A Novel Paradigm for Parameter Optimization of Hydraulic Fracturing Using Machine Learning and Large Language Model

Chunxi Yang[1], Chuanyou Xu[2], Yue Ma[3], Bang Qu[4], Yiquan Liang[5], Yajun Xu[6], Lei Xiao[7], Zhimin Sheng[8], Zhenghao Fan[9], Xin Zhang*[,10]
DownHole Service Company, CNPC XiBu Drilling Engineering Company Limited, Karamay 834000, China[1,2,3,4,5,6,7,8]
School of Civil Engineering, Chongqing University, Chongqing 400045, China[9]
School of Big Data & Software Engineering, Chongqing University, Chongqing 400044, China[10]

***Abstract*—Hydraulic fracturing is a common practice in the oil and gas industry meant to increase the production of oil and natural gas. In this process, appropriate fracturing design parameters are important to maximize the efficiency of fracture propagation. However, conventional fracturing parameter design methods often rely on expert experience or fail to take into account complex geological conditions, resulting in suboptimal parameter design schemes. Therefore, this paper presents PPOHyFrac, a novel paradigm for optimizing hydraulic fracturing parameters with large language model and machine learning, which aims to automatically extract, assess and optimize fracturing parameters. PPOHyFrac uses advanced large language model to perform the extraction of key parameters from hundreds of fracturing design documents, and then refines the extracted data using statistical methods such as missing value imputation and feature normalization. Besides, the techniques in correlation analysis are utilized to identify key influencing factors and finally machine learning methods are implemented to optimize and predict the key influencing factors. This paper also presents a comparative study of five machine learning methods. Experiments show that random forest is the best choice for parameter optimization and can improve the prediction and optimization accuracy of key parameters.**

***Keywords*—Hydraulic fracturing; parameter optimization; large language model; machine learning**

I. Introduction

The global requirement for energy is increasing and never-ending, leading to the increased demand to produce more natural resources, such as oil and natural gas [1]. Hydraulic fracturing, which is a technique that improves oil and gas recovery worldwide, stands to improve high production efficiency since it boosts flow movement of hydrocarbons into low-permeability reservoirs. In this technique, a fluid mixture with extremely high pressure is injected into the reservoir to create fractures, and then proppants are used to keep the fractures open so that oil or natural gas can flow smoothly into the wellbore, and finally achieves the purpose of increasing the production of oil wells [2]. Hydraulic fracturing improves the efficiency in production and recovery rates through the increased permeability of the rocks, which makes it an indispensable technology in modern resource extraction [3].

However, optimizing the parameters from hydraulic fracturing becomes a tough task due to the multiplicity of elements involved, such as geological conditions, the propagation behavior of the fracture and rock mechanics. These components tend to interact with one another frequently in a nonlinear way, which makes it difficult to predict the effects that a given design will have on the system. All of these add up to the need to have a thorough understanding of reservoir dynamics and the ability to sensibly tweak specific designs to particular conditions [4].

In the traditional sense, hydraulic fracturing optimization has largely relied on the experience of professionals and numerical simulation [5], [6], [7], [8]. Although these solutions can provide initial findings under certain conditions, there are limitations in traditional hydraulic fracturing optimization methods. The expert experience-based approaches are usually historical and based on the operator's expertise, while numerical simulation approaches usually take more time to execute and require updating every time new information is input. In recent years, the rapid development of data-driven methods, such as machine learning [9], [10], deep learning [11], [12], and data mining [13], have greatly improved the ability to model complex systems. These methods are good at extracting potential patterns from large-scale datasets and identifying relationships between variables, providing new ways to optimize the production performance and design parameters of hydraulic fracturing [14], [15], [16], [17].

Inspired by the rapid development of data-driven methods, this paper proposes the implementation of a data-driven PPOHyFrac for optimizing hydraulic fracturing parameters using large language model (LLM) [18] and machine learning techniques [19] to systematically extract, analyze, and optimize key fracturing parameters. The locally deployed large language model QWen2.5 enables PPOHyFrac to automatically extract key parameters defined by experts and build a high-quality dataset. Data preprocessing and statistical analysis can help identify and extract key parameters affecting the general design of the overall fracturing scheme. For these extracted key parameters, the model employs five classic machine learning algorithms for prediction and optimization purposes, finally determining the random forest algorithm as the optimum strategy.The main contributions are as follows.

- We proposed a data-driven PPOHyFrac for optimizing hydraulic fracturing parameters, which integrates an LLM with traditional machine learning algorithms to

systematically extract, analyze and optimize key fracturing parameters, thus enhancing oil well fracturing production efficiency.

- Through the dedicated local LLM, a database of hundreds of fracturing design documents from an oilfield in China is constructed. The dataset spans a number of fracturing modes, namely conventional fracturing, repeated fracturing, and multi-stage fracturing, creating a rich basis for subsequent analysis and model development.
- The correlation analysis enables identification of a potential association among the retrieved fracturing parameters. Experimental evidence suggests that Average Proppant-to-Liquid Ratio and Preflush Percentage are the most important parameters affecting fracturing performance.
- A comparative study of five different machine learning techniques, such as neural networks, random forest, linear regression, Bayesian ridge regression, and ridge regression, shows that random forest is better than other techniques, thereby providing the best result along with its predictions for optimizing fracturing parameters.

The remainder of the paper is organized as follows: Section II introduces the related work. Section III describes the methodology of our work. Section IV "Experiment" has detailed the experimental setup and results. And Section V "Conclusion" summarizes our work and explains its practical application.

II. Related Work

Hydraulic fracturing is an important technology to increase the production of aging oil wells, as it improves the flow efficiency of gas and oil by creating fractures in the reservoir rock. To achieve optimal economic benefits and operational performance, it is essential to optimize the key parameters in hydraulic fracturing. This section reviews the literature on hydraulic fracturing parameter optimization. By analyzing the strengths and limitations of these methods, we highlight the motivation to develop PPOHyFrac proposed in this study.

A. Methods Based on Expert Experience

Methods based on expert experience have long been a cornerstone in the optimization of hydraulic fracturing parameters, particularly during the early development of the technology [20]. Commonly, these methods are effective when geological conditions close to the oil well appear to be relatively clear but may fail in cases of greater complexity or greater uncertainty. As noted by Mata and Zhou [21], these approaches usually struggle in scenarios involving complex geological conditions, where they may not be easily configured to the dynamic and diverse character of geological environments, leading to inefficiency and unsatisfactory results.

In addition, the reliance of personal experience and expertise is also evident in the process of selecting parameters, which is the inherent limitation of expert-based methods [22]. Miskimins et al. further pointed out [23] that although expert methods are valuable, they must be complemented by advanced modeling and data analysis to address the challenges of today's unconventional reservoirs. Despite the defects mentioned above, expert-based methods are still an indispensable part of PPOHyFrac, as the final determination of key fracturing parameters still requires in-depth participation of experts.

