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 678-683 https://doi.org/10.38032/scse.2026.4.308

Prediction of Porosity and Permeability Using Well Log and Core Data: A Data-Driven Approach

Md Alamin Islam, Bintun Zaman, Shahria Nayem Ahmed*

Department of Petroleum and Mining Engineering, Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh

ABSTRACT

Accurate prediction of porosity and permeability is very important for reservoir characterization and hydrocarbon extraction. Traditional workflow in the form of empirical correlations is usually difficult, time-consuming, spatially limiting, and entirely dependent on formation geology. The current study examines a different approach, which uses machine learning (ML) regression models based on well-log and core data. Three regression architectures including Random Forest, CatBoost, and K-Nearest Neighbors (KNN) were trained and validated based on a dataset consisting of 340 samples of shaly sand gas reservoirs. The gamma ray (GR), resistivity (RLLD), spontaneous potential (SP), bulk density (RHOB), neutron porosity (NPHI), and depth were used as the input variables with the core-derived porosity (CPHI) and permeability (CKHG) being used as the targets. The quantitative measures of performance of the models included R^2, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The findings indicated that KNN regression was better than its counterparts as it achieved $R^2 = 0.8933$ and $R^2 = 0.9340$ in terms of porosity and permeability prediction, respectively, and more acceptable metrics of errors showed. Comparatively, the traditional empirical methods showed a significantly lower accuracy rate. The findings highlight that machine learning has the potential to provide precise, scalable, and low-cost predictions of the reservoir properties which could lead to better choices for exploration and production activities.

Keywords: Porosity, Permeability, Reservoir Characterization, Machine Learning

Copyright @ All authors

This work is licensed under a Creative Commons Attribution 4.0 International License.

1. Introduction

In studying the potential geological formation, reservoir porosity and permeability are known as the fundamental rock properties that represent the amount of fluid present in a reservoir and its ability to flow easily through a rock [1]. These properties need to be assessed because they help to make effective reservoir characterization, well location optimization, and enhanced oil recovery (EOR) techniques. Thus, an accurate evaluation of these properties can enhance the production performance of the reservoir [2]. Usually, the porosity and permeability are estimated using conventional methods such as core analysis and log interpretation. The main problem with the core analysis technique is that precise measurements are difficult and costly techniques that are also time-intensive and limited to specific depths [3].

Several correlations have been developed between porosity to wireline readings, including the sonic transit time and density logs. Porosity estimation can also be possible with a combination of two or three porosity logs. However, the equivalent porosity is not trivial, which is measured from the sonic transit time and density log [1]. Porosity derived from well logs often needs to be corrected due to various factors (especially shale volume) can affect the porosity raw value [4].

Permeability is a complex function of porosity, lithology, and pore-fluid composition. Besides the core analysis, several traditional methods have been offered to estimate permeability from well log data [5]. All of the traditional methods contain different coefficients which is most likely dependent on the formation geology. Deep-seated rock structure is usually distinct and rife, and it is difficult to apply conventional approaches to determine permeability effectively [6]. The Geology may be oversimplified with these older methods, and the results are not always accurate. As well, other traditional methods such as core analysis are costly and as well as time-consuming [7].

Recently, machine learning (ML) has emerged as a driving engineering resource in reservoir engineering because it provides a data-driven approach to forecast petrophysical properties with greater accuracy. ML models can explore intricate relationships within extensive datasets, lessening reliance on empirical correlations and augmenting predictive capabilities. ML techniques can be used as a quick and cost-effective solution to evaluate reservoir parameters from complex coupled relations to indirect measurements, in this case represented by well logs [8].

This study aims to utilize different machine learning models to predict porosity and permeability, to compare those models' performance with traditional methods based on model evaluation metrics, and to determine the most appropriate predictive model.

