SciEn

Proceedings of 8th International Conference on Mechanical Engineering and Renewable Energy 2025
10-12 December, 2025, Chattogram, BANGLADESH

SciEn Conference Series: Engineering Vol. 4, 2026, pp 105-110 https://doi.org/10.38032/scse.2026.4.57

Operational Threshold-Embedded Machine Learning Models for Accurate Hydraulic Fracturing Monitoring

Fardin Ahmed[1,] and Md. Nahul Rahman[2]*

[1]Department of Petroleum & Mining Engineering, Military Institute of Science & Technology, Dhaka-1216, Bangladesh
[2]Department of Computer Science & Engineering, Military Institute of Science & Technology, Dhaka-1216, Bangladesh

ABSTRACT

Hydraulic fracturing is critical process in unconventional reservoir development, where accurate real time monitoring directly impacts operational safety, job efficiency and financial aspect. This study presents a hybrid framework integrating domain specific operational thresholds with supervised machine learning for anomaly detection and multi parameter prediction. In anomaly detection, XGBoost achieved the highest accuracy of 0.99 and F1-score of 0.98, outperforming random forest (0.97), logistic regression (0.65), KNN (0.90) across traditional evaluation metrics. In predictive model, gradient boosting algorithm achieved the highest R^2 score of 0.97 and the lowest RMSE, outperforming both linear and ridge regression model used for base model. The inclusion of operational constraints into the machine learning model enhanced reliability, prediction accuracy, robustness ensuring practical and actionable results for hydraulic fracturing processes.

Keywords: Hydraulic Fracturing, Machine Learning, Anomaly Detection, Multi-Parameter Prediction, Operational Thresholds.

1. Introduction

Hydraulic fracturing (HF) is a revolutionary technique that has unlocked vast reserves of hydrocarbon all over the world from unconventional formations such as shale, tight gas and coalbed methane. In the United States alone, HF is applied to over 95% of new oil and gas wells, contributing to approximately two-thirds of the nation's gas production and more than 50% of crude oil output [1], [2]. This technology has been central to the shale revolution, permitting economically viable production from formations with too low permeabilities [3]. Globally, HF has been instrumental in stabilizing energy supplies, with major operations in North America, Argentina's Vaca Muerta Shale, and China's Sichuan Basin [4], [5].

An HF job involves pumping a carefully engineered mixture of base fluid, proppant, and chemical additives at high pressure to initiate and extend fractures in the reservoir. Key operational parameters routinely monitored during the job include Percent HF Job, Mass Ingredient, Total Base Water Volume, and True Vertical Depth [6], [7]. In addition, slurry rate, surface treating pressure, proppant concentration, fluid viscosity, bottomhole pressure, tubing and casing pressure, and pump rate are tracked also [8], [9], [10]. These diversify in parameters can help to track the fracture propagation efficiency, formation response or operational integrity.

Real time assessment of these metrics is essential. For example, a sudden increase in surface treating pressure may signal a near screen out condition, requiring immediate adjustment of pump rates or proppant concentration [11]. Conversely, sudden reduction in slurry rate without a corresponding control action could indicate mechanical failure or fluid loss to unintended zones. The cost of an HF operation can range from USD 2-10 million per well, even the short periods of inefficiency or downtime can have substantial economic consequences [12], [13].

Generally, anomaly detection in HF has relied on fixed operational thresholds derived from the historical field experience and engineering best practices [14]. While these rule-based approaches are simple and interpretable, they are often static, unable to adapt to variable reservoir conditions, stage-specific designs, or evolving operational strategies [15]. This rigidity increases the risk of false alarms and missed detections, especially in multi-stage HF programs where operating parameters shift over time.

Machine learning (ML) has emerged as a powerful complement to these traditional methods, capable of analyzing large volumes of multi-sensor data, identifying non-linear relationships between parameters, and providing predictive insights in real time [16], [17], [18]. In HF, ML has been applied to fracture geometry estimation [19], production forecasting [3], and frac-hit prediction in multi-well pads [5]. Advanced algorithms such as gradient boosting machines (GBMs), random forests (RF), extreme gradient boosting (XGBoost), and deep neural networks (DNNs) have shown the ability to improve predictive accuracy and can influence operational decisions.

