PROCEEDINGS, 51st Workshop on Geothermal Reservoir Engineering
Stanford University, Stanford, California, February 9-11, 2026
SGP-TR-230

# Machine learning Based Prediction of Porosity from Well Log Data

Emmanuel Gyimah[1], Shari Kelley [1], Adewale Amosu [2], Kwamena Opoku Duartey [3], Emmanuel Agyei [3]

1 New Mexico Bureau of Geology and Minerals Resources

2 Petroleum Research and Recovery Center, New Mexico Institute of Mining and Technology

3 Petroleum Engineering Department, New Mexico Institute of Mining and Technology

**Keywords:** Porosity Prediction, Feature Importance, Root Mean Square Error, Machine Learning Algorithms

**ABSTRACT**

Accurately quantifying porosity in reservoirs is essential for optimizing geothermal resource exploration and subsurface resource evaluation. Conventional predictive methods, such as density porosity models, often have limited accuracy, which can impede effective reservoir characterization. To overcome these limitations, this study leverages advanced machine learning (ML) techniques including AdaBoost, XGBoost, Gradient Boosting, LightGBM (LGBM), and ExtraTrees to predict porosity using well log data. ML offers a powerful alternative by leveraging data-driven approaches to uncover complex, nonlinear relationships between well log responses and porosity. Feature selection techniques, including correlation analysis from heat maps and feature importance are employed to identify well log data relationships. A comprehensive comparative analysis of well log data from two wells in North Dakota demonstrates that the comparative ML-based approach could be utilized to predict porosity effectively. The robustness of the data-driven models is validated through 5-fold cross-validation to confirm its reliability. Additionally, blind testing on a second well further verifies the model's generalization capability and practical applicability. The results highlight the strong potential of machine learning in enhancing porosity estimation for geothermal reservoirs. By providing more precise and efficient predictions, this ML-driven framework can support better decision-making in geothermal exploration and development, subsurface characterization, reservoir modeling uncertainty, ultimately contributing to more sustainable and cost-effective energy extraction.

## 1. INTRODUCTION

### 1.1 Porosity Prediction

Petrophysical properties such as porosity, permeability and water saturation, remain fundamental to characterizing a reservoir. These properties are used in geothermal energy assessments such as heat in place (total geothermal resources), aquifer geothermal resources and producible geothermal resources. Well log data such as density, neutron, sonic and resistivity logs can be used to indirectly measure petrophysical properties such as porosity, permeability and water saturation. The traditional way to validate well log predictions for petrophysical properties is to have experimental core data and perform calibration. The presence of shale, which contains bound water and clay minerals, can cause anomalous readings. Wyllie (1942, 1956, 1958) established empirical equations to determine other petrophysical properties including porosity, although most of his work was based on sandstone, limestone and unconsolidated sands. Although core sampling provides accurate data, it is unable to represent the entire reservoir (Neal, et al. 2023). Also, traditional porosity predictions are limited by geological heterogeneity and incomplete logging data hence the need for a data-driven model which captures complex data patterns and non-linear relationships.

### 1.2 Machine Learning Algorithms

Machine Learning (ML) techniques serve as a cost effective and highly accurate solution for estimating petrophysical properties. ML methods have demonstrated significant potential in addressing limitations with porosity predictions (Anderson et al, 2022). In a typical workflow, with limited and expensive core data available but an abundance of well log data, ML models are trained to predict core-derived measurements (e.g., porosity, permeability) from a suite of well logs (e.g., gamma ray, density, neutron, resistivity, sonic) in the cored intervals. Once trained and validated, these algorithms can be deployed to predict properties continuously across un-cored intervals or in entire un-cored wells, optimizing costs and maximizing the value of limited core data. The fundamental advantage of ML lies in its ability to learn complex, non-linear relationships directly from data without requiring explicit physical models. Machine Learning represents a paradigm shift in petrophysical analysis by leveraging abundant log data to augment scarce core data.

## 2. METHODOLOGY

Core and well log data from two wells in North Dakota were used in this investigation. This study utilizes the Adaboost, XGBoost and Gradient Boosting algorithms (best performing algorithms) for the first well prediction of porosity and the LGBM, Extra Trees and XGBoost algorithms ((best performing algorithms)) for the second well prediction of porosity. These algorithms were selected based on cross-validation and model performance. The model performance is evaluated using Final Test R-squared ($R^2$), Final Test Root Mean Squared Error (RSME), Cross Validation R-squared ($R^2$) and Cross Validation Root Mean Squared Error (RSME). Feature importance is further used to determine how much each type of well log data (training feature) contributes to porosity prediction. Lastly, a blind test is performed to assess the general performance of the algorithm.

