Petroleum and Coal

Article **Open Access**

Accurate Prediction of Post-Fracturing Productivity Index Using Advanced Machine Learning Models: A Field-Validated Framework

Ahmed Gamal[1], Samuel Nashed[2], Oluchi Ejehu[2], Emad Nabil[3] and Rouzbeh Moghanloo[2]*

[1] *Faculty of Petroleum and Mining Engineering, Suez University, Suez, Egypt*
[2] *Mewbourne School of Petroleum and Geological Engineering, Mewbourne College of Earth and Energy, The University of Oklahoma, Norman, OK, USA*
[3] *General Petroleum Company, Cairo, Egypt*

Received May 8, 2025; Accepted September 19, 2025

Abstract

Forecasting the post-fracturing productivity index (PI) accurately remains essential as it enables both hydraulic fracturing optimization and better well performance as well as financial sustainability. The prediction of post-fracturing productivity index through traditional methods such as empirical correlations along with analytical models and numerical simulations and well testing results in both time inefficiencies and high costs and restricted capabilities in complex reservoirs. The objective of this research focuses on creating advanced machine learning models that deliver accurate predictions for post-fracturing PI. This research used standard hydraulic fracturing and reservoir parameters to develop nine advanced machine learning model systems. The trained developed models received their training data from actual post-fracturing PI measurements taken from pressure gauges deployed in wells and traditional production test results. Machine learning algorithms were developed by analysing a large dataset from 258 wells. The most efficient algorithms for post-fracturing PI prediction yielded excellent results, with RMSE values of 0.019 for AdaBoost, 0.021 for Gradient Boosting, 0.021 for Neural Network, and 0.047 for Random Forest. In a field application of AdaBoost, the model demonstrated a 3.2% prediction error, which closely matched the actual PI results. The study established that ML methods deliver precise PI predictions while offering both budget-friendly features and high operational speed, making them viable substitutes for traditional prediction tools in hydraulic fracturing design applications and reservoir management needs.

Keywords: *Hydraulic fracturing; Production forecasting; Stimulation; Productivity index; Formation damage.*

1. Introduction

1.1. The importance of predicting productivity index

Hydraulic fracturing optimization relies on an essential predictive tool that calculates the productivity index (PI) by dividing the predicted production rate by pressure drawdown [1]. Successful hydraulic fracturing should result in a cost-effective and sustained fold-of-increase (FOI) that compares PI values before and after fracturing [2]. Additionally, the success of fracturing design and optimization greatly depends on accurately estimating well productivity for stimulated wells [3]. The required PI after hydraulic fracturing helps fracture engineers determine the necessary fracture geometry and conductivity values [4]. However, these parameters, along with proppant, fracture fluid systems, hydraulic horsepower, and pumping techniques, introduce complexity, making optimization a significant challenge for fracture engineers [5]. Furthermore, precise forecasting capabilities enable engineers to improve decision-making regarding artificial lift systems and reservoir production strategies [6-7]. Accurate predictions would allow operators to establish a reliable economic case, which is crucial for vali-

Downloaded from SHAREOK. Persistent Handle URL: https://hdl.handle.net/11244/342971

dating the expenditure on hydraulic fracturing against production gains [8]. Additionally, precise estimation of post-fracturing production indices is crucial for tight formations and deep unconventional reservoirs, as it maximizes asset recovery and supports optimized well spacing design, thereby enhancing the sustainability of field development [9].

1.2. Traditional calculation methods

The prediction of the post-fracturing productivity index relies on various traditional methods, each with its own limitations. Current prediction approaches include empirical correlations, analytical models, numerical simulations, and well testing procedures.

1.2.1. Empirical correlations

The post-fracture productivity index (PI) estimation, based on production data, wellbore geometry, and reservoir characteristics, is widely used in conjunction with empirical correlations. These correlations, as developed by McGuire and Sikora [10], Prats [11], Nolte [12], and Barree *et al.* [13], are quick estimates based on field history and observational data. These methods have the advantage of computational efficiency and convenience. Still, they cannot address the complex reservoir characteristics of real-world reservoirs, as they struggle to capture the full impact of dynamic fracture conductivity, multiphase flow effects, and reservoir heterogeneities [14]. Empirical models are also limited to specific conditions and must then be calibrated with numerous parameters to avoid inaccuracy [15].

1.2.2. Analytical models

The description of fluid flow in fractured reservoirs is based on simplified mathematical equations that form the analytical models. The Perkins-Kern-Nordgren (PKN) [16], Khristianovic-Geertsma-de Klerk (KGD) [17], Barree [18], and Raghavan [19] types of models are based on homogeneity of reservoir properties, uniform fracture geometry, and single-phase flow to account for the effects of fracturing on PI. Because these models provide a theoretical basis for PI, they do not consider fracture complexity, complex proppant distribution, or stress-dependent permeability variations [20]. In addition, analytical models assume uniform fracture conductivity and symmetrical fracture growth, which are rarely observed in field applications [21]. Therefore, the predicted PI values are usually inaccurate [22].

1.2.3. Numerical models

Numerical reservoir simulations are highly effective for calculating the post-fracturing productivity index (PI) by modelling fluid flow and pressure distribution within the reservoir [23]. These simulations effectively combine wellbore conditions, fracture geometry, and fracture conductivity with reservoir parameters to predict well performance post-stimulation [24]. Furthermore, numerical approaches such as pseudo-3D (P3D) and planar 3D (PL3D) models are used for modelling hydraulic fracture geometry and conductivity [25]. P3D models assume fracture height growth based on the stress barriers, which simplifies calculations but loses accuracy [26]. On the contrary, PL3D models provide a complete three-dimensional view of fractures, taking into account changes in height, width, and length, but at a costly computational expense [27]. While having their advantages, numerical models require a significant amount of field data and laboratory tests, as well as high computational power, making them less practical for making rapid decisions [28]. In addition, these models still fail to account for actual complex downhole conditions, such as non-uniform fracture propagation and multi-fracture interaction and do not fully capture the complexities of unconventional reservoirs [29]

1.2.4. Well testing

Post-fracturing well tests, such as production test analysis, pressure transient analysis, and rate-transient analysis, are routinely used to calculate the best estimate of PI by analyzing bottom-hole pressure trends and fluid flow characteristics [30]. These tests provide valuable insights into the skin factor and PI following hydraulic fracturing treatment [31]. However, performing well tests is expensive and time-consuming, as it requires production shut-ins or

the deployment of downhole pressure gauges, which result in a loss of revenue [32]. Furthermore, well testing interpretations are highly sensitive to reservoir heterogeneities and fracture interference, making it challenging to isolate hydraulic fracture effects from natural fractures or formation damage [33]. Additionally, in unconventional reservoirs, where fluid flow is transient and non-Darcy, the traditional well testing method may yield incorrect PI values [34].