However, despite these advances, most ML applications in HF focused on post-job analysis or single-target predictions, with limited emphasis on integrating domain-specific operational thresholds into the modeling process or on multi-output regression for simultaneous prediction of interdependent operational parameters [6], [20]. In addition, real-time anomaly detection etc. remains underrepresented in HF research, leaving a critical gap in operational risk mitigation [7], [21].

Addressing these challenges, this study proposes a hybrid anomaly detection and predictive modeling framework. This framework integrates domain knowledge-based operational thresholds, derived from industry best practices and historical field data, with supervised machine learning classifiers to detect anomalies in critical fracturing parameters. In parallel, a multi-output regression model is employed to predict percent hf job, mass ingredient, total base water volume simultaneously. By uniting domain expertise with data-driven adaptability, the proposed approach aims to deliver more accurate anomaly detection, improved predictive reliability, and enhanced interpretability.

2. Literature Review

Hydraulic fracturing (HF) has been extensively investigated over the past decades, with early works by Hubbert and Willis [22] and Perkins and Kern [23], provided the foundational physics for fracture initiation and propagation, while subsequent models incorporated proppant transport and leak-off behavior. These physics-based models, faced scalability challenges in multi-stage unconventional wells, where complex geo mechanical heterogeneity and operational variability exists [24].

Thresholds based anomaly detection remains a widely applied method in field monitoring where operational limits for parameters such as slurry rate, surface treating pressure, proppant concentration, proppant pump rates, and fluid volume based on prior well performance and equipment ratings are set. Deviations beyond these limits are flagged as potential issues, such as screen outs, premature pressure spikes, or excessive proppant settling. Yang. et al. [25] applied a gradient boosting regression model to optimize fracturing parameters in tight gas reservoirs, using a dataset of over 3200 fracturing stages from multiple wells the model achieved an R^2 of 0.948 and reduced the mean absolute error for fluid volume prediction to 4.7%, outperforming traditional multiple linear regression by more than 30% in accuracy. Similarly, karami et al. demonstrated the practical reliability of ML models in predicting fracture width, with RF model achieving an RMSE of 6.194.[26]

The emergence of machine learning (ML) has brought adoptive and predictive ability to HF data analysis. Methods such as random forests (RF), gradient boosting machines (GBMs), support vector machines (SVMs), and deep neural networks (DNNs) has been employed for tasks ranging from fracture geometry estimation, production forecasting, and frac-hit risk assessment to optimization of stage designs. Ma and Ye [27] showed that field data in ML-based preprocessing, forecasting and real-time optimization in hydraulic fracturing. Morozov et al. [28] introduced a CatBoost based workflow leveraging a large field dataset for production forecasting, enhancing reliability over traditional models. Similarly, wang et al. [15] applied an XGBoost model for bottomhole pressure prediction, achieving a mean absolute percentage error (MAPE) of 6.8%, enabling proactive pump rate adjustments to avoid operational failures.

While these ML driven methods show marked improvements over static thresholding, three major gaps remain. First, most models are purely data driven, lacking domain knowledge based operational rules that can act as physical constraints, thereby risking operationally infeasible predictions [16]. Second, existing studies often target single parameter prediction, such as only proppant concentration or pressure, without accounting for correlations between multiple operational metrics [17]. For instance, field datasets from the eagle Ford Shale show strong correlations between slurry rate, proppant mass, and job completion percentage, indicating interdependencies that single-output models overlook [18]. Third, much of ML work focuses on post job analyses; for example, Ali Mahmudi et al [19] used neural networks to predict fracture half-length after completion, but such retrospective outputs cannot support immediate operational decision-making.

Hybrid approaches that embed operational rules into ML, such as, integrating pressure thresholds within tree-based models, have improved precision, yet multi parameter hybrid modeling in hydraulic fracturing remains underexplored.

3. Methodology

The proposed workflow for hydraulic fracturing anomaly detection and predictive modeling follows a structured, multi-phase pipeline that integrates domain knowledge-based rules with machine learning techniques to ensure both accuracy and interpretability.

