Journal of Petroleum Research and Studies

Open Access
Vol. 15, No. 1, March 2025, pp. 43-61

P- ISSN: 2220-5381
E- ISSN: 2710-1096

DOI: http://doi.org/10.52716/jprs.v15i1.869

Hyperparameter Optimization of Tree-Based Machine Learning (TB-ML) to Predict Permeability of a Heterogeneous Carbonate Oil Reservoir

Alqassim A. Hasan[1*], Ali A. Nimr[1], Yousif T. Yaseen[1], Watheq J. Al-Mudhafar[2], and David A. Wood[3]

[1] Department of Oil and Gas Engineering, Basrah University for Oil and Gas, Basrah, Iraq
[2]Basrah Oil Company, Basrah, Iraq
[3]DWA Energy Limited, Lincoln, United Kingdom
[*]Corresponding Author E-mail: qassimali624@gmail.com

Received 26/12/2023, Revised 28/03/2024, Accepted 07/04/2024, Published 21/03/2025

Abstract

Permeability is a crucial petrophysical attribute to be accurately estimated due to its direct influence on reservoir characterization, heterogeneity assessment, reservoir simulation, and the level of uncertainty in decision-making during field development planning. However, measuring permeability often involves expensive core analysis or well test analysis. It is typically not feasible to conduct such analysis across an entire reservoir involving cores from all wells. Therefore, there is a need to accurately model and predict permeability as a function of routinely obtained, lower cost, well logging data. Machine learning algorithms (ML) have been recently developed to reliably predict permeability by leveraging well logs data. In this research, an efficient tree-based (TB-ML) algorithm incorporating extreme gradient boosting (XGBoost) is employed to predict permeability in the Mishrif carbonate reservoir (Iraq) based on facies and well logging data. The recorded and interpreted well log variables used as input variables include gamma ray, caliper, density, neutron porosity, shallow and deep resistivity, total porosity, spontaneous potential, photoelectric factor, and water saturation. Additionally, core-derived permeability and porosity data is used to calibrate the ML predictions. The discrete reservoir facies are distinguished by applying a k-means clustering algorithm. Subsequently, the TB-ML algorithm is developed using the default and fine-tuned hyperparameters with the aid of two search algorithms: random search and Bayesian optimization. The permeability predictions are evaluated using cross-validation and error quantification metrics, which include the adjusted coefficient of determination (R2) and root mean squared error (RMSE). A comparison of adjusted-R2 and RMSE for the various TB-ML model configurations developed is compared for training and testing subsets to illustrate their permeability prediction performance. These results suggest that the method is sufficiently reliable to be generalized for application in both carbonate and clastic reservoirs in other oil and gas fields.

Keywords: Machine learning, reservoir characterization, permeability prediction, XGBoost, hyperparameter tuning.

Journal of Petroleum Research and Studies

Open Access
Vol. 15, No. 1, March 2025, pp. 43-61

P- ISSN: 2220-5381
E- ISSN: 2710-1096

تحسين المعاملات الفائقة (Hyperparameter) لأساليب التعلم الآلي المعتمدة على خوارزميات (TB-ML) للتنبؤ بالنفاذية في مكمن نفط كربوني غير متجانس

الخلاصة:

النفاذية هي خاصية بترو فيزيائية مهمة يجب احتسابها بدقة وذلك لتأثيرها المباشر على توصيف المكمن، ومحاكاة المكامن، ومستوى عدم اليقين في اتخاذ القرار أثناء التخطيط لتطوير الحقل. ومع ذلك، فإن قياس النفاذية يتم إما عن طريق فحوصات نماذج اللباب الصخري او عمليات فحص الابار. وعادةً، هذه الفحوصات لا تكون متاحة لطبقات المكمن بأكملها ولا في جميع الابار المحفورة. لذلك، هناك حاجة ملحة لنمذجة النفاذية بدقة والتنبؤ بها كدالة لبيانات تسجيل الآبار. تم استخدام خوارزميات التعلم الآلي (ML) مؤخرًا كطريقة واعدة للتنبؤ بالنفاذية من خلال الاستفادة من بيانات سجلات الآبار جنبًا إلى جنب مع السحنات أو الصخارية. في هذا البحث، تم استخدام خوارزمية فعالة وهي XGBoost للتنبؤ بالنفاذية في مكمن مشرف بالاعتماد على بيانات تسجيل الآبار و السحنات. تتضمن بيانات تسجيل الآبار (Gamma Ray, Caliper, Density, Neutron Porosity, Shallow and Deep Resistivity, Total Porosity, Spontaneous Potential, Photoelectric Factor, and Water Saturation) بالإضافة إلى النفاذية والمسامية من نماذج اللباب الصخري. ايضاً، تم الحصول على السحنات المنفصلة عن طريق استخدام خوارزمية k-means clustering. بعد ذلك، تم تحسين أداء الخوارزمية المستخدمة باستخدام خوارزميتين للبحث: Random Search و Bayesian Optimization. تم تقييم توزيعات النفاذية المتوقعة بناءً على:Adjusted R2,, RMSE. و اخيراً، تم عمل مقارنة بين بين النماذج الثلاثة بالنظر إلى مجموعات التدريب والاختبار لتوضيح دقة نهج التعلم الآلي المستخدم للتنبؤ بالنفاذية. النتائج اظهرت ان هذا النهج دقيق بما يكفي ليتم تعميمه للتطبيق في كل من مكامن العراق سواء كانت كربوناتية أو سيليكاتية في حقول النفط والغاز الأخرى.