1.3. Machine learning models for predicting PI

The interest in ML models in the oil and gas industry stems from the fact that these models can address problems that are difficult to solve with traditional techniques [35-37]. Additionally, the process power increase along with the data available for analysis are reasons that justify the growing industry interest in the application of ML algorithms [38-40]. Over the past decades, the oil and gas industry has produced huge quantities of data, but it is still difficult to find valuable patterns in these huge data sets [41-42]. Yet, machine learning algorithms excel in analyzing large and complex data to discover useful correlations that can be used for running forecasting applications and enhancing the process [43]. ML models generalize across different reservoir conditions and can avoid the need for simplified assumptions as compared to empirical correlations [44-45]. ML-based approaches are more adaptable than analytical models, as they learn patterns directly from field data [46-47]. Moreover, ML models are both more computationally efficient and require less effort to provide rapid predictions with fewer computational requirements after training, as compared to numerical models like Planar 3D (PL3D) and Pseudo 3D (P3D) [48-49]. ML-based predictions are faster and less expensive than well testing, and they enable real-time decision-making without interrupting production [50]. However, despite these advantages, ML models are prone to requiring large, high-quality datasets for training and to interpretability challenges [51]. Recent advances in explainable AI (XAI) techniques, namely SHAP values and feature importance analysis, have made ML more interpretable and a serious contender over such empirical correlations, analytical models, numerical models, and well testing [52-53]. To the best of our knowledge, the advanced machine learning approach has not been implemented to predict post-fracturing PI in any of the available studies about stimulation treatments. Given this knowledge gap, this study aims to produce nine advanced ML algorithms that can forecast post-fracturing PI. We envision these ML algorithms as fine and swift alternatives to traditional calculation methods, and that these models will decrease human interaction in the future, allowing engineers to focus on their most important strategic issues.

2. Methodology

In this study, a four-stage methodological framework is presented in Figure 1, where the phases are strategically connected. By having a systematic progression over these sequential stages, the research objectives are aligned coherently.

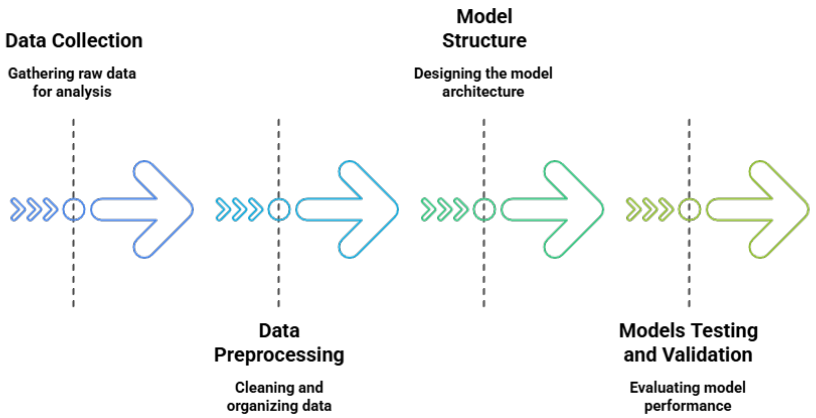


Figure 1. A diagram of the methodology.

2.1 Data collection

Surface production tests and bottomhole pressure gauges are commonly used by operators after hydraulic fracturing treatments for PI calculations. Our team collected a comprehensive dataset of 3,870 observations to perform this research, and each observation captured individual measurement results. The dataset comes from 258 wells across the Egyptian Western Desert. A comprehensive dataset helps machine learning models learn real patterns better than random noise during development. Moreover, analysis of extensive data prevents overfitting and helps the models understand complex data connections. In addition, comprehensive datasets help models handle real-world situations better and decrease bias because they provide accurate representations of all possible scenarios. Our dataset enables us to divide it into proper training and testing segments to make reliable assessments of our models. The dataset includes actual post-frac PI for 258 wells. PI of each well was calculated from actual bottomhole pressure gauges and post-fracturing production tests. Using real datasets for training models lets us deal with actual life data issues and accurately capture their learning patterns. Measuring models on raw actual data shows the actual distribution of data and helps stakeholders trust results, making models more reliable for deployment in critical applications like hydraulic fracturing evaluation. However, hydraulic fracture geometry and dimensionless fracture conductivity were calculated by using a 3D numerical simulator. Table 1 displays the dataset information that includes Solution GOR (GOR), Oil Gravity (OG), Gas Gravity (GG), Reservoir Pressure (RP), Reservoir Temperature (RT), Reservoir Permeability (K), Reservoir Thickness (H), Drainage Area (A), Wellbore Radius (R), Time (T), Reservoir Porosity (PHI), Fracture Height (FH), Fracture Half Length (XF), Dimensionless Fracture Conductivity (FCD), and Productivity Index (PI). Various fluid properties, reservoir characteristics, fracture geometries, and conductivities are presented throughout the data set. A generalized and accurate model will be developed in the following sections of this paper, given the wide range of input parameters.

Table 1. Statistical analysis for the collected database.

Parameter	Unit	MIN	MAX	AVG	Median
Solution GOR	scf/stb	218	3833	989	250
Oil gravity	API	28	45	37	40
Gas specific gravity	-	0.58	0.74	0.65	0.64
Reservoir pressure	psi	1156	5485	2959	3000
Reservoir temperature	deg F	140	282	226	246
Reservoir permeability	md	1	72	6	4
Reservoir thickness	ft	16	89	46	45
Drainage area	acres	50	152	78	70
Wellbore radius	ft	0.281	0.350	0.328	0.350
Time	days	0.3	10	4	2
Reservoir porosity	-	0.10	0.20	0.13	0.12
Fracture height	ft	16	89	46	45
Fracture half length	ft	50	441	229	180
Dimensionless fracture conductivity	-	0.90	5	2.11	1.40
Productivity index	stb/day/psi	0.05	16.50	1.92	1.32

Figure 2 illustrates a violin plot showing how all parameters in our study distribute their values. This visualization combines box plots and kernel density estimates to depict the data distribution shapes for each parameter. The graphs reveal distinct value ranges and patterns that vary significantly between measurements. Some parameters, like RT, display wide spreads, whereas others, such as PI and K, remain more condensed. Black dots indicating median values cluster around 0.8-1.0 in some cases, while in others, they cluster around 0.2-0.4. The size of each violin indicates the number of values at each position, revealing hidden patterns within the dataset.

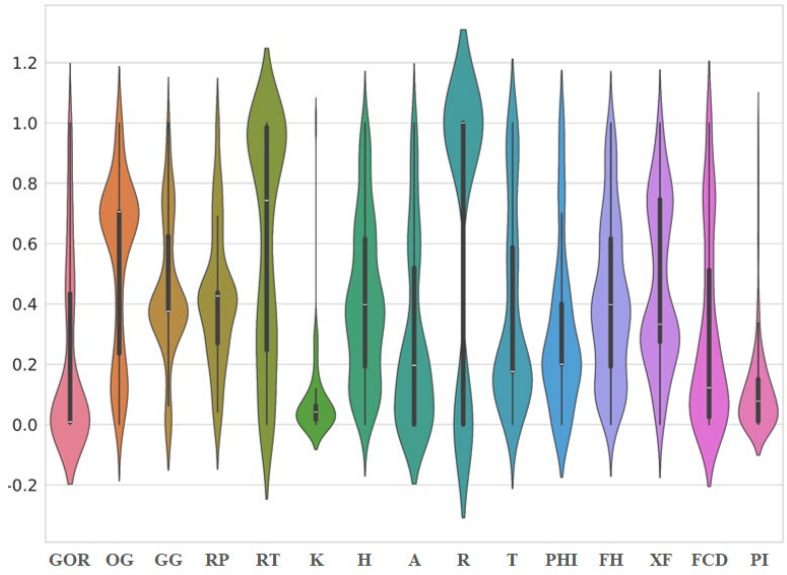


Figure 2. Violin plots for each parameter in the dataset.