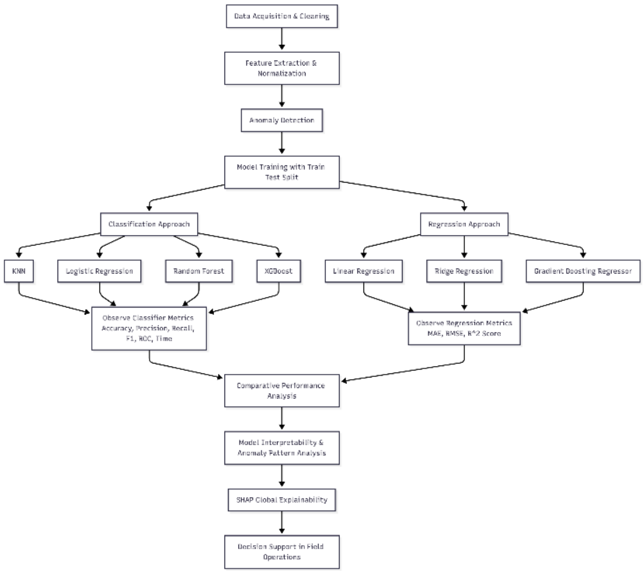

Fig.1 Methodological workflow

The overall framework is illustrated in Figure 1, showing the development of anomaly detection and predictive modeling, with their eventual integration for final interpretation. The methodology consists of five main phases, described as follows:

3.1 Phase 0: Data Set, Collection and Preprocessing

The dataset includes detailed operational parameters related to hydraulic fracturing stages. Initial preprocessing involved removing duplicates entries and handling missing data through imputation: mean or median methods were selected based on feature distribution skewness. Feature names were standardized to maintain consistency across data records. In order to mitigating the effect of erroneous measurements, outliers clearly arising from data entry errors (e.g. negative fluid volume) were removed prior to anomaly detection. Continuous variables were normalized using Min-Max scaling to ensure compatibility with distance-based algorithms such as k-Nearest Neighbors (KNN).

3.1.1 Data Set

The dataset utilized in this study comprises operational records from hydraulic fracturing stages, which gone through preprocessing and curation for different modeling tasks. The essential characteristics of the dataset summarized below:

- Total Number of samples: 173920
- Total number of features: 11

3.1.1.1 Anomaly Detection Model:

- Input features: 4 operational parameters, such as Percent HF Job (percentage of job completed), Mass Ingredient (total additives used in pounds), Total Base Water Volume (base fluid volume in gallons), True vertical depth (TVD in feet).
- Output features: Binary anomaly labels derived from domain specific operational thresholds

3.1.1.2 Predictive Modeling:

- Input features: 6 parameters (3 operational and 3 derived), such as CAS Number (chemical abstracts service registry number), TVD (true vertical depth in ft), purpose (reason behind using chemical), water per foot, mass per foot, mass per gallon
- Output features: Total Base Water Volume, Mass Ingredient, Percent HF Job

3.2 Phase 1: Rule Based Anomaly Detection

An initial anomaly detection layer was established based on domain knowledge driven operational thresholds collected from industrial practices and established literatures. These thresholds delineate typical ranges for key fracturing parameters and serve as criteria to flag anomalous data points.

Notably, percentage of HF job done (PercentHFJob) values outside the range of 0.01% to 2.0% were flagged as anomalies, reflecting typical fracturing job proportions reported in several industrial report. For mass of ingredient used (MassIngredient), accepted operational values lies between 4.53592 kg and 4535920 kg (10 -10000000lbs), with deviation considered as anomalous. Similarly, required total base water (TotalBaseWaterVolume) values were expected within 1893 m^3 to 37854 m^3 (500000 to 10000000 gallons) with deviation flagged as anomalous as according to industry practice. Lastly, true vertical depth (tvd) values were scrutinized based on ranges commonly reported in multiple literature reviews, typically between 304 m to 6096 m (1000ft to 20000ft).

After flagging based on individual parameters, condition is applied to generate the final flag, leading to create ground truth labels for subsequent machine learning model training.

3.3 Phase 2 Anomaly Detection Model

The anomaly detection model was trained using four supervised classification algorithms (3 models are traditional & 1 model is boosting algorithm): Random Forest, k-Nearest Neighbors (KNN), Logistic Regression and Extreme Gradient Boosting (XGBoost). These models aimed to classify data points as normal or anomalous based on labeled data generated during preprocessing by applying domain based operational thresholds. The training uses 80% of the dataset for training and remaining 20% dataset for model validation or testing. Key operational features such as Percent HF Job, Mass Ingredient, Total Base Water Volume and TVD were included as inputs on the models, Model performance was assessed on the test set using standard classification metric including accuracy (overall proportion of correctly classified cases), precision (fraction of predicted anomalies that are correct), recall (fraction of actual anomalies correctly detected) and F1-score (the harmonic mean of precision and recall), ROC-AUC (discriminative ability across classification thresholds). These metrics helped to understand the ability of each model to detect anomalies along with controlling false positives and false negatives.