1. Introduction:

To maximize recovery and optimize output, reservoir engineers need accurate reservoir permeability forecasts. Numerous techniques, such as empirical correlations, analytical models, and machine learning algorithms, are extensively applied to predict permeability. [1] used the group method of data handling (GMDH) and gene expression programming (GEP) algorithms to model permeability based on petrophysical properties. Various studies have combined well log and core data to provide reservoir quality and productivity indicators. [2] conducted a lithofacies and permeability study comparing the performance of support vector machine (SVM), back propagation neural network (BPNN), and general regression neural network (GRNN) algorithms. [3] used artificial neural networks (ANN) and XGBoost for permeability prediction. However, to achieve optimal performance, it required hyperparameter tuning which is critical to achieving accurate facies classification and permeability prediction, [4] applied five algorithms with optimization tools including artificial neural networks (ANN), fuzzy decision tree (FDT), and least squared support vector machine (LSSVM).

In this study, an integrated workflow is developed for permeability prediction of all uncored intervals of the Mishrif formation, applying various configurations of the tree-based, extreme gradient boosting (XGboost) algorithm. The XGBoost hyperparameters of the algorithms are tuned with random search and Bayesian optimization algorithms. Three

Open Access
Vol. 15, No. 1, March 2025, pp. 43-61

P- ISSN: 2220-5381
E- ISSN: 2710-1096

normalization methods were applied to the permeability variable of the studied dataset: log transformation, Box-Cox, and normal score transformation. Normalization was conducted to transform the target variable of the training subset only into a normal distribution to improve the permeability prediction performance of the models. The prediction performance was evaluated using the root mean square error (RMSE) and adjusted R^2 metrics. The results were recorded and compared for each normalization method and hyperparameter optimization algorithm applied.

1.1. Reservoir Description

The Mishrif Formation is the largest, oil-productive, carbonate reservoir in Iraq. It accounts for more than 30% of the country's confirmed oil reserves in 32 different structures. The Mishrif carbonates vary in age from Cenomanian-Early Turonian era [5]. The Khasib Formation, which is composed of fine-grained marls and gray/green shales, sits on top of the Mishrif. In certain places, the Rumaila Formation's marls and limestones underlie the Mishrif. Based on petrophysical characteristics derived from log analysis, the Mishrif can be divided into five stratigraphic units MA (youngest/uppermost), MB1, MB2.1, MB2.3, and MC1 (oldest/lowermost). The units MA, MB2.1, and MC1 are responsible for most oil production from the Mishrif Formation [6]. Reservoir heterogeneity manifests itself as substantial variations in porosity, water saturation, and fluid volumes both vertically and horizontally through the stratigraphic units [7]. Figure (1) illustrates the stratigraphic column of Majnoon Field with the stratigraphic units of the Mishrif Formation distinguished.

1.2. Reservoir depositional environment and sedimentary facies

Reservoir lithofacies are typically categorized using rock-core data to identify whether or not they are capable of producing oil and gas based on measures of fluid saturation, porosity, and permeability. Such information is also used to identify barriers and non-productive zones within reservoirs [8]. The Mishrif Formation's type section has been characterized using such data, revealing that it consists of a heterogeneous sequence of organic detrital limestone, algal beds, rudist, and coral-reef limestone, with limonitic freshwater limestone as the capstone [7]. Four facies can be distinguished: (1) shallow open marine, (2) shoal/reef, (3) back-shoal/reef, and (4) restricted marine. Each of these main lithofacies are comprised of several microfacies, with some gradation between then. Figure (2) provides a diagrammatic illustration of the depositional

Open Access
Vol. 15, No. 1, March 2025, pp. 43-61

P- ISSN: 2220-5381
E- ISSN: 2710-1096

environment and the juxtaposition of its lithofacies based on the interpretation of multiple wellbore cores [8].

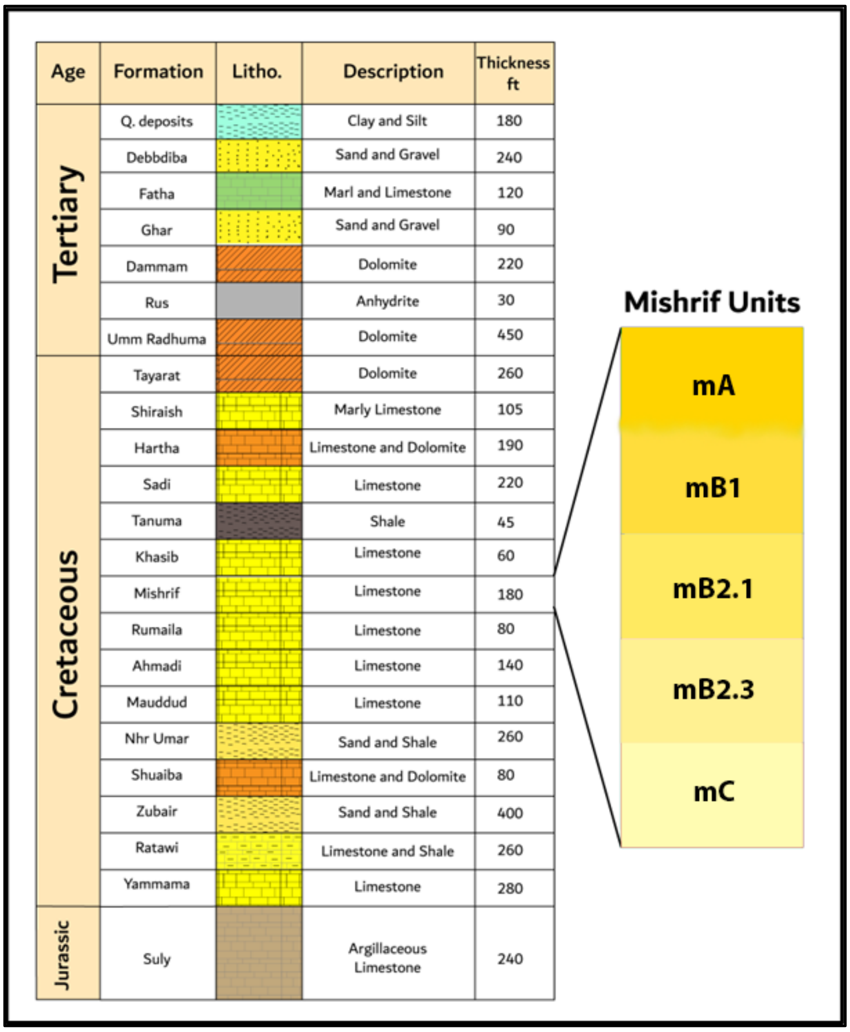

Fig. (1): The stratigraphic column of the Majnoon oil field distinguishing the Mishrif reservoir units. [9].