2. Literature Review

In previous times, many studies have been done to predict porosity and permeability by applying various machine learning methods. These techniques consistently exceed the

*Corresponding Author Email Address: alaminkst31@gmail.com

performance of conventional empirical correlations and regression analyses, particularly in complex reservoirs such as shaly sand and carbonate formations. Aminian & Ameri [2] and Aïfa et al. [6] indicated that hybrid neuro-fuzzy models and ANN-based flow unit classification enhance the accuracy of permeability predictions. Another author, Trisna et al. [8] showed that the XGBoost and Random Forest models improve the precision of predictions, while Aïfa et al. [6] effectively combined electrofacies classification with nonparametric regression methods. Comparative research, like that of Lim et al. [7] and Tam et al. [9], validated that ANNs are more effective than linear regression in capturing the complexities of reservoir behavior. The results highlight the significance of ML models, which include greater accuracy in reservoir characterization, diminished dependence on expensive core analysis, and improved strategies for hydrocarbon recovery. Moreover, these ML models facilitate quicker and more efficient decision-making regarding well completions and water injection initiatives. Nevertheless, issues such as overfitting, limitations posed by small datasets, and challenges in feature selection persist.

3. Methodology

Data pre-processing is an important step in machine learning and data-driven reservoir characterization. It is the process of converting raw data into a format suitable for ML modelling. It involves data handling, data cleaning, feature selection, and dataset splitting[10].Effective data pre-processing enhances the precision, dependability, and generalization capability of ML models. A flowchart of the functioning of ML algorithms is given in Fig. 1. In this study, 80% of the dataset is considered as training dataset, which is used to train the model. The rest of the data is used to model evaluation for predicting porosity and permeability. Hyperparameters tuning are required when the desired performance is not achieved, and this process continues until meets the desired performance.

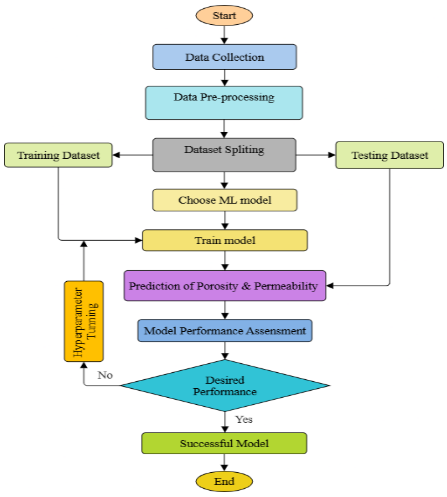

Fig.1 Workflow diagram of ML model

3.1 Random Forest Regression

The random forest regression is a supervised learning approach that is based on bootstrap aggregation (bagging) . It involves the creation of multiple regression trees based on a variety of bootstrapped samples that form part of the original data. The trees are trained separately as well, and the results of them are averaged to get the final prediction. Key stages of the Random Forest model are bootstrap sampling, random feature choice, tree construction, and prediction aggregation [11]. The forest makes T predictions $f_t(x)$ based on single trees given an input x. This ensemble is the average of the predictions of all trees:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(x) \quad (1)$$

where T is the number of trees in the forest and $f_t(x)$ is the prediction from t[th] decision tree.

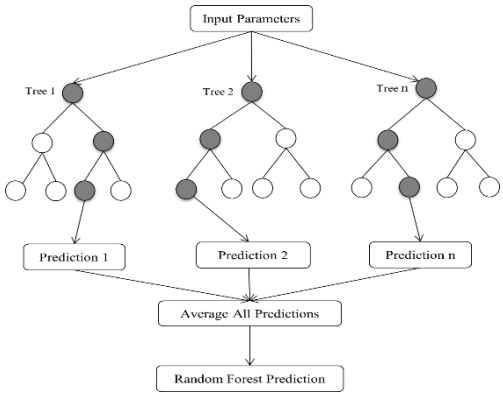

Fig. 2 Basic structure of Random Forest Regression

3.2 Categorical Boosting (CatBoost Regression)

CatBoost regression relies on Gradient boosting concepts, where the learning approach is to collect the predictions of many weak learners and then compose them to make a strong predictor [12]. Any split condition is equally used at all levels of the tree to all data points characterized by oblivious trees so that the tree is balanced and symmetric. And this is done by each tree in the ensemble has an effect in the eventual prediction, and the trees are grown in sequence [13].