3.4 Phase 3 Continuous Predictive Modeling

In this study of continuous predictive modeling, the motive was to estimate three critical operational parameters simultaneously. Gradient Boosting Regressor algorithm, is used for multi-output, was trained on the clean dataset using the same 80:20 train-test split. Three field parameters and three derived parameters were used as input for the training to increase the model performance. In order to establish base line performance, Linear regression and Ridge Regression model were also trained and examined under identical input. For more efficient and stable model training categorical variables were transformed using One Hot Encoder to create binary indicator features and continuous variables were standardized using Standard Scaler to ensure zero mean along with unit variance. Model predictions were examined through Mean Absolute Error (MAE), which measures the average magnitude of prediction errors; Root Mean Squared Error (RMSE), which handles the large errors more severely; and the coefficient of Determination (R^2), which quantifies the proportion of variance in the target variables explained by the model. Combination of evaluation matrix provide a rigorous quantitative assessment of the predictive models used in this study.

4. Result & Discussion

Interpretation of classification and regression model outcomes shows distinct performance patterns and operational insights relevant to fracturing job evaluation. Importantly, model comparisons expose trade-offs between predictive accuracy and computational efficiency, while interpretability tool highlight key drivers of anomaly detection aligned with practical operational parameters.

4.1 Anomaly Detection Performance

Among the model tested, XGBoost and Random Forest constantly produced better predicted results than other classification model. These two models gained almost ideal scores in precision, recall, F1 Score and ROC AUC criteria. This high degree of precision in all category depicts that dataset's abnormal scenarios were robustly identified.

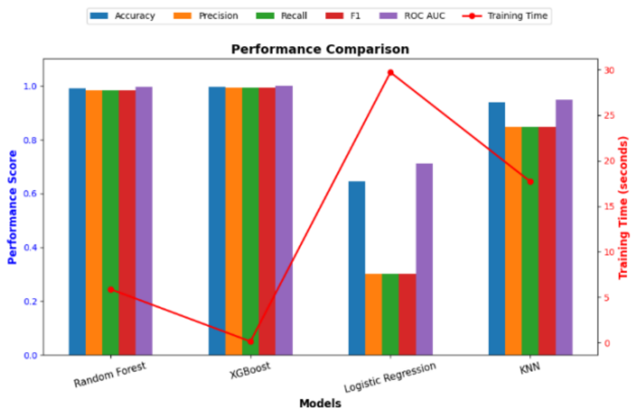

Fig. 2 Performance comparison of 4 anomaly detection model

However, XGBoost needs less time compare to Random Forest model. Despite having reasonable accuracy rate, KNN showed lower recall which defines that it occasionally fails to detect anomalies. This could lead to missed early warnings in operational settings. Logistic Regression did poorly on recall and F1 Score, which shows that this model can't handle the variance and non-linear interactions of different fracturing parameters.

4.2 Accuracy & Error Features of Regression Model

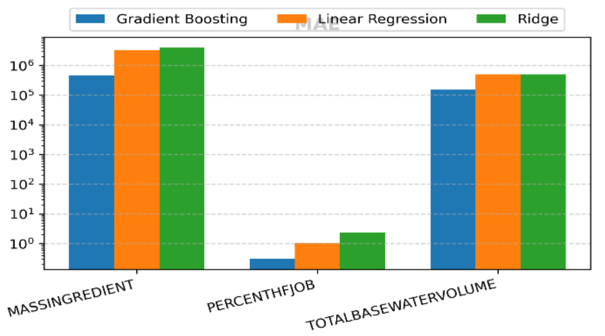

Fig.3 MAE results of 3 models

Gradient Boosting model outperformed linear baseline model in both error metrics and explained variance in regression tasks predicting continuous operational variables. Its advantages were most exhibited for parameters showing greater variability and non-linearity, such as Percentage of HF job, total base water volume required, where R^2 values reached nearly 0.97. This indicates the gradient boosting (GB) effectively captured complex relationships and interaction influencing these outputs.