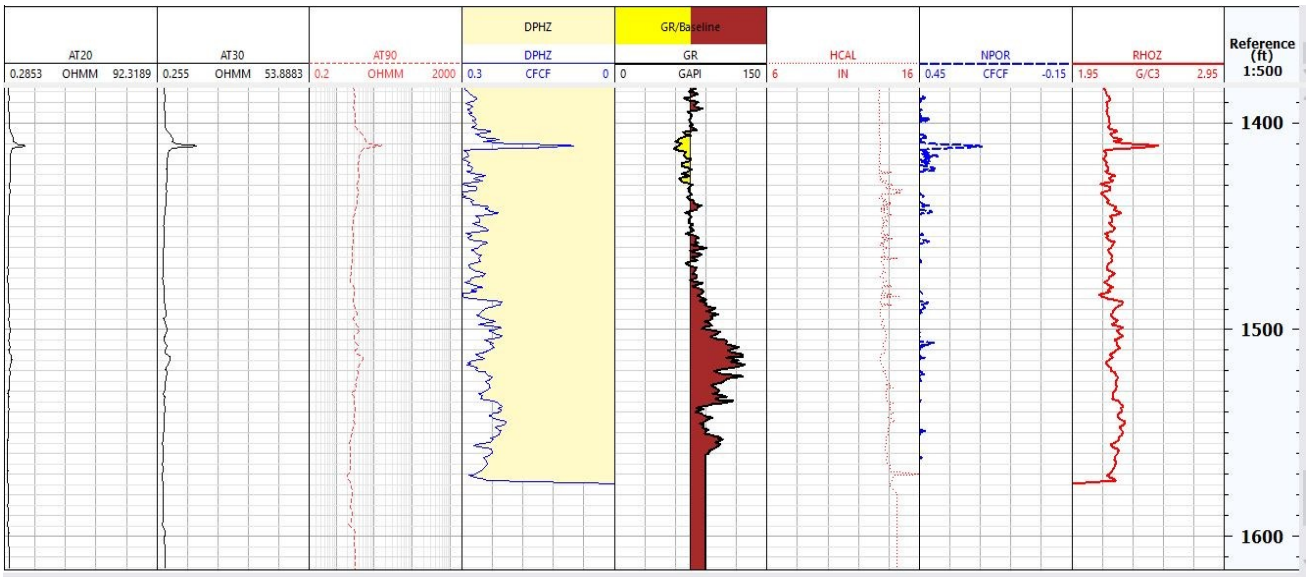


**Figure 1: Well log data utilized for study**

### 2.1 Data Pre-processing

This step is necessary to ensure high-quality data for analysis. This step involves the removal of outliers, redundant entries, and inconsistent data. The data distribution can be summarized using histogram chart (Figs. 2 and 3) which are effective for detecting outliers, skewness, and measures of central tendency. Furthermore, heat maps as represented in Figures 4 and 5, provide a visual representation of variable relationships via color gradients, aiding significantly in the analysis of feature correlations for both data analysis and machine learning applications.