Open Access
Vol. 15, No. 1, March 2025, pp. 43-61

P- ISSN: 2220-5381
E- ISSN: 2710-1096

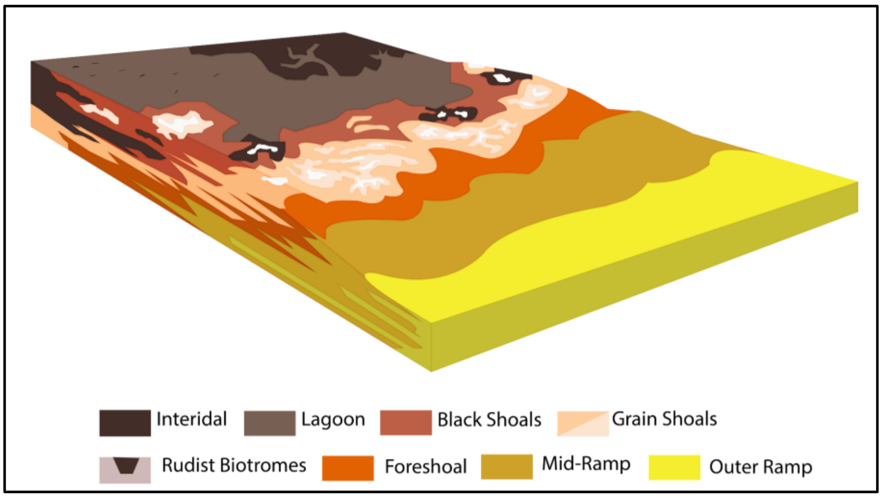

Fig. (2): Diagrammatic illustration of the depositional environments and associated lithofacies based on the interpretation of cored sections recovered from the Mishrif Formation, [8].

1.3. Data Description and Analysis

The well-log data used in this study was recorded in nine wellbores and compiles gamma ray (GR), spontaneous potential (SP), caliper (CALI), deep resistivity (RES DEP), shallow resistivity (RES SLW), neutron (NEU), bulk density (DEN), sonic transient time (DT) and photoelectric factor (PEF) log curves. Data from computer-processed interpretations (CPI) of these data also provide porosity and water saturation data for the studied intervals. In addition, the dataset includes laboratory-measured permeability and porosity values from the available well cores. The well-log data was processed with a k-means clustering algorithm to classify the depth section studied into four lithofacies groups (F1 to F4). Figure (3) shows the visualization of the datasets (Well Logs and Core). Tables (1) show a statistical description for the data.

Table (1): Statistical summary of the well-log and core data variables in the studied dataset. The dataset is comprised of 399 recorded well log and core data points ranging in depth from 2720m to 3030 m.

feature	mean	std	min	25%	50%	75%	max
DT	66.78	6.57	48.51	62.59	65.76	69.75	88.64
GR	37.40	10.70	19.56	29.12	36.01	44.65	83.70
NEU	0.10	0.03	0.02	0.07	0.09	0.11	0.22
DEN	2.49	0.08	2.20	2.44	2.51	2.55	2.67
PEF	5.10	0.15	4.29	5.04	5.12	5.20	5.45
RES_DEP	35.76	63.07	5.74	11.84	20.15	33.96	858.78

Journal of Petroleum Research and Studies

Open Access
Vol. 15, No. 1, March 2025, pp. 43-61

P- ISSN: 2220-5381
E- ISSN: 2710-1096

RES_SLW	22.86	35.41	3.76	8.90	13.88	21.91	462.17
SP	23.01	3.21	14.00	20.77	22.76	25.27	32.38
PHIT	0.11	0.04	0.02	0.08	0.10	0.13	0.26
SW	0.37	0.29	0.02	0.16	0.27	0.49	1.00
CORE_PORE	0.13	0.03	0.06	0.10	0.12	0.15	0.22
CORE_PERM	17.57	43.09	0.01	0.23	0.98	8.89	285.50

