diverse fields and drilling operations. Through optimisation efforts, the Multigene Genetic Model has been fine-tuned to achieve data reduction, universal prediction capability, and compatibility with pre-existing platforms. Consequently, it emerges as a versatile and scalable solution for forecasting lost circulation occurrences. The study further highlights the significance of seven surface drilling parameters that are readily accessible in each well. These parameters exert a significant influence on the prediction accuracy of the model, thereby offering opportunities to enhance drilling practices and diminish the likelihood of lost circulation incidents. The achieved results demonstrated a mean absolute error of 1.33, a root

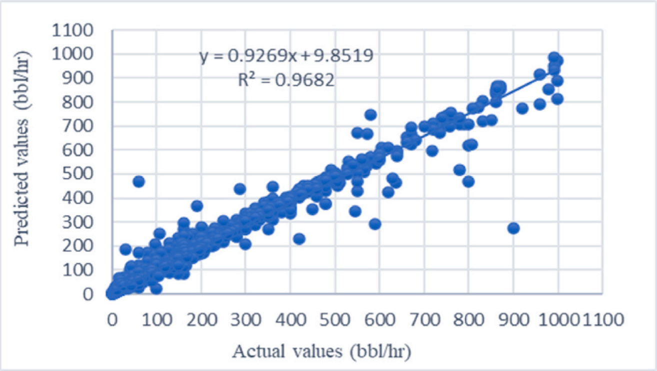


Fig. 15. Validation of the proposed Multigene Genetic Model for onset loss circulation prediction.

Table 8
Model prediction performance indices.

References	Model	Input parameters	No. of data points	RMSE	R^2
This Study	Multigene Genetic Programming	7	16,970	2.58	0.968
[32]	Genetic Algorithm-multi-Layer Perception (GA-MLP)	19	1900	0.137	0.826
[20]	MLP-GA	10	2820	0.930	0.930
[31]	Artificial neural network (ANN)	18	32 wells	–	0.82
[37]	Artificial neural network (ANN)	15	589	–	0.76
[38]	Adaptive Neuro-Fuzzy Inference System (ANFIS)	17	42,948	0.154	0.937
[36]	Data Mining; ANFIS LSSVM; CNN; COA-MELM; PSO-MELM; GA-MELM; COA-LSSVM; PSO-LSSVM; GA-LSSVM	9	2783	1.634	0.95

mean square error of 2.58, and a coefficient of determination of 0.968. Such outcomes positioned the developed model favourably when compared with other existing loss circulation prediction models, further validating its efficacy.

In conclusion, this study provides a solid groundwork for future investigations pertaining to the application of MGGP for optimising drilling operations. It underscores the potential of such technologies to ameliorate drilling challenges, including fluid loss, by curtailing costs, improving safety, and mitigating the environmental impact of drilling activities.

5.1. Recommendations

This study offers several recommendations, which can be summarised as follows:

- The results of feature ranking are specific to the Marun field, and different combinations of priority input variables may be more appropriate for other oil fields and geological contexts. Therefore, when applying similar predictive models to other fields, it's essential to adapt the feature selection process to the unique characteristics of the specific site.
- The study highlights the challenge of data quality in drilling operations due to uncertainties caused by factors like human error and equipment malfunction, especially in harsh environmental conditions like temperature fluctuations and mechanical shocks. To address this, establishing a recalibration equipment cycle is crucial to ensuring the accuracy and reliability of the collected data. Accurate data is essential for meaningful interpretation and decision-making by the drilling team.
- The datasets used in this research were collected from various rigs with different acquisition systems operating at different frequencies. To enhance future studies, it is recommended to implement a standardised digital data acquisition system across different wells. This system should be capable of collecting critical surface parameters from drilling operations and mud system characteristics at consistent frequencies. This approach would enable more accurate solutions and predictions related to the occurrence of onset-loss circulation events.

Funding

This research did not receive any specific grants from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

List of symbols and Abbreviations

AI	Artificial Intelligence
ANFIS	Adaptive Neuro-Fuzzy Inference System
ANN	Artificial Neural Networks
AZI	Azimuth
bbl/hr	Barrels Per Hour
CBR	Case-Based Reasoning
CNN	Convolutional Neural Network
COA	Cuckoo Optimisation Algorithm
cP	Centipoise
DDRs	Daily Drilling Reports
DLT	Drilling Time
DMRs	Daily Mud Reports
DT	Decision Trees
EC	Evolutionary Computing
EWRs	End-Of-Well Reports
FR	Flow Rate
ft	Feet
G1 G2 G3	Genes
GA	Genetic Algorithm
CNN	convolutional neural network
GP	Genetic Programming
HPHT	High-Pressure High Temperature
NFRs	Naturally Fractured Reservoirs
LCM	Loss Circulation Materials
MGGP	Multigene Genetic Programming
mins	Minutes
ML	Machine Learning
MLP	Multi-Layer Perceptron
GA-MLP	Genetic Algorithm – Multi-Layer Perceptron
RMS	Root mean square
NPT	Non-Productive Time
OBMs	Oil-Based Muds
psi	Pounds Per Square Inch
PSO	Particle Swarm Optimisation
PSO-MNN	Particle Swarm Optimisation – Modular Neural Network
R^2	Coefficient of Determination/Regression Coefficient
LOT	leak-off Test
PSD	particle size distribution
RMSE	Root Mean Square Error
ROP	Rate of Penetration
RPM	Rotations/Revolution Per Minute
SQRT	Square Root
SVM	Support Vector Machines
USD	United States Dollars
WBMs	Water-Based Muds
WOB	Weight on Bit
m^3/d	Cubic Meter Per Day
m^3/hr	Cubic Meter Per Hour
MAE	Mean Absolute Error
MCP	Mud Circulating Pressure
MDN	Mixture Density Network
MELM	Multilayer Extreme Learning Machine