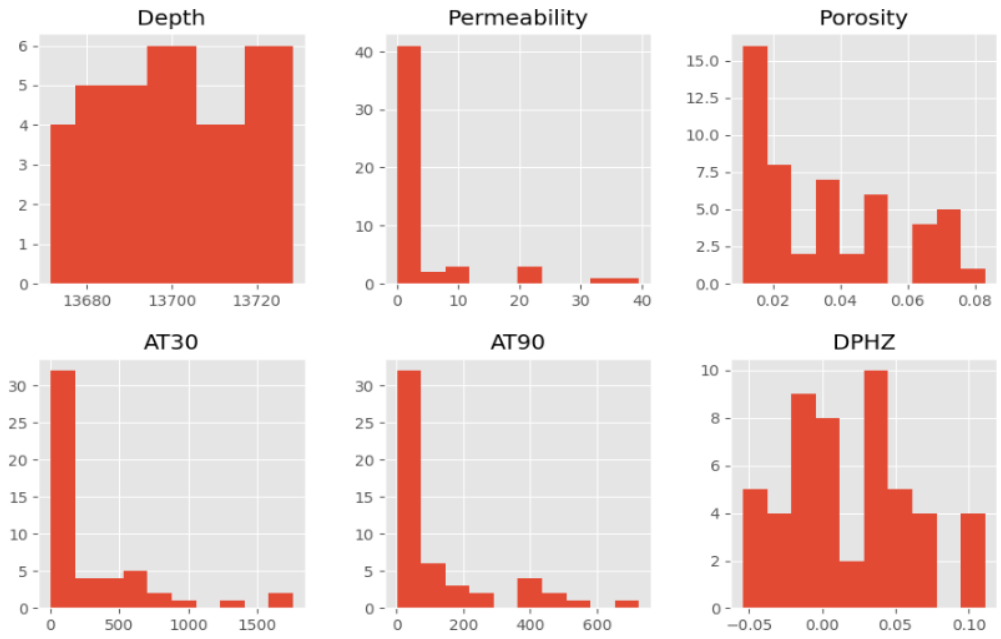


**Figure 2: Histogram of six selected features (W18631)**

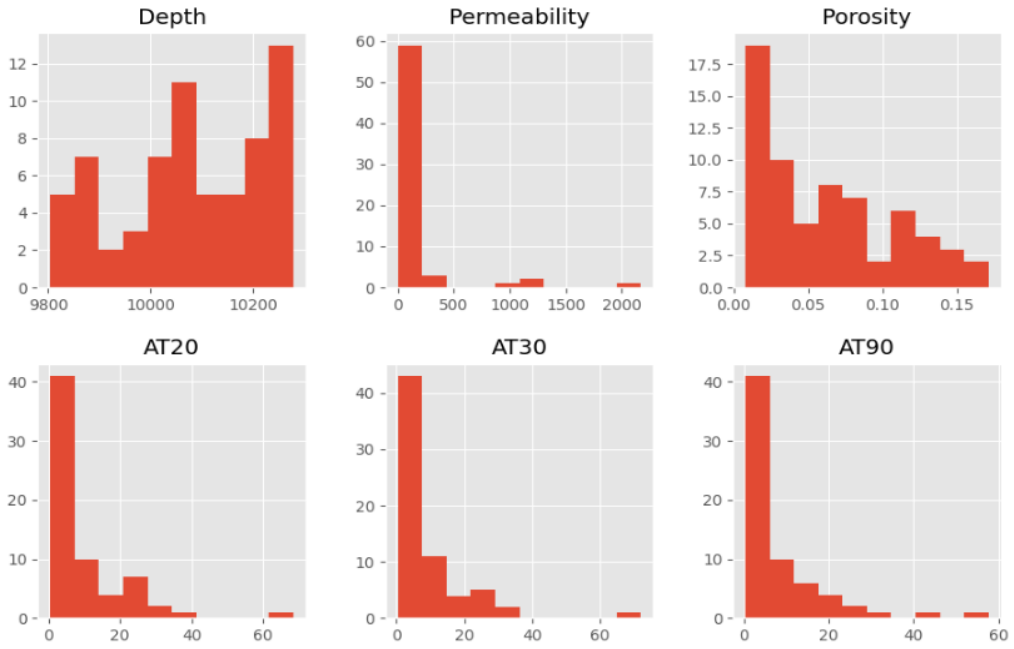


**Figure 3: Histogram of six selected features (W37380)**

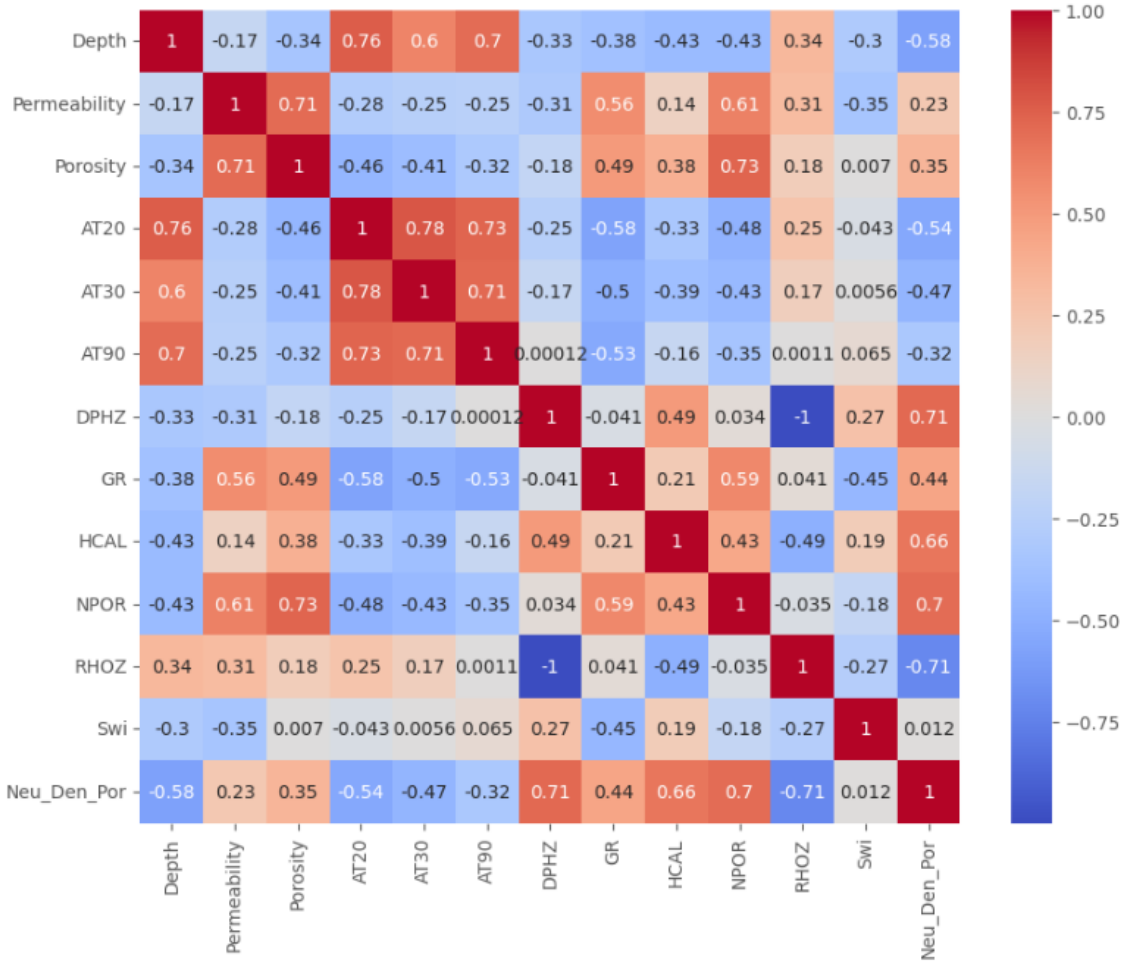


**Figure 4: Heat map of well data from W18631**

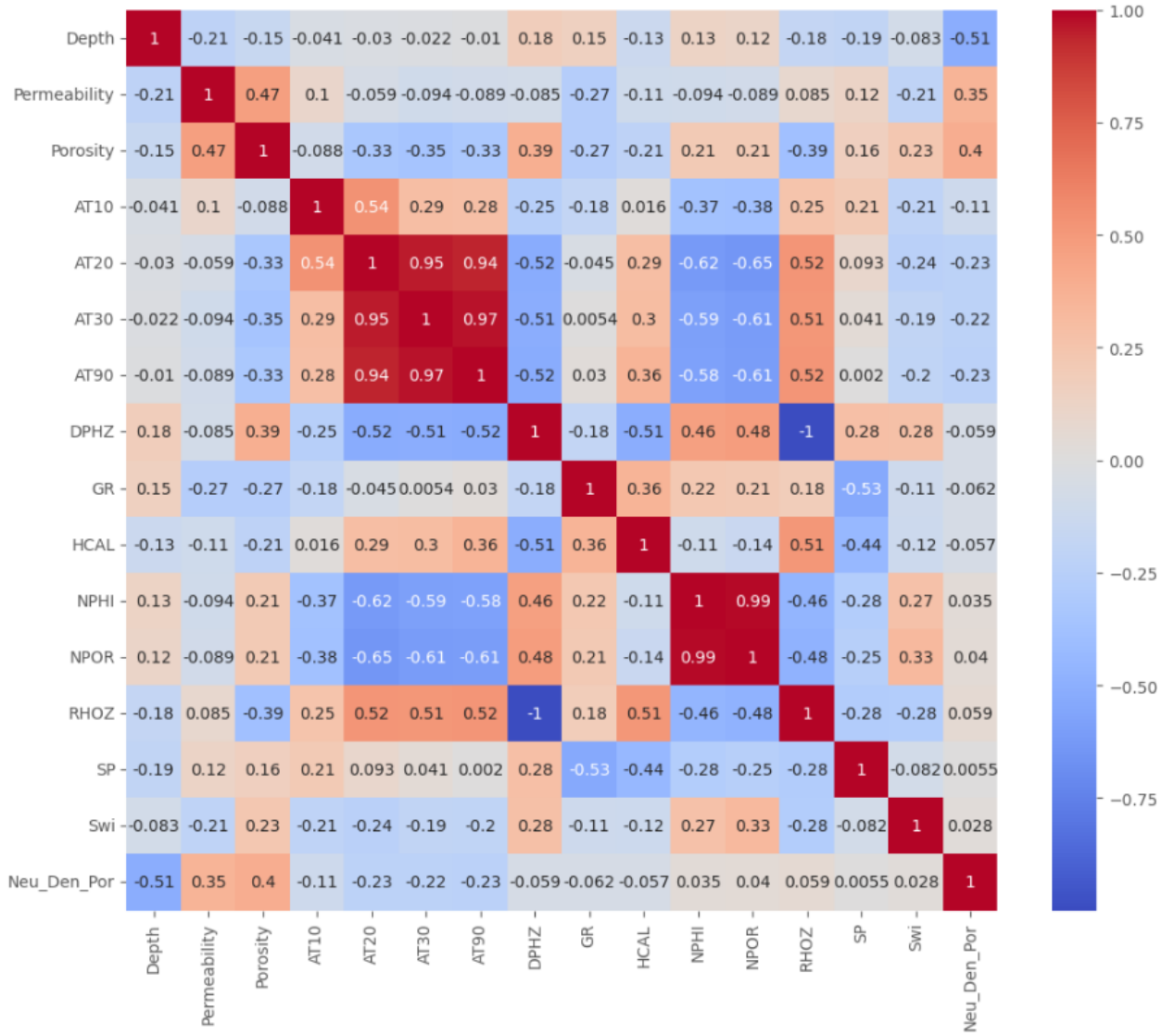


**Figure 5: Heat map of well data from W37380**

## 2.2 Supervised Machine Learning

### 2.2.1 Adaboost Regressor

Also known as adaptive boosting, it is a pioneering boosting algorithm that creates a strong classifier by combining multiple weak learners (typically shallow decision trees). It works by iteratively reweighting the training data, giving more weight to misclassified samples so subsequent learners focus on harder cases. It uses weighted combinations rather than gradient descent and particularly sensitive to noisy data and outliers (Freund, Y., and Schapire, R. E. 1996, 1997).

### 2.2.2 Gradient Boosting Regressor (GBR)

GBR is a powerful supervised ensemble learning method that builds sequential decision trees to correct errors from previous trees, optimizing prediction accuracy (Friedman, 2001). GBR is often one of the best-performing algorithms on structured data, but training sequentially can be slow and resource-intensive. GBR uses boosting, iteratively improving predictions by focusing on residual errors. It starts with a weak initial model (e.g., a single decision tree).

2.2.3 Extreme Gradient Boosting (XGBoost)

XGBoost is an optimized distributed gradient boosting algorithm that provides a highly efficient and scalable implementation of the gradient boosting framework. It enhances traditional GBR through several key innovations: a regularized model objective to prevent overfitting, second-order gradient optimization for more accurate tree learning, and sophisticated techniques for handling sparse data and finding optimal splits (Chen, T., and Guestrin, C., 2016).

2.2.4 Light Gradient Boosting Machine (LGBM)

