Petroleum 11 (2025) 516–532

Contents lists available at ScienceDirect

Petroleum

journal homepage: www.keaipublishing.com/en/journals/petlm

Full Length Article

A novel approach to modeling breakdown pressure dynamics using machine learning

Subhan Aliyev, Talal Al Shafloot *, Murtada Saleh Aljawad **, Abdulazeez Abdulraheem, Salaheldin Elkatatny

Department of Petroleum Engineering, King Fahd University of Petroleum & Minerals (KFUPM), Dhahran, 31261, Saudi Arabia

ARTICLE INFO

Article history:
Received 21 February 2025
Received in revised form
19 June 2025
Accepted 12 July 2025

Keywords:
AI
Machine learning
Hydraulic fracturing
Breakdown pressure
Geomechanics

ABSTRACT

This study presents a novel machine learning (ML)-based approach for predicting breakdown pressure (BP) in hydraulic fracturing using experimental data. Unlike traditional analytical models that rely on simplified assumptions, ML models can capture complex nonlinear relationships between BP and its influencing factors. However, a key limitation in BP prediction stems from dataset constraints, particularly the scale differences between experimental setups and real-world formations. To mitigate these limitations, this research utilizes a unique dataset of 144 BP data points, incorporating various rock mechanical properties, injection parameters, and fluid properties. Additionally, a separate analysis of pressurization rate, based on 32 additional experimental data points, was conducted to better understand its effect on fracture initiation—an aspect often overlooked in ML-based studies. The dataset includes critical parameters such as injection rate, confining pressure, tensile strength, Young's modulus, permeability, unconfined compressive strength, Poisson's ratio, porosity, wellbore radius, and fracture geometry ratio. Five ML models—LightGBM, CatBoost, XGBoost, Kolmogorov-Arnold Network (KAN), and TabNet—were trained and evaluated. TabNet achieved the highest predictive performance (R^2 = 0.94) due to its attention-based feature selection and deep-learning-based representation learning. Model performance was assessed using mean absolute error (MAE) and mean squared error (MSE) to ensure robustness. To further enhance model interpretability, SHapley Additive exPlanations (SHAP) and TabNet's attention mechanism were used to explicitly assess feature importance, providing insights into the relative influence of different parameters on BP predictions. Additionally, advanced feature-handling techniques were employed to address categorical variables automatically, ensuring minimal preprocessing bias. The findings demonstrate the scalability of ML models for BP prediction using experimental data, reducing reliance on costly and time-consuming laboratory testing. By incorporating advanced interpretability techniques, systematic pressurization rate analysis, and robust ML architectures, this research provides a more accurate, data-driven approach for optimizing hydraulic fracturing designs.

1. Introduction

Hydraulic fracturing plays a crucial role in the development of unconventional reservoirs, which include shale, tight sandstone, and tight carbonate formations, due to their extremely low permeability and porosity, and elevated in situ stresses [1,2]. To overcome these limitations, multistage hydraulic fracturing improves hydrocarbon recovery by creating fractures that enhance contact area and decrease fluid flow barriers [3]. Achieving the necessary breakdown pressure (BP), however, is often difficult due to varying in-situ stress conditions and the constraints of available surface and downhole equipment [4].

The hydraulic fracturing process involves injecting fluids under high pressure to create fractures in the rock, with breakdown pressure (BP) being the peak pressure required to propagate these fractures [5]. In some cases, this operation is repeated throughout the lifespan of a well to maintain and boost production. The choice of fracturing fluids, which may include slick-water, polymers,

* Corresponding author.
** Corresponding author.
E-mail addresses: shafloot@kfupm.edu.sa (T. Al Shafloot), mjawad@kfupm.edu.sa (M.S. Aljawad).
Peer review under the responsibility of Southwest Petroleum University.

https://doi.org/10.1016/j.petlm.2025.07.008

foams, or surfactants, is tailored to the specific rock properties to maximize efficiency [6,7]. Understanding the BP is crucial for designing effective fracturing treatments, as it directly influences both the cost-effectiveness and safety of the procedure [8]. The BP is determined by several factors, including the rock's tensile strength, its elastic properties and in-situ stress [9,10]. Despite extensive research on predicting formation pore pressure using well log data, there remains a gap in studies focused on forecasting BP using rock mechanical properties [11,12].

Historically, researchers have relied on experimental, analytical, and numerical methods to estimate the BP [13,14]. Nonetheless, the BP remains a complex parameter influenced by factors like the pressurization rate of the wellbore, the viscosity of the fracturing fluid, and the tensile strength of the rock [15]. Estimating BP from computational models is inherently limited by simplifying assumptions and generalizations, making it challenging to capture the full complexity of real-world reservoir conditions [16,17]. While analytical approaches provide valuable insights, they suffer from scale limitations, as stress distribution [18,19]. Laboratory experiments, although accurate, are often costly and time-consuming, requiring core samples to simulate reservoir conditions [20]. As a result, machine learning models have become a promising alternative, offering a faster and more cost-effective way to predict BP and minimize the need for extensive laboratory testing [21,22].

To ensure our ML model overcomes dataset biases, it is trained on a unique and diverse dataset, reducing the risk of feature imbalance. Additionally, state-of-the-art techniques are incorporated to handle categorical variables automatically, eliminating manual preprocessing errors and improving predictive stability. Each feature's impact on BP estimation was explicitly examined using interpretability techniques, ensuring that the model's predictions align with real-world fracture mechanics. These enhancements allow our ML approach to not only surpass traditional estimation methods but also provide more reliable BP predictions for practical field applications.

1.1. Theoretical approaches for estimation of BP

Predicting the breakdown pressure (BP) in geological formations is essential for effective hydraulic fracturing operations. Various theoretical models have been developed to estimate the BP under different operational conditions, each incorporating specific rock properties and stress regimes. However, many of these models assume standard geothermal and pressure gradients. Recent studies suggest that under HPHT conditions, thermal stress becomes a critical factor influencing fracture pressure predictions, particularly in deepwater reservoirs [23].

Tensile strength models: one foundational approach is the tensile strength model [24], which posits that a fracture initiates when the circumferential stress (σ_θ) around a pressurized wellbore reaches the rock's tensile strength. This model is particularly applicable to vertical wells in a normal faulting regime, where the vertical stress (σ_V) exceeds the maximum horizontal stress (σ_H), which in turn is greater than the minimum horizontal stress (σ_h) ($\sigma_V > \sigma_H > \sigma_h$). The breakdown pressure (P_b) can be expressed as:

$$P_b = 3\sigma_h - \sigma_H + T_0 - P_0 \quad (1)$$

where T_0 represents the tensile strength of the rock formation P_0 denotes the pore pressure.

Poroelasticity considerations: to account for poroelastic effects [25], the above equation can be modified to the following:

$$P_b = 3\sigma_h - \sigma_H + T_0 - 2\gamma\frac{P_0}{2(1-\gamma)} \quad (2)$$

where the poroelastic parameter γ is defined as:

$$\gamma = \alpha\frac{(1-2\nu)}{2(1-\nu)} \quad (3)$$

with ν represents Poisson's ratio, while α indicating Biot's coefficient.

Fracture mechanics approaches: fracture mechanics models [26] consider the propagation of fractures under applied stresses. These models often utilize the stress intensity factor and fracture toughness to predict the BP. The stress intensity method determines the breakdown pressure (BP) using the critical stress intensity factor (K_{IC}) and the fracture geometry, as described as:

$$P_b = \frac{K_{IC}}{\sqrt{\pi a}} \quad (4)$$

where a is crack length.

1.2. Experimental observations of factors influencing BP

Breakdown pressure (BP) is primarily influenced by confining pressure [27–29]. Fjar et al. [30] found that increasing confining pressure raises the BP by applying additional stress to the rock, making fracture initiation more difficult. The recent study [31] demonstrated that in ultra-deep carbonate reservoirs, variations in fault integrity, in-situ stress fields, and rock elastic properties strongly influence the fracture initiation pressure and stress distribution. Laboratory studies [32,33] demonstrated that the high permeability formations have lower BP due to greater fluid leak-off, which reduces the pressure near the fracture tip. Similarly, high porosity rocks exhibit lower BP [34,35], because they absorb more fluid before fracture initiation, reducing resistance to fracturing. Injection rate also plays a crucial role in BP value [36,37], as higher injection rates increase the fluid pressure buildup within the rock, leading to higher BP. Generally, higher Poisson's ratio increases ductility, which enhances a rock's ability to deform under stress, often leading to higher BP [38]. However, in some cases, increased ductility can lead to lower BP because the rock deforms more easily under stress, dissipating energy that would otherwise contribute to fracture initiation [39]. Additionally, experimental studies [40,41], showed that tensile strength is directly proportional to BP, meaning stronger rocks require greater pressure to initiate fractures.

Morita et al. [17] highlighted that the BP was affected by Young's modulus, as rocks with higher Young's modulus requires greater pressure to initiate hydraulic fractures due to their resistance to deformation. Additionally, the wellbore geometry significantly affects BP, as studies [42,43] have determined that larger wellbore diameters result in lower BP, likely due to the increased surface area reducing localized stress concentration at the fracture initiation point. Wei [44] emphasized that both micro-level mineral composition differences and macro-scale discontinuities, such as bedding planes and natural fractures, significantly influence fracture morphology and growth dynamics. This is further supported by large-scale 3D physical simulation experiments in interbedded tight sandstone formations, where fracture growth exhibited strong dependence on bedding thickness, rock layering, and mineral contrasts [45]. Tuzingila et al. [46] reviewed how mineral composition affects BP, highlighting that higher quartz content tends to lower the BP due to increased brittleness, while higher clay content results in greater BP because of the enhanced

ductility. Similarly, fluid properties can influence the BP; an experiment by Wang et al. [47] showed that higher viscosity fluids increase the BP by creating greater resistance to flow and reducing fluid penetration into the surrounding rock.

1.3. Machine learning applications in BP prediction

The integration of machine learning (ML) into BP prediction has gained traction, offering substantial improvements in accuracy and efficiency. According to the study conducted by Tariq et al. [48], ML algorithms are capable of processing extensive datasets, allowing them to uncover complex, non-linear relationships among geological, petrophysical, and operational variables that traditional models may fail to identify. For instance, the study applied ML techniques to predict the BP in various unconventional rock types, demonstrating that ML could expedite the prediction process and improve accuracy by analyzing large datasets of rock properties and fracturing parameters.

Almani and Khan [22] applied neural network models and compared them with practical models to evaluate their predictive accuracy for estimating the BP. However, the effectiveness of these models was often constrained by the availability of high-quality datasets and the challenge of interpreting complex results. Many models could not account for all key parameters affecting BP, such as rock brittleness, fluid viscosity, and in-situ stress conditions.

Hybrid models that combine numerical simulations with ML techniques have shown enhanced the predictive performance by leveraging the strengths of both methods [11]. However, these models can be computationally intensive and may still be limited by incomplete datasets and the exclusion of critical variables.

In addition, unique data points are trained and tested to enhance the model's robustness and predictive accuracy. This approach ensures that the model is not biased by redundant or correlated data, improving generalizability and preventing overfitting [49]. It is worth clarifying that a unique data point refers to a single instance derived from a dataset of experiments, helping to eliminate inconsistencies and enhance the reliability of predictions by ensuring a diverse representation of influencing factors. Furthermore, this method helps mitigate class imbalance by ensuring that the dataset contains a well-distributed representation of various conditions [50,51], preventing the model from being overly influenced by dominant classes and improving overall predictive performance.