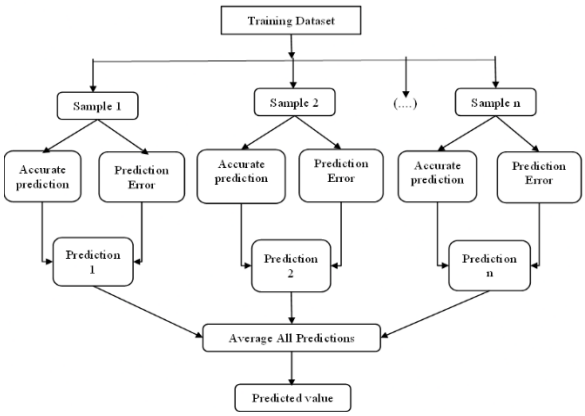

Fig. 3 Basic structure of CatBoost Regression

The model during each of the steps, attempts to reduce the residual error of preceding ensemble. Combine all trees to form the final prediction:

$$\hat{y} = F_0(x) + \sum_{t=1}^{T} \eta . f_t(x) \quad (2)$$

where, η= learning rate and $f_t(x)$= output of the t[th] tree

3.3 K-Nearest Neighbors (KNN) Regression

K-Nearest Neighbors (KNN) regression is a non-parametric model-based technique that uses local learning,

where the algorithm predicts a value of the target variable based solely on the k most similar data points in the train data [14]. In KNN regression, the estimation of a new input point is done by evaluating the outputs of the k nearest neighbors of the input point in the feature space. Sui X et al. [14] describe KNN regression in following way:
A distance is computed between the new observation and every one of the training points, the Euclidean distance being most frequently selected as the criterion.

$$d(x_i, x_{new}) = \sqrt{\sum_{j=1}^{d} w_j \left(x_{new,j} - x_{i,j}\right)^2} \quad (3)$$

Here, $x_{new,j}$ and $x_{i,j}$ is the j[th] feature of the new and training samples respectively, and w_j is the weight assigned to that respective feature. After the identification of the k closest neighbors, a kernel function can be used in unsupervised learning tasks as well as in semi-supervised and map every neighboring instance to a sparse weight with regard to its distance to the query point.

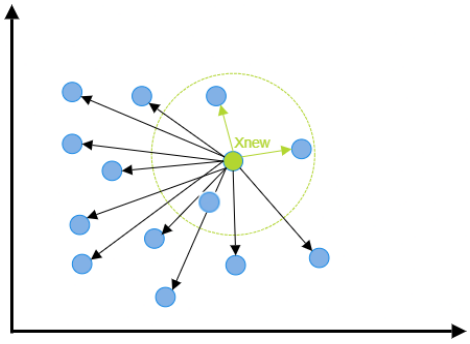

Fig. 4 Basic structure of KNN Regression

The predicted response is then computed as a weighted average of the neighbors' responses:

$$\hat{y}_{new} = \frac{\sum_{i=1}^{k} K(x_{new}, x_{(i)}) y_{(i)}}{\sum_{i=1}^{k} K(x_{new}, x_{(i)})} \quad (4)$$

where, $K(x_{new}, x_{(i)})$ = kernel function, $y_{(i)}$ = known response of the i[th] neighbor, $\hat{y}_{new}$ = new predicted value.

3.4 Model Evaluation Metrics

Model evaluation is the process of comparing a machine learning model's performance by applying different evaluation metrics. In this study, we used coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE) as evaluation metrics.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=n}^{n}(y_i - \bar{y})^2} \quad (5)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (6)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (7)$$

4. Results and Discussions

In this study, 340 data points of a reservoir collected from an open-access platform were used to predict porosity and permeability [15]. This dataset contains well log parameters, including gamma ray (GR), spontaneous potential (SP), resistivity laterolog deep (RLLD), neutron porosity (NPHI), bulk density (RHOB), core porosity (CPHI), and core horizontal permeability (CKHG) by depth.