2.2. Data preprocessing

Data preprocessing creates the essential foundation for training strong machine learning models by making raw inputs ready for analysis. The initial part of data preparation included thorough checks to remove wrong measurements and abnormal readings. We resolved missing data values either by removing the entire measurement record or by filling gaps with straight-line predictions based on surrounding points. In addition, we used box plot statistics to identify and delete abnormal data points from our study because these points could weaken training accuracy. The final stage involved preparing the data for algorithm consumption by normalizing and structuring input for effective processing. This comprehensive preprocessing framework ensured the generation of top-quality training data that made good starting ground for future modeling work.

Data normalization is a statistical data preprocessing technique that deals with resizing numerical variables to a defined zone, in most cases 0 and 1 or -1 and 1, by applying mathematical transformations such as min-max scaling and standardization of z-scores. This procedure is essential for machine learning model development, where it reduces scale-dependent patterns within features, avoiding characteristics to dominate the learning process due to their scale. Normalization speeds up the convergence of gradient descent, reduces the computational complexity, and diminishes the risk of getting an unstable training model. In addition, it provides fair feature comparison and also enhances model performance by making sure all the features contribute equally and proportionally to learning process especially in the algorithms which are sensitive to the feature scaling such as Neural Networks, K-Nearest Neighbors, and Support Vector Machines.

The Min-Max Scalar normalization method was adopted here to change all numerical data points into values between 0 and 1 using minimum and maximum values. The Min-Max Scaler normalization preserves the original data distribution while scaling features to a fixed range, making it ideal for algorithms sensitive to absolute magnitudes like neural networks and k-NN. Unlike Z-score standardization, it ensures all values remain within a bounded interval, improving interpretability and preventing large-scale features from dominating model training. By using the following equation, the data transformation occurs:

The Min-Max Scalar normalization technique was used here to convert all the numerical data into values ranging from 0 to 1 utilizing the minimum and maximum values. The Min-Max Scaler normalization maintains the original data distribution but scales the features up to a consistent spread, which is best suited for algorithms that are sensitive to the absolute

magnitude, such as neural networks and kNN. Unlike Z-score normalization, it keeps all values within a bounded range and hence improves interpretability and prevents domination of features by the large-scale values. Through the following equation, the data transformation is performed:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

where: X is the original data point; X_{min} is the minimum value in the feature; X_{max} is the maximum value in the feature; X_{scaled} is the normalized value within the [0,1] range.

At last, the preprocessed data was split (80% training, 20% testing). We chose this ratio since it strikes a good balance between enough training data & reliable test set results. This ratio gives sufficient training data to uncover underlying structures and specifications amongst the dataset while having enough testing set for evaluation models. The 80/20 split is quite good for datasets in the middle of the spectrum (usually 1,000-100,000 samples) as it guarantees the test set is large enough to get meaningful performance metrics with decent confidence intervals. This ratio also helps to prevent both underfitting (with a sufficient amount of training data) and overfitting (with the remaining test data). In addition, the 80/20 ratio produced the best performance in our research with the use of cross-validation and hyperparameter optimization.

2.3. Model structure

The development and evaluation of nine models with Python 3.10.12 required a specific set of hyperparameter settings as outlined in Table 2.

The operational learning mechanism of machine learning models depends on hyperparameters, which consist of key parameters including learning rate and model architecture. Data scientists use regularization methods such as L1, L2, and Elastic Net to prevent overfitting through penalty terms added into the loss function. The L1 (Lasso) technique selects features through coefficient zeroing, while L2 (Ridge) controls coefficient sizes. Regulation methods show their most value when analyzing large datasets and limited training samples since they find optimal complexity levels between model complexity and prediction accuracy.

Table 2. Summary of the machine learning models used and their algorithm hyperparameters.

Model	Algorithm parameters
GB	Method: xgboost.
	Number of trees is 100.
	Learning rate is 0.099.
	Regularization is 0.6
	Limit depth is 6.
AB	The number of estimators is 50.
	Learning rate is 1.
	Regression loss function is linear.
RF	Number of trees in the forest is 10.
	Minimum subset size is 5.
SVM	SVM Cost is 1.
	Regression loss epsilon: 0.1.
	Kernel type is radial basis function.
	Numerical tolerance: 0.001
	Iteration limit is 100.
DT	Minimum instances in leaves are 19.
	Minimum subset size: 9.
	Maximal tree depth is 100.
	The stopping point is at 95% of the majority.
KNN	Number of nearest neighbors is 5.
	Metric is Euclidean.
	Weights are uniform.
LR	None.

Model	Algorithm parameters
NN	Neurons per hidden layer are 500.
	Activation is ReLu.
	Solver is L-BFGS.
	Regularization is 0.04.
	Maximum iterations are 500.
SGD	Loss function is hinge.
	Regularization is Lasso (L1).
	Learning rate is constant.
	Initial learning rate is 0.01.
	Number of iterations is 1000.

Feeding a multiple regression model such as Gradient Boosting (GB), Adaptive Boosting (AB), Random Forest (RF), Support Vector Machines (SVMs), Decision Tree (DT), K-Nearest Neighbors (KNN), Logistic Regression (LR), Neural Networks (NN), and Stochastic Gradient Descent (SGD) all on the same data set provides a full cross-comparison of performance. Each model has its own individual mathematical logic and assumptions: Neural networks extract high-dimensional, complex, non-linear relationships; random forests and gradient boosts handle feature interactions so well; linear regression is very useful for interpretable coefficients; SVM is effective in huge dimensional data; tree models treat categorical features very well; SGD excels at handling large datasets efficiently; and KNN is well suited to handle local patterns. This broad approach enables the determination of which model is most suited to the nature of our dataset, brings out different facets of the underlying relationships, and will lead to more accurate predictions.

Figure 3 illustrates the Pythagorean Forest visualization, which gives a view of a collection of 10 decision trees from a Random Forest model, each tree arranged as a Pythagorean tree shape having distinct branching patterns. The visual encoding relies on geometry to encode important model insights: branch width and brightness stand for decision path importance, and colors denote the range of regression predictions. By using sophisticated algorithms in distance metrics, Random Forest resulted in the best stability of the model, effectively managed the bias-variance trade-off, and minimized overfitting. This upgraded framework outperforms the standard black-box methods, such as traditional neural networks, in capturing the relationships of features and keeping transparency. This combined visual-algorithmic strategy of the Pythagorean Forest allows for a close inspection of the decision borders, variable influence and prediction reliability all at once across the full tree ensemble.