2. Methods

2.1. Data collection from the literature

In this study, the dataset for breakdown pressure (BP) was compiled from previously published literature and publicly available laboratory experiment data. The assembled dataset encompasses fracturing fluid properties, rock properties, and experimental conditions, providing a comprehensive overview of factors influencing BP predictions.

The fracturing fluid properties considered include fluid phase and viscosity (η), both of which are crucial in determining flow behavior and pressure dynamics during hydraulic fracturing. The rock properties include rock type (e.g., sandstone, shale), unconfined compressive strength (UCS), tensile strength (σ_t), Poisson's ratio (ν), Young's modulus (E), porosity (ϕ), and permeability (k). Additionally, confining pressure (P_c), though not a rock property, was included as an experimental parameter to assess its effect on rock behavior under stress conditions. Parameters such as the fracture geometry ratio, which typically describes the relationship between incipient fracture dimensions—such as length and width—were also considered to capture geometric variations, while cylindricity was analyzed separately to assess the overall shape and mechanical characteristics of the rock samples.

The experimental conditions comprise essential factors like the injection rate (Q), confining pressure, and hole radius (R). These features provide a realistic representation of the fracturing environment. Furthermore, the sample size (s) was considered to account for the variations of BP and to ensure a comprehensive dataset. This holistic dataset was utilized to develop and validate the machine learning models. Table 1 summarizes some of the key references from which the BP data were obtained.

In this table, the diameter of the rock specimen is denoted as D, while the height of the rock sample is represented as H. The diameter of the sample's hole is indicated by d, and the length of the sample's hole is represented as h.

Several limitations in the methodology should be noted. First, the dataset used for training the machine learning models is limited, with only 144 data points. While this dataset provides valuable insights, its relatively small size may constrain the generalization of the model to larger, real-world datasets. Second, the data used in this study is sourced from multiple literature studies, leading to variability in experimental conditions, rock types, and other factors. This variability can make it challenging to establish consistent correlations between the input features and breakdown pressure (BP), potentially introducing biases into the model's predictions. Lastly, filling missing data by imputation or K-nearest neighbour algoritm for some features in certain data points posed an additional challenge.

2.2. Data collection for pressurization rate

In hydraulic fracturing, accurately determining the pressurization rate is essential [60], because it directly influences the BP and fracture initiation dynamics. Recent study [36] have highlighted that relying on the injection rate alone can lead to inaccurate predictions of the BP, especially when dealing with compressible fluids. This oversight stems from the fact that injection rate does not account for how quickly pressure builds up in the wellbore [61], particularly in cases where fluid compressibility significantly affects the pressure propagation. Therefore, measuring and analyzing the pressurization rate, which describes the rate of pressure increase in the wellbore prior to breakdown, provides a more accurate representation of the conditions leading to fracture initiation.

In this study, we collected 32 data points of pressurization rates immediately preceding the corresponding breakdown pressure during hydraulic fracturing experiments with aforementioned features for sandstone, shale and carbonate formations. The breakdown pressure (BP) range for the pressurization rate dataset spans from 2000 psi to 9200 psi, reflecting the corresponding pressures measured during the experimental conditions. A linear model was fitted to this subset of data to better understand and quantify the relationship between pressurization rate of different rocks and BP. It is important to note that these data points were collected and analyzed separately from the larger dataset of 144 BP data points due to availability limitations. Analyzing pressurization rates prior to breakdown is crucial [48], as it helps assess the relationship between pressure increase and rock fracturing. The study by Wu et al. [62] demonstrated that higher pressurization rates reduce the distance from the fracture initiation point to the borehole wall—approximately 34 mm at 0.01 MPa/s and 1.74 mm at 5 MPa/s—resulting in increased stress concentration. This highlights the need to control pressurization rates to ensure safe and effective fracture propagation.

Table 1
Summary of some datapoints gathered from various literature sources on BP across different rock.

Rock types	Injected fluids	Core dimensions (mm)	Fluid injection rates (cm^3/min)	Permeability (mD)	References
Scioto sandstone	Supercritical CO_2, water	$D = 50.5$, $H = 50.5$, $d = 6$, $h = 20$	5	–	[52]
Sandstone	Water	$D = 35$, $H = 75.1$, $d = 4.8$, $h = 26.4$	2.5–10	–	[53]
Carbonate (Limestone)	Supercritical CO_2, water	$D = 44$, $H = 80.4$ $d = 10.2$, $h = 55$	0.5–5	0.1–1	[54]
Sandstone	Liquid CO_2, water, nitrogen	$D = 38.1$, $H = 76.2$, $d = 6.35$, $h = 25.4$	0.1–10	100–200	[55]
Carbonate (Dolomite)	Supercritical CO_2, water	$D = 50$, $H = 100$, $d = 10$, $h = 50$	0.5–5	0.01–0.1	[56]
Shale (Marcellus)	Nitrogen, water	$D = 20.8$, $H = 55.3$, $d = 4.18$, $h = 21.2$	0.05–1	0.001–0.01	[57]
Sandstone	Supercritical CO_2, water, nitrogen	$D = 30.1$, $H = 66.4$, $d = 5.54$, $h = 23.2$	0.1–10	100–200	[58]
Carbonate (Limestone)	Supercritical CO_2, water	$D = 50$, $H = 100$, $d = 10$, $h = 50$	0.5–5	0.1–1	[59]

2.3. Exploratory data analysis

The collected dataset for this study incorporates key parameters such as injection rate (Q), confining pressure (p_c), rock tensile strength (σ_t), unconfined compressive strength (UCS), Young's modulus (E), Poisson's ratio (ν), permeability (k), porosity (ϕ), sample size (s), hole radius (R), fracture geometry ratio (FGR), and pressurization rate (PR). Fluids used in hydraulic fracturing—-water, supercritical CO_2, and nitrogen (N_2)—were represented in the dataset primarily through their viscosity (η) values. Since fluid viscosity directly influences fracture propagation, pressure buildup, and breakdown pressure (BP), it was included as a key feature in the ML models. Cylindricity was excluded from the statistical analysis due to its categorical nature, as it represents a binary condition and does not contribute numerically to the model inputs. The statistical measures include mean, standard deviation (SD), minimum, 25th percentile (25%), median (50%), 75th percentile (75%), and maximum values, which provide insights into the distribution of the input features used for model training as shown in Table 2.

Fig. 1 illustrates the box plots for the non-categorical variables in the dataset, providing a concise summary of their statistical distributions. Each box plot highlights key metrics, including the median, interquartile range (IQR), and potential outliers. The central box represents the IQR, which captures the range between the 25th (Q1) and 75th (Q3) percentiles, while the whiskers extend to show the variability outside this range. Any points beyond the whiskers indicate outliers, emphasizing deviations from the overall distribution. Box plots are particularly useful for comparing variables side-by-side, as they provide insights into the symmetry, spread, and presence of skewness in the data [63]. For instance, wider IQRs suggest greater variability, whereas shorter whiskers imply concentrated values.

2.4. Evaluation metrics

Various standard error metrics were employed to evaluate the predictive performance of our machine learning models, each offering unique insights into model accuracy and error. The mean absolute error (MAE) provides a straightforward measure of the average magnitude of errors, ignoring their direction, making it useful for understanding the overall accuracy of predictions. The mean squared error (MSE), by squaring each error, emphasizes larger deviations, which helps in identifying models that poorly handle significant outliers. The root mean square error (RMSE), derived from MSE, represents the typical magnitude of prediction errors, offering a more interpretable measure in the same units as the target variable.

The average absolute percentage error (AAPE) expresses errors as a percentage of the actual values, facilitating comparison across different scales. Standard deviation (SD) reflects the variability in the prediction errors, with a lower SD indicating more consistent predictions. Finally, the coefficient of determination (R^2) measures how well the model explains the variance in the observed data, with values closer to 1 representing better predictive accuracy.

2.5. Data preprocessing and feature encoding

The analysis commenced by splitting the dataset of 144 unique data points into training and testing sets, using a 75:25 ratio to

Table 2
Statistical analysis of the total dataset.

Parameter	Mean	Standard deviation	Minimum value	25th percentile	Median	75th percentile	Maximum value
Q (mL/min)	6.2	4.8	0.1	4	6	8	30
η (cP)	1.1	215	0.02	1	1.1	1.2	1.5
p_c (psi)	1300	2100	15	0	600	1600	8500
σ_t (MPa)	7.2	2.5	3.5	5.6	7	8.2	14.5
UCS (MPa)	120	65	28	55	85	170	220
E (GPa)	29.5	16.2	12.5	20	27	47	59
ν	0.26	0.03	0.195	0.235	0.265	0.275	0.305
k (mD)	1.1	28.31	0.0001	0.015	0.5	4.5	50.25
ϕ (%)	7.2	6.1	0.6	1.2	7	13	19
s (mm)	25	5	15	20	25	30	40
R (mm)	5.5	2	2.5	4	5.5	7	10
FGR	0.8	0.15	0.4	0.7	0.8	0.9	1.2
PR (MPa/s)	1	1.5	0.01	0.1	0.5	2	5
BP (psi)	2900	2600	120	1100	1800	4200	15800

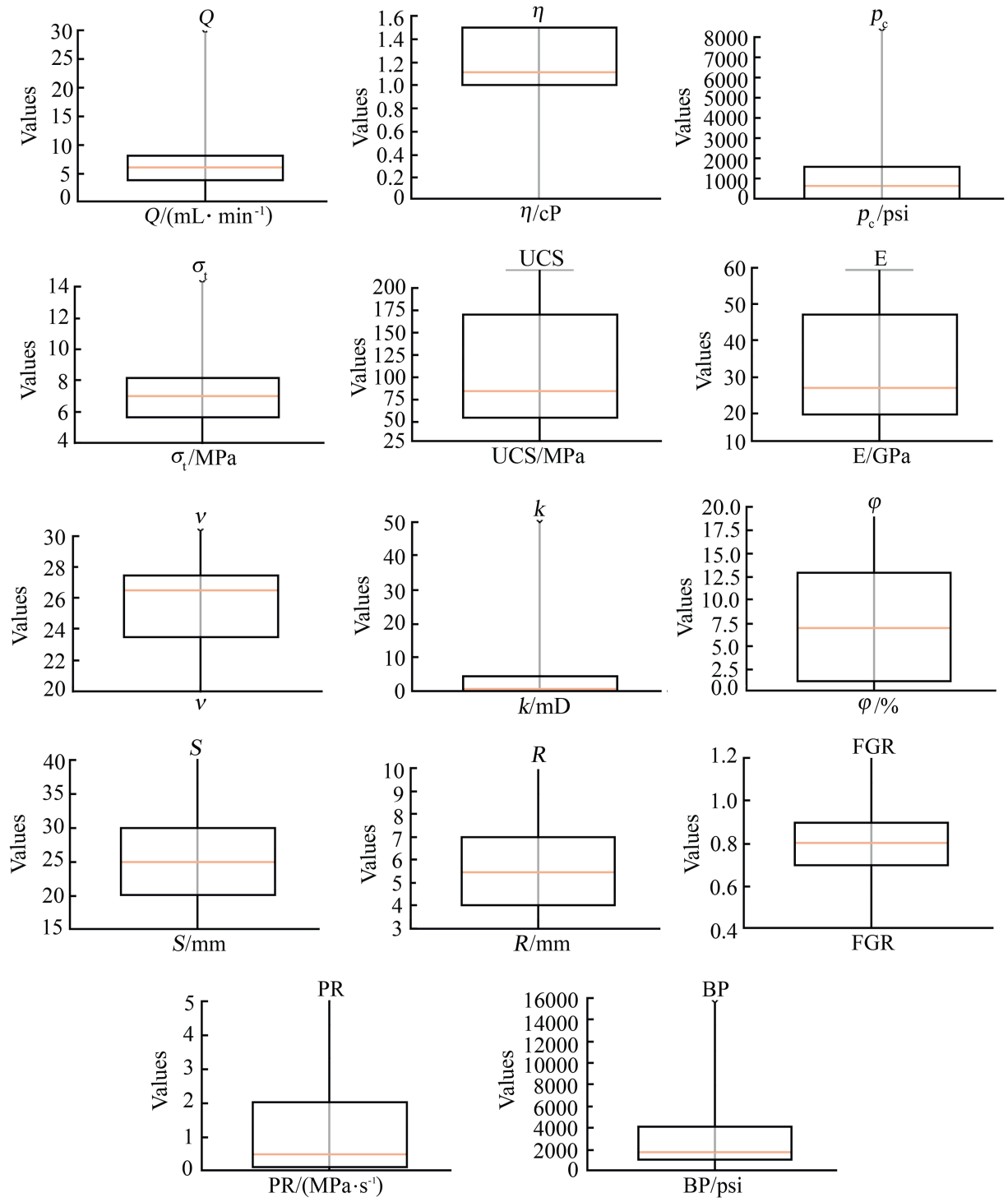

Fig. 1. The box-plot distribution of entire dataset.