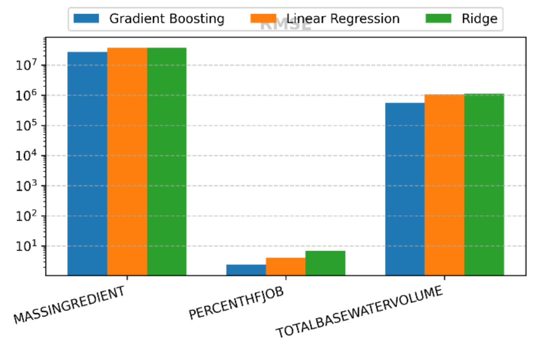

Fig.4 RMSE results of 3 models

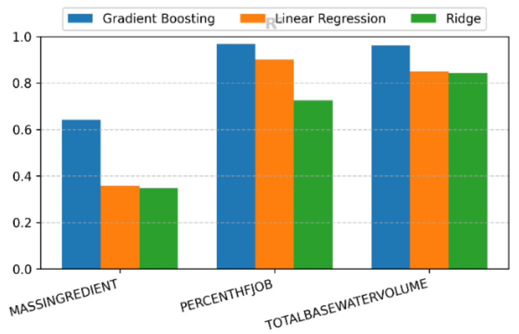

Fig.5 R^2 results of 3 models

On the other hand, linear and ridge regression models is trained for base model evaluation. Both models showed reasonable performance on stable variables but clearly struggled with high variance objects, depicting increased MAE and RMSE.

The divergence between MAE and RMSE across different models suggests the presence of occasional prediction errors (outliners) more effectively mitigated by Gradient Boosting. However, despite its superior accuracy, Gradient Boosting required noticeably higher computational demand than linear model, which make this model less suitable for such cases where predictions need to be refreshed frequently within a very short period of time.

4.3 Model Interpretability and Anomaly Pattern Analysis

This subsection presents both global and anomaly specific interpretability analyses to connect model predictions with real world operational patterns.

4.3.1 Global Anomaly Patterns and Model Drivers

To assess the alignment between the operational patterns and the model's predictive behavior, anomaly co-occurrence frequencies were compared with feature importance derived from SHAP analysis.

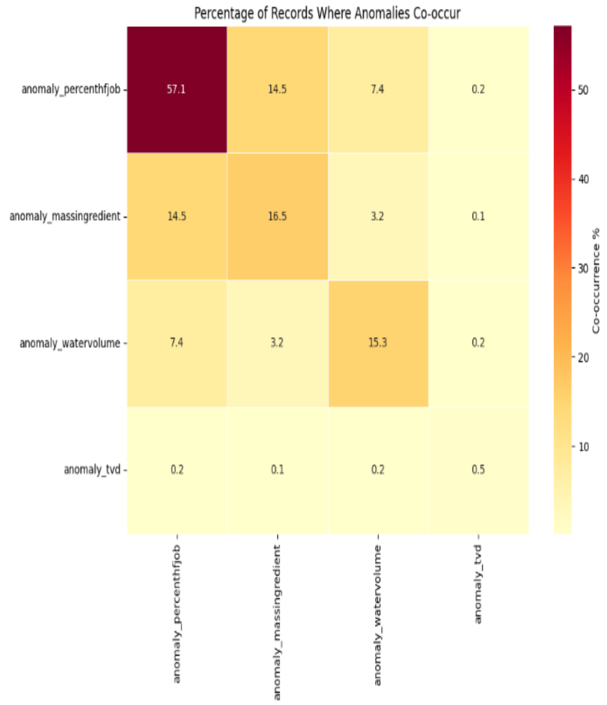

Fig. 6 Anomaly co-occurrence heatmap

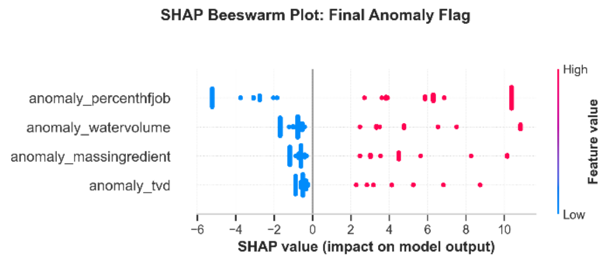

Fig.7 Beeswarm Plot of anomaly detection.