B. Methods Based on Numerical Simulation

Numerical simulation methods simulate fluid flow, rock deformation and fracture propagation through computational models to predict fracture behavior [24]. These methods take a wide range of geological variables into account, and thus provide better predictability than traditional methods [25].

Early numerical simulation methods relied on classical models, such as the Kristianovich-Geertsma-de Klerk (KGD) model and the Perkins-Kern-Nordgren (PKN) model [26]. These models usually perform well under relatively simple geological conditions. However, they are based on some oversimplified assumptions, such as linear elastic fracture mechanics (LEFM), which assumes the formation is homogeneous, isotropic and exhibits in the linear way. But according to Yang et al. [27], the actual formations are generally heterogeneous and anisotropic, which greatly limits the scope of applications of these methods.

In recent years, with the continuous advancement of computer hardware and numerical algorithms, advanced numerical simulation methods such as the extended finite element method (XFEM) [28] and discrete element method (DEM) [29] have been widely developed and applied to hydraulic fracturing simulation, which has significantly improved the simulation accuracy. These models overcome the limitations of traditional models in representing complex geological conditions, making the numerical results more representative of the actual environment. However, these methods also demand more powerful computing resources and processing time, which still poses challenges in practical application.

C. Data Driven Approach

Recent advancements in deep learning and machine learning have brought new solutions to hydraulic fracturing optimization. These data-driven approaches are especially good at capturing complex relationships between parameters, which indicates a promising prospect for optimizing fracturing parameters [30].

Lizhe et al. [31] proposed a method that integrates numerical simulation with machine learning to optimize the production performance of hydraulic fracturing. They designed a novel neural network (NN) structure to predict the net present value (NPV) of fracture parameters through a pre-NN, and transferred the learned weights to the main-NN to predict the NPV of the treatment parameters. Morozov et al. [32] constructed a digital database containing data from more than 5,000 multi-stage hydraulic fracturing operations in western Siberia, and applied the CatBoost algorithm to develop a production performance prediction model, achieving an R^2 accuracy of 0.815, which builds a crucial foundation for further optimizing hydraulic fracturing design parameters.

Despite these successes, data-driven methods still face challenges. Many existing methods are limited to certain aspects

of the optimization process of the hydraulic fracturing design. In addition, the comprehensive integration of data acquisition, processing, and parameter optimization into a single workflow remains a challenging task.

D. Summary of Limitations and Research Gap

Expert-based methods, while crucial in providing information, generally tend to be subjective and not scalable. Numerical simulations, on the other hand, improve the accuracy of the forecast but are based on some oversimplified assumptions and require excessive computational resources. Modern data-driven approaches are indeed more promising, but still tend to address narrow aspects of the optimization space. The limitations mentioned highlight the fact that there is a need for an overarching framework that supports the efficient extraction of data, detailed statistical analysis, and advanced machine learning strategies intended to improve hydraulic fracturing parameters for diverse operating environments. This identified requirement catalyzes the creation of our suggested framework, PPOHyFrac, which utilizes a locally implemented large language model (LLM) combined with traditional machine learning strategies to extract, analyze, and optimize key fracturing parameters systematically.

III. Methodology

The proposed solution has streamlined hydraulic fracturing optimization by extracting parameters, analyzing key parameters, and predicting significant results using machine learning algorithms, as represented in Fig. 1.

A. Parameter Extraction

1) Data Acquisition: Fracturing design documents, which usually exist in unstructured formats, hold plenty of crucial information essential for optimizing hydraulic fracturing operations. These documents are the foundation upon which most data-driven methodologies are built, but the unstructured and heterogeneous nature of these documents makes it difficult to apply traditional data extraction methods, which forces the use of advanced natural language processing techniques to automate and streamline data extraction processes.

To this end, a locally deployed QWen2.5-7B large language model was employed to ensure both data security and scalability for efficient access. The model extracted six key parameters from the unstructured fracturing design documents described in Table I. The reasons for choosing QWen2.5-7B are as follows:

1) Although models with more parameters usually offer higher accuracy, they also require more resources and time. QWen2.5 with 7B parameters has shown enough accuracy for content extraction without too much resource crunch, time, and effort;
2) QWen 2.5 - 7B can accurately follow instructions, generate long text, understand unstructured data format, such as docx, and produce structured formats like JSON, and thereby ensure an exhaustive and organized parameter extraction;
3) The robustness of QWen2.5-7B against different types of tasks has enabled it to sustain a high level of processing performance across diverse document structures and formats;
4) The considerable extent of the context length supported by QWen2.5-7B guarantees that complex documents can be processed as a whole, preserving contextual information and improving parameter extraction accuracy.

TABLE I. Extracted Parameters from Hydraulic Fracturing Design Documents and Description

Parameter	Description
Total Fluid Volume	The total volume of fluid injected during the fracturing process, which typically includes water, chemicals, and other additives. It is a key factor influencing the fracture propagation and overall efficiency of the fracturing job.
Average Proppant-to-Liquid Ratio	The ratio of proppant (sand or other materials) to the *Total Fluid Volume*. This ratio determines the effectiveness of the fracture in terms of proppant transport, fracture conductivity, and the ability to keep fractures open under pressure.
Preflush Percentage	The proportion of fluid used before the main fracturing fluid, typically designed to help improve proppant transport or clean the formation. It is crucial in optimizing the overall fluid performance and enhancing fracture efficiency.
Fracturing Fluid Type	The composition of the fluid used in the fracturing process, which can vary from water-based to oil-based or gel-based fluids. The fluid type affects fracture fluid properties such as viscosity, temperature stability, and proppant suspension ability.
Proppant Type	The material used to prop open fractures, typically sand, ceramic beads, or other engineered materials. The choice of *Proppant Type* influences fracture conductivity, proppant flowback, and long-term fracture performance.
Pumping Rate	The rate at which fracturing fluid is injected into the wellbore. It influences fracture initiation, propagation, and the overall pressure profile within the reservoir. A high *Pumping Rate* may lead to more extensive fractures, but careful management is required to prevent damage to the formation.

As illustrated in Fig. 1, the LLM was equipped with well-designed templates of instructions, which could systematically mark the target parameters across various document types such as free-text, tables, and mixed layouts.

Apart from targeting the extraction of critical parameters from a wide variety of unstructured fracturing design documents, the PPOHyFrac also does this in a reliable way and thus establishes a solid groundwork for future data analysis and machine learning model development.