ensure sufficient data for model training. This split was conducted in a stratified manner, meaning that the proportions of categorical features were maintained in both the training and testing sets. Stratification ensures that the model is exposed to a balanced distribution of data during training and testing, preventing biases and improving generalization. To further avoid data leakage and enhance model reliability, the testing data points were carefully selected to fall within the range of the training data.

One-hot encoding, illustrated in Fig. 2, was applied for categorical features. Rock types (sandstone, shale, and carbonate) were transformed using a binary system (0 for absence, 1 for presence). Additionally, the feature "cylindricity" was encoded similarly (0 for False, 1 for True) since it represents a binary categorical variable. However, for models like CatBoost and TabNet, which can handle categorical features natively, explicit encoding was not necessary.

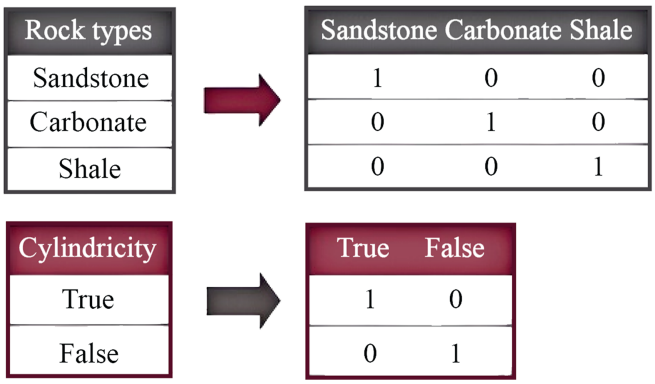

Fig. 2. The transformation of categorical variables to numerical.

2.6. Machine learning models

2.6.1. LightGBM

Light gradient boosting machine (LightGBM) is a highly efficient and scalable gradient boosting framework designed to tackle large-scale data with complex feature interactions. Introduced by ref. [64], LightGBM employs decision trees as base learners, training them sequentially to minimize residual errors from previous models. Unlike traditional gradient boosting frameworks, which grow trees level-wise, LightGBM utilizes a leaf-wise growth strategy. This method splits the leaf with the highest loss reduction, leading to more optimized trees. Such an approach reduces computational complexity and improves model accuracy, especially when working with high-dimensional datasets typical in geomechanical contexts such as hydraulic fracturing.

The LightGBM framework optimizes the following objective function for regression tasks:

$$L(\theta) = \sum_{i=1}^{n} l(y_i, f(x_i;\theta)) + \Omega(f) \tag{5}$$

where $L(\theta)$ represents the total loss, $l(y_i, f(x_i;\theta))$ is the loss function, typically the mean squared error (MSE), and $\Omega(f)$ is a regularization term that penalizes model complexity. This structure encourages the development of models that balance accuracy with generalization, thereby reducing overfitting.

In the context of hydraulic fracturing, where predicting breakdown pressure (BP) requires the integration of numerous geological and operational parameters, LightGBM offers several advantages. The histogram-based technique employed by LightGBM discretizes continuous features into bins, reducing both memory consumption and training time. This is particularly beneficial for handling large datasets efficiently, as seen in geomechanical modeling [65], where the number of features and the volume of data are substantial. The ability to handle both categorical and continuous variables within the same model makes it an ideal choice for geologically diverse datasets [66].

In this study, LightGBM was selected over random forest primarily due to its computational efficiency and superior handling of large datasets. Random forest, while robust and easy to implement, tends to suffer from slower training times and higher memory consumption when dealing with large-scale data [67].

2.6.2. CatBoost

CatBoost is an advanced gradient boosting algorithm that stands out for its ability to handle categorical features without requiring extensive preprocessing. Developed by Yandex [68], CatBoost incorporates ordered boosting and target-based encoding, which are specifically designed to minimize overfitting when dealing with categorical data. CatBoost operates on the principle of creating an ensemble of decision trees, with each tree designed to reduce the residual errors of its predecessor. The primary objective function used in CatBoost for regression tasks is the following:

$$L(\theta) = \sum_{i=1}^{n} l(y_i, f(x_i;\theta)) + \lambda \sum_{j=1}^{T} ||w_j||^2 \tag{6}$$

where $L(\theta)$ is the total loss, $l(y_i, f(x_i;\theta))$ represents the loss function (usually mean squared error), λ is a regularization parameter to avoid overfitting, T represents the number of trees, w_j and are the weights for the j-th tree. This architecture reduces the risk of overfitting by employing ordered boosting, a technique that prevents data leakage by ensuring that only past observations are used to predict the current one.

2.6.3. XGBoost (Extreme gradient boosting)

Developed by Chen and Guestrin [69], XGBoost is a highly regarded gradient boosting framework that sequentially builds decision trees, with each tree correcting the residual inaccuracies of the preceding trees. XGBoost incorporates various enhancements such as regularization, efficient handling of missing data, and the ability to parallelize tree construction, making it highly efficient for large-scale and high-dimensional datasets.

The objective function for XGBoost in regression is expressed as:

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{j=1}^{T} \Omega\left(f_j\right) \tag{7}$$

where $L(\theta)$ represents the total loss. The term $l(y_i, \hat{y}_i)$ denotes the loss function, such as the mean squared error. $\Omega(f_j)$ corresponds to the regularization term for each of the T trees to control overfitting, and f_j represents the individual trees in the ensemble.

XGBoost is particularly known for its ability to reduce overfitting by applying L_1 (Lasso) and L_2 (Ridge) regularization techniques to control the complexity of the model. Additionally, XGBoost optimizes both gradient boosting and tree pruning techniques by using a greedy algorithm that selects the best split points based on maximum loss reduction. Furthermore, one key advantage of XGBoost is its ability to parallelize tree construction, which significantly improves computational efficiency, particularly on large datasets. Its ability to optimize across multiple cores makes it one of the fastest gradient boosting implementations [70].

2.6.4. TabNet

TabNet is a deep learning architecture designed specifically for tabular data, introduced by Arik and Pfister [71]. Unlike traditional machine learning models such as gradient boosting, which focus on sequential learning using trees, TabNet employs a combination of attention mechanisms and sequential decision steps to process tabular data in a more adaptive and interpretable manner. The core innovation of TabNet lies in its use of self-supervised learning and soft feature selection through an attention mechanism that dynamically decides which features to attend to at each decision step.

TabNet's objective function can be expressed as:

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \lambda_{spars} \sum_{t=1}^{T} R\left(Mask^t\right) \tag{8}$$

where L represents the total loss. The term $l(y_i, \hat{y}_i)$ denotes the loss function, typically cross-entropy for classification tasks and mean squared error (MSE) for regression tasks. $Mask_t$ is the feature mask at decision step t, controlling which features are attended to during learning.

TabNet processes data in a sequential manner, where each decision step uses a soft attention mechanism to create a feature mask that determines which features will be selected and passed to the next layer. This attention-based mechanism makes the model more interpretable, as it highlights which features were critical to making each decision. TabNet also includes a sparse regularization term in its objective function, which penalizes the model for attending to too many features, encouraging it to focus only on the most relevant ones.

2.6.5. K-Nets

Kolmogorov-Arnold networks (K-Nets) are a neural network architecture rooted in the Kolmogorov-Arnold representation theorem, which guarantees that any multivariate continuous function can be decomposed into a finite sum of univariate functions and their compositions [72]. Unlike conventional neural networks that rely on dense interconnections across layers, K-Nets leverage this decomposition to represent complex functions with a structured and parameter-efficient architecture. The core innovation of K-Nets lies in their ability to approximate high-dimensional mappings using a combination of univariate transformations, making them both interpretable and computationally efficient.

The general objective function for K-Nets can be estimated as:

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \lambda \cdot \Omega(\Phi, \psi) \tag{9}$$

where $L(\theta)$ represents the total loss. The term $l(y_i, \hat{y}_i)$ denotes the loss function, commonly the mean squared error for regression tasks, and λ is a regularization coefficient. $\Omega(\Phi, \psi)$ refers to a regularization term that enforces smoothness or sparsity on the univariate functions Φ_q (outer functions) and ψ_{pq} (inner functions).

K-Nets decompose the input data into univariate inner functions ψ_{pq} (x_p) that are aggregated and transformed by outer functions Φ_q, resulting in an interpretable and flexible representation of complex multivariate functions. This structured approach, derived from theoretical principles [73], enhances generalization and reduces overfitting risks. The inclusion of regularization ensures that the learned functions are smooth and stable, further improving their performance in practical applications.

2.7. Sensitivity analysis methodology

In this study, we conducted a comprehensive sensitivity analysis to assess the impact of various input features on the prediction of breakdown pressure (BP). Several methods were employed, including preliminary correlation analysis, advanced neural network feature weight extraction, deep learning attention mechanisms, correction using the Frisch-Waugh-Lovell theorem, and SHapley Additive exPlanations (SHAP) for ensemble models.

2.7.1. Correlation analysis: Pearson and Spearman coefficients

Initially, we used Pearson and Spearman correlation coefficients to measure the linear and monotonic relationships between input features and BP.

(1) Pearson correlation coefficient: This coefficient measures the linear relationship between two variables x and y, defined as:

$$r = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2}\sqrt{\sum (y_i - \overline{y})^2}} \tag{10}$$

here, x_i and y_i represent the values of the two variables, and $\overline{x}$ and $\overline{y}$ denote the mean values of x and y, respectively.

(2) Spearman rank correlation coefficient: This non-parametric metric evaluates the rank correlation (monotonic relationship) between features given as:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{11}$$

where d_i is the difference between the ranks of x_i and y_i, and n represents the number of data points.

These metrics provided an initial insight into the relationships, highlighting features such as permeability, injection rate, and porosity as having strong correlations with BP.

2.7.2. Kolmogorov-Arnold neural network (KAN) weights

The Kolmogorov-Arnold neural network approximates complex functions and allows us to extract feature importance based on its learned weights. The optimal neuron configuration was determined through validation. Feature importance is calculated as:

$$W_j = \frac{\left| \sum_{i=1}^{n} w_{ij} \right|}{\sum_{j=1}^{m} \left| \sum_{i=1}^{n} w_{ij} \right|} \tag{12}$$

where W_{ij} represents the importance weight for feature j, w_{ij} denotes the connection weight from input i to the neuron associated with feature j. Additionally, n refers to the total number of neurons in the hidden layer, and mmm is the total number of features.

