

DOI: <http://doi.org/10.52716/jprs.v14i1.777>

Advanced Machine Learning application for Permeability Prediction for (M) Formation in an Iraqi Oil Field

Noor A. K. Mohammed*, Ghanim M. Farman

Department of Petroleum Engineering, College of Engineering, University of Baghdad, Baghdad, Iraq.

*Corresponding Author E-mail: noor.alhuda2108m@coeng.uobaghdad.edu.iq

Received 06/07/2023, Revised 10/08/2023, Accepted 17/08/2023, Published 20/03/2024

This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

Abstract

Permeability estimation is a vital step in reservoir engineering due to its effect on reservoir's characterization, planning for perforations, and economic efficiency of the reservoirs. The core and well-logging data are the main sources of permeability measuring and calculating respectively. There are multiple methods to predict permeability such as classic, empirical, and geostatistical methods. In this research, two statistical approaches have been applied and compared for permeability prediction: Multiple Linear Regression and Random Forest, given the (M) reservoir interval in the (BH) Oil Field in the northern part of Iraq. The dataset was separated into two subsets: Training and Testing in order to cross-validate the accuracy and the performance of the algorithms. The random forest algorithm was the most accurate method leading to lowest Root Mean Square Prediction Error (RMSPE) and highest Adjusted R-Square than multiple linear regression algorithm for both training and testing subset respectively. Thus, random Forest algorithm is more trustable in permeability prediction in non-cored intervals and its distribution in the geological model.

Keywords: Machine learning, Random forest, Multiple linear regression, Permeability prediction.

تطبيق التعلم الآلي المتقدم للتنبؤ بالنفاذية لتكوين (M) في احد حقول النفط العراقية

الخلاصة:

يعتبر تقييم النفاذية خطوة حيوية في هندسة المكامن بسبب تأثيرها على كل من توصيف المكامن والتخطيط للتقيب والكفاءة الاقتصادية للمكامن. يعتمد توقع نفاذية المكامن على بيانات سجل البئر و تحليل اللباب. هناك طرق متعددة للتنبؤ بالنفاذية مثل الطرق الكلاسيكية، الطرق التجريبية والطرق الإحصائية باستخدام التعلم الآلي. في هذا البحث تمت استخدام ومقارنة خوارزمية التعلم الآلي (الغابات العشوائية Random Forest) و (الانحدار الخطي المتعدد Multiple Linear Regression) للتنبؤ بالنفاذية في مكامن مودود في حقل (BH) شمال العراق. كانت خوارزمية الغابات العشوائية Random Forest هي الأكثر دقة ولديها خطأ تنبؤ مربع متوسط الجذر (RMSPE) أقل من خوارزمية الانحدار الخطي المتعدد مع قيمة معامل ترابط احصائي عالية تبلغ 0.965 في مجموعة بيانات التدريب و 0.962 في مجموعة بيانات الاختبار. نتيجة لذلك، توفر خوارزمية Random Forest طريقة مثالية للتنبؤ بالنفاذية في المناطق غير مأخوذ فيها لباب non-core interval و توزيعها في النموذج الجيولوجي.

1. Introduction:

Enhancing hydrocarbon production required accurate understanding of reservoir characteristics such as permeability that affects fluid transport in porous rock. Permeability is considered an essential metric in reservoir management because it is impact on perforation design, reservoir characterization and flow unit identification[1]. Permeability is commonly given the symbol (k), and the units of permeability are often represented as darcy or millidarcy (D or mD) units [2]. Generally, the permeability of rocks depends on many factors such as the pore types, porosity, and the presence of fractures or other permeable pathways. permeability is controlled by the size of the connecting passage between pores which give better estimation of permeability when combined with porosity [3]. An accurate description of the reservoir in the formation evaluation process is very important as prediction of permeability is an essential key to a good description. There are multi sources and scales for permeability calculation and prediction such as core analysis, well logging, and well testing [4]. Usually, due to their time-consuming and high cost, these methods are not always obtainable in all the wells in the region or at all the desired intervals. Sometimes the accuracy of some of these methods is lower, causing to avoid using their results.

Therefore, identifying a model that predicts the permeability values of a reservoir can provide insight into how to act better in various branches such as reserve estimation, production and developing a field plan. Permeability prediction is a common task in various fields including geology, material science, and chemical engineering. Machine learning techniques can be effectively used to predict petrophysical parameters such as permeability, porosity, and water saturation based on input features in different branches of petroleum engineering [5] and [6]. The objective of this research ids to model and estimate the reservoir permeability as a function of raw well logging data and core permeability for the cored section in the Mauddud Formation in Bai Hassan Oil field in order to estimate the permeability at non-cored intervals and wells.

2. Machine Learning

A part of artificial intelligence (AI) is machine learning. The oil and gas industry has become more dependent on machine learning to reservoir characterization activities, forecast future production, drilling, stimulation and formation assessment[8]. Machine learning encompasses both supervised and unsupervised learning approaches.

Supervised learning includes training a model based on labeled data, where each data point has

corresponding label or target value. The purpose for the model to learn a mapping between the input features and the desired output. In 2000s ensemble methods such as random forests and gradient boosting emerged as effective techniques for supervised learning. Random forests combine multiple decision trees to make predictions while gradient boosting builds an ensemble of weak learners iteratively. In recent years, with the advent of deep learning and the availability of massive computing resources, supervised learning has experienced significant progress. Convolutional neural networks and recurrent neural networks are types of deep neural networks that have achieved breakthrough success in natural language processing, image recognition, and many other domains [9].

Unsupervised learning includes training models based on unlabeled data without any specific target values. The purpose for the model to find structures, relationships or hidden patterns within the data. More recently, unsupervised learning has seen advancements in generative modeling with techniques like GANs (generative adversarial networks) and VAEs (variational autoencoders). Ian Goodfellow and his colleagues introduced GANs (Generative Adversarial Networks), are composed of a generator and a discriminator network. GANs can generate new data samples by learning the underlying distribution of the training data [10].

They have been effectively used for tasks such as text generation, style transfer and image synthesis. VAEs (Variational Autoencoders) are a type of generative model that combines concepts from autoencoders and Bayesian inference. They enable the generation of new data samples and facilitate learning latent representations of the data [10].