2) Missing Value Imputation: While LLM can successfully automate parameter extraction, the final dataset has missing values that result from inconsistent initial design documents. Therefore, a non-parametric algorithm, the K-Nearest Neighbors (KNN) imputation technique [33], is used to fill in the missing parts. The KNN imputation technique estimates the values of the missing parts based on the similarity of the observations, which enables the filling values to have the same distribution as the original data. It works in this way:

1) For each observation with missing values, calculate the Euclidean distance to all other observations using the available data. In an n-dimensional feature space, the distance between two observations

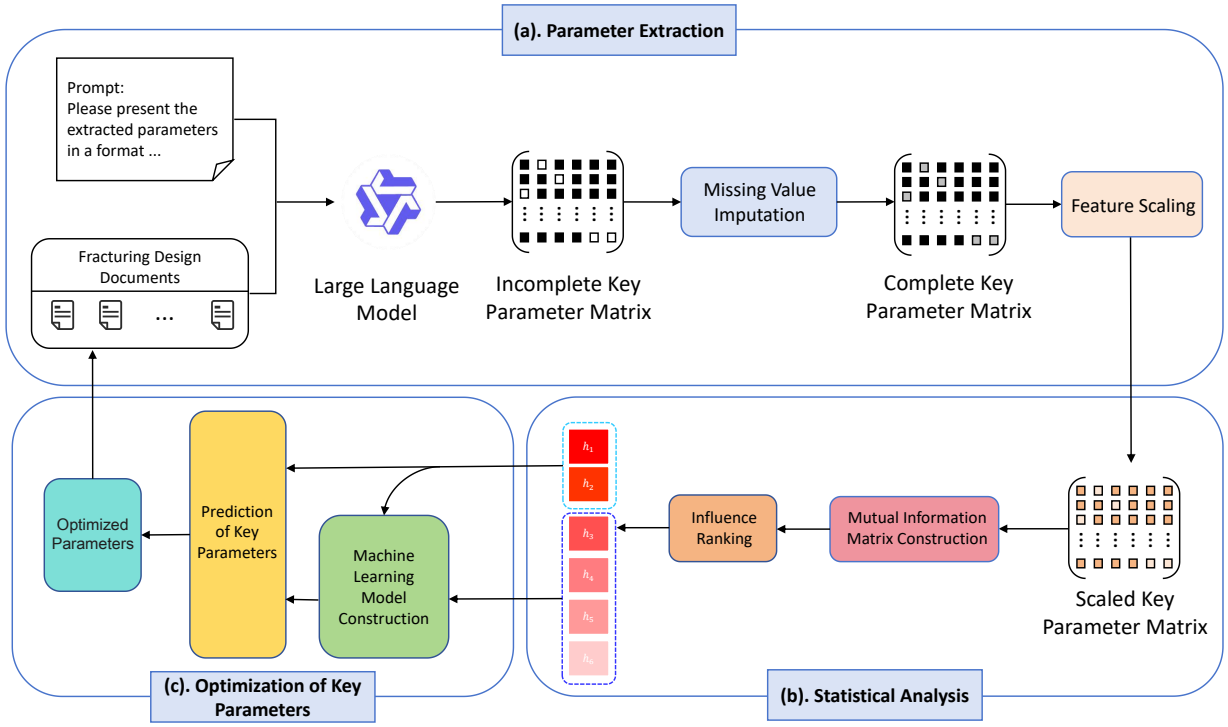


Fig. 1. Schematic workflow: (**a**) A locally deployed LLM automatically extracts parameters identified by experts in geology and hydraulic fracturing, followed by imputation and scaling performed. (**b**) We utilize Mutual information to analyze the parameters and identify the most influencial parameters. (**c**) We utilize random forest to predict these parameters, optimizing the whole design.

$\mathbf{a} = (a_1, a_2, \ldots, a_n)$ and $\mathbf{b} = (b_1, b_2, \ldots, b_n)$ using Euclidean method is defined as:

$$d(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{j=1}^{n} (a_j - b_j)^2}, \tag{1}$$

where $d(\mathbf{a}, \mathbf{b})$ represents the Euclidean distance between $\mathbf{a}$ and $\mathbf{b}$;

2) For every observation $\mathbf{o}$ with missing data, the KNN imputation algorithm identifies the k observations $(\mathbf{n_1}, \mathbf{n_2}, ..., \mathbf{n_k})$ that have the smallest Euclidean distances to $\mathbf{o}$. These nearest neighbors share a similar distribution with the missing values, which is important in statistical analysis.

3) For the missing feature x_{missing} in observation $\mathbf{o}$, the KNN imputation algorithm uses a weighted average of the corresponding feature values from the k-nearest neighbors to estimate its value. The estimation is performed as follows:

$$x_{\text{missing}} = \frac{\sum_{i=1}^{k} w_i x_i}{\sum_{i=1}^{k} w_i}, \tag{2}$$

where x_i is the missing value from the i-th nearest neighbor $\mathbf{n_i}$, and $d_i = d(\mathbf{o}, \mathbf{n_i})$ indicates the i-th nearest neighbor's distance of the target $\mathbf{o}$ from $\mathbf{n_i}$. The weight $w_i = \frac{1}{d_i}$ is expressed as $w_i = \frac{1}{d_i}$, which means that it is inversely related to the distance from $\mathbf{o}$. With this weighted method, closer neighbors are given higher influence on the missing value, which in turn improves the imputation accuracy.

The KNN imputation approach is effective in dealing with missing values through the use of inherent patterns and similarities as it is stored in the dataset, which guarantees the completeness and validity of the retrieved parameters.

3) Feature Scaling: Normalization and standardization are performed during feature scaling with an aim to minimize the effect of different magnitudes and the value range on the optimization results. Data normalization refers to scaling the input data within a uniform range, and this not only maintains the relative size relationship between parameters but also makes the algorithm treat all input features with the same weight. The Min-Max normalization formula is given as follows:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \tag{3}$$

where x is the original value, and $x_{\min}$ is the smallest value, $x_{\max}$ is the maximum value of the variable, and x' is the normalized value.

The data standardization method transforms the distribution of the input data into a standard normal distribution with a mean of zero and a standard deviation of one. The formula of standardization is given as follows:

$$z = \frac{x' - \mu}{\sigma}, \tag{4}$$

where x' is the normalized value of the feature, and μ is the mean of the normalized feature, and σ is the standard deviation of the normalized feature, and z is the standardized value.

The *Fracturing Fluid Type* and *Proppant Type* are written in the categorical manner, while most machine learning algorithms can only handle the numerical variables, thus they are processed in the one-hot encoding manner. One-hot encoding is a way of creating a new, binary-valued feature for every category, the presence of a category is encoded by 1, while the absence of a category is encoded by 0.

The combination of advanced techniques in natural language processing with systematic data preprocessing, not only ensures precise points in the analysis and modeling stages, but lays a solid foundation for section Statistical Analysis and Optimization of Key Parameters.