LGBM is a high-performance gradient boosting framework that prioritizes training speed and memory efficiency while maintaining high accuracy. It achieves this through two innovative techniques: Gradient-based One-Side Sampling and Exclusive Feature Bundling which allows it to handle large-scale data much faster than traditional gradient boosting methods. LGBM is particularly well-suited for large datasets with many features (Ke et al., 2017).

2.2.5 Extra Trees Regressor

The Extra Trees Regressor (Extremely Randomized Trees Regressor) is a powerful ensemble machine learning algorithm that belongs to the family of bagging methods. For each candidate feature in the random subset, the algorithm chooses a split point completely at random. It then selects the best of these randomly generated splits as the splitting rule for the node. The algorithm is similar to the Random Forest Regressor but with additional randomization and works by building multiple decision trees and averaging their predictions to improve accuracy and reduce overfitting (Geurts et al., 2006). Since Extra Trees split randomly, it reduces computation time compared to Random Forest.

## 3. Results And Discussion

### 3.1 Feature Importance

Feature importance provides details of how much each training feature contributes to porosity prediction. It generates a numerical score for each well log data importance to the output. Permeability which is determined from core measurements, was the main dominant feature, and this was consistent across all algorithms (Figs. 6 - 9). The phenomenon reveals a significant geological relationship corresponding by studies of Andersen (2022) and Mabiala (2025). For the first well, W18631, permeability had over 70% feature importance and is the most predictive feature of porosity. Also, for the second well, W37380, permeability had almost 50% feature importance for all three algorithms. This shows a direct relationship between porosity and permeability, and since each algorithm learns differently and wights distinctly, the other well logs vary with its importance.