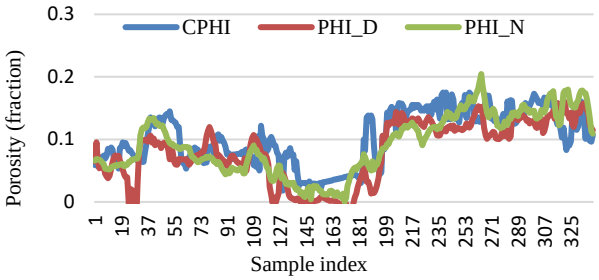

Fig. 5 Comparison between CPHI, PHI_D, and PHI_N

Fig. 5 shows that the estimated porosity from traditional methods always varies with actual core data. This difference indicates that the estimated porosity is not willing to have more shale effect is not corrected for the assessment of shaly formation.

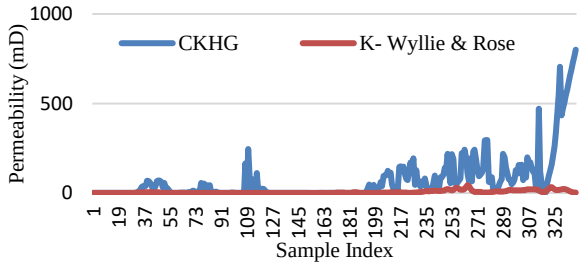

Fig. 6 Comparison between CKHG and K-Wyllie & Rose

Porosity from the neutron method is further used in estimating permeability by Wyllie & Rose method. The results obtained from this method are shown in Fig. 6. This suggests that estimating permeability from traditional method is not reliable in this case, as permeability value is different from most of the core permeability.

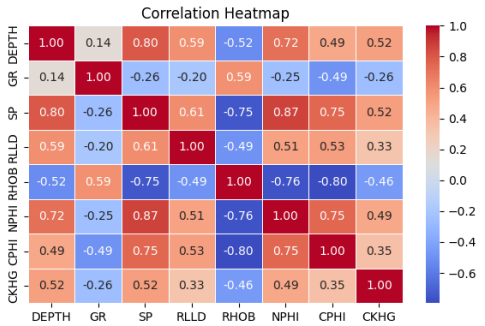

Fig. 7 Correlation heatmap of input parameters

In Fig. 7, a heat map is provided, which is a visual representation of the Pearson correlation coefficients between every possible pair of variables with the span between -1 to +1. The correlation matrix shows that CPHI is strongly and positively correlated with NPHI (0.75) and SP (0.75), and also demonstrates a significant negative correlation with RHOB (-0.80). In the case of CKHG, there is a moderate positive relation with DEPTH (0.52), SP (0.52), and NPHI (0.49), which denotes that the formation depth, porosity, and electrochemical properties are related to compressive strength at CKHG.

Table 1: Performance evaluation of each model

Parameter	Metrics	Random Forest Regression	CatBoost Regression	KNN Regression
Porosity	R^2	0.8670	0.8772	0.8933
	RMSE	10.61	8.95	3.43
	MAE	7.38	6.98	2.90
Permeability	R^2	0.9273	0.9224	0.9340
	RMSE	5.83	6.03	3.33
	MAE	3.18	3.33	2.44

In the following, Table 1 shows a comparative analysis of three machine learning models in terms of their porosity and permeability prediction performance.

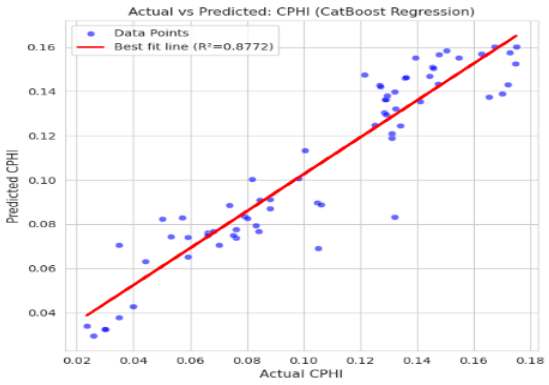

Fig. 8 Scatter plot of actual vs predicted porosity (Random Forest Regression Model)