B. Statistical Analysis

The relationships among hydraulic fracturing variables are inseparable from prioritizing the model's most predominant factors for predictive modeling. Mutual information (MI) is a statistical concept that is used to measure the mutual dependence between two random variables. MI can clearly show the relationship between different variables of hydraulic fracturing. It is also beneficial from the perspective of selecting the relevant parameters which have a major effect on the hydraulic fracturing process.

MI [34] is a measure of the quantity of information shared between two variables, also presenting a nonlinear measure of their dependence. In contrast to linear correlation coefficients, MI contains the account for both linear and nonlinear relationships, making it more useful for examining hydraulic fracturing data of complex nature. Traditional methods, such as Pearson correlation [35], are only capable of detecting linear dependencies. Whereas mutual information is able to depict a wider range of relationships, which proves it is an efficient tool in this study, where fracturing happens in a nonlinear manner.

However, MI is particularly suitable for discrete variables. Given the fact that the extracted parameters include both discrete and continuous variables, the next step is to categorize the continuous variables before obtaining the mutual information matrix. Quantile binning is a specific discretization method. In this method, the data matrix will be weighted in k bins, where each bin covers the same number of observations. This procedure guarantees that each bin has equal frequency, which is a great advantage, especially for databases with skewed distributions.

Quantile binning allows the transformation of continuous variables into discrete intervals such that one may easily compute the mutual information matrix between all pairs of parameters, whether they are naturally discrete or continuous. This step of discretization is a very important preliminary step for the accurate capture of the relationships between parameters and therefore influences the efficiency of the following MI analysis in selecting the most informative variables with regards to hydraulic fracturing optimization. Considering two fracturing parameters, respectively represented by Z_i and Z_j, mutual information is defined as:

$$I(Z_i; Z_j) = \sum_{z_i \in Z_i} \sum_{z_j \in Z_j} p(z_i, z_j) \log \left(\frac{p(z_i, z_j)}{p(z_i)p(z_j)} \right) \quad (5)$$

where $p(z_i, z_j)$ is the joint probability distribution of Z_i and Z_j, while $p(z_i)$ and $p(z_j)$ are the marginal probability distributions of Z_i and Z_j, respectively. Fig 2 depicts the results of the mutual information among all parameters.

Each element of the mutual information matrix will be summed row by row to pick up parameters that have the most influence on others, and the results are listed in Table II. This approach gives a measure of the total influence of each parameter on all other parameters in the system and clearly shows the parameters that have the most impact on fracturing performance. According to Table II, the Preflush Percentage and Average Proppant-to-Liquid Ratio have relatively higher total MI scores. As a result, it is reasonable to believe that they play a more important role in the process of hydraulic fracturing optimization. Therefore, it makes the optimization work focus on the most impactful parameters to leverage a better fracturing design with an enhanced overall efficiency and success.

TABLE II. PARAMETER IMPORTANCE BASED ON SUM OF ROWS IN MUTUAL INFORMATION MATRIX

Parameter	Value
Preflush Percentage	3.125211
Average Proppant-to-Liquid Ratio	3.103264
Total Fluid Volume	2.988999
Fracturing Fluid Type	2.918993
Pumping Rate	1.886995
Proppant Type	1.642624

C. Optimization of Key Parameters

At the data analysis stage, all key parameters that need to be optimized are identified methodically according to their effect on fracking performance. Such target parameters are those that will be predicted from other parameters as input in the optimization step. According to such intrinsic patterns, a good modeling of the relationship between input parameters and target parameters will be done to make sure that the learned relationships reflect the real-world successful fracturing schemes.

To model these relationships, we employed five machine learning algorithms: neural networks [36], random forest [37], linear regression [38], Bayesian ridge regression [39], and ridge regression [40]. Among them, the best performance, according to the overall performance comparison, was obtained using a random forest algorithm for the prediction of the target parameter.

Random forest is a kind of ensemble learning approach that joins the predictions from several decision trees to obtain more accurate as well as stable results. In a random forest, each decision tree is developed with a bootstrapped subset of the training data, where samples are drawn with replacement

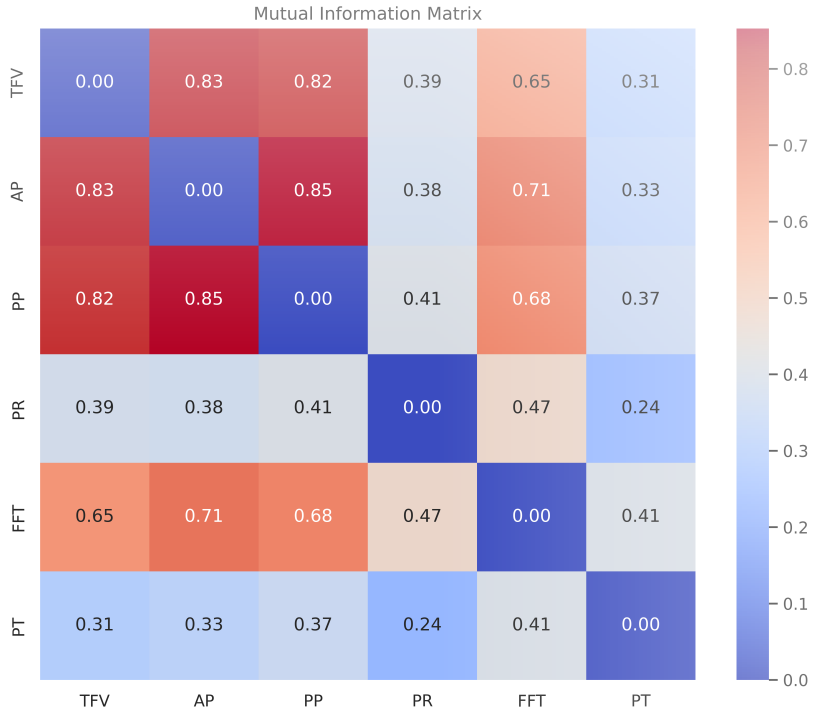


Fig. 2. Mutual Information Matrix within *Total Fluid Volume* (TFV), *Average Proppant-to-Liquid Ratio* (AP), *Preflush Percentage* (PP), *Pumping Rate* (PR), *Fracturing Fluid Type* (FFT), and *Proppant Type* (PT).

and independently of each other. Also, at each split of a decision tree, only a random subset of features is contemplated when determining the next best split. This technique leads to a decreasing correlation between the individual trees and, consequently, a better performance of the model model as far as generalization is concerned.

The input to the random forest is given as Z, and Z is a 372 by 4 matrix in this study. The output $\hat{Z}$ represents the predicted mean values of the target parameter. Apart from classification, random forest can also be adapted to the regression task, and it takes the average of all leaf nodes' outputs in the regression task as the final prediction:

$$\hat{Z} = \frac{1}{T}\sum_{t=1}^{T} f_t(Z), \tag{6}$$

where T denotes the total number of trees in the forest, and $f_t(Z)$ is the forecast made by the t-th decision tree.