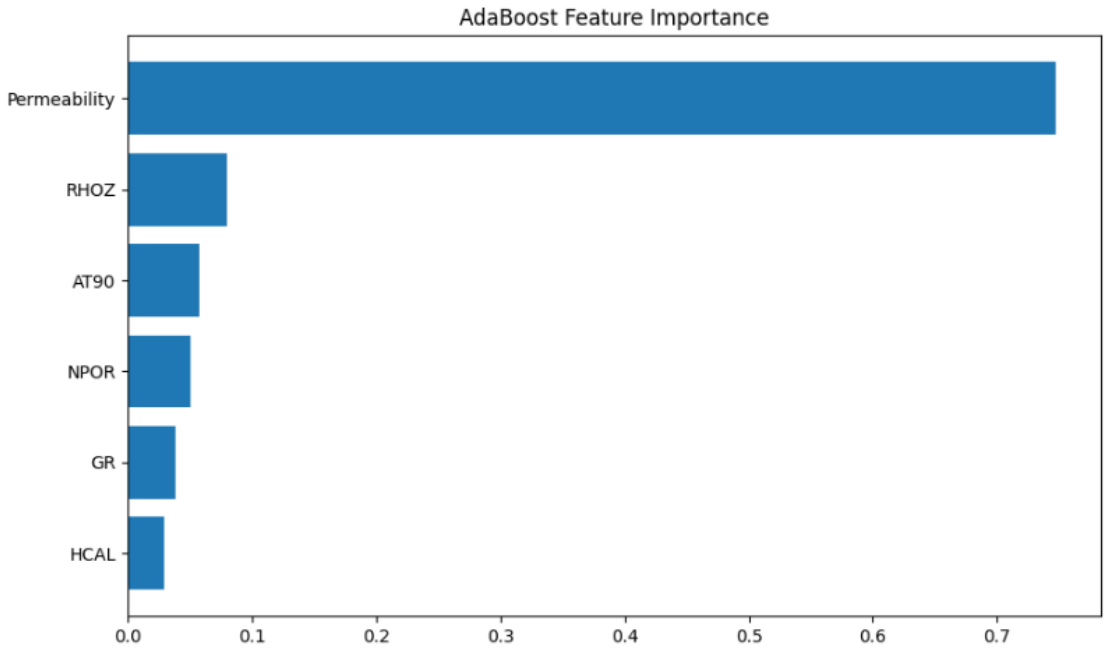


**Figure 6: Adaboost Feature Importance Plot for W18631**

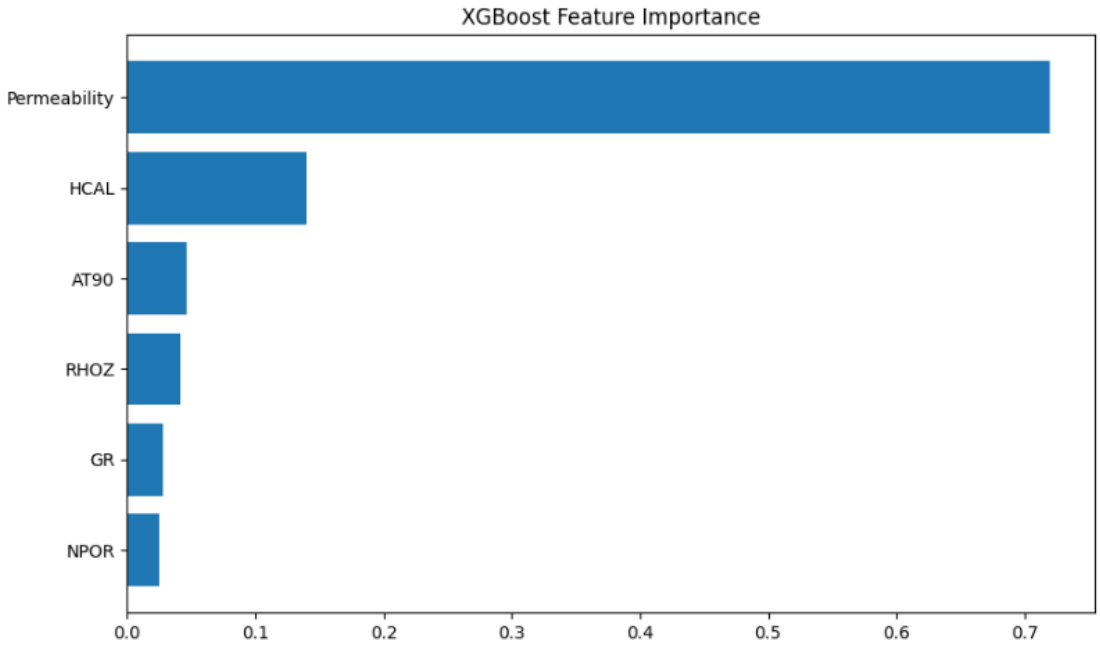


**Figure 7: XGboost Feature Importance Plot for W18631**

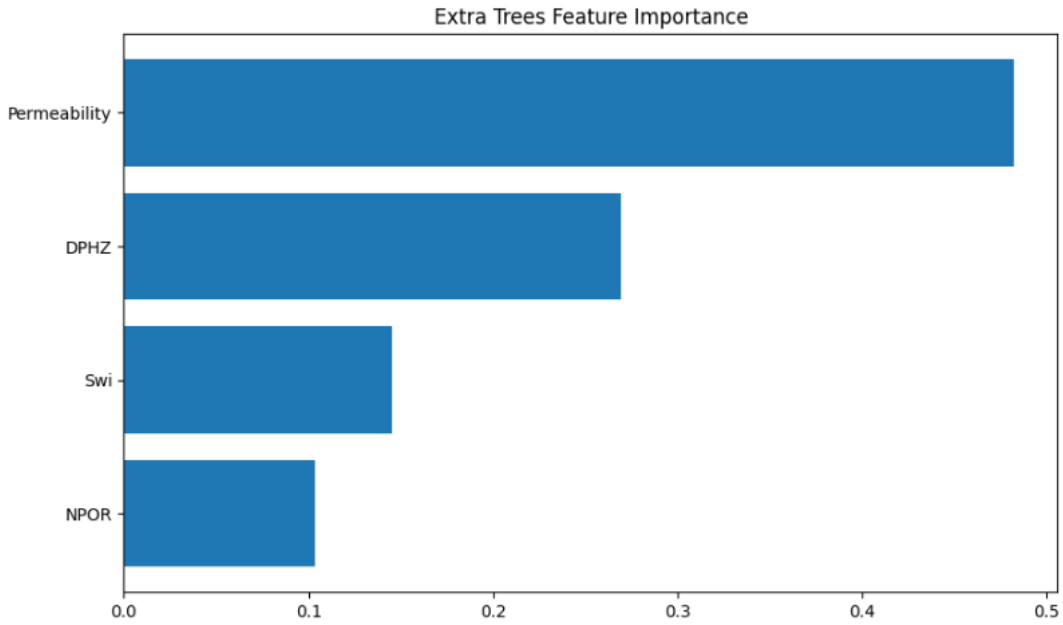


**Figure 8: Extra Trees Regressor Feature Importance Plot for W37380**

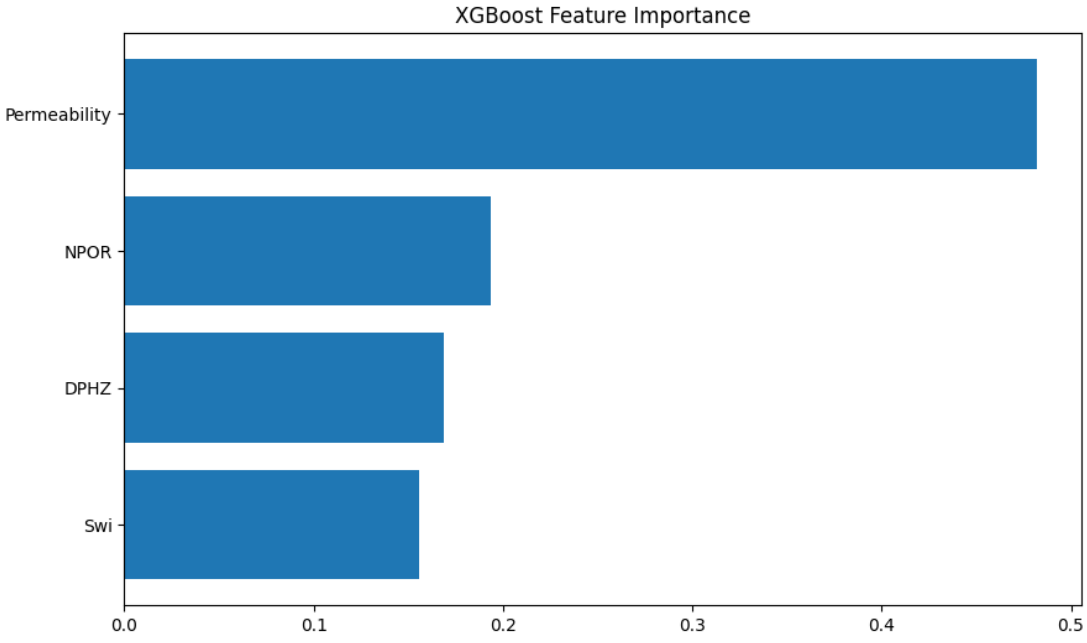


**Figure 9: LGBM Feature Importance Plot for W37380**

### 3.2 Supervised Machine Learning

The three selected algorithms (Adaboost, XGBoost and Gradient Boosting) were able to predict porosity with high accuracy and low error for the first well (W18631). Also, good results were achieved for the second well (W37380) after working with these three algorithms (LGBM, Extra Trees and XGBoost). All models were cross-validated to have the optimum model for future blind tests. Table 1, Table 2 and Figures 10 - 13 summarize all the model testing performance and cross-validation performance. Generally, testing performance and cross-validation performance for W18631 were higher than that of W37380, and this tells us that W18631 had cleaner data, and the well log data (training features) were strongly correlated to porosity. For W18631, XGBoost is the best choice if the priority is ultimate predictive accuracy but if model robustness and consistency are more important then, Gradient Boosting is a more stable alternative. For W37380, LGBM provides the best outright accuracy, whereas Extra Trees is the most generalizable model, as its performance is consistent across validation folds.

**Table 1: Model Performance for W18631**

| ML Algorithm | Final Test R-Squared | Final Test RSME | Cross-Validation R-Squared | Cross-Validation RSME |
|---|---|---|---|---|
| AdaBoost Regressor | 0.9711 | 0.0037 | 0.8554 | 0.0076 |
| XGBoost Regressor | 0.9779 | 0.0032 | 0.8278 | 0.008 |
| GradientBoosting Regressor | 0.9667 | 0.004 | 0.8698 | 0.0073 |

**Table 2: Model Performance for W37380**

| ML Algorithm | Final Test R-Squared | Final Test RSME | Cross-Validation R-Squared | Cross-Validation RSME |
|---|---|---|---|---|
| LGBM Regressor | 0.8792 | 0.0136 | 0.7276 | 0.0237 |
| ExtraTrees Regressor | 0.8380 | 0.0158 | 0.7710 | 0.0217 |
| XGBoost Regressor | 0.8629 | 0.0145 | 0.7894 | 0.0207 |