This method highlighted permeability and injection rate as the most influential features.

2.7.3. TabNet deep learning sensitivity analysis

TabNet uses an attentive feature selection mechanism, leveraging masks that learn to prioritize different features during training. The importance of each feature is quantified by its attention score can be expressed as:

$$A_j = \frac{\exp\left(Mask_j X\right)}{\sum_{j-1}^{m} \exp\left(Mask_j X\right)} \tag{13}$$

here, A_j is the attention score for feature j, $Mask_j$ represents the learned mask weight for feature j, and X denotes the input feature vector.

2.7.4. Frisch-Waugh-Lovell (FWL) theorem for pressurization rate adjustment

The FWL theorem is used to control for the impact of pressurization rate while accounting for its interaction with other

features. This correction ensures a more accurate estimation of feature effects estimated as:

$$Y_{adj} = Y - \beta_{rate} X_{rate} \tag{14}$$

where Y_{adj} represents the adjusted target (BP), β_{rate} denotes the regression coefficient for the pressurization rate, and X_{rate} refers to the pressurization rate feature.

This adjustment process reduces the bias introduced by the pressurization rate, allowing for a clearer interpretation of the effects of other variables.

2.7.5. SHapley additive exPlanations (SHAP) for ensemble models

SHAP values employ a game-theoretic approach to break down and attribute the effect of each feature on the outcome generated by the model. This method was applied to ensemble models (XGBoost, LightGBM, CatBoost) to interpret their complex decisions.

SHAP Value estimated by the following way:

$$\varphi_j = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{j\}) - f(S)] \tag{15}$$

where φ_j denotes the SHAP value for feature j, N represents the set of all features, S is any subset of features excluding j, and $f(S)$ corresponds to the model output for subset S.

3. Results and discussion

3.1. Correlation analysis

Pearson and Spearman correlation coefficients were initially used to evaluate feature importance as shown in Fig. 3. Pearson correlation assesses linear relationships, while Spearman correlation examines monotonic trends. The results highlighted confining pressure (p_c), injection rate (Q), rock tensile strength (σ_t), permeability (k), and Young's modulus (E) as the top five influential features. However, features like Poisson's ratio (ν) exhibited low correlation, indicating negligible impact on breakdown pressure (BP), despite some studies suggesting otherwise.

3.2. Model training and cross-validation

A comprehensive 5-fold cross-validation (Fig. 4) was employed for each model to ensure robust hyperparameter optimization.

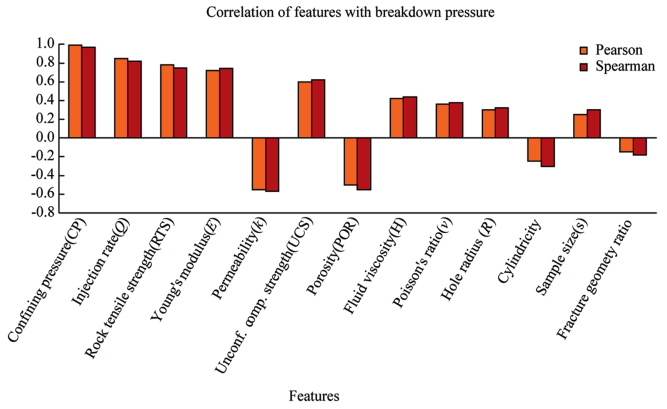

Fig. 3. Distribution of Spearman and Pearson correlations for breakdown pressure features.

The best configurations identified through grid search (Fig. 5) include the following:

(1) TabNet: batch size of 6, learning rate optimized during cross-validation.
(2) LightGBM: N_estimators set to 60, with maximum depth adjusted based on validation metrics.
(3) CatBoost: N_estimators set to 70, leveraging its native support for categorical features.
(4) XGBoost: N_estimators tuned to 55 with a finely adjusted learning rate.

N_estimators refers to the number of trees or boosting rounds in the ensemble models, such as LightGBM, CatBoost, and XGBoost. Increasing this value typically improves model performance by reducing bias, but if set too high, it can lead to overfitting, where the model becomes overly complex and fails to generalize well on new data. Batch size, on the other hand, defines the number of samples processed together in one iteration during model training, particularly for deep learning models like TabNet. Smaller batch sizes provide more accurate updates to the model's parameters, albeit with increased training time, while larger batch sizes accelerate the training process but may result in less stable convergence due to noisier updates.

The Kolmogorov-Arnold network (KAN) model was subjected to a neuron sensitivity assessment, identifying 32 as the optimal number of neurons based on minimized average absolute percentage error (AAPE) and maximized R^2 score (Fig. 6). The optimal neuron configuration effectively balanced complexity and predictive power. The structure of KAN model with 2 hidden layers and 14 input features is shown in Fig. 7.

The hyperparameters of these models were optimized, and the optimal configurations are shown in Table 3.

The results for model training and testing are shown in Fig. 8. TabNet achieved the highest R^2 scores, with 0.94, demonstrating excellent generalization capabilities. This highlights TabNet's robust performance in capturing complex patterns within the dataset while avoiding overfitting. The Kolmogorov-Arnold network (KAN) followed closely, with an R^2 of 0.95 for training and 0.92 for testing (Fig. 9), indicating strong predictive accuracy but also suggesting potential overcomplexity in its architecture, which may require further optimization. CatBoost and LightGBM showed comparable performance, with CatBoost slightly outperforming LightGBM. XGBoost had the lowest R^2 scores, likely due to its sensitivity to hyperparameter tuning, emphasizing the importance of careful model configuration.

3.3. Ranking and sensitivity analysis

The feature importance analysis using SHapley Additive exPlanations (SHAP) values for XGBoost, LightGBM, and CatBoost models (Fig. 10) consistently identified confining pressure, injection rate, permeability, rock tensile strength, and Young's modulus as the most influential factors.

These insights were further validated by TabNet's attention mechanism (Fig. 11), where similar rankings were observed. Notably, TabNet assigned the highest attention score to confining pressure (14.6%), followed by injection rate (13.4%) and permeability (11.7%), underscoring its ability to capture complex, nonlinear feature interactions.

A detailed examination of the KAN weight distribution revealed (Fig. 12) that the model assigns the highest weights to confining pressure (p_c), injection rate (Q), and permeability (k). This aligns with the top features identified in correlation analysis, showcasing

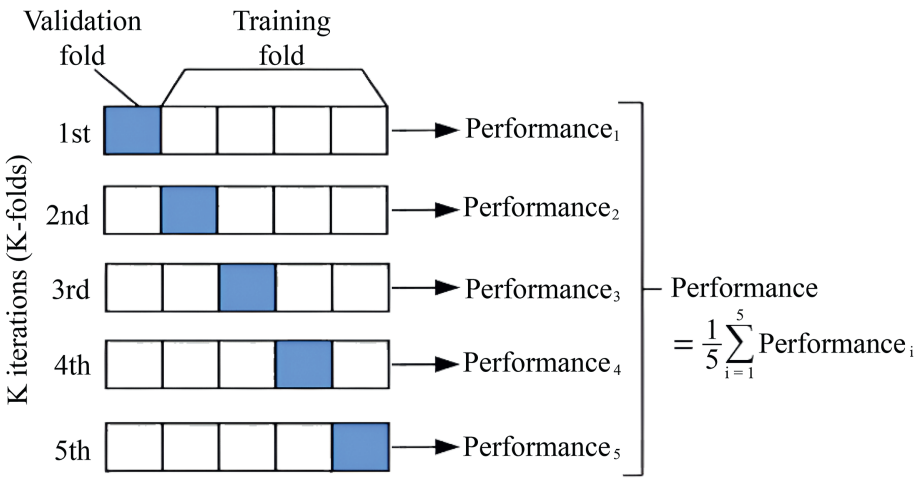

Fig. 4. A 5-Fold Cross-Validation approach applied in the research.

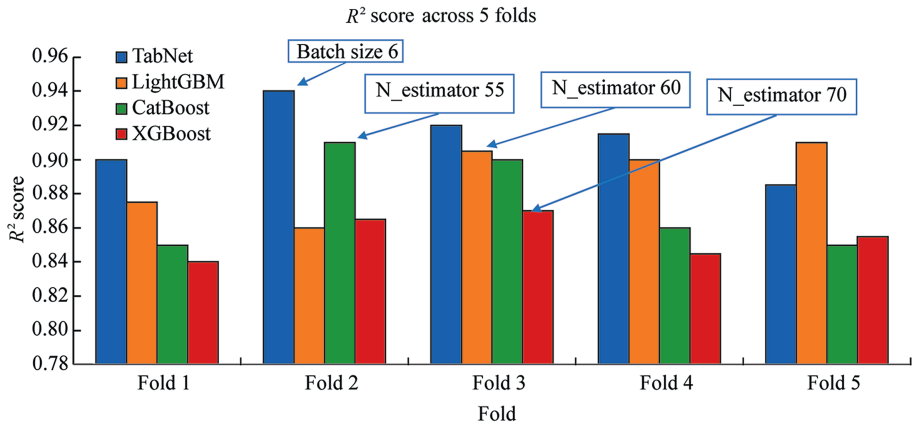

Fig. 5. Estimated optimal parameters for Ensemble and TabNet models.

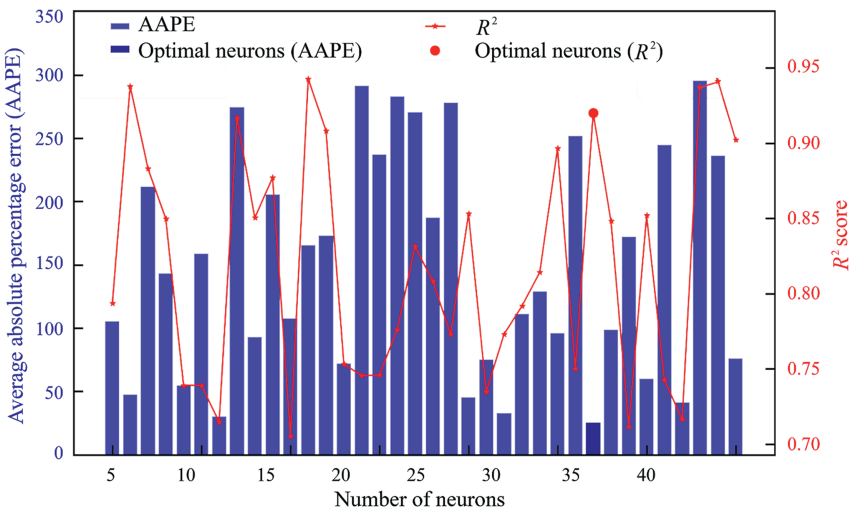

Fig. 6. Neuron number sensitivity for Kolmogorov-Arnold network.