There are many uses for machine learning in petroleum engineering as following:

1. Reservoir characterization: Machine learning algorithms can analyze vast amounts of data, such as seismic data, well logs, and production history, to identify patterns and relationships. This helps in understanding reservoir properties, predicting reservoir behavior, and optimizing production strategies[11].
2. Production optimization: Machine learning can be used to develop predictive models that optimize production rates, manage equipment maintenance schedules, and identify potential production issues. These models can leverage real-time data and historical information to improve production efficiency[11][12].
3. Drilling and completion operations: Machine learning algorithms can analyze drilling data and

sensor measurements to optimize drilling parameters, detect anomalies, and predict well integrity issues. This can lead to improved drilling performance, reduced non-productive time, and enhanced wellbore stability[13] and [14].

4. Data-driven decision-making: Machine learning techniques enable engineers to analyze large datasets and make informed decisions. For example, machine learning can be used to predict equipment failures, estimate reservoir performance under various scenarios, and optimize production strategies based on economic and environmental factors[15].

5. Field development planning: Machine learning can aid in optimizing field development plans by integrating geological, geophysical, and engineering data. It can help identify the most prospective drilling locations, optimize well placement, and estimate reserves more accurately[16].

6. Production forecasting: Machine learning models can predict future production rates by analyzing historical data, reservoir characteristics, and production trends. This assists in estimating reserves, planning budgets, and optimizing production schedules[17].

7. Enhanced oil recovery (EOR): Machine learning algorithms can aid in optimizing EOR techniques by analyzing reservoir and production data. They can identify the most effective EOR methods, determine injection strategies, and monitor the progress of the EOR process[12].

8. Safety and risk assessment: Machine learning can be used to analyze safety-related data, such as incident reports, equipment failure logs, and environmental monitoring data. It can help identify potential risks, predict equipment failures, and enhance safety measures in oil and gas operations[12].

3.Methodology

3.1.Random Forest algorithm (RF)

The Core Permeability were modeled using advanced machine learning algorithm Random Forest (RF). The Random Forest algorithm is a widely used and theoretically simple supervised machine learning method that belongs to the ensemble learning family. It is used for both regression, classification and feature selection tasks in various domains. Breiman proposed the idea of random forests in general to combines multiple decision trees to create a powerful predictive model. Random forest algorithm generated a large number of decision trees by using data training to lower the variance compared to that of a single decision tree. The generalization error of a forest of tree

classifiers is determined by the strength of each individual tree inside the forest and the correlation between them[18]. The random forest algorithm is used in several domains. Random Forest can be utilized for reservoir characterization tasks such as predicting reservoir properties or identifying hydrocarbon-bearing zones. By training the algorithm on data from well logs, seismic data, and production data, it can learn the relationships between various features and the target variables of interest. This can aid in understanding the properties of subsurface reservoirs and making informed decisions about drilling, production, and field development. The Random Forest algorithm can be utilized to predict permeability based on various geological and petrophysical attributes. The Random Forest approach was used to classify the lithofacies in a cored well and predicted their distribution in additional non-cored wells[19] ,[20]and[21].

The basis for the work of the random forest algorithm shown in Table (1).

Table (1) The basis for the work of the random forest algorithm.

Step	Description
1	Data Preparation: Random Forest requires a labeled dataset, meaning a dataset where the target variable (the variable to be predicted) is known. The dataset is divided into two parts: the features (input variables) and the labels (output variable).
2	Random Sampling: The algorithm randomly selects a subset of the original dataset with replacement. This process is called bootstrapping. The selected subset is used to train each decision tree in the Random Forest.
3	Decision Tree Construction: For each decision tree, a random subset of features is selected from the original feature set. This subset is used to construct the decision tree using a method like the CART (Classification and Regression Trees) algorithm. The decision tree is built by recursively splitting the data based on the selected features until a stopping criterion is met .
4	Ensemble Creation: Once all the decision trees are built, the Random Forest algorithm combines their predictions to make the final prediction.
5	Prediction: The Random Forest algorithm uses the ensemble of decision trees to predict the target variable for new, unseen data points. Each decision tree in the ensemble independently makes a prediction, and the final prediction is determined based on the majority vote (for classification) or averaging (for regression).

3.2. Multiple Linear Regression Algorithm

Multiple Linear regression (MLR) is a statistical technique that expands the applications of linear regression by integrating extra explanatory variables. This technique used to model the relationship

between a dependent or criterion variable and multiple independent or predictor variables [22]. Multiple linear regression assumes linearity (linear relationship between the dependent variable and the independent variables) and independence of errors. Additionally, other regression techniques, such as nonlinear regression or machine learning algorithms, can be explored if the relationship between the variables is more complex or if there are interactions between them. It is commonly used in various domains, including economics, finance and permeability prediction as function of other core and well log data in geosciences and petroleum engineering. The success of any regression model heavily depends on the quality and representativeness of the data, as well as the appropriate selection and engineering of independent variables[23].

4. Result and Discussion

Data Review: In the first step, the dataset has to be called to the R interface for preparation and visualization for further processing. The computer-saved data set is called BH dataset and includes input well logs. The input variables involve Gamma rays (GR), neutron porosity (NPHI), bulk density (RHOB), and compressional slowness (DT) with routine core analysis (Core Porosity, Core Permeability). The input variables into the permeability model were the raw logs (GR, NPHI, RHOB, DT) and not the Computer Processing Interpretation (CPI) interpretations like porosity, shale volume, water saturation, etc. As a result, the built model may be utilized without any interpretation to guess the permeability of the raw logs. That makes permeability prediction in non-cored wells simpler and helps geologists and engineers reduce uncertainty. After that, the necessary R packages are installed in order to carry out the various modeling processes. The data set was subdivided randomly using machine learning into two main groups as a cross-validation tool before adopting permeability prediction. This group includes the "Training" set with 75% (91 data samples) and the "Testing set with 25% (31 data samples). Figure (2) clarified scatterplot of response (Core Permeability) and predictors (Raw Well Log). The histogram of both core permeability (CKHA) and core porosity (CPOR) given permeability shown in Figures (3) and (4).