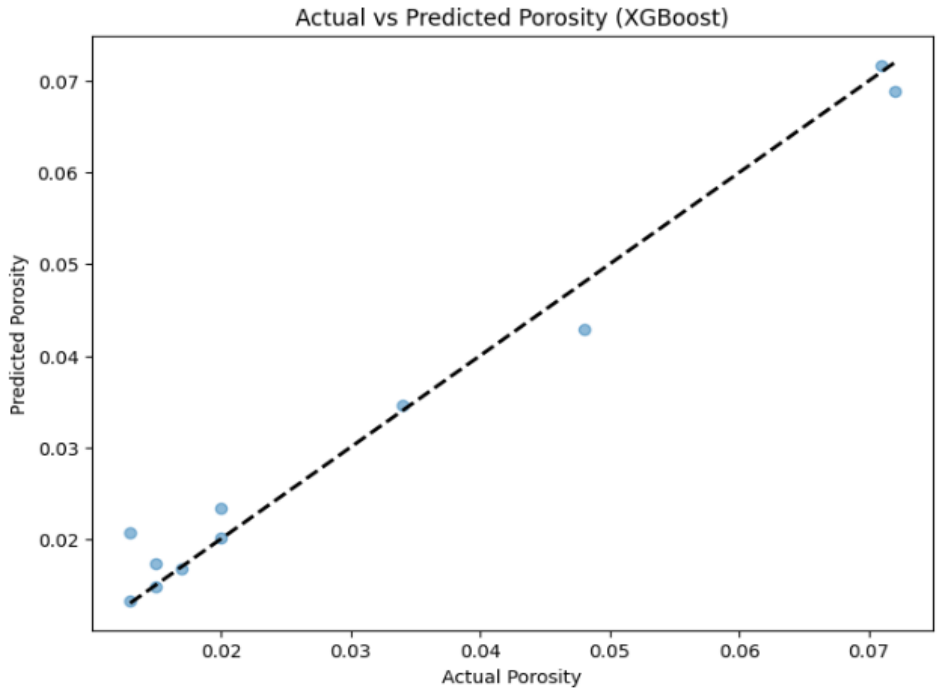


**Figure 10: Scatter Plot of Actual Porosity (core measurement) vs Predicted Porosity for W18631**

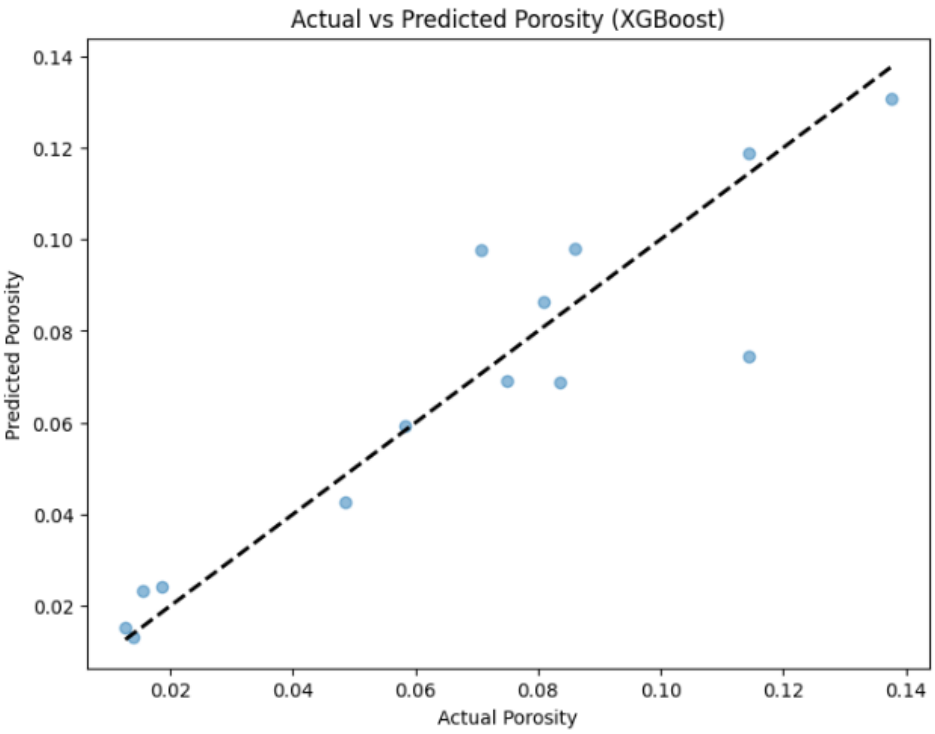


**Figure 11: Scatter Plot of Actual Porosity (core measurement) vs Predicted Porosity for W37380**

### 3.3 Cross-Validation

Cross-validation evaluates the algorithm's performance and generalizability of the algorithms. This provides an idea of how models perform using previously unseen data. The red dashed horizontal line represents the average RSME across the 5 cross validation folds. The 5-fold cross-validation utilized for this study preserves the percentage of samples for each class in every fold to prevent overfitting and ensure better use of data. Gradient Boosting and Extra Trees had the best cross validation results for W18631 and W37380 respectively. Hence, these algorithms provide a true estimate of model performance.

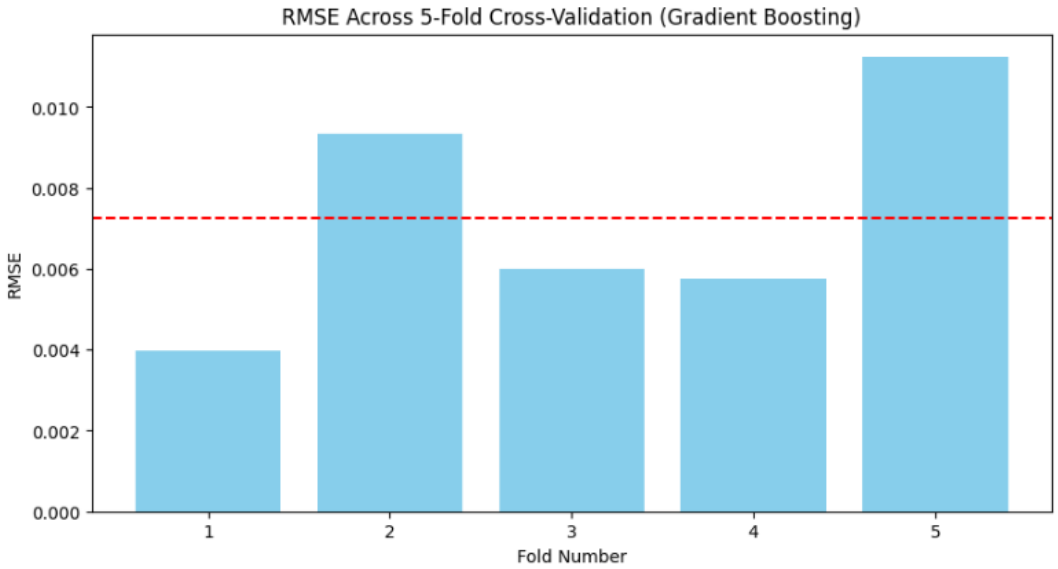


**Figure 12: Bar Chart of 5-Fold Cross-Validation for W18631**

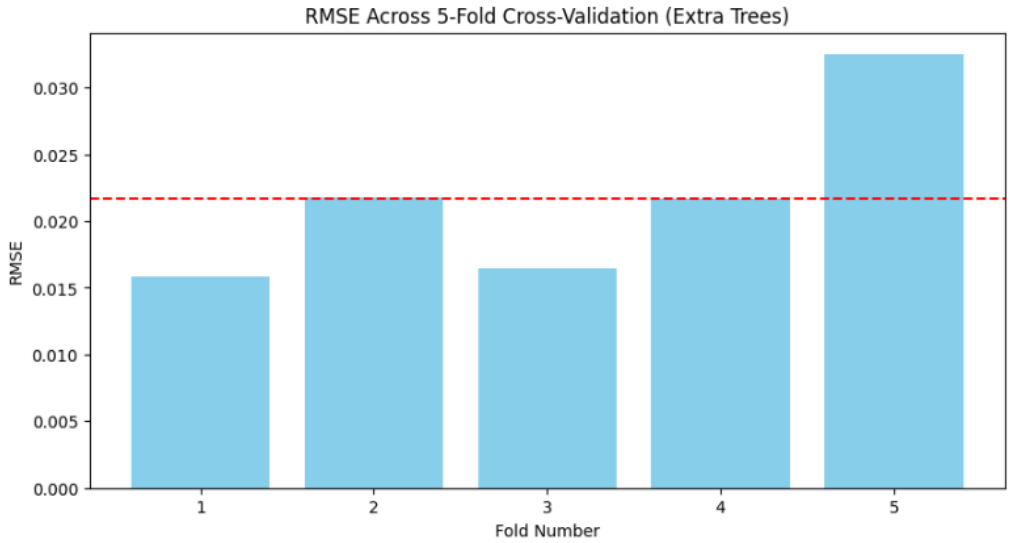


**Figure 13: Bar Chart of 5-Fold Cross-Validation for W37380**

### 3.4 Blind Testing

Blind Testing is also referred to as hold-out testing and it is the practice of evaluating a final model on a dataset never seen before during either training or development. This is the best way to get an unbiased estimate of the generalization error (Figs. 14 and 15). XGBoost Regressor is trained on W37380 to predict W18631. The XGBoost Regressor was selected because it was one of the best cross-validation algorithms.