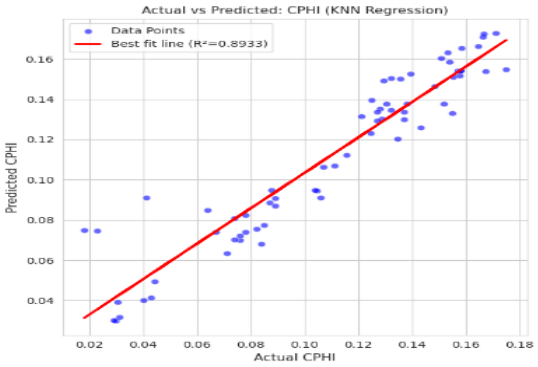

Fig. 9 Scatter plot of actual vs predicted porosity (CatBoost Regression Model)

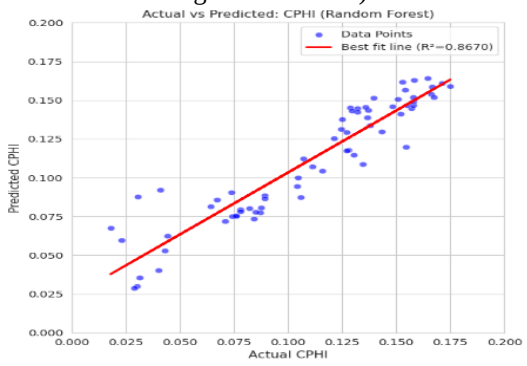

Fig. 10 Scatter plot actual vs predicted porosity (KNN Regression Model)

KNN recorded the best $R^2 = 0.8933$ in the porosity prediction with lowest RMSE of 3.43% and MAE of 2.9%. KNN regression also emerged as the best in predictions of permeability. It had the highest value of R^2 =0.9340, the lowest RMSE (3.41%), and the lowest MAE (2.44%).

Fig. (8-10) shows all three models' capability to predict porosity accurately. Each blue dot represents a single data sample, plotting the actual CPHI value (x-axis) against its corresponding predicted value (y-axis) from each machine learning model. The red line represents linear regression best fit to the data points. which indicates how well the predicted value matches the actual value. The close concentration of the blue dots on the red line indicates that the KNN regression model has a strong approximation of CPHI values.

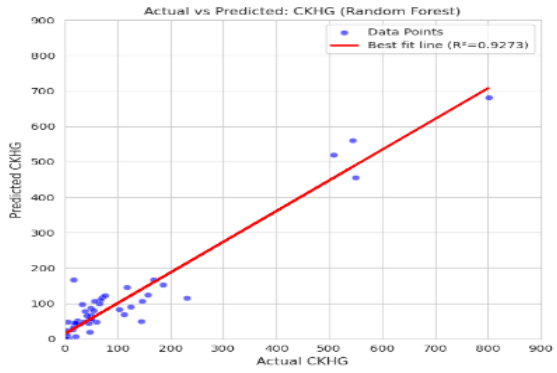

Fig. 11 Scatter plot of actual vs predicted permeability (Random Forest Regression Model)

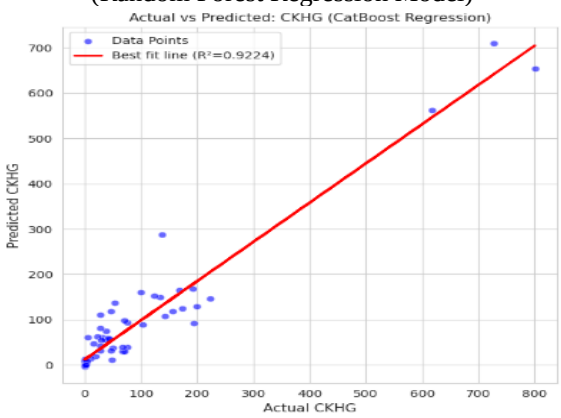

Fig. 12 Scatter plot of actual vs predicted permeability (CatBoost Regression Model)

Similarly, Fig. (11-13) demonstrates the correlation between actual and predicted CKHG and indicates that three models predict permeability very accurately. The KNN Regression yields the highest prediction accuracy for CKHG among all the models, demonstrated by the highest R^2 value (0.9340), closely followed by Random Forest (R^2 = 0.9273) and CatBoost (R^2 = 0.9224), as seen by how closely the predicted values align with the actual ones in each plot.