MFVIS	Marsh Funnel Viscosity
MOS	Magnesium Oxysulphate
MVA	Majority Voting Algorithm
LSSVM	Least-Squares Support Vector Machines
MNN	Modular Neural Network
PP	Pore pressure
FR	Fracture pressure

References

[1] S. Krishna, et al., Conventional and intelligent models for detection and prediction of fluid loss events during drilling operations: a comprehensive review, J. Petrol. Sci. Eng. (2020) 195. Google Scholar.

[2] J. Sun, et al., Research progress and prospect of plugging technologies for fractured formation with severe lost circulation, Petrol. Explor. Dev. 48 (2021) 732–743. Google Scholar.

[3] F. Ahammad, S. Mahmud, S.Z. Islam, Computational Fluid Dynamics Study of Yield Power Law Springer, 2019. Google Scholar.

[4] A. Hamza, et al., Polymeric formulations used for loss circulation materials and wellbore strengthening applications in oil and Gas Wells: a review, J. Petrol. Sci. Eng. 180 (2019) 197–214. Google Scholar.

[5] S. Mardanirad, D.A. Wood, H. Zakeri, The Application of Deep Learning Algorithms to Classify Subsurface Drilling Lost Circulation Severity in Large Oil Field Datasets - SN Applied Sciences, SpringerLink. Springer International Publishing, 2021. Google Scholar.

[6] M.I. Magzoub, et al., Gelation kinetics of PAM/PEI based drilling mud for lost circulation applications, in: J. Petrol. Sci. Eng., Elsevier B. V, 2021, p. 200.

[7] Klungtvedt, et al., Preventing drilling fluid induced reservoir formation damage, in: SPE/IADC 202187 Middle East Drilling Technology Conference and Exhibition Held in Abu Dhabi, UAE, 2021, pp. 25–27 May21.

[8] A. Lavrov, Mechanisms and Diagnostics of Lost Circulation, Gulf Professional Publishing, 2016. Google Scholar.

[9] R. Caenn, H.C.H. Darley, G.R. Gray, Introduction to Drilling Fluids - Composition and Properties of Drilling and Completion Fluids, seventh ed., Gulf Professional Publishing, 2017. Google Scholar.

[10] O.E. Agwu, et al., Settling Velocity of Drill Cuttings in Drilling Fluids: A Review of Experimental, Numerical Simulations and Artificial Intelligence Studies, Powder Technology. Elsevier, B. V, 2018. Google Scholar.

[11] E. Fidan, T. Babadagli, E. Kuru, Use of Cement as Lost-Circulation Material: Best Practice, 2004. Google Scholar.

[12] K.-X. Cui, et al., Preparation and properties of magnesium oxysulfate cement and its application as lost circulation materials, Petrol. Sci. 18 (2021) 1492–1506. Google Scholar.

[13] A.T. Al-hameedi, et al., Real-time lost circulation estimation and mitigation, Egypt. J. Petrol. 27 (4) (2018) 1227–1234. Google Scholar.

[14] H.H. Alkinani, A.T.T. Al-Hameedi, S. Dunn-Norman, Data–driven decision–making for lost circulation treatments: a machine learning approach, Energy and AI 2 (2020) 100031. Elsevier. B. V. Google Scholar.

[15] A.K. Abbas, N.A. Al-Haideri, A.A. Bashikh, Implementing artificial neural networks and support vector machines to predict lost circulation, Egypt J. Petrol. 28 (4) (2019) 339–347. Elsevier. Google Scholar.

[16] M. Anemangely, et al., Machine learning technique for the prediction of shear wave velocity using petrophysical logs, J. Petrol. Sci. Eng. 174 (2019) 306–327. Google Scholar.

[17] M.G. De Giorgi, M. Quarta, Hybrid multigene genetic programming - artificial neural networks approach for dynamic performance prediction of an aeroengine, Aero. Sci. Technol. 103 (2020) 105902. Elsevier Masson. Google Scholar.

[18] H. Citakoglu, B. Babayigit, N.A. Haktanir, Solar Radiation Prediction Using Multi-Gene Genetic Programming Approach - Theoretical and Applied Climatology, SpringerLink. Springer Vienna, 2020. Google Scholar.