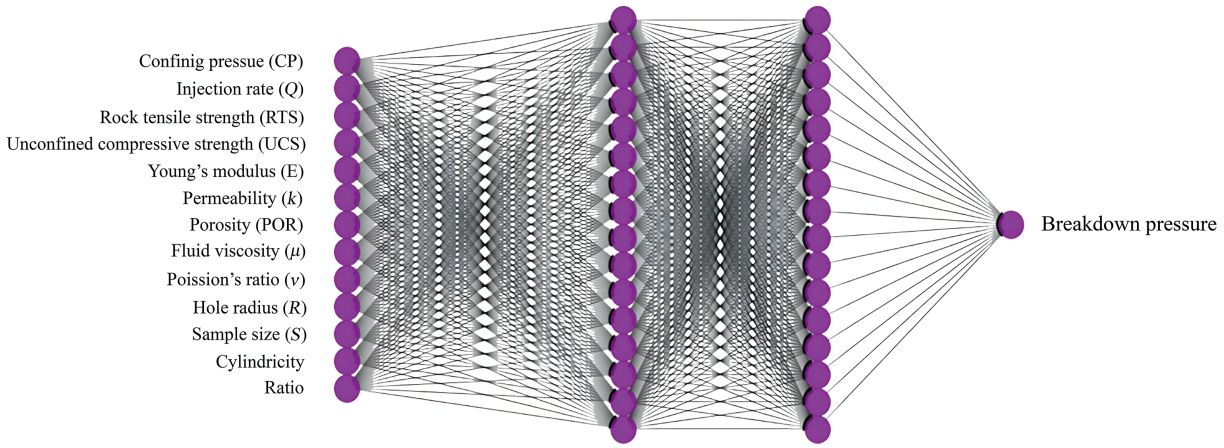

Fig. 7. Final structure of Kolmogorov-Arnold network.

Table 3
Optimized hyperparameter ranges utilized for training machine learning algorithms.

Model	Hyperparameters	Ranges explored	Selected values
TabNet	Virtual batch size	4–10	6
	Number of decision steps	3–10	5
	Learning rate	0.01–0.5	0.2
	Batch size	10–50	16
	Number of epochs	50–300	150
LightGBM	Number of estimators (n_estimators)	50–100	60
	Maximum depth (max_depth)	5–15	10
	Learning rate	0.01–0.3	0.1
	Number of leaves	20–50	30
CatBoost	Number of estimators (n_estimators)	50–100	55
	Maximum depth	5–10	8
	Learning rate	0.01–0.2	0.15
	L2 regularization	1–10	3
XGBoost	Number of estimators (n_estimators)	50–100	70
	Maximum depth	5–10	6
	Learning rate	0.01–0.3	0.1
	Subsample ratio	0.5–1.0	0.8
Kolmogorov-Arnold network (KAN)	Number of neurons (hidden layer)	5–40	32
	Activation function	Relu, Tansig, Sigmoid	Relu
	Learning rate	0.1–0.5	0.15
	Optimization algorithm	Gradient descent, Adam, RMSProp	Adam

the model's capability to accurately capture the relationships between features and BP.

Several studies [74–77] in the literature have suggested that parameters such as cylindricity, sample size, hole radius, and especially Poisson's ratio significantly influence breakdown pressure (BP) during hydraulic fracturing experiments. Cylindricity, representing the geometric uniformity of the borehole, was expected to impact stress distribution, while sample size and sample (hole) radius were believed to influence fracture initiation by affecting boundary conditions. Similarly, Poisson's ratio, which

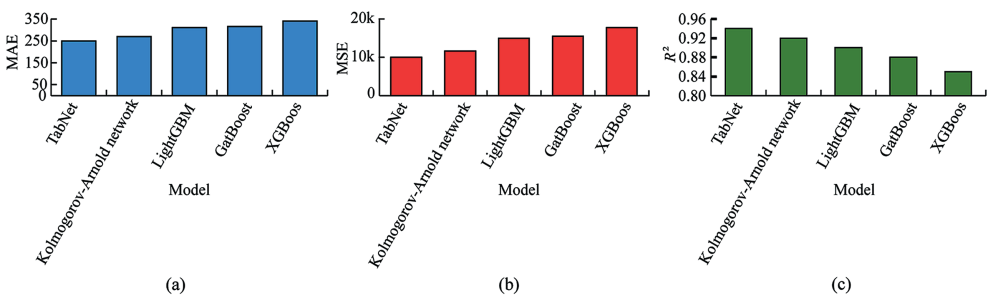

Fig. 8. Metrics comparison for trained models.

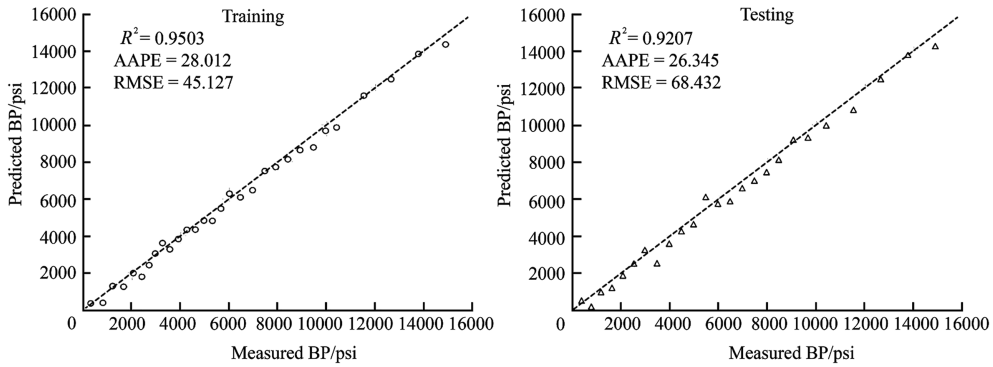

Fig. 9. Training and testing metrics for Kolmogorov-Arnold network.

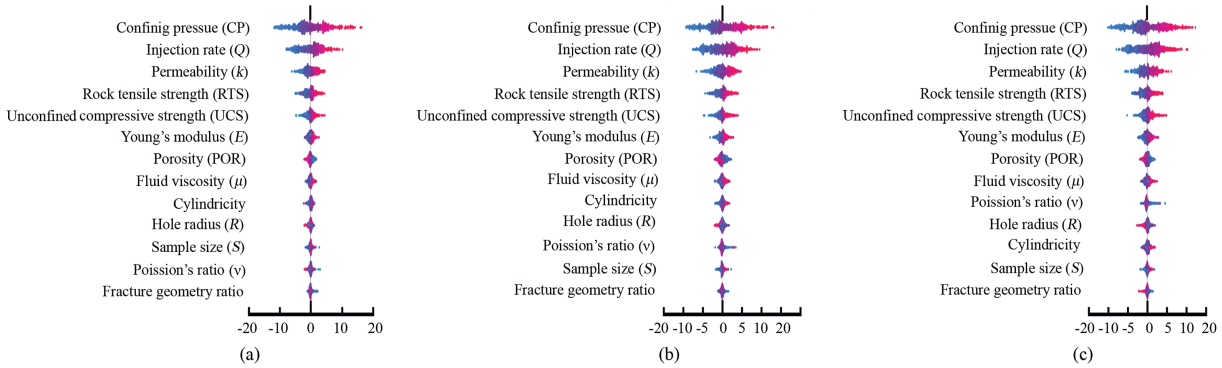

Fig. 10. SHAP Analaysis per feature for XGBoost, LightGBM, CatBoost.

reflects the material's ductility, has been linked to rock deformation behavior under stress, suggesting its potential influence on BP. However, in the current analysis, these parameters exhibit negligible correlation with BP. This discrepancy may stem from differences between laboratory-scale experimental conditions, highlighting the need for further investigation and possible upscaling methodologies to bridge the gap between experimental and real-world scenarios.

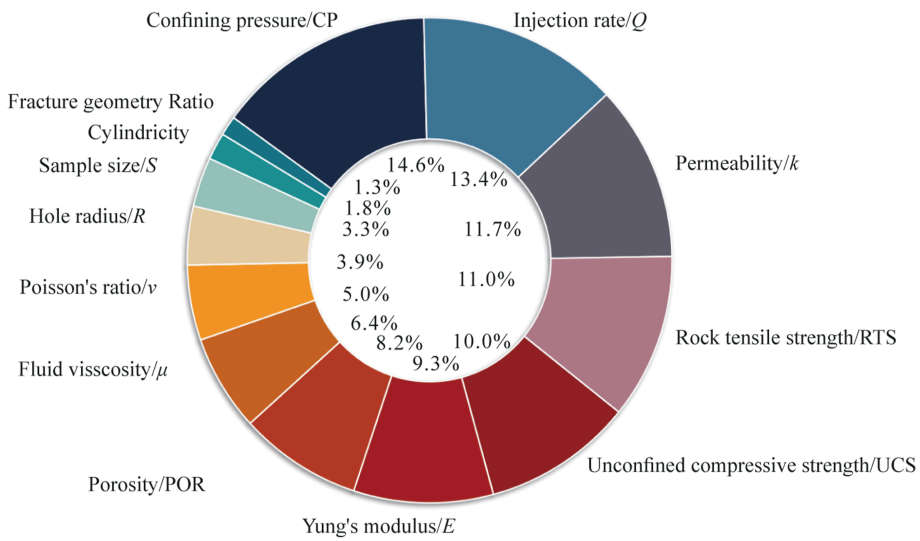

Fig. 11. Attention score per feature for TabNet model.

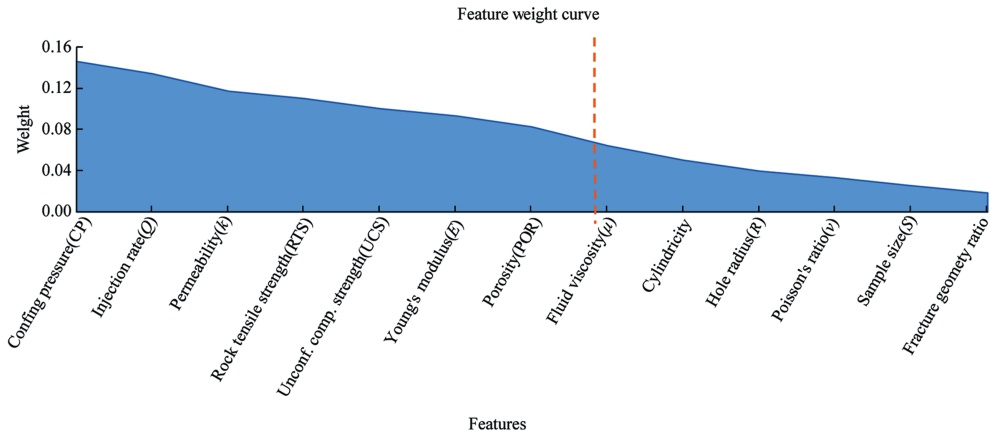

Fig. 12. Distribution of features according Kolmogorov-Arnold network.

Although Pearson and Spearman correlations provided useful initial rankings, their reliance on linear or monotonic assumptions can be misleading. Sensitivity analysis and SHAP values toffered a deeper understanding of feature importance by accounting for non-linear relationships and feature interactions. For instance, Poisson's ratio had a low correlation value but showed a nuanced impact when analyzed with SHAP and attention scores, highlighting the limitations of relying solely on correlation metrics.