Fig. 14 and Fig.15 demonstrate a 2D line plot comparing the actual data and the predicted data for porosity and permeability. These two figures show that the predicted data closely matches with actual data in many points. KNN Regression shows good relationships between actual and

predicted values for both parameters (porosity and permeability) due to the high R^2 value.

Fig. 16 and 17 provide a comparison of the machine learning model in terms of evaluation metrics for porosity and permeability prediction. It shows KNN model performs better than its competitors in both cases.

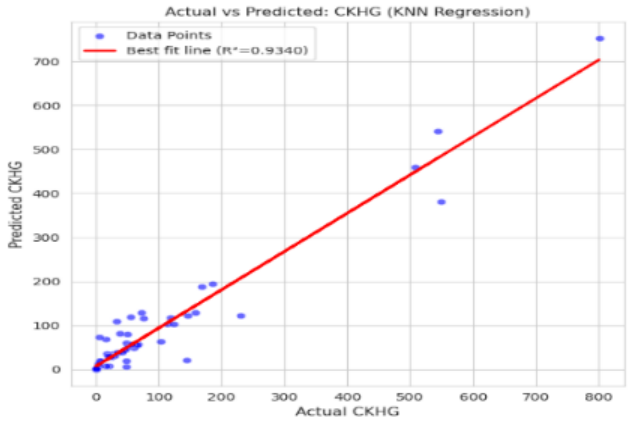

Fig. 13 Scatter plot of actual vs predicted permeability (KNN Regression Model)

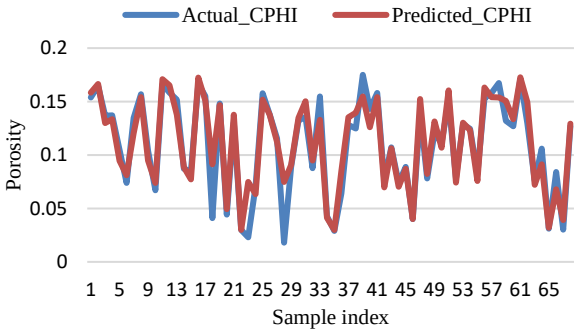

Fig.14 Comparison of actual vs predicted porosity (KNN Regression Model)

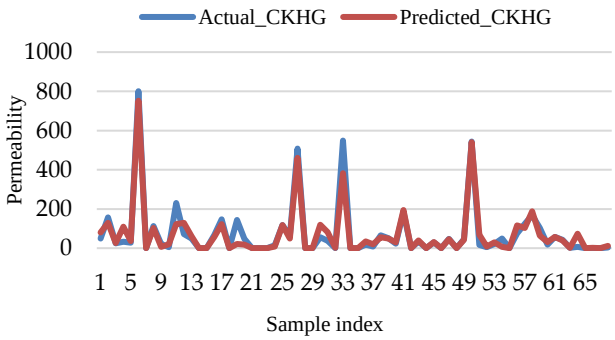

Fig.15 Comparison of actual vs predicted permeability (KNN Regression Model)

Three machine-learning models were compared with empirical methods of their effectiveness in predicting the reservoir properties based on a multidimensional dataset indicating the well-log records. Fig. 18 shows that these three machine learning models have good predictive scores compared with traditional empirical methods like neutron porosity and density porosity methods. In Fig. 19, a comparative analysis of the predictive abilities of Wyllie and Rose correlation method and three different machine learning models in the prediction of permeability of well log data is systematically provided. It is seen that the Wyllie and Rose correlation approach is weaker than the more advanced machine learning models, which means that in most situations, data-driven approaches are not only resilient but even competitive with the standard traditional correlation methods.