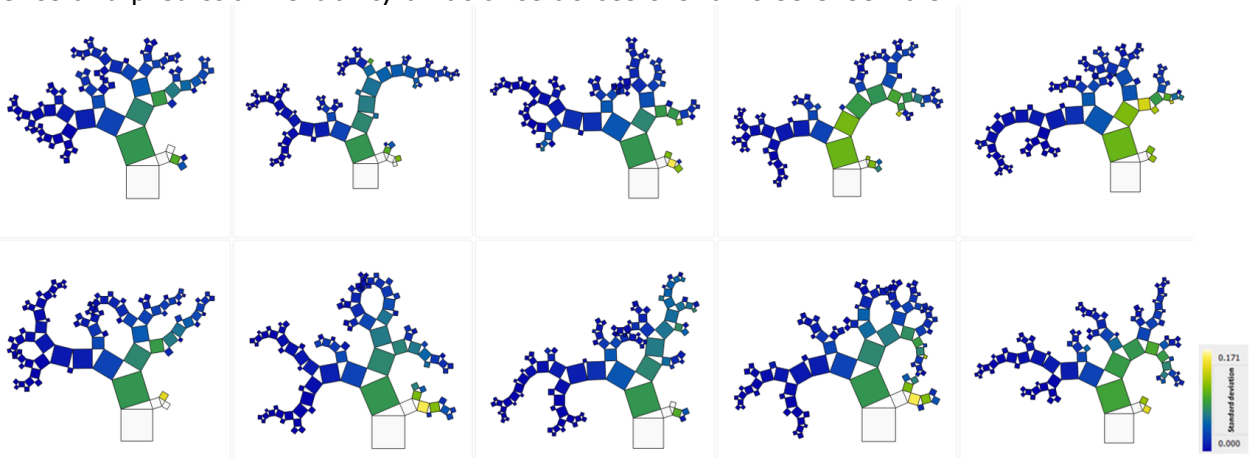


Figure 3. Pythagorean forest shows all learned decision tree models from the RF model.

3. Results and discussion

3.1. Model results

The analysis results from models by machine learning appear in Figure 4. MSE (Mean Squared Error), RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), and R^2 (Coefficient of Determination) are four commonly used metrics to measure one's regression model performance, and each one gives a different kind of insight into the prediction's accuracy. MSE is the average of the squared difference between predicted and actual values, heavily punishing big errors because of squaring but having no direct interpretation on the original unit scale. RMSE, the square root of MSE, keeps the same source unit scale while keeping the same square penalty, making it more understandable than MSE. MAE finds the average absolute difference of predictions from actuals, more robust against outliers than MSE/RMSE but possibly down weighting of big errors as all deviations are considered linear. The value 0 to 1 associated with R^2 indicates proportion of the variance in the dependent variable that is explained by the model, with higher numbers representing closer fit. Smaller MSE, RMSE, and MAE values represent better model performance and are closer to zero, whereas larger R^2 values (closer to 1) indicate better fit of the model. By considering these metrics as a whole series, it offers a coherent framework to assess model performance: MSE / RMSE give an importance to large offsets and to model's stability, MAE give an information about the average accuracy of prediction which is less sensitive to outliers, R^2 provide a dimensionless measure of model's explanatory power comparable across different dataset. This multi-metric approach allows us to evaluate both absolute and relative model performance, to know error distributions, and to conduct informed decisions about model choice based on error tolerance.

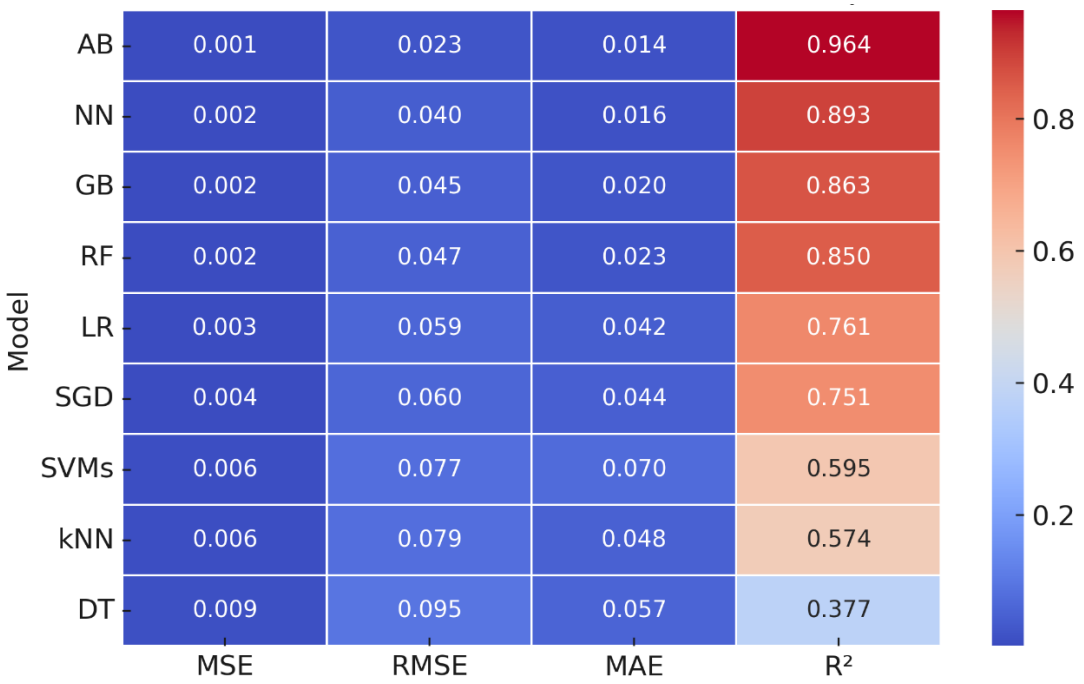


Figure 4. Developed models performance metrics heatmap.

Figure 5 presents the normalized anticipated productivity index measurements after hydraulic fracturing from the output of AB that match actual productivity index measurements obtained from the actual bottomhole pressure gauge and production test at the surface. While lying near the 45° straight line in results, the predictions made from the machine learning model stay equally distributed. This indicates that the built AB resulted in accurate predictions of PI values spanning the whole spectrum of the parameters in the data set.

The most significant factors associated with PI prediction by the AB model are presented as SHAP (SHapley Additive exPlanations) in Figure 6. SHAP is a game-theoretic approach for explaining machine learning model predictions by computing feature contributions as a cooperative game. Unlike conventional feature importance methods, SHAP evaluates each feature for an individual prediction by examining all possible feature combinations and their corresponding marginal contributions. The SHAP summary plot shows feature impacts through a beeswarm plot, where each point represents the Shapley value for a feature and an instance, color is used for feature value (red for high and blue for low) and the position indicates the magnitude and direction of the effect on the model output. SHAP has several benefits over traditional feature importance methods, such as permutation importance, LIME, and feature importance plots from tree-based models. SHAP provides consistent local and global interpretability, handles interactions between the features, and is more accurate in attributing the feature while taking into account non-linear relationships and complex feature interactions. Among the examined parameters, K, FH, H, and XF have the highest effect on predicted PI predictions, but other features have a smaller impact.