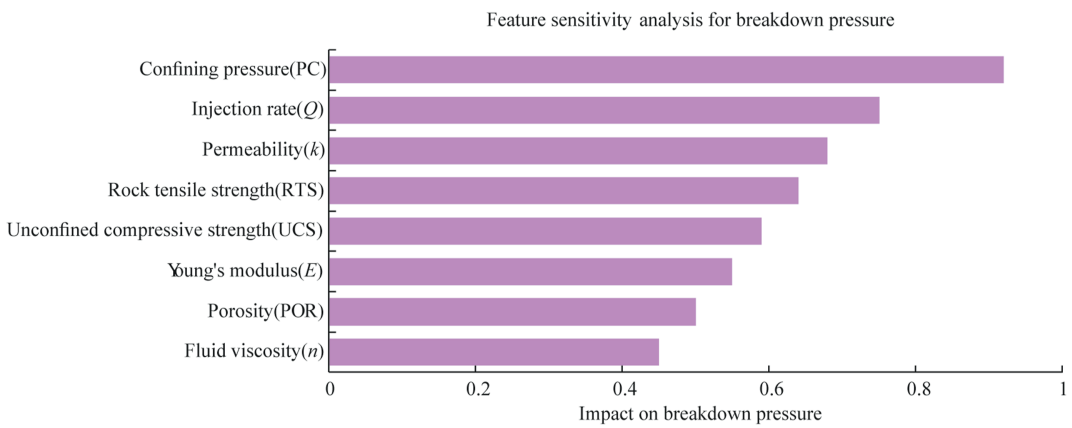

Fig. 13. Top chosen features for Re-training.

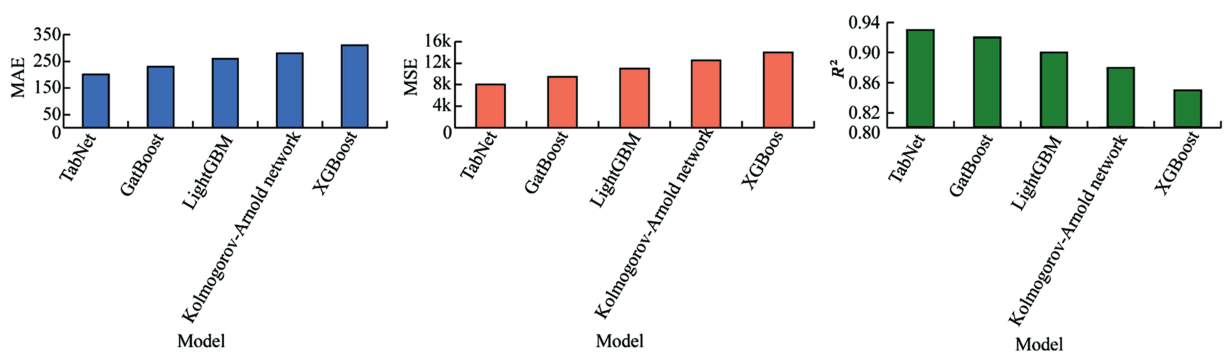

Fig. 14. Evaluation metrics for Re-trained models.

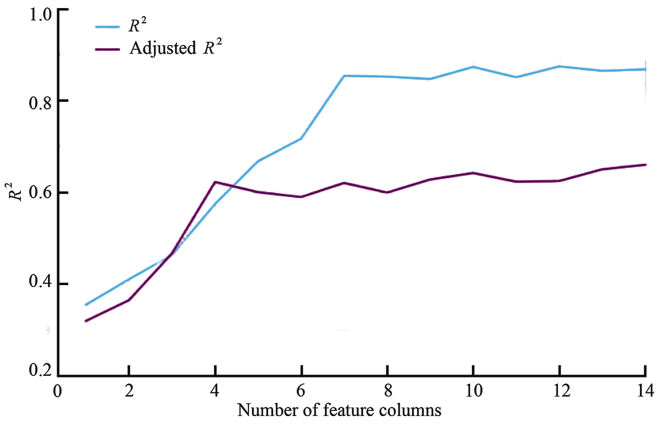

Fig. 15. Application of Frisch-Waugh-Lovell for pressurization rate dataset.

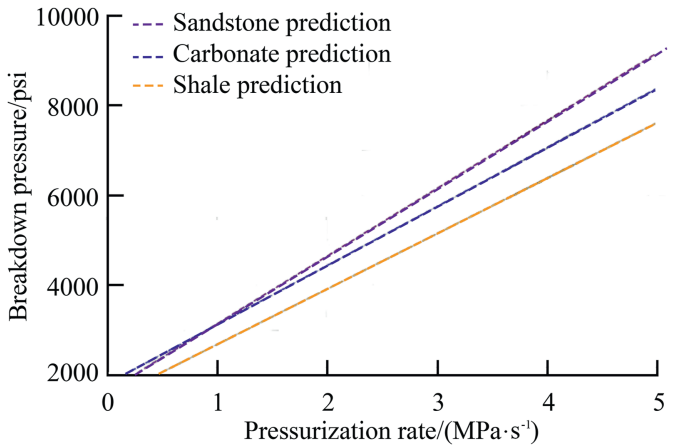

Fig. 16. Modeled relationship between pressurization rate and breakdown pressure for various rock types.

3.4. Re-training with top features

The models were re-trained using the top eight features identified from the sensitivity and SHAP analysis (Fig. 13).

The re-training results (Fig. 14) demonstrate that TabNet achieved superior performance in terms of mean absolute error (MAE) and mean squared error (MSE), highlighting its robustness in managing complex feature interactions. TabNet and CatBoost showed a slight reduction in R^2 values (0.93 for TabNet and 0.92 for CatBoost), potentially due to the exclusion of lower-ranked features that may still contribute marginal information. LightGBM maintained stable performance, while KAN exhibited a noticeable decline in prediction accuracy. This decline may indicate the model's sensitivity to feature diversity, suggesting the need for a potential restructuring of the neural network architecture to avoid overcomplexity.

3.5. Pressurization rate analysis

The analysis of pressurization rate as a predictor for breakdown pressure initially indicated a high positive correlation, which was unexpected and suggested potential anomalies in the dataset. This anomaly may be attributed to the limited availability of diverse pressurization rate data points, leading to a biased estimation in the initial rough analysis. To address this, the Frisch-Waugh-Lovell (FWL) technique was employed to better isolate the true effect of pressurization rate on breakdown pressure. The FWL method adjusts the correlation by controlling for the influence of other features, resulting in an adjusted R^2 value that better reflects the true relationship. As shown in the accompanying Fig. 15 while the standard R^2 score was inflated, the adjusted R^2 0.66 indicates a more realistic relationship that aligns well with previous studies on hydraulic fracturing and geomechanical modeling [15,78]. This result aligns with findings in recent literature that highlight the secondary influence of pressurization rate, compared to primary factors like confining pressure, injection rate, and rock mechanical properties.

Furthermore, the presented Fig. 16 illustrates the relationship between pressurization rate and breakdown pressure for three distinct rock types: sandstone, carbonate, and shale. The trend depicted in the graph indicates a linear increase in breakdown pressure as the pressurization rate rises across all rock types, aligning with previously reported patterns observed in study related to injection rate [79]. The strong linear relationship, as demonstrated by the parallel nature of the prediction lines, suggests that pressurization rate, similar to injection rate, significantly influences the breakdown pressure.

4. Conclusions

This study systematically explored the relationships and predictive capabilities of advanced machine learning models for estimating breakdown pressure (BP) based on diverse rock and experimental parameters. The following key conclusions can be drawn from the Results and Discussion:

(1) Model performance: among the evaluated models, TabNet demonstrated the highest R^2 score (0.94), effectively capturing complex feature interactions while maintaining robust generalization. KAN closely followed but showed signs of overcomplexity, while CatBoost and LightGBM achieved competitive accuracy. XGBoost displayed comparatively lower performance, emphasizing its sensitivity to hyperparameter optimization.
(2) Feature importance: confining pressure, injection rate, permeability, rock tensile strength, and Young's modulus emerged as the most influential factors for BP prediction. SHAP analysis and TabNet's attention scores validated these findings, offering a nuanced understanding of feature interactions.
(3) Correlation analysis: pearson and Spearman correlation methods provided valuable initial rankings but were limited by linearity assumptions. SHAP values and sensitivity analysis highlighted the complex, non-linear impacts of features such as Poisson's ratio, underlining the need for multi-faceted approaches in feature evaluation.

(4) Re-training outcomes: re-training models with the top features identified from sensitivity analysis and SHAP values confirmed TabNet's robustness, achieving the lowest MAE and MSE. However, slight reductions in accuracy for models like KAN suggested the potential need for architectural adjustments.
(5) Pressurization rate insights: analysis of pressurization rate revealed an initially inflated correlation, corrected by the Frisch-Waugh-Lovell method. The adjusted R^2 score (0.66) provided a more realistic representation of its role, reaffirming its secondary influence compared to primary factors like confining pressure and injection rate.
(6) Less feature influence: while parameters such as cylindricity, sample size, hole radius, and especially Poisson's ratio were initially hypothesized to have a significant impact on BP, their relatively low importance in the sensitivity analysis indicates that their correlation may be negligible in practice.

These findings emphasize the effectiveness of advanced ML techniques in modeling BP, highlighting the importance of careful feature selection and validation. The results also provide actionable insights into the critical parameters influencing BP, which can aid in optimizing hydraulic fracturing designs and operations.

CRediT authorship contribution statement

Subhan Aliyev: Writing - Original Draft, Methodology, Visualization, Software, Formal analysis, Data curation. **Talal Al Shafloot:** Supervision, Conceptualization, Writing - Review & Editing, Methodology. **Murtada Saleh Aljawad:** Supervision, Writing - Review & Editing, Validation. **Abdulazeez Abdulraheem:** Conceptualization, Methodology, Validation. **Salaheldin Elkatatny:** Supervision, Methodology, Writing - Review & Editing.

Funding

College of Petroleum Engineering & Geosciences at King Fahd University of Petroleum & Minerals.

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.

Acknowledgement

The authors would like to acknowledge the support of the College of Petroleum Engineering & Geosciences at King Fahd University of Petroleum & Minerals.

Nomenclature and Abbreviations

ML	Machine learning
BP	Breakdown pressure
UCS	Unconfined compressive strength
POR	Porosity
PR	Pressurization rate
FGR	Fracture geometry ratio
p_c	Confining pressure
Q	Injection rate
η	Fluid viscosity
E	Young's modulus
ν	Poisson's ratio
k	Permeability
σ_t	Rock tensile strength
s	Sample size
r,R	Hole radius
KAN	Kolmogorov-Arnold network
LightGBM	Light gradient boosting machine
CatBoost	Categorical boosting algorithm
XGBoost	Extreme gradient boosting
TabNet	Tabular neural network
n_estimators	Number of estimators
max_depth	Maximum depth
MAE	Mean absolute error
MSE	Mean squared error
RMSE	Root mean squared error
R^2	Coefficient of determination
AAPE	Average absolute percentage error
SD	Standard deviation
FWL	Frisch-Waugh-Lovell Theorem
SHAP	SHapley Additive exPlanations

Appendix

Table A1 presents the weights and biases of the fully connected neural network used in the Kolmogorov-Arnold network (KAN) model for predicting breakdown pressure. These weights represent the importance of each feature in influencing the model's output, while the biases adjust the model's predictions by introducing additional flexibility.

Table A1
Weights and biases of the fully connected neural networks.