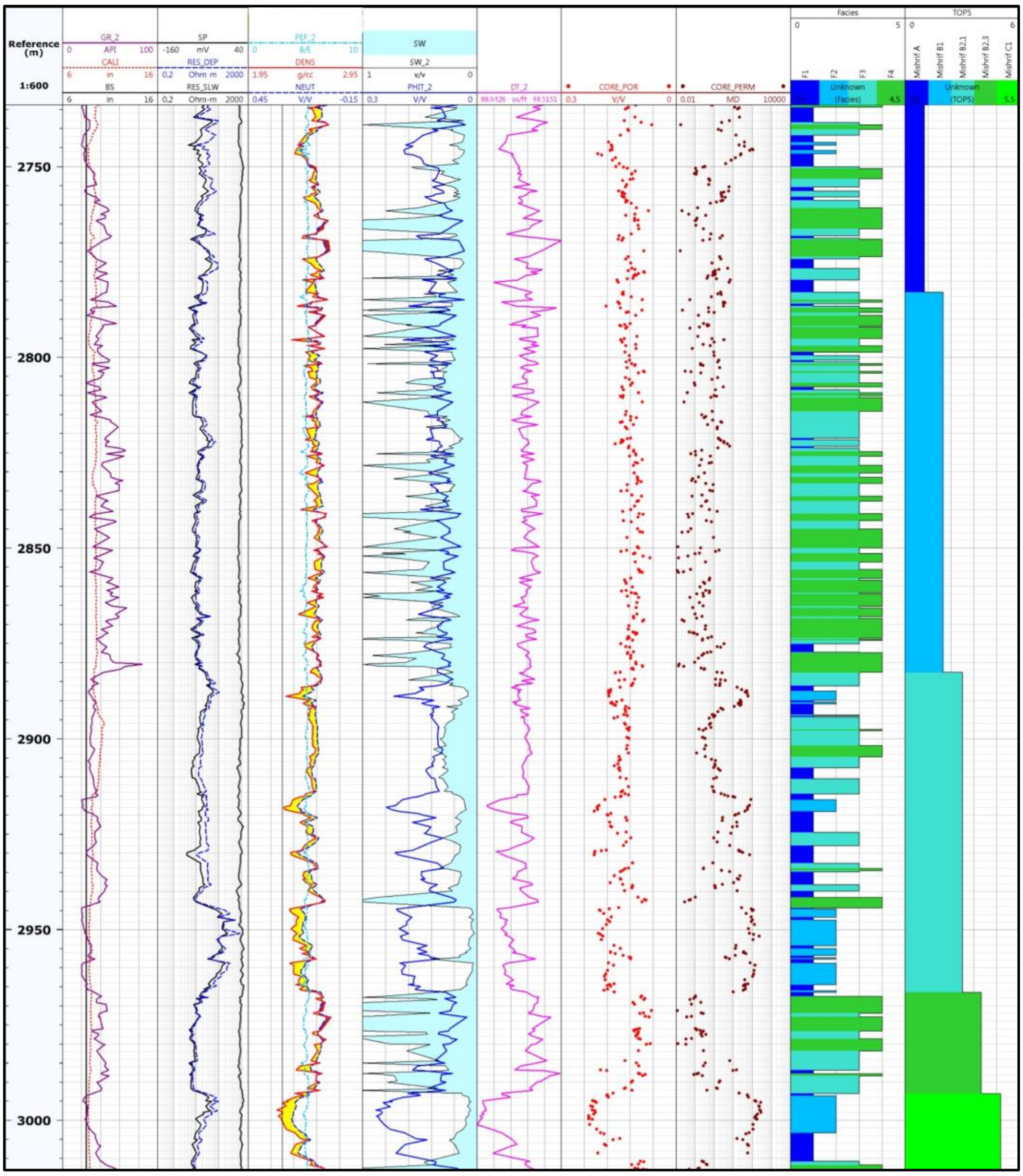

Fig. (3): Visual representation of the available Mishrif reservoir well logs and core data for the studied dataset.

Open Access
Vol. 15, No. 1, March 2025, pp. 43-61

P- ISSN: 2220-5381
E- ISSN: 2710-1096

2. Material and methods

Figure (4) describes the workflow components involved in predicting permeability for the compiled Mishrif reservoir section with different configurations of the tree-based XGBoost model.

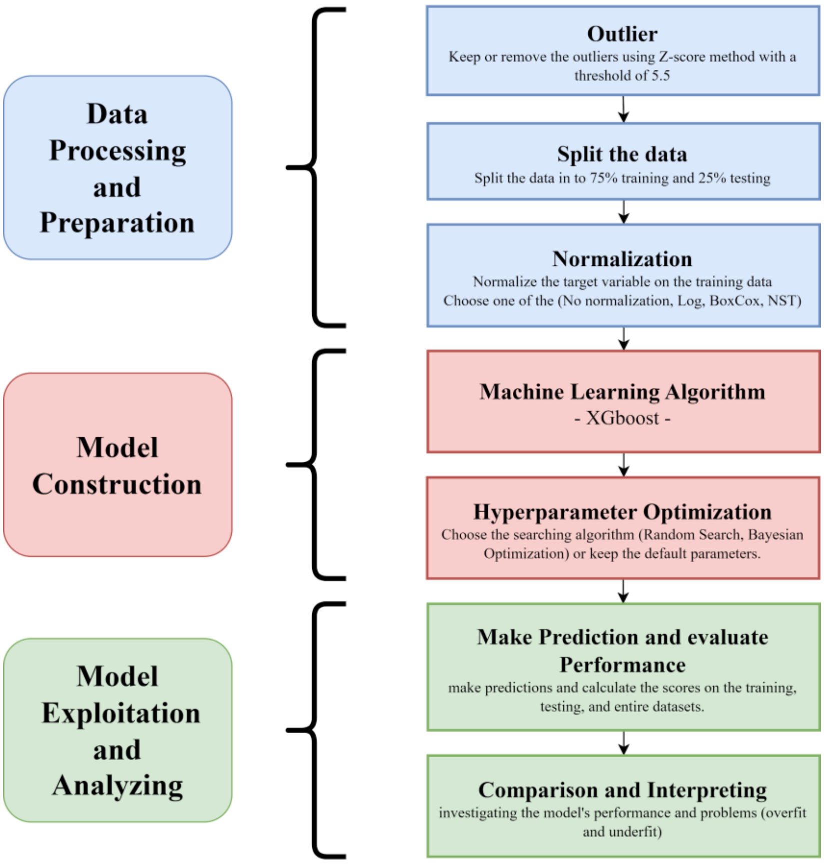

Fig. (4): Workflow diagram for conducting permeability predictions for the studied dataset involving XGBoost hyperparameter optimization and target variable normalization.