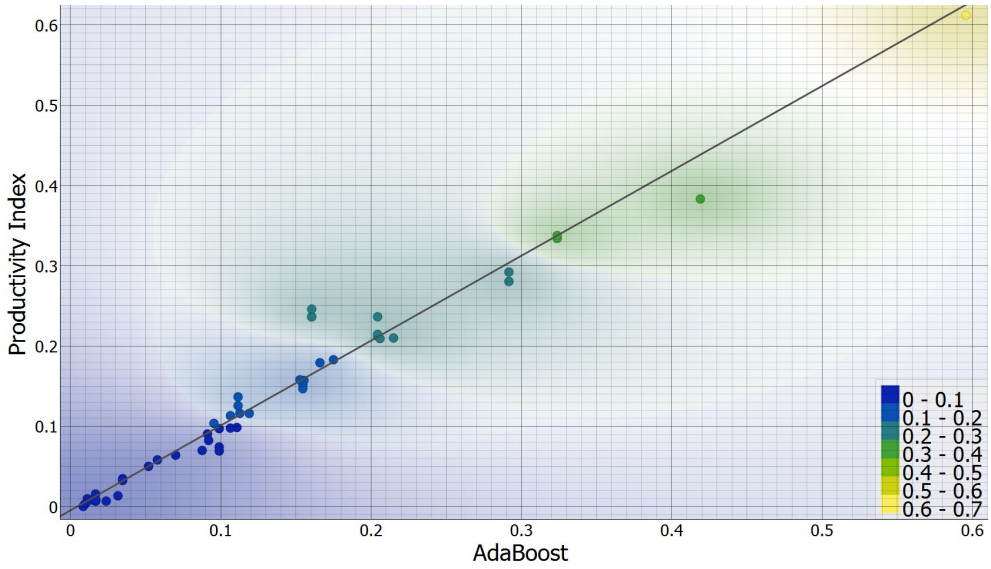


Figure 5. Predicted vs. actual normalized productivity index using AdaBoost model.

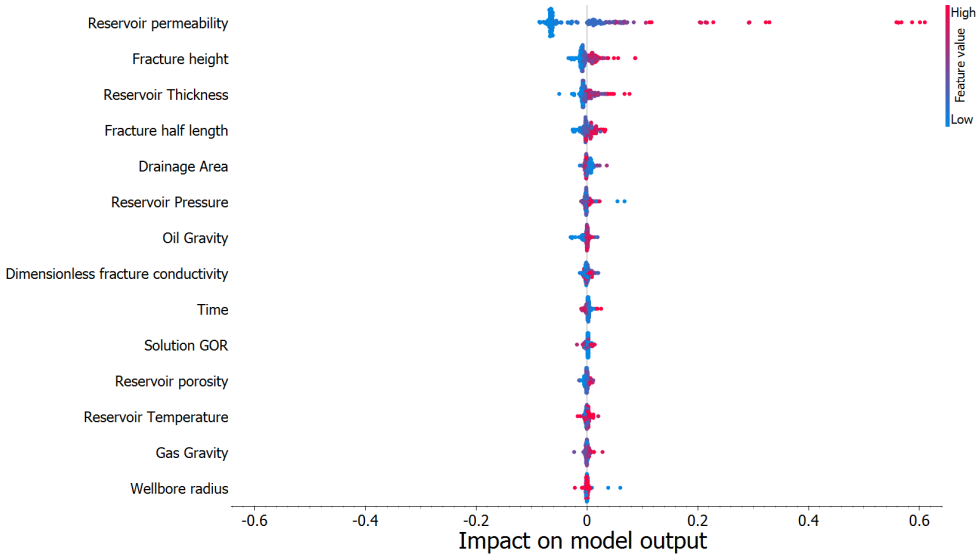


Figure 6. SHAP plot of the AdaBoost model.

3.2. Model testing and validation

The assessment of the accuracy of the machine learning model is done by two major methods. The first technique is the K-fold cross-validation, which is a robust model validation that divides the dataset into k equal parts, such that k-1 parts are trained, and the prediction is performed on the left-out section for all possible folding combinations. This repetitive cycle results in k distinct models each one used of a different validation fold groove a more robust measurement of the model's performance by counteracting sampling bias and variance. The method is especially important for tackling overfitting issues, as well as for evaluating model generalization properties since it allows for performance analysis across a variety of data correspondences distributed within the collection of the data. The general model performance metrics are usually averaged over all k iterations, giving a more reliable and representative indication of the model's robustness compared to a single train-test split. Figure 7 shows the results of the K-fold cross-validation, which was implemented with 10 folds due to evaluation.

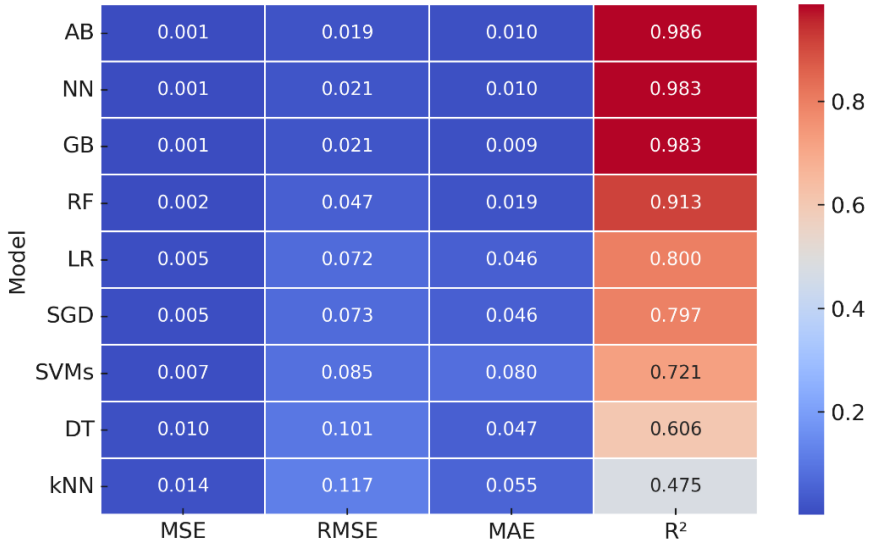


Figure 7. K-fold cross-validation models performance metrics heatmap.

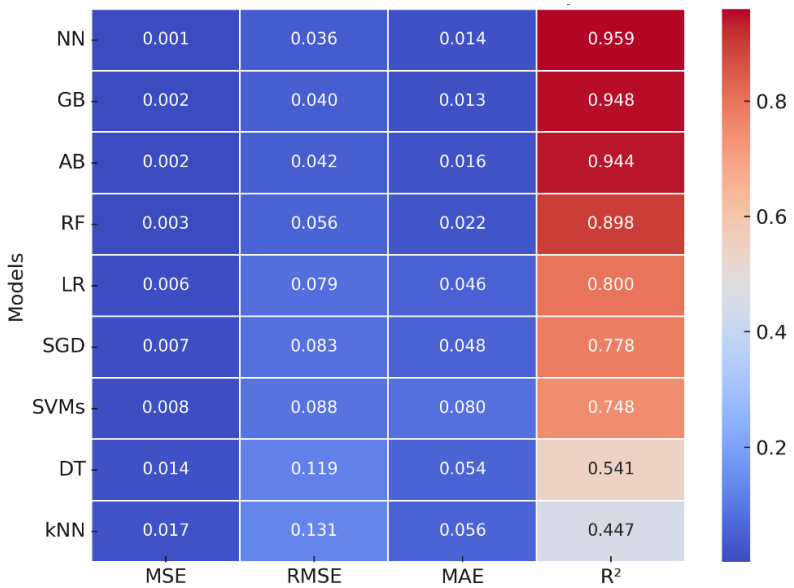


Figure 8. Random sampling cross-validation models performance metrics heatmap.