N	Sandstone	Shale	Carbonate	Q	η	P_c	σ_t	UCS	E	ν	k	ϕ	Ratio	Cylindricity	R	Weights (ω_2)	Bias (b_1)	Bias (b_2)
1	−0.17	0.38	−0.65	−0.26	0.01	0.49	−0.46	−0.33	−0.48	−0.07	−0.9	0.16	0.76	0.3	0.34	−0.98	0.03	−0.04
2	−0.04	−0.12	−0.8	−0.14	0.68	−0.6	0.65	−0.95	−0.51	0.94	0.57	0.27	−0.15	0.24	0.12	−0.06	−0.14	−0.61
3	−0.96	−0.49	0.32	−0.12	0.47	0.92	0	−0.95	0.81	0.22	−0.6	0.6	−0.68	−0.3	−0.78	−0.4	−0.37	0.22
4	−0.48	0.68	0.53	0.23	0.08	−0.26	−0.85	0.66	−0.5	−0.3	−0.48	−0.21	−0.97	0.68	−0.11	0.2	−0.13	−0.44
5	0.52	−0.92	−0.47	0.89	0.18	−0.35	−0.88	−0.45	−0.46	−0.77	−0.67	0.83	0.12	−0.06	−0.08	−0.41	0.55	−0.59
6	−0.73	0.8	−0.96	−0.52	0.02	−0.7	−0.33	0.04	0.52	−0.7	−0.34	0.07	0.05	0.96	0.73	−0.4	0.2	0.03
7	0.07	−0.08	−0.84	−0.76	−0.4	−0.39	0.57	−0.4	−0.1	−0.55	0.51	−0.68	0.44	0.27	0.09	0.49	0.79	−0.99
8	−0.57	0.27	0.94	−0.61	0.13	0.75	0.42	0.88	0.55	−0.5	0.04	0.39	0.78	−0.75	−0.24	−0.9	−0.11	−0.98
9	−0.98	0.32	−0.41	0.77	0.38	0.99	0.58	−0.48	−0.87	0.7	−0.59	0.59	−0.84	0.35	0.95	0.81	0.21	−0.56
10	−0.52	0.79	0.54	0.29	0.75	−0.26	0.03	−0.14	−0.02	0.12	0.76	−0.37	0.46	−0.35	−0.78	0.7	0.26	−0.93
11	0.95	0.27	0.25	−0.43	0.27	−0.1	−0.12	0.75	−0.93	0.05	0.76	0.71	−0.63	0.37	−0.15	0.34	0.18	−0.78
12	0.6	0.23	−0.24	0.63	0.52	0.44	−0.71	0.68	−0.87	−0.77	0.74	0.81	0.72	−0.86	−0.92	0.19	0.41	−0.32
13	0.92	−0.87	−0.59	0.72	−0.68	0.77	−0.34	−0.63	0.81	0.72	−0.52	−0.45	0.64	−0.65	0.48	0.78	−0.53	0.61
14	−0.02	0.04	−0.76	0.69	−0.08	0.19	−0.13	0.61	−0.72	0.45	−0.1	0.97	0.08	0.71	0.84	−0.63	0.02	0.14
15	−0.78	−0.7	0.23	0.84	−0.98	−0.22	−0.82	−0.08	0.06	−0.86	0.97	−0.72	0.42	−0.55	−0.44	−0.84	−0.79	0.03
16	0.1	0.47	0.55	−0.5	−0.51	−0.17	−0.56	−0.03	−0.18	0.42	0.54	−0.6	−0.37	0.67	0.72	−0.52	−0.23	−0.41
17	−0.09	0.02	0.29	0.51	0.45	0.39	0.2	−0.73	−0.31	0.09	−0.95	−0.63	−0.06	−0.44	−0.42	0.59	−0.02	0.86
18	0.69	0.36	0.06	−0.08	0.98	−0.99	0.47	−0.84	0.8	−0.84	−0.87	0.79	0.64	0.29	0.82	−0.93	0.3	−0.21
19	−0.8	−0.92	−0.92	0.68	−0.8	0.24	0.24	0.46	−0.96	−0.08	−0.07	0.31	−0.08	0.39	0.51	0.17	0.9	−0.83
20	−0.02	−0.83	0.94	0.46	−0.2	−0.29	0.87	−0.01	0.33	−0.03	0.82	−0.7	−0.28	0.03	0.61	0.99	0.2	0.23
21	−0.7	0.43	0.6	0.55	0.6	0.59	0.29	−0.13	0.93	−0.67	0.08	−0.12	−0.01	−0.39	−0.96	0.71	0.49	−0.77
22	−0.35	−0.86	−0.41	0.31	−0.59	−0.81	−0.16	0.46	0.12	0.89	0.2	0.23	0.66	−0.57	0.93	0.04	0.01	−0.31
23	0.47	−0.86	0.96	−0.65	0.11	0.18	0.27	0.53	0.87	0.7	−0.79	−0.83	−0.33	−0.93	0.45	−0.87	0.27	0.01
24	−0.05	−0.98	0.2	0.09	0.47	−0.04	0.57	−0.68	−0.9	0.34	0.31	0.76	−0.65	−0.39	−0.39	0.66	−0.86	0.75
25	−0.25	0.91	0.16	0.97	0.23	0.28	−0.76	0.22	−0.16	−0.08	0.64	0.61	0.42	0.31	0.66	0.2	−0.49	−0.01
26	−0.21	0.48	0.5	0.87	−0.62	−0.87	−0.18	−0.73	−0.48	−0.18	−0.24	0.01	0.65	0.88	−0.44	−0.77	−0.28	0.4
27	−0.08	−0.29	0.62	−0.91	−0.29	0.16	0.68	0.5	0.46	0.3	0.55	0.93	−0.8	0.74	0.75	−0.81	−0.06	0.99
28	0.57	−0.41	0.31	−0.67	0.57	0.12	−0.23	0.31	0.96	0.09	0.93	−0.16	−0.52	0.53	−0.77	0.82	−0.91	−0.74
29	0.78	−0.3	−0.74	−0.74	0.11	0.12	0.14	0.91	−0.49	−0.88	−0.59	0.97	−0.72	0.58	0.41	0.34	−0.72	−0.45
30	0.91	0.55	−0.32	0.45	−0.99	0.21	0.18	−0.86	0.31	0.03	0.05	0.34	−0.3	0.33	0.08	0.66	−0.45	−0.21
31	0.57	0.32	0.86	0.64	0.52	0.35	−0.63	−0.89	−0.6	0.61	−0.43	0.27	−0.1	−0.48	−0.81	0.76	0.94	−0.16
32	−0.37	−0.63	−0.55	−0.57	−0.93	0.61	−0.28	−0.44	0.13	−0.08	0.59	−0.67	0.52	0.11	−0.52	0.14	−0.34	−0.18

References

[1] J.C. Glorioso, A. Rattia, Unconventional reservoirs: basic petrophysical concepts for shale gas, in: SPE/EAGE European Unconventional Resources Conference and Exhibition, Spe, 2012. SPE-153004.

[2] Q. Li, H. Xing, J. Liu, A review on hydraulic fracturing of unconventional reservoir, Petroleum 1 (1) (2015) 8–15.

[3] J. Guo, Y. Liu, A comprehensive model for simulating fracturing fluid leakoff in natural fractures, J. Nat. Gas Sci. Eng. 21 (2014) 977–985.

[4] G.A. Al-Muntasheri, A critical review of hydraulic-fracturing fluids for moderate-to ultralow-permeability formations over the last decade, SPE Prod. Oper. 29 (4) (2014) 243–260.

[5] X. Zhao, Mechanical behavior of rocks under complex stresses and high temperatures: effects on breakdown pressure, Int. J. Rock Mech. Min. Sci. 86 (2016) 90–102.

[6] E.K. Watkins, C.L. Wendorff, B.R. Ainley, A new crosslinked foamed fracturing fluid, in: SPE Annual Technical Conference and Exhibition?, SPE, 1983. SPE-12027.

[7] S. Al-Hajri, B.M. Negash, M.M. Rahman, M. Haroun, T.M. Al-Shami, Perspective Review of polymers as additives in water-based fracturing fluids, ACS Omega 7 (9) (2022) 7431–7443.

[8] M.J. Almarri, M.J. AlTammar, K.M. Alruwaili, S. Zheng, Numerical feasibility of near-wellbore cooling as a novel method for reducing breakdown pressure in hydraulic fracturing, J. Nat. Gas Sci. Eng. 102 (2022) 104549.

[9] B. S, B.M. Aadnoy, Elasto-thermo-poroelastic model for stress, pore pressure and temperature around a wellbore, J. Petrol. Sci. Eng. 60 (1) (2008) 1–13.

[10] I. Berchenko, E. Detournay, Deviation of hydraulic fractures through poroelastic stress changes induced by fluid injection and pumping, Int. J. Rock Mech. Min. Sci. 34 (6) (1997) 1009–1019.

[11] H. Yang, et al., Breakdown pressure prediction of tight sandstone horizontal wells based on the mechanism model and multiple linear regression model, Energies 15 (19) (2022) 6944.

[12] J. Zhang, Pore pressure prediction from well logs: methods, modifications, and new approaches, Earth Sci. Rev. 108 (1–2) (2011) 50–63.

[13] W. Zhu, Z. Chen, X. He, Z. Tian, M. Wang, Numerical investigation of influential factors in hydraulic fracturing processes using coupled discrete element-lattice Boltzmann method, J. Geophys. Res. Solid Earth 128 (9) (2023) e2023JB027292.

[14] L. Huang, X. Liao, M. Fan, S. Wu, P. Tan, L. Yang, Experimental and numerical simulation technique for hydraulic fracturing of shale formations, Adv. Geo-Energy Res. 13 (2) (2024) 83–88.

[15] R. Carbonell, E. Detournay, Modeling fracture initiation and propagation from a pressurized hole: a dislocation-based approach, in: ARMA US Rock Mechanics/Geomechanics Symposium, ARMA, 1995. ARMA-95.

[16] F. Guo, N.R. Morgenstern, J.D. Scott, Interpretation of hydraulic fracturing breakdown pressure, in: International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts, Elsevier, 1993, pp. 617–626.

[17] N. Morita, A.D. Black, G.-F. Fuh, Borehole breakdown pressure with drilling fluids—I. Empirical results, in: International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts, Elsevier, 1996, pp. 39–51.

[18] Y. He, et al., Investigation of low water recovery based on gas-water two-phase low-velocity Non-Darcy flow model for hydraulically fractured horizontal wells in shale, Petroleum 9 (3) (2023) 364–372.

[19] H. Peng, et al., Influence of supercritical CO_2 on the physical property of tight sandstone, Petroleum 10 (3) (2024) 520–526.

[20] T. Ishida, et al., Acoustic emission monitoring of hydraulic fracturing laboratory experiment with supercritical and liquid CO_2, Geophys. Res. Lett. 39 (16) (2012).

[21] Y. Gong, M. Mehana, I. El-Monier, F. Xu, F. Xiong, Machine learning for estimating rock mechanical properties beyond traditional considerations, in: Unconventional Resources Technology Conference, Unconventional Resources Technology Conference (URTeC); Society of, Denver, Colorado, 2019, pp. 466–480, 22-24 July 2019.

[22] T. Almani, K. Khan, Predicting Formation breakdown pressures for arbitrary wellbore orientation using machine learning techniques, in: ARMA/DGS/SEG International Geomechanics Symposium, ARMA, 2023 (ARMA-IGS).

[23] X. Xuesong, Y. Junliang, L.I. Zhonghui, S.U.N. Chong, Z. Yi, Determination of formation fracture pressure under high temperature and high pressure in deep water of the South China Sea, Petrol. Drill. Tech. 51 (6) (2023) 18–24.

[24] M.K. Hubbert, D.G. Willis, Mechanics of hydraulic fracturing, J. Petrol. Technol. 9 (6) (1957) 153–168, 1957.

[25] B. Haimson, C. Fairhurst, Initiation and extension of hydraulic fractures in rocks, Soc. Petrol. Eng. J. 7 (3) (1967) 310–318.

[26] X. Jin, S.N. Shah, J.-C. Roegiers, B. Hou, Breakdown pressure determination-a fracture mechanics approach, in: SPE Annual Technical Conference and Exhibition?, SPE, 2013 D011S006R006.