2.1. Data processing and preparation

Prior to applying the XGBoost prediction model, the quality of the compiled data was thoroughly checked for outliers and errors. The raw recorded well logs and the core laboratory measurements were subjected to different methods of preprocessing, as described in the following subsections.

Open Access
Vol. 15, No. 1, March 2025, pp. 43-61

P- ISSN: 2220-5381
E- ISSN: 2710-1096

2.1.1. Outlier detection

An outlying data point is one that has values that do not fall within the main data range defined by other data points. The Z-score method is used to determine the number of standard deviations a data point deviates from the mean of a normally distributed dataset. A data point is considered an outlier if its Z-score is above a certain threshold value, typically greater than 3 [12]. The Z-score is calculated using Eq. (1):

$$Z = (X - \mu)/\sigma \qquad (1)$$

Where X is a specific data point value, μ is the mean value of the entire dataset, and σ is the standard deviation of the entire dataset.

The Z-score method, with a threshold value of 5.5 was utilized to detect and eliminate outliers. This threshold value was adopted after trial-and-error testing of permeability prediction accuracy using different threshold values. Figure (5) distinguishes the data points of the compiled Mishrif dataset that are considered as outliers by the Z-score method.

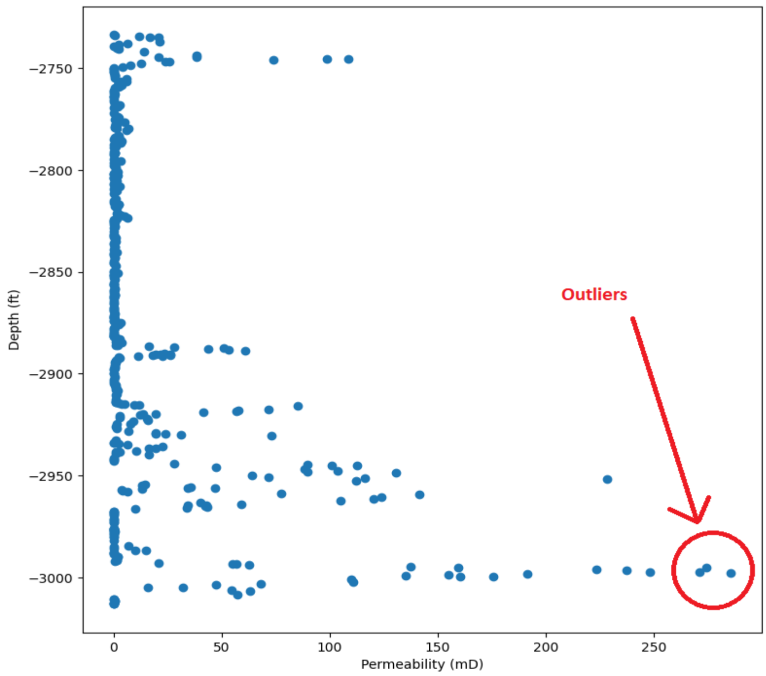

Fig. (5): Identified outlying data points in the compiled Mishrif dataset based on a Z -score threshold of 5.5 standard deviations.

Open Access
Vol. 15, No. 1, March 2025, pp. 43-61

P- ISSN: 2220-5381
E- ISSN: 2710-1096

2.1.2. Data Splitting

Based on trial-and-error analysis, the data records of the compiled dataset were split randomly into two subsets: 75% for XGBoost model training, and 25% for testing the trained model.

2.1.3. Normalization

Some machine learning algorithms and statistical techniques assume a normal (Gaussian) distribution of the target variable being predicted. If this is not the case, a transformation of the data can be applied to impose a normal distribution on the target variable distribution [11]. There are several statistical tests available for checking whether a variable is normally distributed or not. The D'Agostino and Pearson's test combines skew and kurtosis measurements of a variable distribution to provide a normality test. This test calculates the probability (p) of the null hypothesis that a sample comes from a normal distribution.

Three normalization transformations were used:

- Log Transformation

$$\hat{y} = log_{10}\, y \tag{2}$$

- Box-Cox Transformation

$$\hat{y} = \left(y^{\lambda} - 1\right)/\lambda \tag{3}$$

Where λ is a transformation parameter, the value of which can be set to maximize the p value of the normality test.

- Normal score transformation (NST)

The NST method transforms a dataset to a normal distribution by applying Eq. (4) [12].

$$\hat{y} = F^{-1}\left(\frac{rank(y) - 1/2}{n}\right) \tag{4}$$

Where n is the number of observations (samples), the $rank$ component is a function that assigns each element in a distribution a corresponding index position number in the sorted $[y_1, y_2, \ldots, y_n]$ list of distribution values, and F represents the cumulative density function (CDF) of a Gaussian distribution as expressed in Eq.(5).

$$F(x) = \int_{-inf}^{x} \frac{e^{-t^2/2}}{\sqrt{2\pi}}\, dt \tag{5}$$

Open Access
Vol. 15, No. 1, March 2025, pp. 43-61

P- ISSN: 2220-5381
E- ISSN: 2710-1096

Each of these transformations is bijective (i.e. both injective (one-to-one) and surjective (onto)), Figure (6) displays the unnormalized and normalized permeability distribution using the three transformation techniques described.