Random sampling cross-validation (also called Monte Carlo cross-validation) divides the data into random training and validation sets repeatedly, without guaranteeing that every observation is used for validation. It is essential for the prediction of model generalization performance and overfitting minimization. In contrast to k-fold cross-validation, random sampling has the advantage of allowing more flexibility in terms of the split fractions and could be computationally effective on big data. However, it has greater variance in performance estimates and is not able to validate every single observation. K-fold cross-validation avoids using any single observation for validation multiple times, making the results more reliable and stable, but demanding k runs of model training and heavy computation for the large dataset and complex models. The repeated random sampling technique produced its results through 10 individual trials, which are summarized in Figure 8. Results include MSE, RMSE, MAE and R^2 for every model based in the evaluation.

3.3. Field application

Well-X is a producing oil well in the Egyptian Western Desert. Hydraulic fracturing treatment was considered the best method to enhance PI for the Formation Y. So, a fracture string with bottomhole pressure gauges was run in the well to record flowing bottomhole pressure during a production test at the surface. The actual PI was 2.2 stb/day/psi based on measured flowing bottomhole pressure and surface fluid production rate. A summary of all parameters available is shown in Table 3. The data from Well-X was used as input into running the developed AB model. The model's performance was very impressive when the prediction was against the actual PI of 2.2 stb/day/psi. The model predicted a PI of 2.13 stb/day/psi with only a 0.07 difference in absolute terms from the recorded pressure. This translates to an extremely low percentage error of 3.2% which is proof without a doubt of the accuracy of the model's predictions.

The AB model’s accurate field application at Well-X demonstrates its potential for accurate PI prediction of individual wells. This innovation not only allows hydraulic fracturing treatment optimization, but it also enables optimization of artificial lift selection and reservoir management approaches. Besides, it also offers a solid economic justification for conducting the treatment, as companies have to justify the cost of hydraulic fracturing against forecasted production increases. These developments in predictive modeling are essential tools for optimizing the efficiency of stimulation treatments and increasing production economies. This is particularly critical in the complex reservoirs like Egypt's Western Desert where traditional PI prediction methods are typically costly, time consuming and unreliable.

Table 3. Well-X input parameters.

Parameter	Unit	Value
Solution GOR	scf/stb	500
Oil gravity	API	40
Gas specific gravity	-	0.65
Reservoir pressure	Psi	2600
Reservoir temperature	Deg f	276
Reservoir permeability	Md	3
Reservoir thickness	Ft	50
Drainage area	Acres	70
Wellbore radius	Ft	0.354
Time	Days	1
Reservoir porosity	-	0.13
Fracture height	Ft	50
Fracture half length	Ft	362
Dimensionless fracture conductivity	-	1.2
Productivity index	Stb/day/psi	2.2

4. Conclusion

The results of this research show that advanced machine learning models are very effective in accurately forecasting the post-fracturing productivity index (PI) and can be considered as a promising alternative to the current empirical, analytical, numerical, and well testing methods. The models were trained on data from 258 wells and showed very good predictive capability with the lowest RMSE of 0.019 and the highest R^2 of 0.986 obtained by AdaBoost, followed closely by Neural Networks (RMSE = 0.021, R^2 = 0.983) and Gradient Boosting (RMSE = 0.021, R^2 = 0.983). The Random Forest model, while less accurate (RMSE = 0.047, R^2 = 0.913), was still better than the conventional estimation techniques. These results show that machine learning methods are suitable for accurate post-fracturing PI prediction, which can help in decision making and reduce costs and time of the conventional modelling. Moreover, the successful use of AdaBoost in predicting the PI for Well-X with only 3.2% prediction error shows the effectiveness of the models developed in this study. The incorporation of feature importance analysis, especially through SHAP values, increases the interpretability of machine learning forecasts and identifies important features like permeability, reservoir thickness, and fracture half-length as important input factors that affect the PI estimation. This study has established the capability of applying ML in the optimization of hydraulic fracturing and opens up new possibilities of data-assisted management of reservoirs. It is recommended that future work should focus on the integration of physics-informed ML models and a larger dataset to enhance the predictive capacity and feasibility for field application.

Nomenclature

A	*Drainage Area*
AB	*Adaptive Boosting model*
Ada-Boost	*Adaptive Boosting model*
Adam	*Adaptive Moment Estimation optimization algorithm*
DT	*Decision Trees*
FCD	*Dimensionless fracture conductivity*
FH	*Fracture height*
FOI	*Fold of increase*
GB	*Gradient Boosting*
GG	*Gas Gravity*
GOR	*Solution GOR*
H	*Reservoir Thickness*
K	*Reservoir permeability*
KGD	*Khristianovic-Geertsma-de Klerk*
KNN	*K-Nearest Neighbor*
L-BFGS	*Limited-memory - Broyden - Fletcher - Goldfarb - Shanno optimization algorithm*
LR	*Linear Regression*
MAE	*Mean absolute error*
MAPE	*Mean absolute percent error*
ML	*Machine learning*
MSE	*Mean square error*
NN	*Neural Network*
OG	*Oil Gravity*
P3D	*Pseudo-3D model*
PHI	*Reservoir porosity*

PI	*Productivity Index*
PKN	*Perkins-Kern-Nordgren model*
PL3D	*Planar 3D model*
R	*Wellbore radius*
R^2	*Correlation coefficients*
RF	*Random Forest*
RMSE	*Root mean square error*
RP	*Reservoir Pressure*
RT	*Reservoir Temperature*
SGD	*Stochastic Gradient Descent*
SHAP	*SHapley Additive exPlanations*
SVMs	*Support Vector Machines*
T	*Time*
XAI	*Explainable AI techniques*
XF	*Fracture half length*

Disclosure statement: *No potential conflict of interest was reported by the authors.*

Data availability statement

The data that support the findings of this study are available from the corresponding author, A.G, upon reasonable request.

Funding: *The authors reported there is no funding associated with the work featured in this article.*

References

[1] Rueda J, Alzamil A, Alowaid A, Valbuena J, Alnaimi R. Maximizing Productivity Index by Minimizing Fluid, Proppant and Additive Volumes. In: Middle East Oil, Gas and Geosciences Show. SPE; 2023:D021S070R003. https://doi.org:10.2118/213271-MS

[2] Helmy AG, Salem SK, Elnoby M. Sensitivity Analysis of Hydraulic Fracturing Parameters for Optimum Horizontal Well Spacing in Tight Oil Reservoirs. Arab J Sci Eng. 2025;50(7):4813-4832. https://doi.org:10.1007/s13369-024-09311-z

[3] Zhang S, Zheng D, Yu M, Jakaria M. Development of Hydraulic Fracturing Evaluation Techniques Using Microchip Sensing System. In: 58th U.S. Rock Mechanics/Geomechanics Symposium. ARMA; 2024:D032S043R012. doi: https://doi.org:10.56952/ARMA-2024-0301

[4] Khan AM, Isaev V, Belyakova L. Multimaterial Multiphysics Modelling Coupled with Post-Fracturing Production Flow Simulations: Revamping Hydraulic Fracturing Design Strategy. In: SPE Asia Pacific Oil & Gas Conference and Exhibition. SPE; 2022:D032S008R002. https://doi.org:10.2118/210633-MS

[5] Asadi S, Ghosh A, Chrislianto R. Hydraulic Fracturing Design Optimization Using a Fieldwide Calibrated Geomechanics Model. In: APOGCE 2024. SPE; 2024:D031S024R001. https://doi.org:10.2118/221130-MS