Each tree in the forest splits the data at its nodes according to the best criterion that minimizes the prediction error. The resultant output can effectively exploit the strength of this ensemble to reduce overfitting and variance, thus the predictive accuracy remains comparatively high. The unique integration mechanism of the random forest is reliable in dealing with the complex nonlinear relationship described among fracturing parameters. These parameters then will be used to update the original fracturing parameters:

$$X \leftarrow \hat{Z} \tag{7}$$

IV. Experiment

A. Experimental Setup

This section presents the experimental setup and explores the results derived from the proposed workflow. A complete dataset was processed from 372 fracturing design documents from an oilfield in China with the application of the QWen2.5-7B model. The dataset includes six important hydraulic fracturing design parameters—*Total Fluid Volume, Average Proppant-to-Liquid Ratio, Preflush Percentage, Fracturing Fluid Type, Proppant Type, and Pumping Rate*—which further intern describes on the Table I. These parameters will be used as the basis for predictive modeling and parameter optimization.

To evaluate the relationships among parameters, we applied the mutual information matrix. However, some of the parameters extracted from the documents are continuous and the MI matrix requires discrete variables, thus the continuous parameters were discretized with the *KBins* method with $n = 20$ bins. In particular, this choice aimed to meet the need for the greatest granularity of information while safeguarding robustness against overfitting. Among the analyzed parameters, *Preflush Percentage*, as well as *Average Proppant-to-Liquid*

Fig. 3. Performance of imputation: The KNN method effectively addressed data sparsity by filling missing values in alignment with the existing data structure. The mode and spread of all three parameters remained consistent after imputation.

Ratio have shown to have the strongest relationships with other variables in the database.

Five machine learning models were used to predict the *Average Proppant-to-Liquid Ratio* and *Preflush Percentage*, and the remaining variables were used as input parameters. The neural network architecture used here consisted of three fully connected layers: an input layer with 64 neurons, a hidden layer with 32 neurons, and an output layer equal to the number of target parameters. The random forest model was configured with 100 decision trees to be more predictive and robust. Linear regression is the baseline model because it minimized the residual sum of squares without regularization. Bayesian ridge regression used Gaussian priors for the regularization of model coefficients, and the strength of regularization was adaptively estimated from the data. Finally, the ridge regression used L_2 regularization and set its strength parameter (α) to 1.0, a balanced choice for both training accuracy and generalization. These configurations were chosen with the aim of investigating different modeling strategies.

B. KNN-Based Imputation

In this study, the *Average Proppant-to-Liquid Ratio* and *Preflush Percentage* have been detected of missing values, so the k-nearest neighbor (KNN) technique was used to fill the missing values. Significant improvement before and after imputation, both qualitative and quantitative, may be noticed in Fig. 3. The imputation preserved the central tendency and shape of the original distributions. For the *Average Proppant-to-Liquid Ratio*, the imputation did not affect the generally positively skewed nature of the distribution, and it also smoothed out sparsity in the tail region. The *Preflush Percentage* had a central peak around **25%** and had improved continuity without the introduction of distortions, whereas for the *Preflush Percentage*, centered around **40%**, it maintained its overall spread while filling gaps and enhancing smoothness. KNN method amply resolved the challenge of data sparsity by filling missing values in line with the structure already inherent in the data. These histograms proved that the mode and spread of the two parameters remain the same after imputation. The frequency of values around the central peaks, especially for *Average Proppant-to-Liquid Ratio* and *Preflush Percentage*, has significantly increased, which preserves important statistical properties for further analyses.

C. Experimental Analysis

This section visualizes the distributions of important continuous parameters as violin plots and displays the frequencies of discrete parameters as histograms. This section also builds

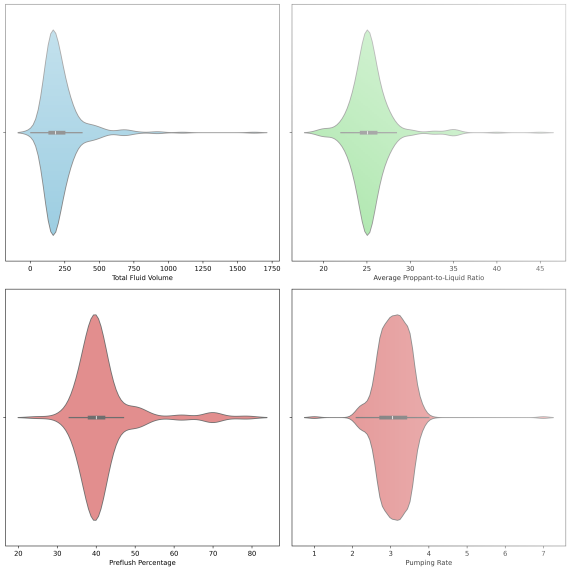


Fig. 4. Violin plot of continuous parameters *Average Proppant-to-Liquid Ratio*, *Preflush Percentage* and *Pumping Rate*.

a mutual information matrix to quantify the dependencies between key hydraulic fracturing parameters.

Fig 4 shows violin plots for the distributions of the following four critical continuous variables in hydraulic fracturing: *Total Fluid Volume*, *Average Proppant-to-Liquid Ratio*, *Preflush Percentage*, and *Pumping Rate*. The *Total Fluid Volume* is right-skewed; most values lie below 300, indicating common operational practices. The *Average Proppant-to-Liquid Ratio* and *Preflush Percentage* are also relatively symmetrically distributed around the center of 25 and 40, respectively, which could indicate consistent design patterns. The *Pumping Rate* reflects a narrower range, clustering around 3 to 4, reflecting its controlled nature in fracturing operations.

As shown in Fig. 5, *Fracturing Fluid Type* and *Proppant Type* frequency histograms are highly concentrated in a few categories. Regarding *Fracturing Fluid Type*, category 2 *Guar Gum Fracturing Fluid* is used most, closely followed by categories 3 *Polymer Fracturing Fluid* and 5 *Low-Polymer Fracturing Fluid*, suggesting dependence on certain types of fracturing fluids that may suit geological conditions and operational requirements. A similar case is *Proppant Type*, dominated by category 0 *Quartz Sand*, reflecting the preference for a given proppant that will provide optimal fracture conductivity and stability. Skewed distributions indicate that, though several options are available, only a few types of fluids and proppants have shown consistent effectiveness through hydraulic fracturing practices, likely due to compatibility with the reservoir conditions and cost efficiency. Understanding these parameters is essential for selecting and optimizing parameters because dominant categories usually represent proven solutions in prior successful fracturing designs.