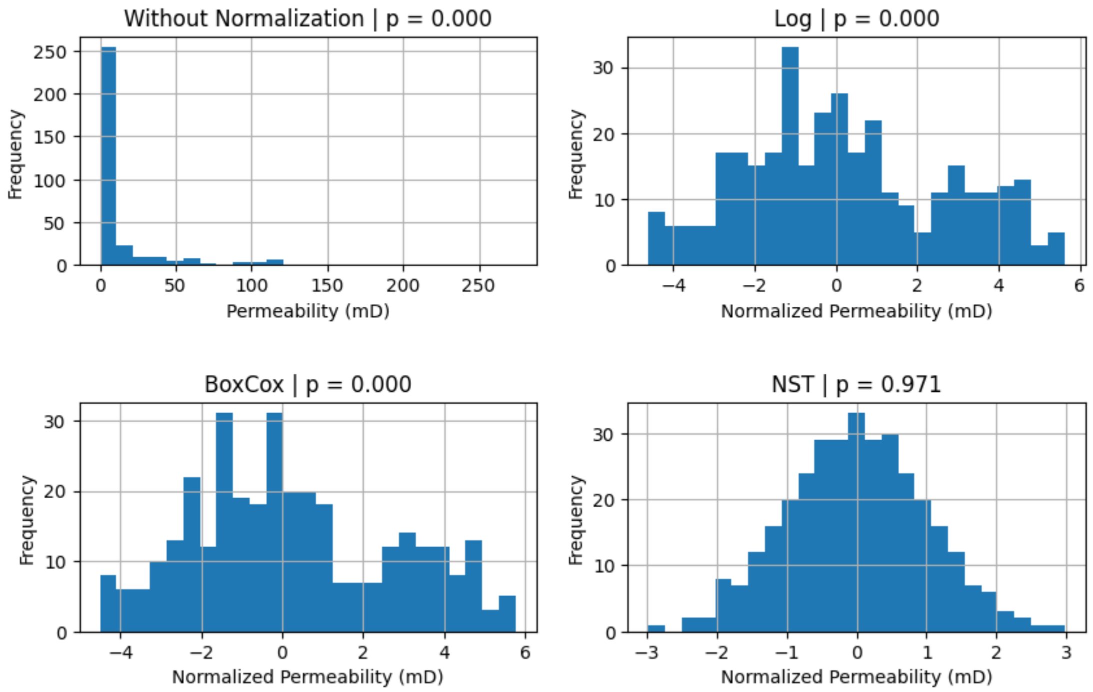

Fig. (6): Histograms of the permeability distribution for the training subset before and after applying normalization, each displaying the p-value of the normality test. The three normalization algorithms applied are Log, Box-Cox, and NST.

2.2. Model construction

2.2.1. Extreme gradient boosting (XGboost)

[13] developed the XGBoost algorithm, an effective, flexible and fast to execute machine learning algorithm for building predictive models. Experimental results show that XGBoost outperforms other tree boosting methods on a variety of benchmarks and real-world datasets.

2.2.2. Cross Validation

Cross-validation (CV) is used in this study to evaluate the permeability prediction performance of the XGBoost models with different combinations of hyperparameters on the training data subset. Such analysis provides a useful estimate of the ability of the trained models to generalize to new data and avoid overfitting and selection bias issues [11].

K-fold CV divides the data into K sets of equal sizes; for a total of K iterations. In the first iteration, the model is trained on k-1 folds and tested on the combined remaining data. This

Open Access
Vol. 15, No. 1, March 2025, pp. 43-61

P- ISSN: 2220-5381
E- ISSN: 2710-1096

process is repeated K times, with a different divided set of data used for training the model each iteration, and then test it using the remaining data not used to train the model. The target variable prediction performance scores from each of the K iterations are then averaged to obtain a single estimate of the model's performance [10].

2.2.3. Hyperparameters Optimization

Two optimization algorithms were implemented to search for the optimum set of hyperparameters:

- **Random search**

Many models are trained and evaluated with cross validation in each iteration of the random search algorithm. Each model is built with a random set of hyperparameters that are sampled from a predefined parameter space. The hyperparameters of the model configuration that generates the best target variable predictions are considered the optimal solution [11].

One of the advantages of this algorithm is that it takes significantly less time to execute than the grid search algorithm, where the search goes through the entire parameter space more systematically.

- **Bayesian Optimization**

This probabilistic model maps the hyperparameters to a measure of the quality of the resulting model. The probabilistic model is then used to guide the search for the best set of hyperparameters by sampling candidate hyperparameters from a probability distribution and evaluating their prediction performance [11].

A key advantage of Bayesian optimization is that it allows the algorithm to update its probability distribution based on the results of each iteration. This approach balances the exploration of the available hyperparameter space with a gradually increased focus on the exploitation of regions of that space that provide the most effective model. Table (2) displays the range of each XGBoost hyperparameters evaluated by the random search and Bayesian optimization algorithms.

Open Access
Vol. 15, No. 1, March 2025, pp. 43-61

P- ISSN: 2220-5381
E- ISSN: 2710-1096

Table (2): XGBoost Hyperparameter ranges explored by the optimization process

Parameter	Range
learning rate	0.001 - 1
min child weight	1 - 10
num parallel tree	1 - 20
num estimators	50 - 2000
subsample	0.1 - 1
max depth	2 - 10
gamma	0 - 2
colsample bytree	0.1 - 1
reg alpha	0 - 1
reg lambda	0 - 2

2.3. Model evaluation

Many models were built, each with its own combination of data preprocessing procedures and hyperparameter selection schemes. The predictions generated by each model on the training and testing subsets, and the entire dataset were evaluated with the two metrics (RMSE and adjusted R^2; Table 3).

Table :(3)Prediction performance metrics used to evaluate permeability predictions of the studied dataset

metric	formula	Name, Range, and Interpretation
RMSE	$\sqrt{\frac{1}{n}\sum_{i+1}^{n}(y_i - p_i)^2}$	Root mean squared error (with dimensions) [0, inf]. Small values indicate accurate predictions
Adjusted R^2	$1-(1-R^2)\frac{(n-1)}{(n-k-1)}$	Adjusted R-squared modified for the number of regressors in the model. less than or equal to R^2

Note to Table 3: yi is the number of observations, pi is the number of predictions, n is the number of observations, and k is the number of independent variables.