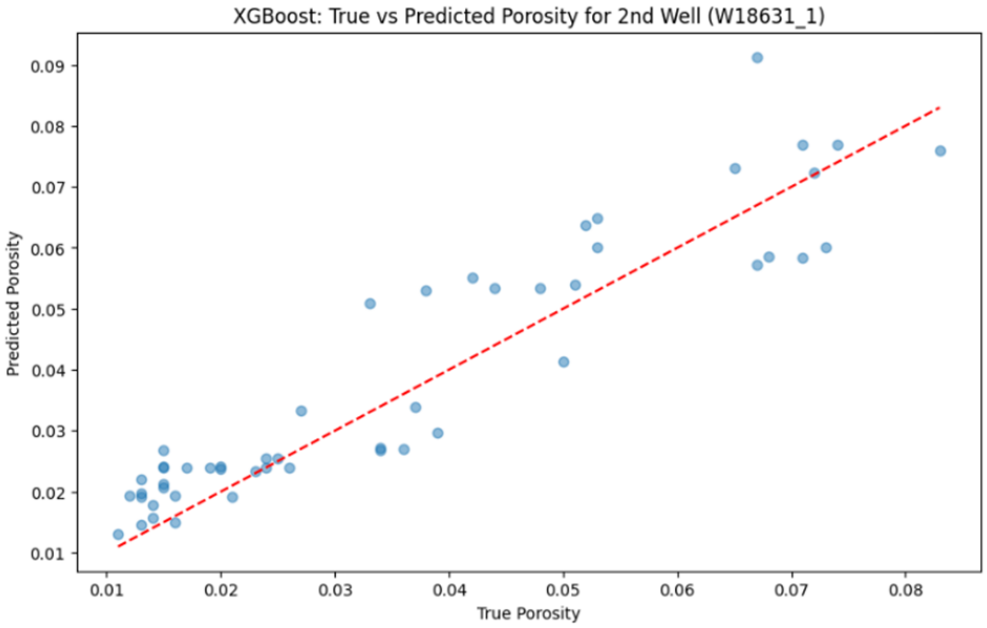


**Figure 14: Cross plot of True Porosity and Predicted Porosity for Well 18631**

```
XGBoost Model Evaluation Metrics on Second Well:
R² Score: 0.8484
Mean Absolute Error: 0.0068
Root Mean Squared Error: 0.0084
```

**Figure 15: Blind Test Metrics for both wells**

**4. Conclusion, Recommendation and Future Work**

- Our approach provides a framework for utilizing ML models for making porosity predictions.
- Each algorithm operates differently utilizing well logs to make predictions, but generally the bulk density log, resistivity log, neutron porosity log, density porosity log, gamma ray log and caliper log are very useful training features.
- Data-driven ML algorithms have the potential to achieve high precision and efficiency in porosity prediction providing valuable guidance for the geothermal industry. More data from multiple wells from different many fields and possibly different basins could be utilized to train the models for better prediction.

**REFERENCES**

Andersen, P.O, Skjeldal, M., and Carita A. (2022). Machine Learning Based Prediction of Porosity and Water Saturation from Varg Field Reservoir Well Logs. SPE EuropEC - Europe Energy Conference, doi: https://doi.org/10.2118/209659-MS

Beardsmore GR, & Cull JP. (2001). Crustal Heat Flow: A Guide to Measurement and Modelling. Cambridge University

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119-139.

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In Proceedings of the Thirteenth International Conference on Machine Learning (pp. 148-156).

Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. Annals of Statistics, pp.1189-1232.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1), 3-42.

Isaka, B. L. A., et al. (2019). Geothermal energy as a major contributor to renewable baseload power. Renewable Energy, 134, 1253–1265

Ke, G., Meng, Q., Finley, T., et al. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Advances in Neural Information Processing Systems (pp. 3146-3154).

Neal A., Ashton M., Bowman M., Hern C., Levell B. (2023) The role of core in twenty-first century reservoir characterization: an introduction https://doi.org/10.1144/sp527-2023-61

Wyllie, M.R.J., Gregory, A.R. and Gardner, L.W. (1956) Elastic Wave Velocities in Heterogeneous and Porous Media. Geophysics, 21, 41-70. http://dx.doi.org/10.1190/1.1438217

Wyllie, M. R. J., Gregory, A. R., & Gardner, G. H. F. (1958). An Experimental Investigation of Factors Affecting Elastic Wave Velocities in Porous Media. Geophysics, 23(3), 459-493. https://doi.org/10.1190/1.1438493

Wyllie, (1949). M.R.J. A Quantitative Analysis of the Electrochemical Component of the S.P. Curve. Journal of Petroleum Technology, https://doi.org/10.2118/949017-G

Mabiala A. P., Cai, Z., Kouassi A. K. F., Zhang H., Mwakipunda G. C., and Abdoulaye S. M. (2025) "Integrating Advanced Machine Learning Models for Accurate Prediction of Porosity and Permeability in Fractured and Vuggy Carbonate Reservoirs: Insights from the Tarim Basin, Northwestern, China." *SPE J.* 30: 3307–3333. doi: https://doi.org/10.2118/226198-PA

**NOMENCLATURE**

AT20 – low resistivity

AT30 – medium resistivity

AT90 – deep resistivity

DPHZ – density porosity log

GR – gamma ray

HCAL – caliper log

NPOR / NPHI – neutron porosity log

RHOZ – bulk density log

SP – spontaneous potential log

Swi – initial water saturation