[6] Bird A, Casero A, Ligocki L, Parada-Machado C, Hoad J, Sesetty V. Production and Hydraulic Fracturing Design Optimization in the Presence of Natural Fractures, Clair Field. In: SPE Hydraulic Fracturing Technology Conference and Exhibition. SPE; 2023:D021S005R005. https://doi.org:10.2118/212356-MS

[7] Thabet SA, Zidan HA, Elhadidy AA, et al. Application of Machine Learning and Deep Learning to Predict Production Rate of Sucker Rod Pump Wells. In: GOTECH. SPE; 2024:D021S035R001. https://doi.org:10.2118/219231-MS

[8] Al Samli H, Al Shueili A, Trejo RS, Kayumov R, Jaboob M. Hydraulic Fracturing Design Optimization Guidelines for Heterogeneous Reservoir. In: SPE International Hydraulic Fracturing Technology Conference & Exhibition. SPE; 2022:D031S011R004. https://doi.org:10.2118/205291-MS

[9] Prasetio MH, Anggraini H, Tjahjono H, et al. Success Story of Optimizing Hydraulic Fracturing Design at Alpha Low-Permeability Reservoir. In: SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition. SPE; 2021:D012S032R060. https://doi.org:10.2118/205703-MS

[10] Crawford HR. Proppant Scheduling and Calculation of Fluid Lost During Fracturing. In: SPE Annual Technical Conference and Exhibition. SPE; 1983:SPE-12064-MS. https://doi.org:10.2118/12064-MS
[11] Salazar F, Vasconez ND, Mayorga CJ, Pozo MF. Hydrocarbon Production Enhancement on the Field Apaika-Nenke by Applying Hydraulic Fracturing as a Tool to Optimize the Extraction of Reservoir M-2 Considered Initially as a Secondary Target. In: SPE Latin American and Caribbean Petroleum Engineering Conference. SPE; 2023:D021S008R003. https://doi.org:10.2118/213110-MS
[12] Zhang S, Tang H, Hurt R, Jayaram V, Wagner J. Joint Interpretation of Fiber Optics and Downhole Gauge Data for Near Wellbore Region Characterization. In: SPE Hydraulic Fracturing Technology Conference and Exhibition. SPE; 2020:D021S004R002. https://doi.org:10.2118/199715-MS
[13] Bhatia K, Pande K. First Application of Nonradioactive Tracer Technology in CSG Unconventional Basin in Central India: Optimization and Evaluation of Fracturing Treatment. In: SPE Asia Pacific Hydraulic Fracturing Conference. SPE; 2016:D021S004R004. https://doi.org:10.2118/181782-MS
[14] Meyer BR, Bazan LW, Brown EK, Brinzer BC. Key Parameters Affecting Successful Hydraulic Fracture Design and Optimized Production in Unconventional Wells. In: SPE Eastern Regional Meeting. SPE; 2013:SPE-165702-MS. https://doi.org:10.2118/165702-MS
[15] Syfan FE, Palisch TT, Dawson JC. 65 Years of Fracturing Experience: The Key to Better Productivity is Not What we Have Learned but What we Have Forgotten and Failed to Utilize! In: SPE Annual Technical Conference and Exhibition. SPE; 2013:D031S037R005. https://doi.org:10.2118/166107-MS
[16] Ding Y, Yang D, Huang H, Wang H. Numerical Simulation of Proppant Transport and Placement in Hydraulic Fractures with the Hybrid Perkins-Kern-Nordgren-Carter (PKN-C) Model and Particle Tracking Algorithm. SPE Journal. 2022;27(06):3914-3937. https://doi.org:10.2118/210561-PA
[17] Majeed AJ, Yaseen DT, Hassan MA, Al-Mukhtar AM. Enhancing Realism in Hydraulic Fracturing Simulation Models: The Evolution of KGD and PKN Models. Procedia Structural Integrity. 2024;66:212-220. https://doi.org:10.1016/j.prostr.2024.11.072
[18] Barree RD. A Practical Numerical Simulator for Three-Dimensional Fracture Propagation in Heterogeneous Media. In: SPE Reservoir Simulation Symposium. SPE; 1983:SPE-12273-MS. https://doi.org:10.2118/12273-MS
[19] Raghavan R, Chen C. Fractional diffusion in rocks produced by horizontal wells with multiple, transverse hydraulic fractures of finite conductivity. Journal of Petroleum Science and Engineering. 2013;109:133-143. https://doi.org:10.1016/j.petrol.2013.08.027
[20] Ismail A, Azadbakht S. A comprehensive review of numerical simulation methods for hydraulic fracturing. International Journal for Numerical and Analytical Methods in Geomechanics. 2024;48(5):1433-1459. https://doi.org:10.1002/nag.3689
[21] Yu G, Aguilera R. 3D Analytical Modeling of Hydraulic Fracturing Stimulated Reservoir Volume. In: SPE Latin America and Caribbean Petroleum Engineering Conference. SPE; 2012:SPE-153486-MS. https://doi.org:10.2118/153486-MS
[22] Rahman MK, Rahman MM, Joarder AH. Analytical Production Modeling for Hydraulically Fractured Gas Reservoirs. Petroleum Science and Technology. 2007;25(6):683-704. https://doi.org:10.1080/10916460500411663
[23] Bakhshi E, Golsanami N, Chen L. Numerical Modeling and Lattice Method for Characterizing Hydraulic Fracture Propagation: A Review of the Numerical, Experimental, and Field Studies. Arch Computat Methods Eng. 2021;28(5):3329-3360. https://doi.org:10.1007/s11831-020-09501-6
[24] Ju Y, Wu G, Wang Y, Liu P, Yang Y. 3D Numerical Model for Hydraulic Fracture Propagation in Tight Ductile Reservoirs, Considering Multiple Influencing Factors via the Entropy Weight Method. SPE Journal. 2021;26(05):2685-2702. https://doi.org:10.2118/205385-PA
[25] Nguyen HT, Lee JH, Elraies KA. Review of pseudo-three-dimensional modeling approaches in hydraulic fracturing. J Petrol Explor Prod Technol. 2022;12(4):1095-1107. https://doi.org:10.1007/s13202-021-01373-1
[26] Nguyen HT, Lee JH, Elraies KA. Pseudo-three-dimensional simulations of a vertical hydraulic fracture with time-dependent injection rates. Petroleum Science and Technology. 2022;40(11):1349-1368. https://doi.org:10.1080/10916466.2021.2020818