The anomaly occurrence heatmap (Figure 6) show that the most prevalent anomaly is anomaly in percent HF job (anomaly_percenthfjob) with 57.1% along with notable co-occurrence rates of anomaly in mass of ingredient (anomaly_massingredient) and water volume anomaly (anomaly_watervolume) is 14.5% and 7.5% respectively.

However, variance in vertical depth (anomaly_tvd) appears in less than 0.5% of records that shows minimal co-occurrence compare to other anomalies. The operational patterns are mirrored in the figure 7, where anomaly_percenthfjob exhibits the largest spread of SHAP values and anomaly_watervolume shows moderate co-occurrence. Conversely, anomaly_tvd shows SHAP values tightly clustered around zero. This alignment between anomaly patterns and model-based feature attribute confirms that both analyses identify the same dominant operational drivers.

5. Conclusion

This study presented a hybrid anomaly detection and multi-parameter framework for hydraulic fracturing operations by integrating domain-based operational thresholds with supervised machine learning models. The anomaly detection component achieved high classification performance, with XGBoost obtaining 99.6% accuracy, 0.994 precision, 0.983 recall, and a Random Forest also performed competitively, while logistic regression showed limited capability in capturing nonlinear operational patterns.

In predictive modeling, gradient boosting outperformed baseline linear and ridge regression models across all target variables. For job percentage and total base water volume model achieved R^2 values of 0.9667 and 0.9605, respectively, with substantial reductions in MAE and RMSE compared to baseline model. For Mass Ingredient parameter, an R^2 values of 0.6425 was obtained, representing an improvement over the baseline model (0.3566) but also indicating sensitivity to high variance and unmeasured operational factors. This finding suggests that including additional features such as slurry density fluctuations, pump efficiency metrics, and stage level blending parameters could improve predictive accuracy.

Hydraulic fracturing remains a cornerstone of unconventional resource development, contributing significantly to global oil and gas supply. Accurate anomaly detection and parameter forecasting directly enhance operational safety, equipment reliability, and simulation effectiveness.

Future work will focus on expanding the feature set with geological and high frequency sensor data to improve robustness, particularly for high variance parameters like proppant mass. Deploying the framework in live field data systems could enable closed loop control, where predictions guide pumping schedules, proppant loading, and fluid design resulting in safer, more efficient and cost-effective operations.

6. References

[1] "Annual Energy Outlook 2023 with Projections to 2050," 2023.

[2] J. R. Bartis, T. LaTourrette, L. Dixon, D. W. Peterson, and G. Cecchine, *Oil Shale Development in the United States: Prospects and Policy Issues*. RAND Corporation, 2005

[3] J. R. Curtis, R. M. Haggerty, and J. E. Spivey, "Hydraulic Fracturing in Shale Reservoirs: A Review of the State of the Art," *Journal of Petroleum Technology*, vol. 69, no. 9, pp. 34–45, 2017.

[4] A. L. English, M. J. McDonald, and S. J. Parker, "Shale Gas Development in North America and Implications for Global Energy Markets," *Energy Policy*, vol. 107, pp. 317–329, 2017.

[5] C. Li, Y. Zhang, and D. Wu, "Overview of Hydraulic Fracturing Development and Operations in China's Sichuan Basin," *Int J Coal Geol*, vol. 210, p. 103305, 2019

[6] M. T. Smith and J. K. Johnson, "Operational Parameters and Real-Time Monitoring in Multi-Stage Hydraulic Fracturing Jobs," in *SPE Hydraulic Fracturing Technology Conference*, 2020.

[7] L. Zhao, H. Liu, and W. Wang, "Key Parameters for Hydraulic Fracturing Monitoring and Control," *J Pet Sci Eng*, vol. 189, p. 106900, 2020

[8] F. Garcia and E. G. Willhite, "Slurry Rate and Proppant Concentration Effects on Fracture Conductivity," *SPE Production & Operations*, vol. 25, no. 1, pp. 35–44, 2010.