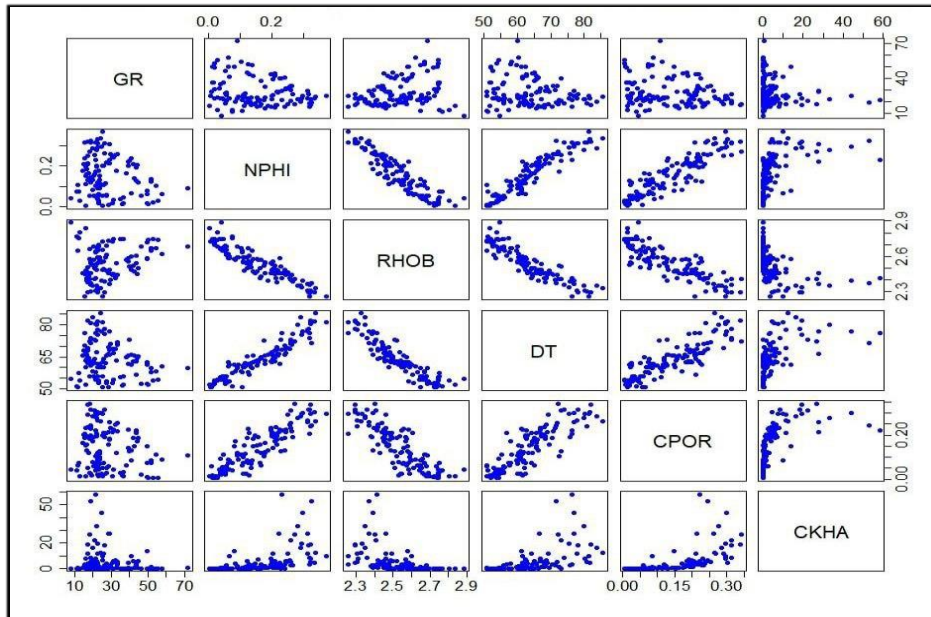


Fig. (2): Scatter matrix plot of the well log records and core data

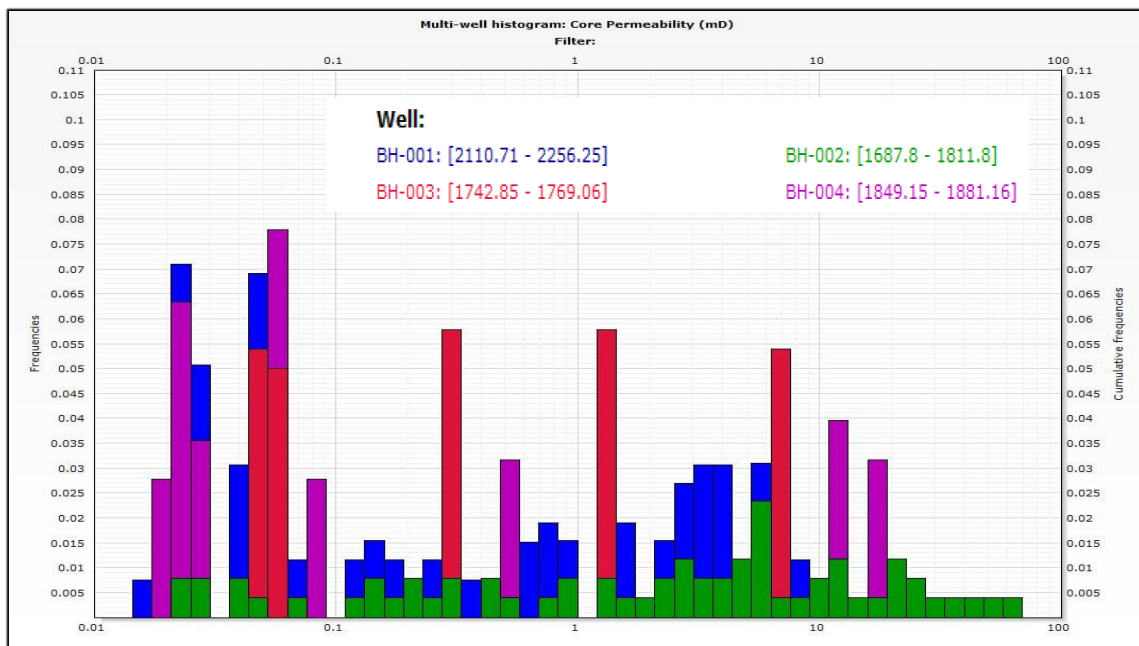


Fig. (3): Histogram of core permeability given the cored wells

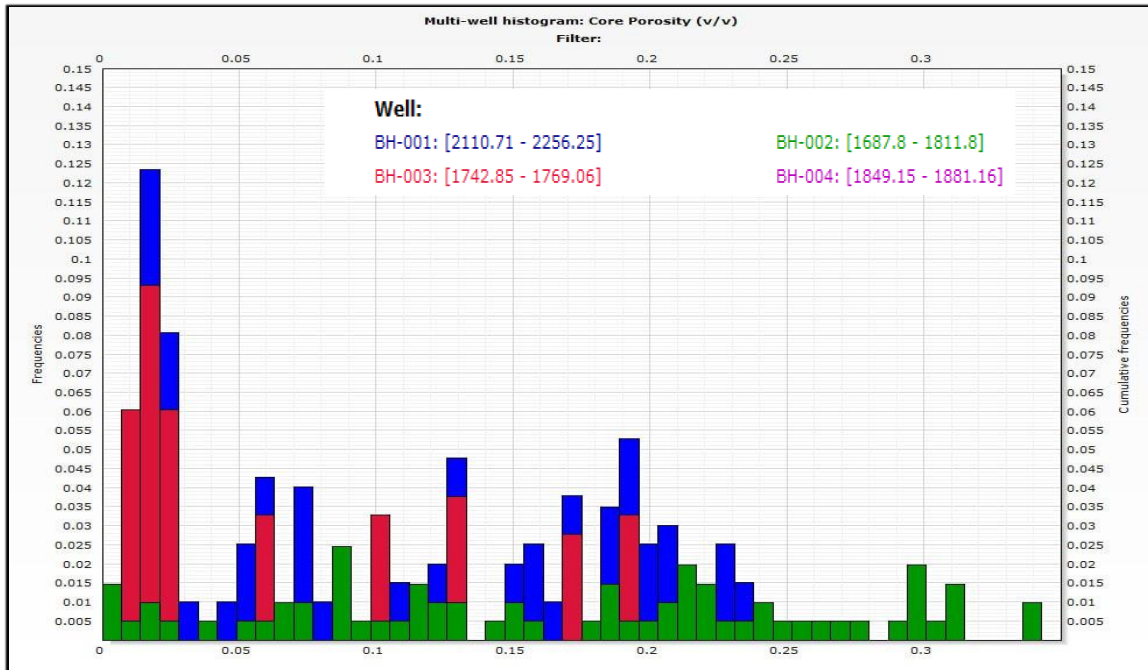


Fig. (4): Histogram of core porosity given the cored wells

4.1. Random Forest algorithm

The Random Forest algorithm was used for permeability prediction. The equation used in this method is shown in Figure (5). Permeability and core permeability by using the Random Forest algorithm for training and testing data sets shown in Figures (6) and (7) respectively. The statistical results achieved an average R value of 0.965 on the training set and 0.962 on the testing set while RMSE (Root Mean Square Prediction Error) was 0.191 for the training set and 0.154 for the testing set as shown in Figures (5) and (6).

```
Call:
randomForest(formula = log10(CKHA) ~ GR + NPHI + RHOB + DT, data = BH, ntree = 1000, proximity = TRUE, mtry = 3)
Type of random forest: regression
Number of trees: 1000
No. of variables tried at each split: 3

Mean of squared residuals: 0.0414801
% Var explained: 96.71
```