[27] Wojnarowski P, Janiga D. Study of the Effectiveness of the Egs/Hdr Fracturing Process with the Application of a 3d Planar Fracture Propagation Model. Social Science Research Network. Preprint posted online January 26, 2024. https://doi.org:10.2139/ssrn.4707654
[28] Wang Y, Chen M, Guo T, et al. Proppant Transport Coupled with Propagation of Three-Dimensional Hydraulic-Natural Fractures. Rock Mech Rock Eng. Published online December 4, 2024. https://doi.org:10.1007/s00603-024-04264-z
[29] Liu B, Ghassemi A. Simulation of Proppant Placement in Multiple Simultaneously Propagating Hydraulic Fractures in Horizontal Wells. Rock Mech Rock Eng. 2024;57(11):9635-9650. https://doi.org:10.1007/s00603-024-04022-1
[30] Chen B, Barboza BR, Sun Y, et al. A Review of Hydraulic Fracturing Simulation. Arch Computat Methods Eng. 2022;29(4):1-58. doi:10.1007/s11831-021-09653-z
[31] Wang Z, Li C, King MJ. Applications of Asymptotic Solutions of the Diffusivity Equation to Infinite Acting Pressure Transient Analysis. SPE Journal. 2024;29(08):4069-4093. https://doi.org:10.2118/180149-PA
[32] Wang Z, Ning Z, Jia Z, et al. Estimation of the Half-Length of Non-Simultaneous-Closed Fracture Through Pressure Transient Analysis: Model and Case Study. In: SPE EuropEC - Europe Energy Conference Featured at the 83rd EAGE Annual Conference & Exhibition. SPE; 2022:D042S011R004. https://doi.org:10.2118/209716-MS
[33] Chu H, Liao X, Wei C, Lee J. Rate/Pressure Transient Analysis of a Variable Bottom Hole Pressure Multi-Well Horizontal Pad with Well Interference. In: SPE Annual Technical Conference and Exhibition. SPE; 2021:D021S023R002. https://doi.org:10.2118/206046-MS
[34] Shamsiev MN, Gadil'shina VR. Pressure-Transient Analysis of Vertical Wells in Low-Permeability Reservoirs. Lobachevskii J Math. 2022;43(5):1195-1200. https://doi.org:10.1134/S1995080222080297
[35] Thabet SA, El-Hadydy AA, Gabry MA. Machine Learning Models to Predict Pressure at a Coiled Tubing Nozzle's Outlet During Nitrogen Lifting. In: SPE/ICoTA Well Intervention Conference and Exhibition. SPE; 2024:D021S013R003. https://doi.org:10.2118/218294-MS
[36] Nashed S, Moghanloo R. Replacing Gauges with Algorithms: Predicting Bottomhole Pressure in Hydraulic Fracturing Using Advanced Machine Learning. Eng. 2025;6(4):73. https://doi.org:10.3390/eng6040073
[37] Gharieb A, Gabry MA, Elsawy M, et al. Data Analytics and Machine Learning Application for Reservoir Potential Prediction in Vuggy Carbonate Reservoirs Using Conventional Well Logging. In: SPE Western Regional Meeting. SPE; 2024:D031S015R006. https://doi.org:10.2118/218913-MS
[38] Li D, You S, Liao Q, Sheng M, Tian S. Prediction of Shale Gas Production by Hydraulic Fracturing in Changning Area Using Machine Learning Algorithms. Transp Porous Med. 2023;149(1):373-388. https://doi.org:10.1007/s11242-023-01935-3
[39] Zhao P, Gray KE. Analytical and Machine-Learning Analysis of Hydraulic Fracture-Induced Natural Fracture Slip. SPE Journal. 2021;26(04):1722-1738. https://doi.org:10.2118/205346-PA
[40] Thabet S, Zidan HM, Elhadidy A, et al. Prediction of Total Skin Factor in Perforated Wells Using Models Powered by Deep Learning and Machine Learning. In: GOTECH. SPE; 2024:D011S004R001. https://doi.org:10.2118/219187-MS
[41] Alarifi SA, Miskimins J. A New Approach To Estimating Ultimate Recovery for Multistage Hydraulically Fractured Horizontal Wells by Utilizing Completion Parameters Using Machine Learning. SPE Production & Operations. 2021;36(03):468-483. https://doi.org:10.2118/204470-PA
[42] Nashed S, Lnu S, Guezei A, Ejehu O, Moghanloo R. Downhole Camera Runs Validate the Capability of Machine Learning Models to Accurately Predict Perforation Entry Hole Diameter. Energies. 2024;17(22):5558. https://doi.org:10.3390/en17225558
[43] Gharieb A, Gabry MA, Algarhy A, et al. Real-Time Identification of Geological Factuals by Integrating Formation Micro Resistivity Imaging Logs With Computer Vision. In: GOTECH. SPE; 2024:D021S019R001. https://doi.org:10.2118/219194-MS
[44] Sprunger C, Muther T, Syed FI, Dahaghi AK, Neghabhan S. State of the art progress in hydraulic fracture modeling using AI/ML techniques. Model Earth Syst Environ. 2022;8(1):1-13. https://doi.org:10.1007/s40808-021-01111-w
[45] Ibrahim AF, Alarifi SA, Elkatatny S. Application of Machine Learning to Predict Estimated Ultimate Recovery for Multistage Hydraulically Fractured Wells in Niobrara Shale Formation. Computational Intelligence and Neuroscience. 2022;2022(1):7084514. https://doi.org:10.1155/2022/7084514

[46] Khan AM, Luo Y, Ugarte E, Bannikov D. Physics-Informed Machine Learning for Hydraulic Fracturing—Part I: The Backbone Model. In: SPE Conference at Oman Petroleum & Energy Show. SPE; 2024:D021S016R002. https://doi.org:10.2118/218562-MS
[47] Alhemdi A, Gu M. Optimizing Unconventional Hydraulic Fracturing Design Using Machine Learning and Artificial Intelligent. In: SPE Western Regional Meeting. SPE; 2022:D031S015R008. https://doi.org:10.2118/209269-MS
[48] Gharieb A, Adel Gabry M, Algarhy A, et al. Revealing Insights in Evaluating Tight Carbonate Reservoirs: Significant Discoveries via Statistical Modeling. An In-Depth Analysis Using Integrated Machine Learning Strategies. In: GOTECH. SPE; 2024:D011S012R003. https://doi.org:10.2118/219199-MS
[49] Khan AM, Ugarte E, Kurniadi S, Al Shueili A, Wang WK. Physics-Informed Machine Learning for Hydraulic Fracturing—Part III: The Transfer Learning Validation. In: International Petroleum Technology Conference. IPTC; 2025:D022S005R007. https://doi.org:10.2523/IPTC-24850-MS
[50] Fasola SL, Brudzinski MR. Machine Learning Reveals Additional Hydraulic Fracture-Induced Seismicity in the Eagle Ford Shale. Journal of Geophysical Research: Solid Earth. 2023;128(2):e2022JB025436. https://doi.org:10.1029/2022JB025436
[51] El Khouly I, Sabet A, El-Fattah MAA, Bulatnikov M. Integration Between Different Hydraulic Fracturing Techniques and Machine Learning in Optimizing and Evaluating Hydraulic Fracturing Treatment. In: International Petroleum Technology Conference. IPTC; 2024:D021S084R003. https://doi.org:10.2523/IPTC-24296-MS
[52] Gharieb A, Elshaafie A, Gabry MA, Algarhy A, Elsawy M, Darraj N. Exploring an Alternative Approach for Predicting Relative Permeability Curves from Production Data: A Comparative Analysis Employing Machine and Deep Learning Techniques. In: Offshore Technology Conference. OTC; 2024:D041S054R003. https://doi.org:10.4043/35037-MS
[53] Barghoth T. M., Elramah S. G., Salem A. M. Pore and Fracture Pressures Real Time Prediction during Drilling Operation Using a Regression Ensemble Machine Learning Model. Petroleum & Coal, 2024; 66(4).

To whom correspondence should be addressed: Ahmed Gamal, Faculty of Petroleum and Mining Engineering, Suez University, Suez, Egypt, E-mail: Ahmed.HeHa@pme.suezuni.edu.eg