Fig. 2 presents a Mutual Information Matrix. This matrix is obtained by calculating the MI between six important hydraulic fracturing parameters, and these six parameters are *Total Fluid Volume* (TFV), *Average Proppant-to-Liquid Ratio* (AP), *Preflush Percentage* (PP), *Pumping Rate* (PR), *Fracturing*

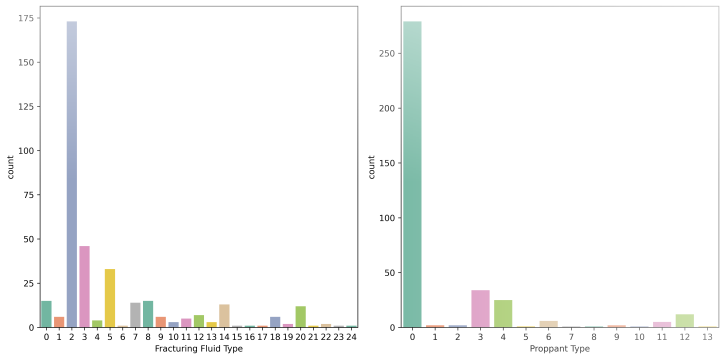


Fig. 5. Frequency histograms of discrete parameters *Fracturing Fluid Type* and *Proppant Type*.

Fluid Type (FFT), and *Proppant Type* (PT). As shown in the matrix, the *Average Proppant-to-Liquid Ratio* and *Preflush Percentage* are more correlated with the other parameters, and the color blocks in the corresponding areas are also darker. For example, the MI between *Average Proppant-to-Liquid Ratio* and *Fracturing Fluid Type* is 0.71, and this number increases to 0.83 when MI is calculated between *Average Proppant-to-Liquid Ratio* and *Total Fluid Volume*. *Proppant Type* follows a similar rule to *Average Proppant-to-Liquid Ratio*. All of the above information shows the dominance of *Average Proppant-to-Liquid Ratio* and *Preflush Percentage* in the fracturing process.

It is also worth noting that both *Average Proppant-to-Liquid Ratio* and *Preflush Percentage* have relatively high MI with *Fracturing Fluid Type*, which actually reflects the influence of fluid choice on proppant behavior. In fact, the efficiency of proppant transport and fracture conductivity is directly related to the different types of fracturing fluids. For example, guar gum and polymer-based fluids have different rheological properties, which significantly affect the proppant behavior. On the other hand, lower cumulative mutual information scores are obtained for *Pumping Rate* and *Proppant Type*, indicating that these parameters depend less on other parameters in this dataset.

In all, the identification of *Average Proppant-to-Liquid Ratio* and *Preflush Percentage* as the most relevant parameters agrees with the basic principles of hydraulic fracturing, in which the optimization of proppant concentration and preflush strategy is of paramount importance in attaining effective fracture propagation and improving the performance of the reservoir.

D. Parameter Optimization

Feature selections were performed based on the results obtained from mutual information analysis for the target parameters to be predicted and optimized in this work, namely *Average Proppant-to-Liquid Ratio* and *Preflush Percentage*. Five different machine learning models were used to predict these target parameters. Performance comparisons are made based on the mean squared error-MSE, the root mean squared error-RMSE, the mean absolute error-MAE, R^2 score, and the maximum absolute error between the true value and the model prediction value-Max Error.

A total of five machine learning models in Table III and Table IV display their performance in predicting *Average Proppant-to-Liquid Ratio* and *Preflush Percentage*, respectively.

TABLE III. PERFORMANCE COMPARISON ON PREDICTING *Average Proppant-to-Liquid Ratio*.

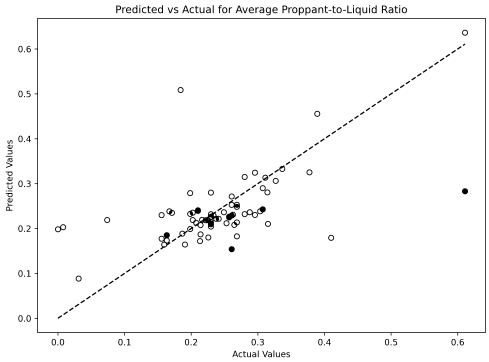
Model	MSE	RMSE	MAE	R^2	Max Error
Neural Network	0.008997	0.094850	0.063098	0.142182	0.339860
Random Forest	0.007582	0.087078	0.053089	0.277015	0.328058
Linear Regression	0.010878	0.104296	0.067682	-0.037174	0.395630
Bayesian Ridge	0.010783	0.103841	0.067177	-0.028151	0.380968
Ridge	0.010804	0.103941	0.067449	-0.030123	0.385983

TABLE IV. PERFORMANCE COMPARISON ON PREDICTING *Preflush Percentage*

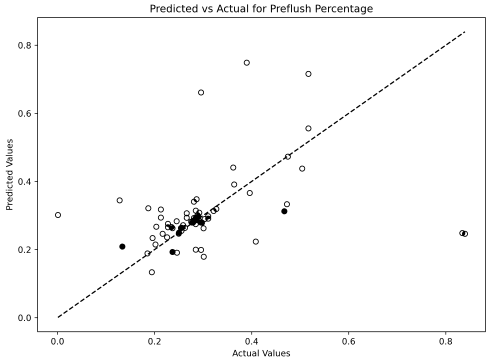
Model	MSE	RMSE	MAE	R^2	Max Error
Neural Network	0.017676	0.132950	0.071928	-0.115064	0.542560
Random Forest	0.018345	0.135443	0.073504	-0.157274	0.595212
Linear Regression	0.014045	0.118512	0.074651	0.113971	0.493211
Bayesian Ridge	0.013969	0.118191	0.074049	0.118776	0.495936
Ridge	0.013683	0.116973	0.071300	0.136836	0.507866

From Table III, the random forest model has the best predictive capability for *Average Proppant-to-Liquid Ratio*, with the lowest MSE of 0.007582 and RMSE of 0.087078, while the R^2 score is high at 0.277015. What's more, its strong performance is further supported by the lowest MAE of 0.053089 and a Max Error of 0.328058, hence it is reliable to capture the underlying relationships of the target variables. The neural network is also doing quite well, with an RMSE of 0.094850 and MAE of 0.063098. However, the R^2 score of 0.142182 shows that it explains less variance in the target compared to the random forest. On the other hand, linear models such as linear regression, Bayesian ridge, and ridge regression have larger errors and negative R^2 scores, which point out their inability to model the nonlinear trends within the data. From Table IV, the ridge regression model yields the best results for the *Preflush Percentage* with an MSE of 0.013683 and an RMSE of 0.116973, while the R^2 score is very high, equal to 0.136836. It follows that the ridge regression greatly balances the prediction accuracy and generalization of the model performance for this parameter. The Bayesian ridge model runs relatively well, with a higher error but still a positive R^2 score at 0.120377. On the other hand, both the neural network and the random forest model underperform. The neural network shows an MSE of 0.17676 and a Max Error of 0.542560, reflecting greater variability and lower reliability in its predictions for this parameter. Taking into account both targets and overall metrics, the random forest model proves to be the most powerful method. It is very consistent when forecasting the *Average Proppant-to-Liquid Ratio*, for which it is ranked first among all models, and delivers competitive results in the *Preflush Percentage*. Its capability to handle nonlinear relationships and keep prediction errors low for different parameters makes it very robust for hydraulic fracturing applications. Meanwhile, the interpretability and robustness of the random forest against overfitting increase its practical value in optimizing key fracturing parameters.