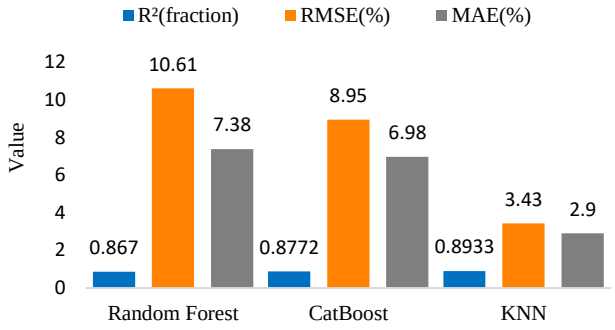

Fig. 16 Comparative analysis of the predictive porosity model based on evaluation metrics

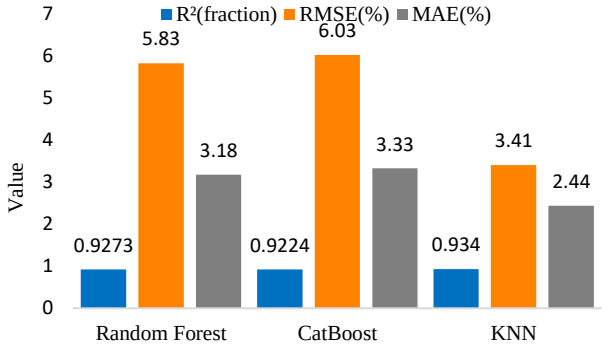

Fig. 17 Comparative analysis of the predictive permeability model based on evaluation metrics

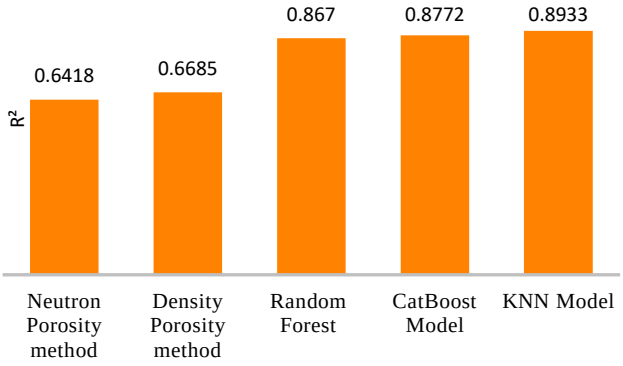

Fig. 18 Comparison of R^2 between traditional empirical methods and ML model for porosity prediction

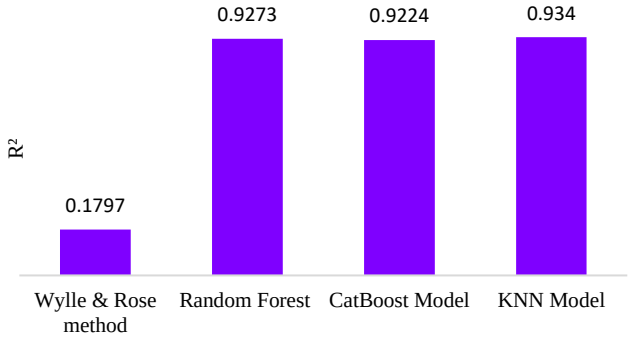

Fig. 19 Comparison of R^2 between traditional empirical methods and ML model for permeability prediction