Fig. (5): The equation used in Random Forest algorithm

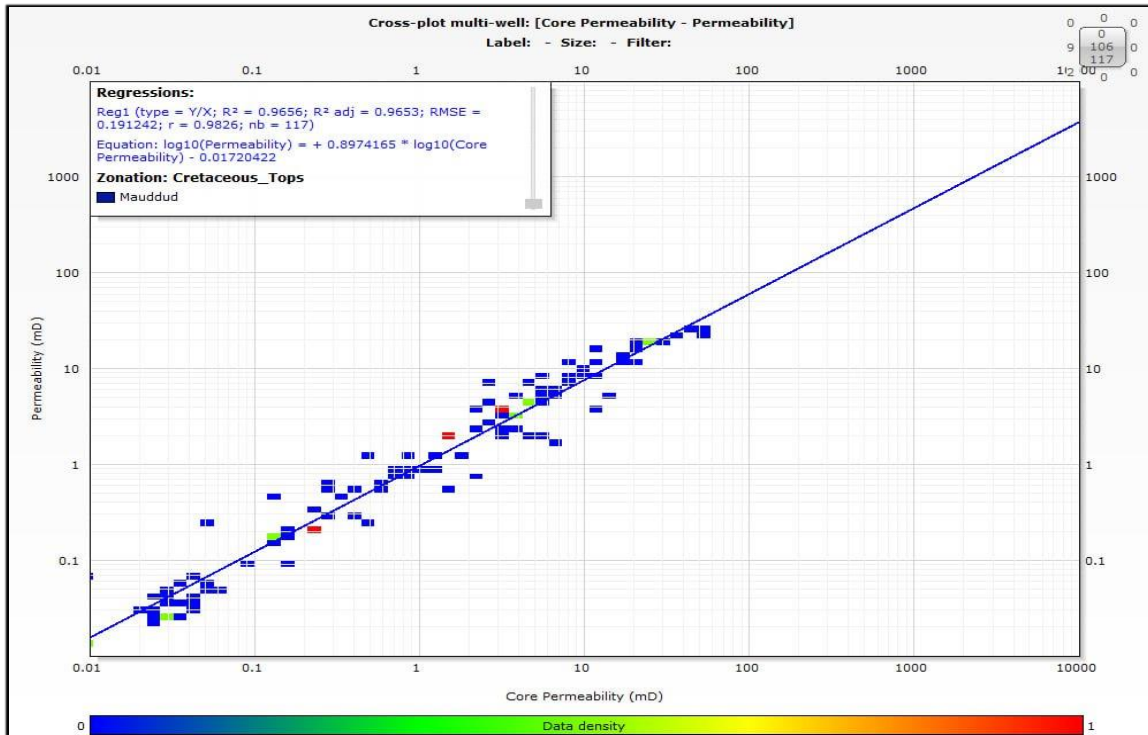


Fig. (6): Permeability vs. core permeability by using RF for training data set

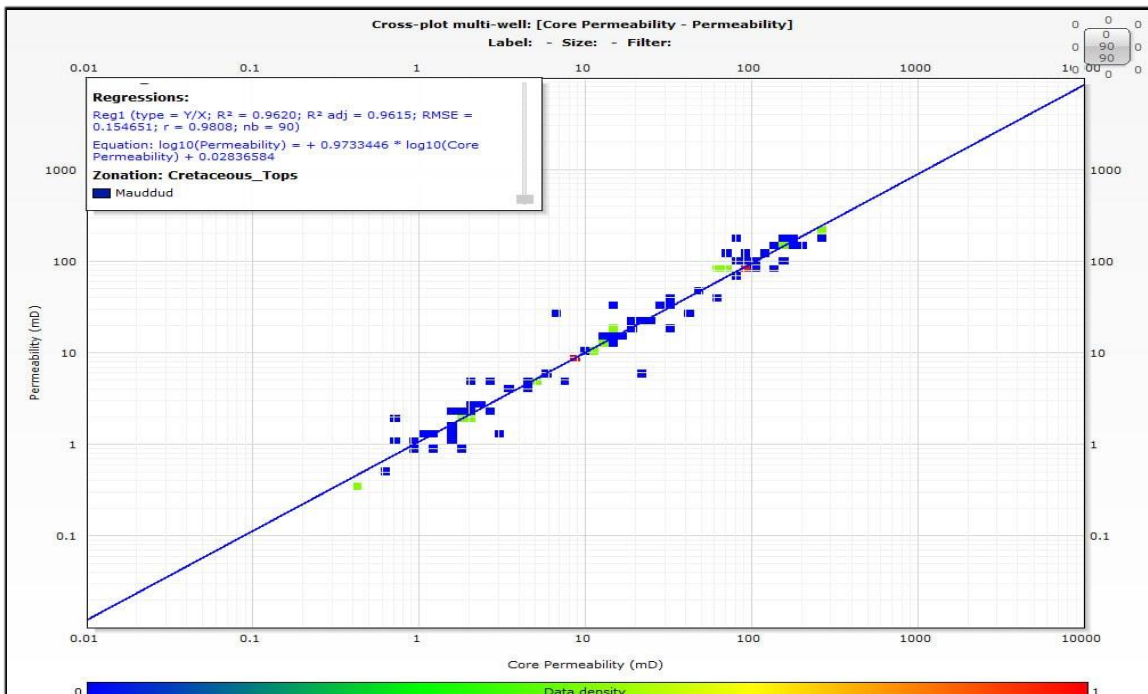


Fig. (7): Permeability vs. core permeability by using RF for testing data set

4.2. Multiple Linear Regression Algorithm

The equation used in this method is shown in Figure (8). Permeability and core permeability by using linear machine learning (MLR algorithm) for training and testing data sets shown in Figures

(8) and (9) respectively. The statistical results achieved an average R value of 0.620 on the training set and the same value on the testing set while RMSE (Root Mean Square Prediction Error) were 0.537 for the training set and 0.561 for the testing set as shown in Figures (9) and (10).

```
Call:
lm(formula = log10(CKHA) ~ GR + NPFI + RHOB + DT, data = BH)

Residuals:
    Min       1Q   Median       3Q      Max
-1.5738 -0.5143 -0.0515  0.4621  1.9365

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.3919268   4.1612067   0.815  0.41665
GR          -0.0006857   0.0057114  -0.120  0.90464
NPFI         6.4129185   2.1980641   2.918  0.00423 **
RHOB        -1.8723356   1.3101649  -1.429  0.15564
DT           0.0025745   0.0238965   0.108  0.91439
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7059 on 117 degrees of freedom
Multiple R-squared:  0.6212,    Adjusted R-squared:  0.6082
F-statistic: 47.97 on 4 and 117 DF,  p-value: < 2.2e-16
```