[9] A. Singh, B. P. Patel, and M. K. Jain, "Dynamic Operational Thresholds for Hydraulic Fracturing Based on Reservoir Heterogeneity," *J Nat Gas Sci Eng*, vol. 45, pp. 239–250, 2017

[10] Nguyen and P. Tran, "Bottomhole Pressure Prediction Using Machine Learning Algorithms in Unconventional Wells," *J Pet Sci Eng*, vol. 184, p. 106533, 2019

[11] Production Forecasting of Shale Gas Wells Using Gradient Boosting Machines," *J Nat Gas Sci Eng*, vol. 63, p. 103788, 2019

[12] N. Sharma and P. Verma, "Frac-Hit Prediction Using Machine Learning Techniques," in *SPE Asia Pacific Oil and Gas Conference*, 2018.

[13] M. E. Williams and J. C. Roberts, "Optimization of Stage Designs for Hydraulic Fracturing Using Predictive Models," in *SPE Hydraulic Fracturing Technology Conference*, 2019.

[14] M. Mahmoud, A. Al-Durgham, and S. K. Al-Jabri, "Random Forest Based Anomaly Detection in Hydraulic Fracturing Operations," *J Pet Explor Prod Technol*, vol. 11, pp. 2975–2987, 2021.

[15] J. Wang, Y. Chen, and L. Liu, "Real-Time Bottomhole Pressure Prediction Using XGBoost for Hydraulic Fracturing Operations," *J Pet Sci Eng*, vol. 201, p. 108363, 2021.

[16] S. A. Kim and J. H. Kim, "Physics-Informed Machine Learning for Hydraulic Fracture Modeling: A Review," *J Pet Sci Eng*, vol. 204, p. 108714, 2021.

[17] D. J. Smith and H. J. Lee, "Multi-Parameter Modeling for Hydraulic Fracturing Operations Using Machine Learning," *International Journal of Rock Mechanics and Mining Sciences*, vol. 129, p. 104327, 2020

[18] K. M. Patel, S. J. Mehta, and R. R. Shah, "Correlation Analysis of Hydraulic Fracturing Operational Parameters in Eagle Ford Shale," *J Nat Gas Sci Eng*, vol. 55, pp. 424–436, 2018.

[19] A. Alimahmudi, S. M. Hosseini, and H. Moradi, "Neural Network Modeling for Post-Job Fracture Half-Length Prediction," in *SPE Hydraulic Fracturing Technology Conference*, 2019.

[20] A. Raghavan, N. Verma, and S. Chakraborty, "Hybrid Domain Knowledge and Machine Learning Based Anomaly Detection in Hydraulic Fracturing," *J Pet Sci Eng*, vol. 196, p. 107703, 2021.

[21] Y. Xu, T. Ma, and H. Zhou, "Integration of Operational Thresholds in Machine Learning Models for Hydraulic

Fracturing Stage Fluid Volume Prediction," *J Nat Gas Sci Eng*, vol. 96, p. 104291, 2021.

[22] M. K. Hubbert and D. G. Willis, "Mechanics of Hydraulic Fracturing," *Petroleum Transactions, AIME*, vol. 210, pp. 153–168, 1957.

[23] T. K. Perkins and L. R. Kern, "Widths of Hydraulic Fractures," Journal *of Petroleum Technology*, vol. 13, no. 9, pp. 937–949, 1961

[24] A. Akbari, A. Karami, Y. Kazemzadeh, and A. Ranjbar, "Evaluation of hydraulic fracturing using machine learning," *Sci Rep*, vol. 15, p. 26926, 2025.

[25] H. Yang *et al.*, "Optimization of tight gas reservoir fracturing parameters via gradient boosting regression modeling," *Heliyon*, vol. 10, no. 5, p. e27015, 2024.

[26] A. Karami, A. Akbari, Y. Kazemzadeh, and H. Nikravesh, "Enhancing hydraulic fracturing efficiency through machine learning," *J Pet Explor Prod Technol*, vol. 15, no. 1, Jan. 2025.

[27] Y. Ma and M. Ye, "Application of Machine Learning in Hydraulic Fracturing: A Review," Mar. 25, 2025, *American Chemical Society*.

[28] A. D. Morozov *et al.*, "Data-driven model for hydraulic fracturing design optimization: focus on building digital database and production forecast," *J Pet Sci Eng*, vol. 194, p. 107504, Nov. 2020.