The plot of Predicted versus Actual Values using random forest model for the two most important parameters *Average Proppant-to-Liquid Ratio* and *Preflush Percentage* is given by Fig. 6. Most predicted values of *Average Proppant-to-Liquid Ratio* in Fig. 6 (a) are very close to the red dashed line, especially within the range from 0.1 to 0.3. That reflects that the model has captured the underlying pattern and relationship quite nicely; hence, predictions in most cases are very accurate and reliable. Although there are minor deviations at higher actual values, the values are very minimal and can be attributed to data sparsity or variability in the higher range. These small discrepancies do not take away much from the overall performance, and the model is really robust to nonlinear relationships.



(a) *Average Proppant-to-Liquid Ratio*



(b) *Preflush Percentage*

Fig. 6. Prediction vs Actual: The random forest model performs commendably for both *Average Proppant-to-Liquid Ratio* and *Preflush Percentage*, accurately capturing dominant patterns and relationships within the data.

Similarly, Fig. 6 (b) shows the model's prediction accuracy for the *Preflush Percentage*. Most of the data points are close to the red line, and the low to medium range is well covered between 0.1 and 0.4. This indicates the strength of the model in general trends and thus it makes fairly reliable predictions. There are a few outliers at higher values, which may be due to class imbalance; that is, these higher values occur less in the dataset. However, its strong alignment with actual values over the majority range makes the model practically applicable and reliable. Overall, the Random Forest model performed very well for both *Average Proppant-to-Liquid Ratio* and *Preflush Percentage*, capturing the dominant patterns and relationships in the data quite well. Its robustness in handling nonlinear dependencies makes it a reliable choice for predicting key hy-

draulic fracturing parameters, with minor deviations providing potential opportunities for further refinement.

V. Discussion

A. Theoretical Implications

Our approach demonstrates that the integration of a large language model with classical machine learning algorithms can improve the efficiency of parameter optimization. By automating parameter extraction and involving statistical analysis, PPOHyFrac implements a systematic framework that streamlines the process of optimizing hydraulic fracturing parameters. This integration proves the worth of data-driven methods in capturing complex nonlinear reservoir dynamics while opposing the simplistic conceptions of conventional models.

B. Practical Considerations

PPOHyFrac is practically applicable as a scalable solution for optimizing hydraulic fracturing parameters in different geological environments. Automated data extraction of the system reduces the effort and subjectivity of manual input. The modularity of the framework also makes it possible to adapt it so that it conforms to region-specific fracturing practices. But its performance would still depend on the quality and consistency of the input documents. Moreover, computational requirements will still have to be taken into consideration, especially for large-scale implementation.

C. Future Research Directions

Given that the dataset we use is collected from a specific region, there may be limitations in its transferability and generalization performance; thus, future efforts should focus on obtaining wider datasets in terms of different types of fracturing design documents so that generalization with the model can be improved. In addition, PPOHyFrac mainly focuses on optimizing key parameters in hydraulic fracturing. Nevertheless, a complete hydraulic fracturing project needs to address several other critical factors, such as wellbore design, drilling optimization, environmental impact mitigation, and operational safety, to maximize production efficiency while minimizing operational risks. While PPOHyFrac is of positive significance in simplifying the design of the fracking process, its current scope does not encompass these broader operational and ecological considerations. To achieve end-to-end optimization, subsequent research could consider incorporating multi-objective optimization methods to balance competing goals, and other techniques such as active learning approaches may also be a good choice for refining designs based on real-time oil filed data.

VI. Conclusion

This paper proposes PPOHyFrac, a data-driven scheme that pairs a locally deployed large language model with classic machine learning techniques to optimize key hydraulic fracturing parameters. This framework consists of automated extraction of key parameters in unstructured documents, and rigorous statistical analysis and machine learning models that aim at predicting and optimizing fracture performance-related parameters. By using the locally deployed LLM, we have constructed a holistic dataset from 372 unstructured fracturing design documents. Subsequent mutual information analysis reveals that *Average Proppant-to-Liquid Ratio* and *Pre-flush Percentage* have relatively higher influence on fracking performance. Comparative experiments demonstrate that random forest is the best choice for the optimization of hydraulic fracturing. In conclusion, PPOHyFrac bridges the gap between the usage of unstructured data and the optimization of hydraulic fracturing and also provides actionable and thoughtful insights for sustainable energy extraction. Since the main focus of PPOHyFrac is parameter optimization, future research will pay more attention to build a comprehensive system that can be applied to areas with complex geological conditions.

References

[1] S. Saraji and D. Akindipe, "The role of the oil and gas industry in the energy transition," in *Sustainability in the Oil and Gas Sector: Adaptation and Mitigation Strategies for Tackling Climate Change*. Springer, 2024, pp. 33–63.

[2] G. Liu, X. Wu, and V. Romanov, "Unconventional wells interference: Supervised machine learning for detecting fracture hits," *Applied Sciences*, vol. 14, no. 7, 2024.

[3] L. Gandossi and U. Von Estorff, *An overview of hydraulic fracturing and other formation stimulation technologies for shale gas production*. Publications Office of the European Union Luxembourg, 2015.

[4] C. Chu and Q. Xie, "Study on capacity evaluation of fractured horizontal wells in tight gas reservoirs," *CPCCS*, vol. 44, pp. 11–13, 2024.

[5] H. L and X. W, "Analytical optimization of hydraulic fracturing," *Journal of Energy and Environmental Sciences*, vol. 2, pp. 1–10, 2024. [Online]. Available: https://doi.org/10.23880/jeesc-16000105

[6] L. Huang, X. Liao, M. Fan, S. Wu, P. Tan, and L. Yang, "Experimental and numerical simulation technique for hydraulic fracturing of shale formations," *Advances in Geo-Energy Research*, vol. 13, no. 2, pp. 83–88, 2024.

[7] H. Wu, N. Zhang, Y. Lou, X. Zhai, B. Liu, and S. Li, "Optimization of fracturing technology for unconventional dense oil reservoirs based on rock brittleness index," *Scientific Reports*, vol. 14, no. 1, p. 15214, 2024.