Fig. (8): The equation used in Multiple Linear Regression algorithm

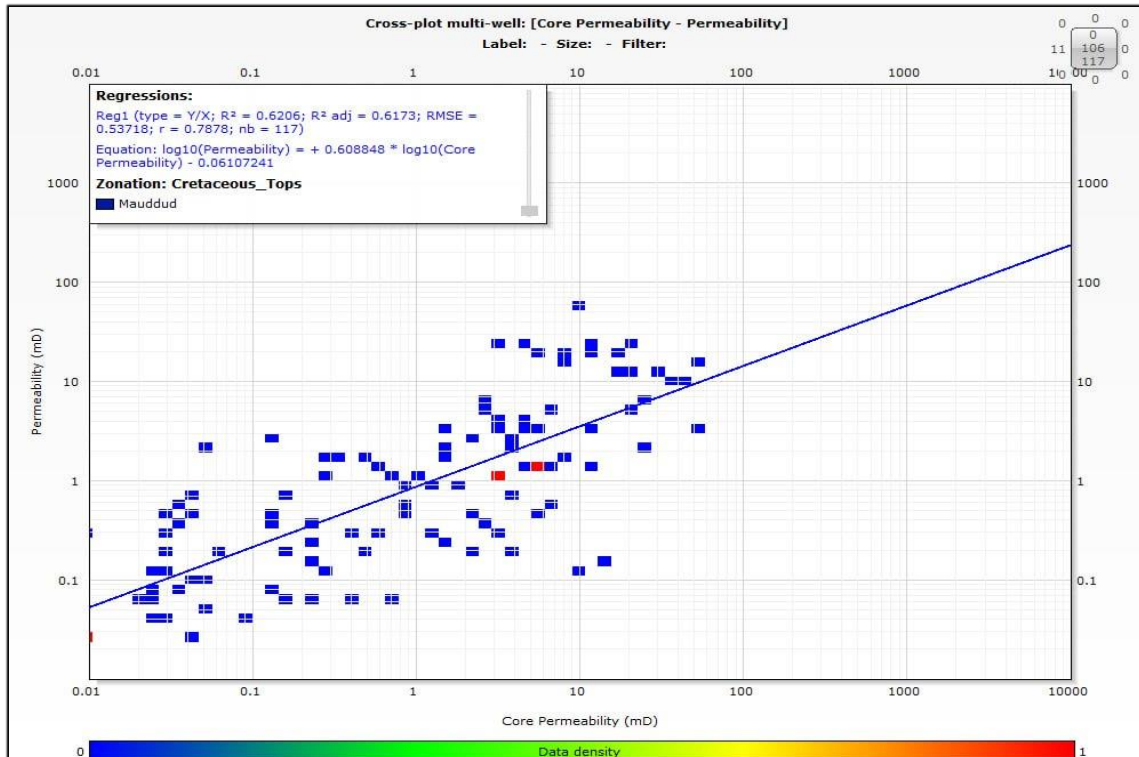


Fig. (9): Permeability vs. core permeability by using MLR for training data set

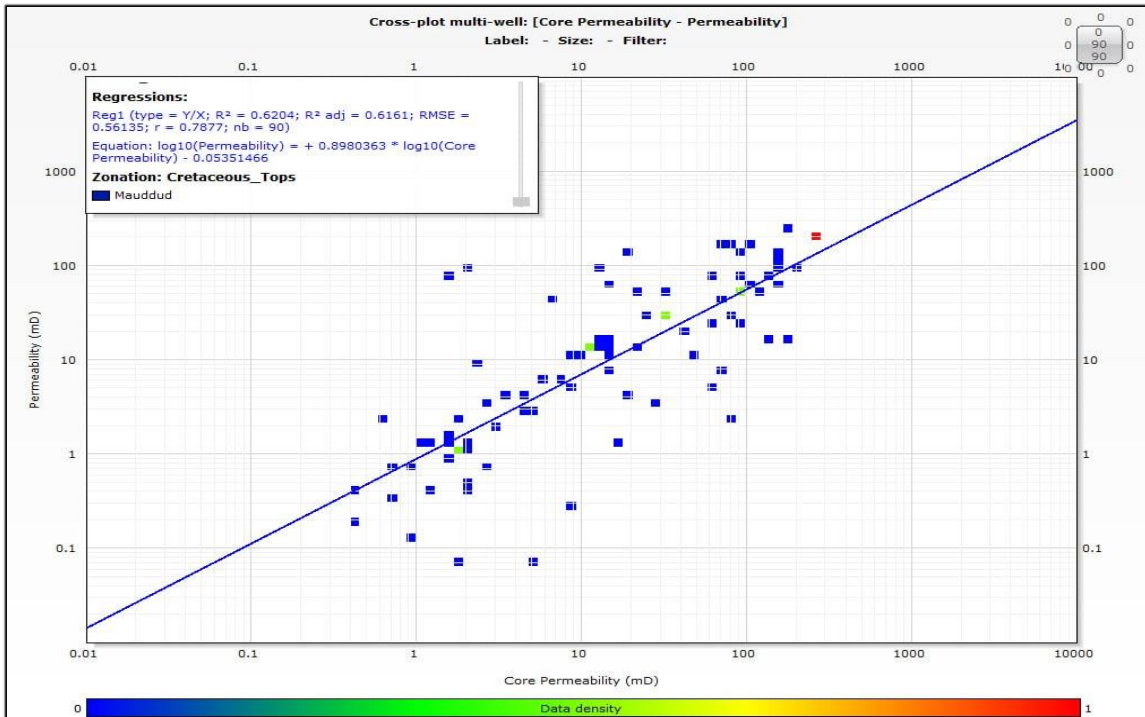


Fig. (10): Permeability vs. core permeability by using MLR for testing data set

The matching of permeability predictions from both algorithms