[19] O. Adeyi, et al., Process integration for food colorant production from Hibiscus Sabdariffa Calyx: a case of multi-gene genetic programming (MGGP) model and techno-economics, Alex. Eng. J. (2022). Elsevier. Google Scholar.

[20] M. Sabah, et al., Hybrid machine learning algorithms to enhance lost-circulation prediction and management in the Marun Oil Field, J. Petrol. Sci. Eng. (2021) 198. Google Scholar.

[21] A.H. Gandomi, A.H. Alavi, A new multi-gene genetic programming approach to nonlinear system modeling, in: Part I: Materials and Structural Engineering Problems - Neural Computing and Applications, SpringerLink. Springer-Verlag, 2012. Google Scholar.

[22] B. Sankar, et al., Application of multi-gene genetic programming technique for modeling and optimization of phycoremediation of Cr(VI) from wastewater, Beni-Suef Univ J Basic Appl Sci 12 (2023) 27. Google Scholar.

[23] W. La Cava, L. Vanneschi, S. Silva, Multi-gene genetic programming: an overview, Genetic Program. Theor. Pract. XVI (2019) 37–55. Google Scholar.

[24] O.E. Agwu, J.U. Akpabio, A. Dosunmu, Modeling the downhole density of drilling muds using multigene genetic programming, Upstream Oil Gas Technol. 6 (2021) 100030. Elsevier. Google Scholar.

[25] F. Pedregosa, et al., 'Scikit-learn: machine learning in Python', J. Mach. Learn. Res. 12 (Oct) (2011) 2825–2830. Google Scholar.

[26] M.P. Deosarkar, V.S. Sathe, Predicting effective viscosity of magnetite ore slurries by using artificial neural network, Powder Technol. 219 (2012) 264–270. Google Scholar.

[27] Mccord, Regional Geology of Asmari Reservoir in Ahwaz–Marun Area, Technical reports in NISOC, Iran, 1974. Google Scholar.

[28] M. Abdideh, S. Hedayati Khah, Analytical and Numerical Study of Casing Collapse in Iranian Oil Field - Geotechnical and Geological Engineering. SpringerLink, Springer International Publishing, 2018. Google Scholar.

[29] Telmadarreiea, S.R. Shadizadeh, B. Alizadeh, An investigation of hydrogen sulfide plume migration in the Asmari Reservoir of the Iranian Marun Oil Field: using repeat formation tests, Energy Sources, Part A Recover, Util. Environ. Eff. 35 (2013), 1991e2001. Google Scholar.

[30] M. Shayesteh, Investigation of Hydrogen Sulfide Contamination of Asmari Reservoir inMarun Oil Field, Geology Department, National Iranian South Oil Company, 2002. Report no. 5207. Google Scholar.

[31] A. Moazzeni, S.G. Jegarluei, N. Nabaei2, Prediction of Lost Circulation Using Virtual Intelligence in One of Iranian Oilfields, Advanced in Petroleum Exploration and Development vpl.1m, 2011, pp. 22–31. Google Scholar.

[32] M. Sabah, et al., Application of decision tree, artificial neural networks, and adaptive neuro-fuzzy inference system on predicting lost circulation: a case study from Marun Oil Field, J. Petrol. Sci. Eng. 177 (2019) 236–249. Elsevier. Google Scholar.

[33] A.K. Abbas, et al., Intelligent Decisions to Stop or Mitigate Lost Circulation Based on Machine Learning, Energy. Pergamon, 2019. Google Scholar.

[34] T. Pham-Gia, V. Choulakian, Distribution of the sample correlation matrix and applications, Open J. Stat. 4 (5) (2014) 330–344. Google Scholar.

[35] A. Jain, D. Zongker, Feature selection: evaluation, application, and small sample performance, IEEE Trans. Med. Imag. 19 (1997) 153–158.

[36] Jafarizadeh, et al., A new robust predictive model for lost circulation rate using convolutional neural network: a case study from Marun Oilfield, Adv. Res. Evolv. Sci. Petrol. J. (2022). Google Scholar.

[37] A. Moazzeni, S.G. Jegarluei, N. Nabaei2, Decision making for reduction of non-productive time through an integrated lost circulation prediction, Petrol. Sci. Technol. 30 (20) (2012) 2097–2107. Google Scholar.

[38] F. Agin, et al., Application of Adaptive Neuro-Fuzzy Inference System and Data Mining Approach to Predict Lost Circulation Using DOE Technique (Case Study: Maroun Oilfield), Petroleum. Elsevier, 2020. Google Scholar.

[39] D.A. Wood, S. Mardanirad, H. Zakeri, Effective prediction of lost circulation from multiple drilling variables: a class imbalance problem for machine and deep learning algorithms, J. Pet. Explor. Prod. Technol. 12 (1) (2022) 83–98. Google Scholar.

[40] T. Chai, R.R. Draxler, Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? – Arguments against Avoiding RMSE in the Literature [online], Geoscientific Model Development, Copernicus GmbH, 2014. Google Scholar.