[27] N. Li, H. Xie, Z. Gao, C. Li, Study on the hydraulic fracturing failure behaviour of granite and its comparison with gas fracturing, Sustainability 14 (21) (2022) 14593.

[28] D. Lockner, J.D. Byerlee, Hydrofracture in Weber sandstone at high confining pressure and differential stress, J. Geophys. Res. 82 (14) (1977) 2018–2026.

[29] Y. Wu, J. Tao, J. Wang, Y. Zhang, S. Peng, Experimental investigation of shale breakdown pressure under liquid nitrogen pre-conditioning before nitrogen fracturing, Int. J. Min. Sci. Technol. 31 (4) (2021) 611–620.

[30] E. Fjar, R.M. Holt, A.M. Raaen, R. Risnes, P. Horsrud, Petroleum Related Rock Mechanics, Elsevier, 2008.

[31] C.A.I. Zhenzhong, et al., Geomechanics modeling of ultra-deep fault-controlled carbonate reservoirs and its application in development, Petrol. Geol. Exp. 46 (4) (2024) 868–879.

[32] D. Zhou, P. He, Major factors affecting simultaneous frac results, in: SPE Oklahoma City Oil and Gas Symposium/Production and Operations Symposium, SPE, 2015. SPE-173633.

[33] T. Almani, Y. Alquhaidan, K. Khan, Sensitivity analysis of the factors affecting formation breakdown pressure using a computational-based approach, in: ARMA/DGS/SEG International Geomechanics Symposium, ARMA, 2023 (ARMA-IGS).

[34] K.E. Gray, C.M. Kim, B.J. Hughes, Stress state, porosity, permeability and breakdown pressure around A borehole during fluid injection, in: ARMA US Rock Mechanics/Geomechanics Symposium, ARMA, 1981. ARMA-81.

[35] T. Ito, Effect of pore pressure gradient on fracture initiation in fluid saturated porous media: rock, Eng. Fract. Mech. 75 (7) (2008) 1753–1762.

[36] A.P. Bunger, A. Lakirouhani, E. Detournay, Modelling the effect of injection system compressibility and viscous fluid flow on hydraulic fracture breakdown pressure, in: ISRM International Symposium on In-Situ Rock Stress, ISRM, 2010 (ISRM-ISRS).

[37] M.J. AlTammar, D.P. Gala, M.M. Sharma, Application of different fluid injection methods to reduce breakdown pressure, in: International Petroleum Technology Conference, IPTC, 2020 D031S077R002.

[38] M. Kong, Z. Zhang, C. Zhao, H. Chen, X. Ma, Y. Zou, Rock mechanical properties and breakdown pressure of high-temperature and high-pressure reservoirs in the Southern Margin of Junggar Basin, Geofluids 2021 (1) (2021) 1116136.

[39] M. Bai, Why are brittleness and fracability not equivalent in designing hydraulic fracturing in tight shale gas reservoirs, Petroleum 2 (1) (2016) 1–19.

[40] Y. Zhang, J. Zhang, B. Yuan, S. Yin, In-situ stresses controlling hydraulic fracture propagation and fracture breakdown pressure, J. Pet. Sci. Eng. 164 (2018) 164–173.

[41] T. Ito, K. Hayashi, Physical background to the breakdown pressure in hydraulic fracturing tectonic stress measurements, in: International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts, Elsevier, 1991, pp. 285–293.

[42] S. Wang, Y. Han, W. Hu, X. Zhao, L. Zhang, T. Wang, Modeling interactions between hydraulic fracture and pre-existing microcracks in crystalline rocks using hydro-grain-texture model, Geoenergy Sci. Eng. 244 (2025) 213459, https://doi.org/10.1016/j.geoen.2024.213459.

[43] C. Jiang, Y. Lei, B. Deng, X. Fan, M. Li, M. Wu, Effect of mesoscopic heterogeneity on hydraulic fracturing characteristics of granite under high temperature, Geoenergy Sci. Eng. 246 (2025) 213626, https://doi.org/10.1016/j.geoen.2024.213626.

[44] Wei Haifeng, Research progress on fracture propagation patterns of hydraulic fracturing in heterogeneous shale, Petrol. Geol. Rec. Eff. 30 (4) (2023) 156–166.

[45] F. Maojun, D.U. Xulin, B.A.I. Yuhu, L.I. Hao, Z. Hao, Z.H.U. Haiyan, Three-dimensional physical simulation experiments on large-scale hydraulic fracturing in multi-thin interbedded tight sandstone reservoirs, Petrol. Geol. Exp. 46 (4) (2024) 786–798.

[46] R.M. Tuzingila, L. Kong, R.K. Kasongo, A review on experimental techniques and their applications in the effects of mineral content on geomechanical properties of reservoir shale rock, Rock Mech. Bull. (2024) 100110.

[47] J. Wang, D. Elsworth, Y. Wu, J. Liu, W. Zhu, Y. Liu, The influence of fracturing fluids on fracturing processes: a comparison between water, oil and SC-CO 2, Rock Mech. Rock Eng. 51 (2018) 299–313.

[48] Z. Tariq, et al., Machine learning-based accelerated approaches to infer breakdown pressure of several unconventional rock types, ACS Omega 7 (45) (2022) 41314–41330.

[49] R. Roelofs, et al., A meta-analysis of overfitting in machine learning, Adv. Neural Inf. Process. Syst. 32 (2019).

[50] K. Ghosh, C. Bellinger, R. Corizzo, P. Branco, B. Krawczyk, N. Japkowicz, The class imbalance problem in deep learning, Mach. Learn. 113 (7) (2024) 4845–4901.

[51] N.U. Niaz, K.M.N. Shahariar, M.J.A. Patwary, Class imbalance problems in machine learning: a review of methods and future challenges, in: Proceedings of the 2nd International Conference on Computing Advancements, 2022, pp. 485–490.

[52] A. Muqtadir, S.M. Elkatatny, M.A. Mahmoud, A. Abdulraheem, A. Gomaa, Effect of saturating fluid on the geomechanical properties of low permeability scioto sandstone rocks, in: ARMA US Rock Mechanics/Geomechanics Symposium, ARMA, 2018. ARMA-2018.

[53] A. Farid Ibrahim, H. Nasr-El-Din, Evaluation of the breakdown pressure to initiate hydraulic fractures of tight sandstone and shale formations, in: SPE Trinidad and Tobago Section Energy Resources Conference?, SPE, 2018 D011S020R002.

[54] A. Mathur, et al., Crude steady-state permeability measurements on carbonate source rocks and effect of water block from fracturing fluid on productivity, in: International Petroleum Technology Conference, IPTC, 2024. IPTC-24536.

[55] Y. Zou, N. Li, X. Ma, S. Zhang, S. Li, Experimental study on the growth behavior of supercritical CO_2-induced fractures in a layered tight sandstone formation, J. Nat. Gas Sci. Eng. 49 (2018) 145–156.

[56] D.T. Nicholson, Pore properties as indicators of breakdown mechanisms in experimentally weathered limestones, Earth Surf. Process. Landf.: J. British Geomorphol. Res. Group 26 (8) (2001) 819–838.
[57] V. Marcon, et al., Experimental insights into geochemical changes in hydraulically fractured Marcellus Shale, Appl. Geochem. 76 (2017) 36–50.
[58] A.F. Siggins, Velocity-effective stress response of CO_2-saturated sandstones, Explor. Geophys. 37 (1) (2006) 60–66.
[59] S. Chatterji, P. Christensen, A mechanism of breakdown of limestone nodules in a freeze-thaw environment, Cement Concr. Res. 9 (6) (1979) 741–746.
[60] B. Haimson, C. Fairhurst, Hydraulic fracturing in porous-permeable materials, J. Petrol. Technol. 21 (7) (1969) 811–817.
[61] I. Song, M. Suh, K.S. Won, B. Haimson, A laboratory study of hydraulic fracturing breakdown pressure in tablerock sandstone, Geosci. J. 5 (2001) 263–271.
[62] F. Wu, X. Fan, J. Liu, X. Li, Analytical interpretation of hydraulic fracturing initiation pressure and breakdown pressure, J. Nat. Gas Sci. Eng. 76 (2020) 103185.
[63] S. Mishra, A. Datta-Gupta, Applied Statistical Modeling and Data Analytics: A Practical Guide for the Petroleum Geosciences, Elsevier, 2017.
[64] G. Ke, et al., Lightgbm: a highly efficient gradient boosting decision tree, Adv. Neural Inf. Process. Syst. 30 (2017).
[65] T. Yang, et al., Quantitative classification and prediction of pore structure in low porosity and low permeability sandstone: a machine learning approach, Geoenergy Sci. Eng. (2025) 213708, https://doi.org/10.1016/j.geoen.2025.213708.
[66] A. Anghel, N. Papandreou, T. Parnell, A. De Palma, H. Pozidis, Benchmarking and Optimization of Gradient Boosting Decision Tree Algorithms, 2018 arXiv preprint arXiv:1809.04559.
[67] A. Choudhury, A. Mondal, S. Sarkar, Searches for the BSM scenarios at the LHC using decision tree-based machine learning algorithms: a comparative study and review of random forest, AdaBoost, XGBoost and LightGBM frameworks, Eur. Phys. J. Spec. Top. (2024) 1–39.
[68] L. Prokhorenkova, G. Gusev, A. Vorobev, A.V. Dorogush, A. Gulin, CatBoost: unbiased boosting with categorical features, Adv. Neural Inf. Process. Syst. 31 (2018).
[69] T. Chen, C. Guestrin, Xgboost: a scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[70] Y. Meng, N. Yang, Z. Qian, G. Zhang, What makes an online review more helpful: an interpretation framework using XGBoost and SHAP values, J. Theoret. Appl. Elect. Com. Res. 16 (3) (2020) 466–490.
[71] S.Ö. Arik, T. Pfister, Tabnet: attentive interpretable tabular learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 6679–6687.
[72] Z. Liu, et al., Kan: Kolmogorov-Arnold Networks, 2024 arXiv preprint arXiv:2404.19756.
[73] Y. Hou, D. Zhang, A Comprehensive Survey on Kolmogorov Arnold Networks (Kan), 2024 *arXiv preprint arXiv:2407.11075*.
[74] X. Wu, et al., Experimental study on cyclic hydraulic fracturing of tight sandstone under in-situ stress, Processes 11 (3) (2023) 875.
[75] S. Jung, M.B. Diaz, K.Y. Kim, H. Hofmann, G. Zimmermann, Fatigue behavior of granite subjected to cyclic hydraulic fracturing and observations on pressure for fracture growth, Rock Mech. Rock Eng. (2021) 1–14.
[76] Z. Guo, S. Tian, Q. Liu, L. Ma, Y. Yong, R. Yang, Experimental investigation on the breakdown pressure and fracture propagation of radial borehole fracturing, J. Pet. Sci. Eng. 208 (2022) 109169.
[77] R.D. Barree, J.L. Miskimins, Calculation and implications of breakdown pressures in directional wellbore stimulation, in: SPE Hydraulic Fracturing Technology Conference and Exhibition, SPE, 2015 D011S001R007.
[78] D. Garagash, E. Detournay, An analysis of the influence of the pressurization rate on the borehole breakdown pressure, Int. J. Solid Struct. 34 (24) (1997) 3099–3118.
[79] B. Lecampion, Hydraulic fracture initiation from an open-hole: wellbore size, pressurization rate and fluid-solid coupling effects, in: ARMA US Rock Mechanics/Geomechanics Symposium, ARMA, 2012. ARMA-2012.