with core permeability in the studied wells is shown in Figures (11) and (12). This figure shows that the random forest algorithm is more accurate and matches the core data better than the multiple linear regression algorithm.

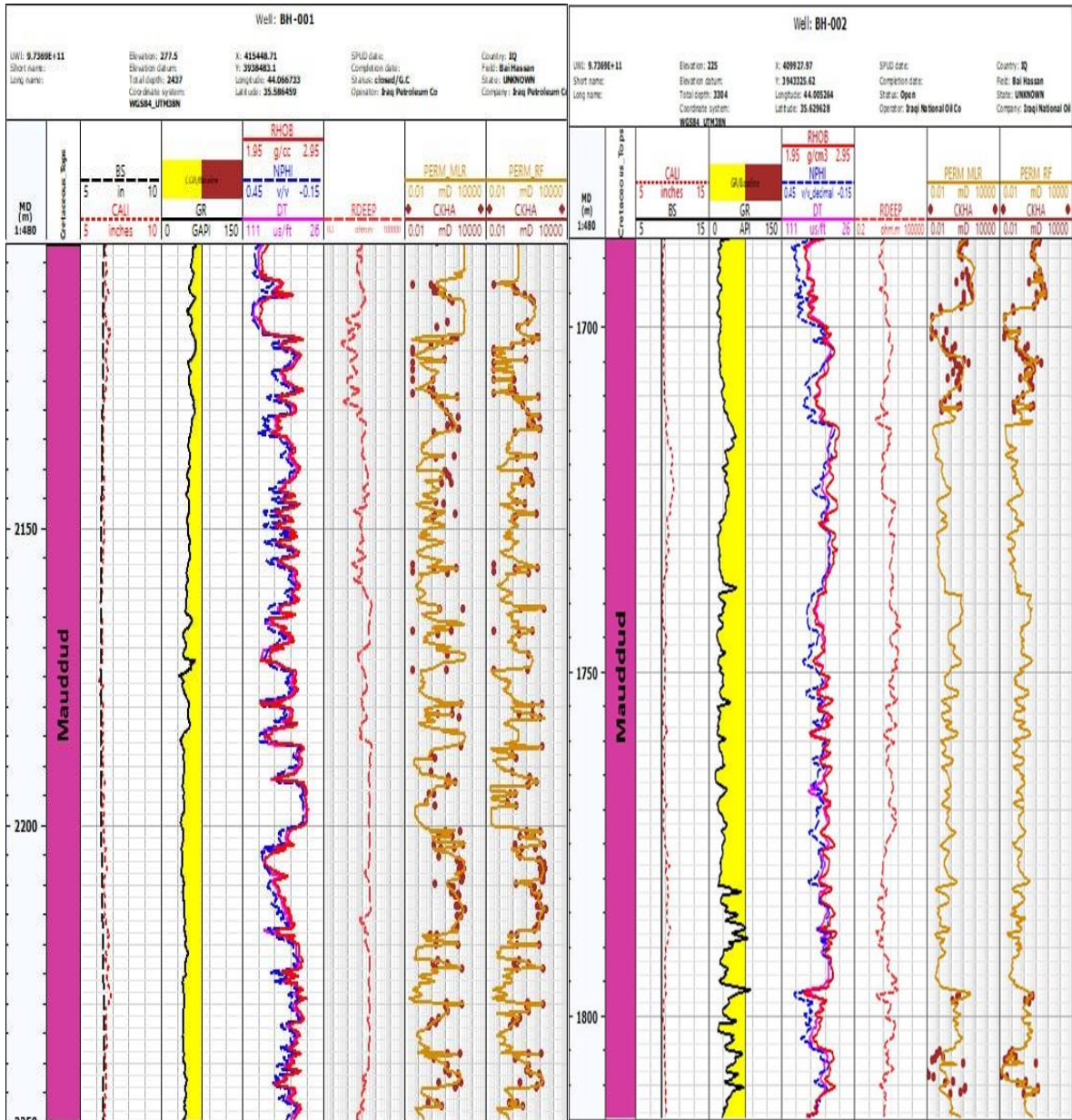


Fig. (11): Matching between prediction permeability and core permeability in well (BH-001) and well (BH-002)