A comparison between the prediction performance of all models was conducted to identify the best performing model most able to mitigate overfitting and data selection bias. That involved a deep investigation of the most effective XGBoost model configuration and dataset pre-

Open Access
Vol. 15, No. 1, March 2025, pp. 43-61

P- ISSN: 2220-5381
E- ISSN: 2710-1096

processing, considering data quality control processes, normalization, search algorithms, and parameter space.

Table (4) lists the Python program libraries used to conduct this study.

Table (4): Programming libraries used in this study

Libraries	Description
pandas	Data analysis and manipulation tool.
numpy	A large collection of high-level mathematical functions that operate on multi-dimensional arrays.
matplotlib	A comprehensive library for creating static, animated, and interactive visualizations
sklearn	Simple and efficient tools for predictive data analysis
scipy	Optimization and data interpolation tools
hyperopt	Distributed Asynchronous Hyper-parameter Optimization
xgboost	Optimized and distributed gradient boosting library. implements machine learning algorithms under the Gradient Boosting framework

3. Results and discussion

The following subsections describe the permeability prediction results for the studied dataset of the three steps in the XGBoost model development.

3.1. The prediction outcomes of hyperparameter optimization

Initially, the XGBoost algorithm was evaluated with default and optimized hyperparameters, where optimization was performed using random search and Bayesian optimization algorithms. Table (5) compares the permeability prediction results, identifying that the optimization procedure employing random search algorithm generated the best results. Nonetheless, the performance resulting from this step has not enhanced the prediction performance significantly compared with the default model. The optimized XGBoost trained model overfits the testing subset in each case.

Open Access
Vol. 15, No. 1, March 2025, pp. 43-61

P- ISSN: 2220-5381
E- ISSN: 2710-1096

Table (5): XGBoost permeability prediction results with default and tuned hyperparameters algorithms without normalization and outlier detection.

Hyperparameter	Adjusted R^2		RMSE	
	Train	Test	Train	Test
Default	1	0.69	0.01	20.64
Random Search	1	0.73	1.35	19.22
Bayesian Random	1	0.69	1.14	20.65

Note to Table 5: An adjusted R^2 values of 1 for the training subset indicates perfect predictions, but the much lower adjusted R^2 values for testing subset are an indication that these models are overfitting the dataset.

3.2. The prediction outcomes of data normalization

In this step, three normalization methods were applied, resulting in nine models. The permeability prediction results reveal that the XGBoost models configured with default hyperparameters continued to overfit the dataset. Table (6) compares the permeability prediction results, identifying that those models with optimized hyperparameters applied to normalized training subsets achieved predictions with reduced overfitting. Nonetheless, the prediction performance of this group of models still results in relatively high errors.

Table (6): XGBoost permeability prediction results applied to training subsets normalized with three transformation techniques: log transformation, Box-Cox and NST.

Hyperparameter	Normalization	Adjusted R^2		RMSE	
		train	test	train	test
Default	Log	1.00	0.61	0.07	23.59
	Box-Cox	1.00	0.60	0.07	23.77
	NST	1.00	0.74	0.07	18.99
Bayesian Random	Log	0.76	0.52	13.56	26.38
	Box-Cox	0.75	0.60	13.82	24.03
	NST	0.85	0.61	10.39	23.50
Random Search	Log	0.95	0.67	6.24	21.35

Open Access
Vol. 15, No. 1, March 2025, pp. 43-61

P- ISSN: 2220-5381
E- ISSN: 2710-1096

	Box-Cox	0.95	0.69	6.22	20.73
	NST	0.91	0.72	8.19	19.61

3.3. The prediction outcomes of outlier detection and removal

In this step, outlier detection was employed with the z-score method, resulting in twelve additional models. The XGBoost model's permeability prediction performance with the studied dataset is further enhanced following the removal of detected outlying data records from the training dataset. Random search optimization of hyperparameters and Box-Cox training subset normalization outperform all the other model configurations in terms of both prediction performance and reducing the model's overfitting tendencies. Table (7) shows the permeability prediction results following the removal of outlying data records removal with different normalization and XGBoost optimization techniques applied.

Table (7): XGBoost permeability prediction results applied to the training datasets following the removal of outlying data records

Hyperparameters	Normalization	Adjusted R^2		RMSE	
		Train	Test	Train	Test
Defaults Parameters	None	1	0.67	0.01	15.07
	Log	1	0.79	0.09	12.01
	Box-Cox	1	0.75	0.1	13.07
	NST	1	0.72	0.12	13.73
Random Search	None	1	0.83	0.29	10.69
	Box-Cox	0.93	0.86	6.9	9.78
	Log	0.87	0.69	9.25	14.53
	NST	0.85	0.63	10.01	16.33
Bayesian Optimization	None	0.85	0.78	10.15	12.12
	Box-Cox	0.88	0.78	9.25	12.24
	Log	0.85	0.83	10.11	10.82
	NST	0.82	0.64	11.07	16.03

Figure (7) compares the observed and predicted permeability values of the studied dataset applying the best performing XGBoost model.