5. Conclusions

This study demonstrates the potential of machine-learning (ML) methods to predict major reservoir characteristics such as porosity and permeability directly based on the well-log data, which is a faster, cheaper, and more reliable alternative to the traditional core-based processes. Traditional empirical correlations often do not capture the complex, nonlinear interactions that are dictated by the geological heterogeneity, with the ML model used, specifically the K -Nearest -Neighbors (KNN) regression, achieving high predictive or precision with a R^2 score of 0.8933 and 0.9340 of porosity and permeability respectively. Random forest achieved lowest R^2 for porosity prediction, while catboost regression is responsible for lowest permeability prediction accuracy compare with others. Despite the limitations of the study due to a small sample size, the restriction to one lithological context (shaly sand), and the absence of hybrid or deep-learning architectures, the study results highlight the strength of ML when it comes to handling sparse or incomplete data. Future research must expand the data to include many wells, consider hybrid or deep-learning networks and include dynamic shale-volume corrections to optimally generalize models and draw the spatial boundaries of the characteristics of the reservoirs, thus aiding in better informed field-development plans.

6. References

[1] Asquith, G.B., Gibson, C.R., Henderson, S. and Krygowski, D., 2004. *Basic well log analysis*. Tulsa: American Association of Petroleum Geologists.

[2] Aminian, K. and Ameri, S., 2005. Application of artificial neural networks for reservoir characterization with limited data. *Journal of Petroleum Science and Engineering*, 49(3–4), pp.212–222.

[3] Urang, J.G., Ebong, E.D., Akpan, A.E. and Akaerue, E.I., 2020. A new approach for porosity and permeability prediction from well logs using artificial neural network and curve fitting techniques: A case study of Niger Delta, Nigeria. *Journal of Applied Geophysics*, 183, p.104210.

[4] Zargari, H., Poordad, S. and Kharrat, R., 2013. Porosity and permeability prediction based on computational intelligence methods including artificial neural networks and adaptive neuro-fuzzy inference systems in a southern carbonate reservoir of Iran. *Petroleum Science and Technology*, 31(10), pp.1066–1077.

[5] Ghargan, H.M., Al-Fatlawi, O. and Bashir, Y., 2024. Reservoir permeability prediction based on artificial intelligence techniques. *Iraqi Journal of Chemical and Petroleum Engineering*, 25(4), pp.49–60.

[6] Aïfa, T., Baouche, R. and Baddari, K., 2014. Neuro-fuzzy system to predict permeability and porosity from well log data: A case study of the Hassi R'Mel gas field, Algeria. *Journal of Petroleum Science and Engineering*, 123, pp.217–229.

[7] Lim, J.S., 2005. Reservoir properties determination using fuzzy logic and neural networks from well data in offshore Korea. *Journal of Petroleum Science and Engineering*, 49(3–4), pp.182–192.

[8] Nugroho, I.D.R., Trisna, M.D., and Saroji, S., 2024. An implementation of XGBoost and Random Forest algorithms to estimate effective porosity and permeability using well log data from the Fajar Field, South Sumatera Basin, Indonesia. *Indonesian Journal of Applied Physics*, 14(2), pp.271–279.

[9] Tam, T.N.T. and Thanh, D.H.T., 2023. Estimation of petrophysical properties using machine learning methods. In: *Environmental Science and Engineering*. Cham: Springer, pp.475–486.

[10] Mahesh, B., 2020. Machine learning algorithms: A review. *International Journal of Science and Research*, 9(1), pp.381–386.

[11] Ho, T.K., 1995. Random decision forests. Technical Report. Murray Hill, NJ: AT&T Bell Laboratories.

[12] Dorogush, A.V., Ershov, V. and Gulin, A., 2018.*CatBoost: Gradient boosting with categorical features support*. Available at: https://arxiv.org/abs /1810.11363 [Accessed 21 July 2025].

[13] GeeksforGeeks, 2025. *CatBoost in machine learning*. Available at: https://www.geeksforgeeks.org/machine-learning/catboost-ml/ [Accessed 5 August 2025].

[14] Sui, X., He, S., Vilsen, S.B., Meng, J., Teodorescu, R., and Stroe, D.I., 2021. A review of non-probabilistic machine learning-based state-of-health estimation techniques for lithium-ion batteries. *Applied Energy*, 300, p.117346.

[15] GeoLOil, 2025. *Import core data*. Available at: https://geoloil.com/importCoreData.php [Accessed 12 July 2025].