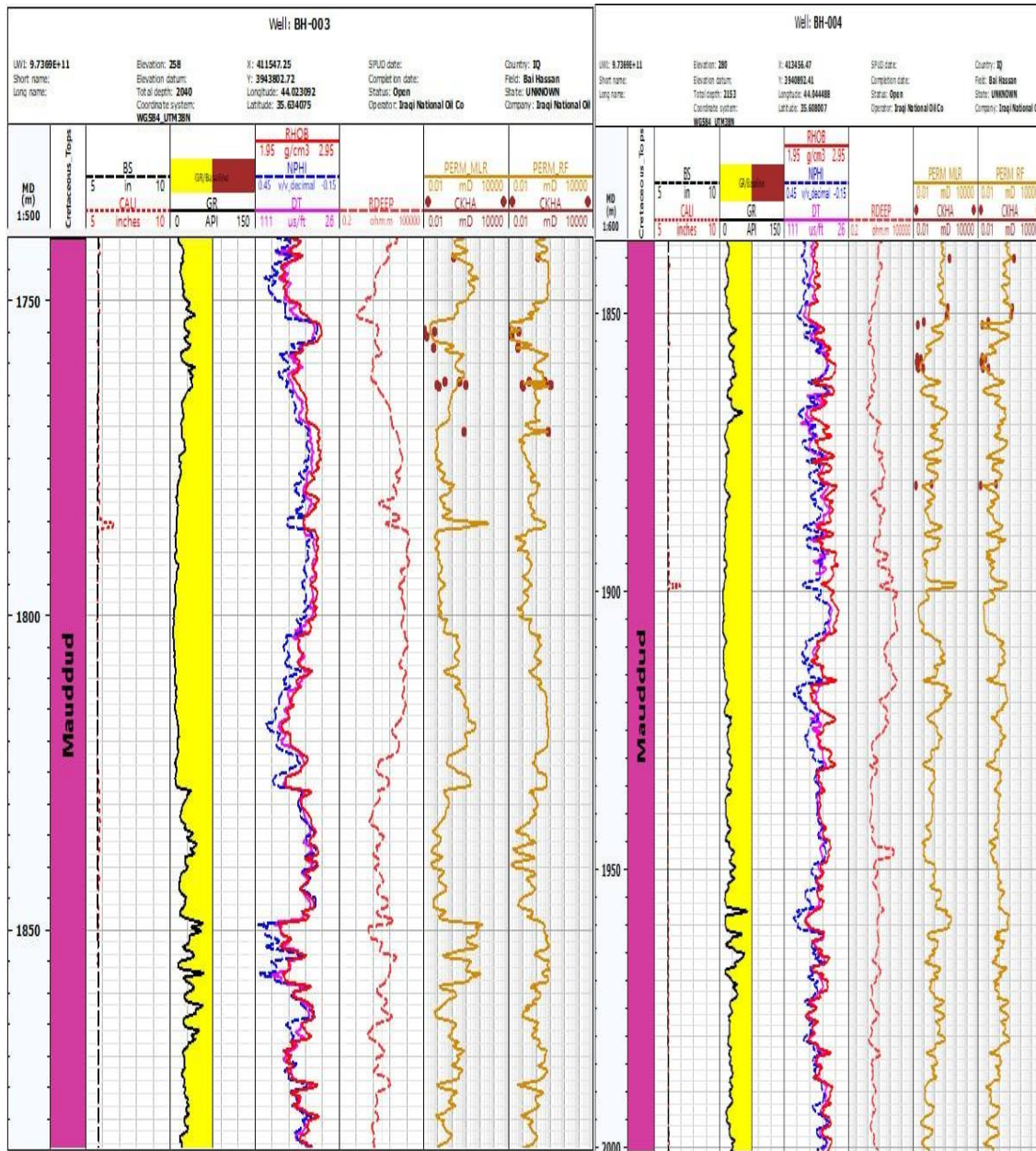


Fig. (12): Matching between prediction permeability and core permeability in well (BH-003) and well (BH-004)

5. Conclusion

Advanced machine learning (ML) approaches were adopted to model and estimate the reservoir permeability as a function of raw well logging data and core permeability at cored intervals in order to predict their sequence in the other wells. The raw well logs involve Gamma rays (GR), neutron porosity (NPHI), bulk density (RHOB), and compressional slowness (DT) were used as inputs for permeability modeling. Two supervised machine learning algorithms were adopted to achieve the prediction process: the Random Forest algorithm and the Multiple Linear Regression algorithm.

The modeling was achieved based on the training dataset, while the prediction was constructed based on the training and the testing subsets after sampling the entire dataset (cross-validation). The comparison of these machine learning algorithms performance and accuracy was done using the root mean square prediction error, and adjusted R-square which reflects how well the predicted and observed core permeability match. The advanced algorithm (Random Forest) was most accurate and had less Root Mean Square Prediction Error (RMSPE) than the linear algorithm (Multiple Linear Regression). Random Forest achieved excellent matching between prediction permeability and core permeability, with a high average R value of 0.965 in the training dataset and 0.962 in the testing dataset. As a result, the Random Forest algorithm provides a perfect approach to permeability prediction in non-cored intervals and its distribution in the geological model.

Symbols:

Symbol	Description
BH	Bai Hassan
AI	Artificial Intelligence
GANs	Generative Adversarial Networks
VAEs	Variational Autoencoders
RF	Random Forest
RMSPE	Root Mean Square Prediction Error

References:

- [1] Z. A. Mahdi and G. M. Farman, "A Review on Models for Evaluating Rock Petrophysical Properties", Iraqi Journal of Chemical and Petroleum Engineering, vol. 24, no. 1, pp. 125–136, 2023. <https://doi.org/10.31699/IJCPE.2023.1.14>
- [2] R. A. Hashim and G. M. Farman, "Evaluating Petrophysical Properties of Sa'di Reservoir in Halfaya Oil Field", Iraqi Geological Journal, vol. 56, no. 2D, pp. 118–126, 2023. <https://doi.org/10.46717/igj.56.2D.9ms-2023-10-15>
- [3] X. Yang, Y. Mehmani, William A. Perkins, A. Pasquali, M. Schönherr, K. Kim, M. Perego, M. L. Parks, N. Trask, M. T. Balhoff, M. C. Richmond, M. Geier, M. Krafczyk, L. Luo, A. M. Tartakovsky, and T. D. Scheibe, "Intercomparison of 3D pore-scale flow and solute transport simulation methods", Adv. Water Resour., vol. 95, pp. 176–189, 2016. <https://doi.org/10.1016/j.advwatres.2015.09.015>.
- [4] H A. H. Tali, S. K. Abdulridha, L. A. Khamees, J. I. Humadi, G. M. Farman, and S. J. Naser, "Permeability estimation of Yamama formation in a Southern Iraqi oil field, case study", Second International Conference on Innovations in Software Architecture and Computational Systems (ISACS 2022), 2023-01-05 | Conference paper. <https://doi.org/10.1063/5.0163281>
- [5] D. A. Otchere, T. O. A. Ganat, R. Gholami, and M. Lawal, "A novel custom ensemble learning model for an improved reservoir permeability and water saturation prediction," J. Nat. Gas Sci. Eng., vol. 91, 103962, 2021. <https://doi.org/10.1016/j.jngse.2021.103962>.
- [6] N. A. Kadhim and G. M. Farman, "Advanced Geostatistical Techniques for Building 3D Geological Modeling: A Case Study from Cretaceous Reservoir in Bai Hassan Oil Field", Iraqi Geological Journal, vol. 56, no. 2F, pp. 214-227, 2023. <https://doi.org/10.46717/igj.56.2F.14ms-2023-12-20>
- [7] G. Zeynalov, Muhammed A. Ismail, and F. Al-Beyati, "Petro physical Modeling of the Tertiary Reservoirs in the Small Giant Bai Hassan Oil", Int. J. Enhanc. Res. Sci. Technol. Eng., vol. 5, no. 12, pp. 80–89, 2016.
- [8] U. S. Alameedy, A. T. Almomen, and N. Abd, "Evaluating Machine Learning Techniques for Carbonate Formation Permeability Prediction Using Well Log Data", Iraqi Geol. J., vol. 56, no. 1D, pp. 175–187, 2023. <https://doi.org/10.46717/igj.56.1D.14ms-2023-4-23>
- [9] M. W. Berry, A. Mohamed, and B. W. Yap, "Supervised and unsupervised learning for data science", Springer, 2019. <https://doi.org/10.1007/978-3-030-22475-2>
- [10] A. Singh, N. Thakur, and A. Sharma, "A review of supervised machine learning algorithms", in

- 2016 3rd international conference on computing for sustainable global development (INDIACom), Ieee, pp. 1310–1315, 2016.
- [11] H. Wang and S. Chen, “Insights into the Application of Machine Learning in Reservoir Engineering: Current Developments and Future Trends”, *Energies*, vol. 16, no. 3, 1392, 2023. <https://doi.org/10.3390/en16031392>
- [12] A. Sircar, K. Yadav, K. Rayavarapu, N. Bist, and H. Oza, “Application of machine learning and artificial intelligence in oil and gas industry”, *Pet. Res.*, vol. 6, no. 4, pp. 379–391, 2021. <https://doi.org/10.1016/j.ptlrs.2021.05.009>
- [13] H. Gamal, A. Alsaihati, and S. Elkatatny, “Predicting the rock sonic logs while drilling by random forest and decision tree-based algorithms”, *J. Energy Resour. Technol.*, vol. 144, no. 4, 2022. <https://doi.org/10.1115/1.4051670>
- [14] H. Gamal, A. Alsaihati, S. Elkatatny, S. Haidary, and A. Abdulraheem, “Rock strength prediction in real-time while drilling employing random forest and functional network techniques”, *J. Energy Resour. Technol.*, vol. 143, no. 9, 2021. <https://doi.org/10.1115/1.4050843>
- [15] M. A. Mirza, M. Ghoroori, and Z. Chen, “Intelligent Petroleum Engineering”, *Engineering*, vol. 18, pp. 27–32, 2022. <https://doi.org/10.1016/j.eng.2022.06.009>
- [16] A. Kumar, “A Machine Learning Application for Field Planning,” in *Offshore Technology Conference*, Offshore Technology Conference, Paper Number: OTC-29224-MS, Houston, Texas, 2019. <https://doi.org/10.4043/29224-MS>
- [17] M. Cordeiro-Costas, D. Villanueva, P. Eguía-Oller, and E. Granada-Álvarez, “Machine learning and deep learning models applied to photovoltaic production forecasting”, *Appl. Sci.*, vol. 12, no. 17, 8769, 2022. <https://doi.org/10.3390/app12178769>
- [18] L. Breiman, “Random forests”, *Mach. Learn.*, vol. 45, pp. 5–32, 2001. <https://doi.org/10.1023/A:1010933404324>
- [19] W. J. Al-Mudhafar, “Applied Geostatistical Reservoir Characterization in R: Review and Implementation of Rock Facies Classification and Prediction Algorithms-Part I”, in *Offshore Technology Conference*, OnePetro, 2016. <https://doi.org/10.4043/26947-MS>
- [20] S. Bhattacharya and S. Mishra, “Applications of machine learning for facies and fracture prediction using Bayesian Network Theory and Random Forest: Case studies from the Appalachian basin, USA”, *J. Pet. Sci. Eng.*, vol. 170, pp. 1005–1017, 2018. <https://doi.org/10.1016/j.petrol.2018.06.075>

- [21] W. J. Al-Mudhafar, “Advanced supervised machine learning algorithms for efficient electrofacies classification of a carbonate reservoir in a giant southern iraqi oil field”, in Offshore Technology Conference, OnePetro, 2020. <https://doi.org/10.4043/30906-MS>
- [22] J. L. L. Pereira, "Permeability prediction from well log data using multiple regression analysis", West Virginia University, 2004.
- [23] B. A. Al-Baldawi, “Permeability Estimation of Khasib Formation in Amara Oil Field from Well logs Data Using Multilinear Regression Technique and Empirical”, J. Pet. Res. Stud., vol. 6, no. 1, 2016. <https://doi.org/10.52716/jprs.v6i1.134>.