[8] O. Kolawole, M. Wigwe, I. Ispas, and M. Watson, "How will treatment parameters impact the optimization of hydraulic fracturing process in unconventional reservoirs?" *SN Applied Sciences*, vol. 2, no. 11, p. 1865, 2020.

[9] B. Mahesh, "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR).[Internet]*, vol. 9, no. 1, pp. 381–386, 2020.

[10] C. Baccouch and C. Bahar, "Advanced machine learning approaches for accurate migraine prediction and classification," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 1, 2025. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2025.0160101

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[12] F. Liu, "A data-driven deep machine learning approach for tunnel deformation risk assessment," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 11, 2024. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2024.0151127

[13] H. Hassani, X. Huang, and E. Silva, "Digitalisation and big data mining in banking," *Big Data and Cognitive Computing*, vol. 2, no. 3, 2018.

[14] A. Alake and E. Oyedeji, "Systematic analysis of novel machine learning techniques for hydraulic fracturing optimization," *Preprints*, April 2024.

[15] A. Johar, *Hydraulic Fracturing Treatment Optimization Using Machine Learning*. West Virginia University, 2023.

[16] Z. Dong, L. Wu, L. Wang, W. Li, Z. Wang, and Z. Liu, "Optimization of fracturing parameters with machine-learning and evolutionary algorithm methods," *Energies*, vol. 15, no. 16, 2022.

[17] C. Lu, H. Jiang, J. Yang, Z. Wang, M. Zhang, and J. Li, "Shale oil production prediction and fracturing optimization based on machine learning," *Journal of Petroleum Science and Engineering*, vol. 217, p. 110900, 2022.

[18] M. U. Hadi, R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, S. Mirjalili *et al.*, "Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects," *Authorea Preprints*, 2023.

[19] K. Sharifani and M. Amini, "Machine learning and deep learning: A review of methods and applications," *World Information Technology and Engineering Journal*, vol. 10, no. 07, pp. 3897–3904, 2023.

[20] Z. Wu, C. Cui, P. Jia, Z. Wang, and Y. Sui, "Advances and challenges in hydraulic fracturing of tight reservoirs: A critical review," *Energy Geoscience*, vol. 3, no. 4, pp. 427–435, 2022.

[21] D. Mata, W. Zhou, Y. Zee Ma, and V. Gonzales, "Chapter 8 - hydraulic fracture treatment, optimization, and production modeling," in *Unconventional Oil and Gas Resources Handbook*, Y. Z. Ma and S. A. Holditch, Eds. Boston: Gulf Professional Publishing, 2016, pp. 215–242.

[22] M. Zhao, "Field experiments and main understanding of shale oil hydraulic fracturing," *Frontiers in Earth Science*, vol. 12, p. 1410524, 2024.

[23] J. L. Miskimins, S. A. Holditch, and J. Veatch, Ralph W., "Preface," in *Hydraulic Fracturing: Fundamentals and Advancements*. Society of Petroleum Engineers.

[24] B. Chen, B. R. Barboza, Y. Sun, J. Bai, H. R. Thomas, M. Dutko, M. Cottrell, and C. Li, "A review of hydraulic fracturing simulation," *Archives of Computational Methods in Engineering*, pp. 1–58, 2022.

[25] A. Ismail and S. Azadbakht, "A comprehensive review of numerical simulation methods for hydraulic fracturing," *International Journal for Numerical and Analytical Methods in Geomechanics*, vol. 48, no. 5, pp. 1433–1459, 2024.

[26] A. J. Majeed, D. T. Yaseen, M. A. Hassan, and A. M. Al-Mukhtar, "Enhancing realism in hydraulic fracturing simulation models: The evolution of kgd and pkn models," *Procedia Structural Integrity*, vol. 66, pp. 212–220, 2024.

[27] K. Yang and D. Gao, "Numerical simulation of hydraulic fracturing process with consideration of fluid–solid interaction in shale rock," *Journal of Natural Gas Science and Engineering*, vol. 102, p. 104580, 2022.

[28] J. Zhang, H. Yu, W. Xu, C. Lv, M. Micheal, F. Shi, and H. Wu, "A hybrid numerical approach for hydraulic fracturing in a naturally fractured formation combining the xfem and phase-field model," *Engineering Fracture Mechanics*, vol. 271, p. 108621, 2022.

[29] L. Huang, E. Dontsov, H. Fu, Y. Lei, D. Weng, and F. Zhang, "Hydraulic fracture height growth in layered rocks: Perspective from dem simulation of different propagation regimes," *International Journal of Solids and Structures*, vol. 238, p. 111395, 2022.

[30] A. Erofeev, D. Orlov, D. Perets, and D. Koroteev, "Ai-based estimation of hydraulic fracturing effect," *SPE Journal*, vol. 26, no. 04, pp. 1812–1823, 2021.

[31] L. Lizhe, Z. Fujian, Z. You, C. Zhuolin, W. Bo, Z. Yingying, and L. Yutian, "The prediction and optimization of hydraulic fracturing by integrating the numerical simulation and the machine learning methods," *Energy Reports*, vol. 8, pp. 15 338–15 349, 2022.

[32] A. D. Morozov, D. O. Popkov, V. M. Duplyakov, R. F. Mutalova, A. A. Osiptsov, A. L. Vainshtein, E. V. Burnaev, E. V. Shel, and G. V. Paderin, "Data-driven model for hydraulic fracturing design optimization: Focus on building digital database and production forecast," *Journal of Petroleum Science and Engineering*, vol. 194, p. 107504, 2020.

[33] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng, "Learning k for knn classification," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 8, no. 3, pp. 1–19, 2017.

[34] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *International conference on machine learning*. PMLR, 2018, pp. 531–540.

[35] P. Schober, C. Boer, and L. A. Schwarte, "Correlation coefficients: appropriate use and interpretation," *Anesthesia & analgesia*, vol. 126, no. 5, pp. 1763–1768, 2018.

[36] S. Schmidgall, R. Ziaei, J. Achterberg, L. Kirsch, S. Hajiseyedrazi, and J. Eshraghian, "Brain-inspired learning in artificial neural networks: a review," *APL Machine Learning*, vol. 2, no. 2, 2024.

[37] H. A. Salman, A. Kalakech, and A. Steiti, "Random forest algorithm overview," *Babylonian Journal of Machine Learning*, vol. 2024, pp. 69–79, 2024.

[38] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[39] M. E. Khan and H. Rue, "The bayesian learning rule," *Journal of Machine Learning Research*, vol. 24, no. 281, pp. 1–46, 2023.

[40] M. Rajan, "An efficient ridge regression algorithm with parameter estimation for data analysis in machine learning," *SN Computer Science*, vol. 3, no. 2, p. 171, 2022.