Open Access
Vol. 15, No. 1, March 2025, pp. 43-61

P- ISSN: 2220-5381
E- ISSN: 2710-1096

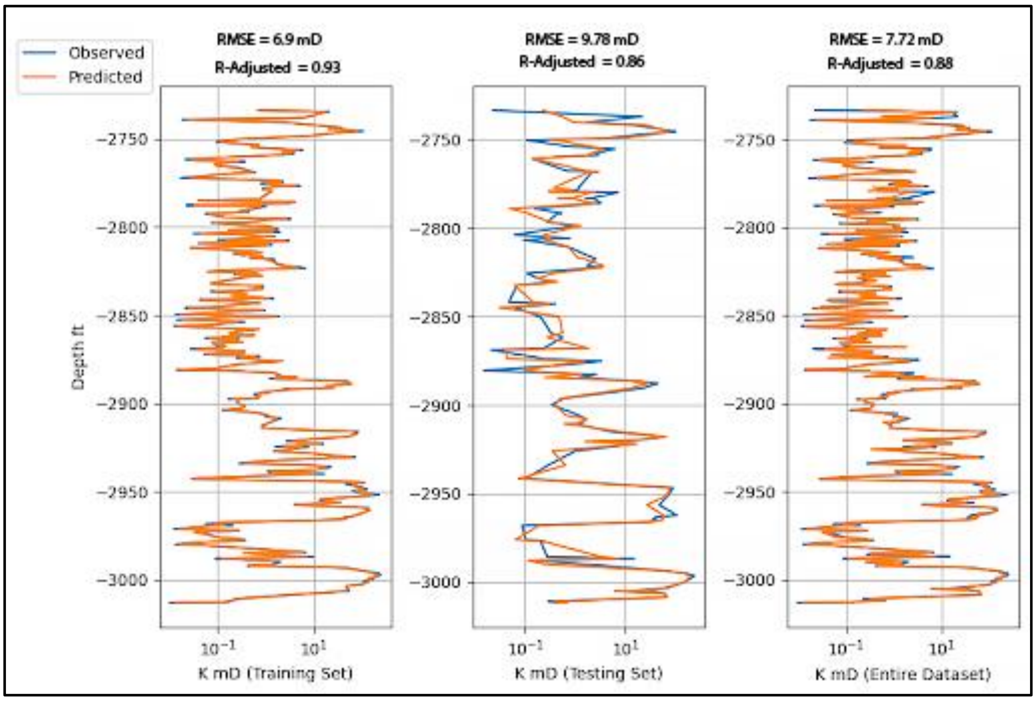

Fig. (7): The best XGBoost model, involving random search optimization and Box-Cox data normalization after the removal of outlying data records from the training subset. Results are displayed for the training and testing subsets, and the entire dataset.

4. Conclusions

Multiple tree-based machine-learning models using the extreme gradient boosting model (XGBoost) are configured and developed to provide reliable and robust permeability predictions for the Mishrif carbonate reservoir using well-log data variables calibrated with core data. Based on the prediction results obtained, the following conclusion can be drawn:

1. The permeability distribution of the core and well-log dataset compiled for the Mishrif reservoir of the Majnoon oil field is highly asymmetrical and positively skewed.
2. The XGBoost models applied to this raw dataset are prone to severe overfitting, unless the target (permeability) variable in the studied dataset is transformed to a more symmetrical distribution.
3. Comparisons of the twenty-four configured models evaluated reveal that the XGBoost model applied to the training subset adjusted with the Box-Cox normalization algorithm, and with its hyperparameter tuned using the random search algorithm outperforms other models. That model achieves permeability predictions for the studied dataset of RMSE =

Open Access
Vol. 15, No. 1, March 2025, pp. 43-61

P- ISSN: 2220-5381
E- ISSN: 2710-1096

6.9mD and 9.782mD, and adjusted R^2 = 0.9294 and 0.8554, for training and testing subsets, respectively.

4. The methodology proposed in this study should be applicable to other reservoirs which display highly skewed permeability distributions.

Acknowledgment

The authors would like to thank the Department of Oil and Gas Engineering of Basrah University for Oil and Gas for their support to participate in this conference.

Open Access
Vol. 15, No. 1, March 2025, pp. 43-61

P- ISSN: 2220-5381
E- ISSN: 2710-1096

References

[1] M. Mahdaviara, A. Rostami, and K. Shahbazi, "State-of-the-art modeling permeability of the heterogeneous carbonate oil reservoirs using robust computational approaches", *Fuel*, 268, p. 117389, 2020. https://doi.org/10.1016/j.fuel.2020.117389

[2] A. Al-Anazi, I. D. Gates, "A support vector machine algorithm to classify lithofacies and model permeability in heterogeneous reservoirs", *Engineering Geology*, vol. 114, no. (3–4), pp. 267–277, 2010. https://doi.org/10.1016/j.enggeo.2010.05.005

[3] R. Qalandari, R. Zhong, C. Salehi, N. Chand, R. L. Johnson, G. Vazquez, J. Mclean-Hodgson, and J. Zimmerman, "Estimation of rock permeability scores using machine learning methods", Paper presented at *the SPE Asia Pacific Oil & Gas Conference and Exhibition*, Adelaide, Australia, October 2022. https://doi.org/10.2118/210711-MS

[4] M. A. Ahmadi, and Z. Chen, "Comparison of machine learning methods for estimating permeability and porosity of oil reservoirs via Petro-physical logs", *Petroleum*, vol. 5, no. 3, pp. 271–284, 2019. https://doi.org/10.1016/j.petlm.2018.06.002

[5] T. A. Mahdi, A. A. M. Aqrawi, A. D. Horbury, and G. H. Sherwani "Sedimentological characterization of the mid-Cretaceous Mishrif Reservoir in southern Mesopotamian Basin, Iraq", *GeoArabia*, vol. 18, no. 1, pp. 139–174, 2013. https://doi.org/10.2113/geoarabia1801139

[6] W. J. Al-Mudhafar, M. A. Abbas, and D. A. Wood, "Performance evaluation of boosting machine learning algorithms for lithofacies classification in heterogeneous carbonate reservoirs", *Marine and Petroleum Geology*, vol. 145, p. 105886, 2022. https://doi.org/10.1016/j.marpetgeo.2022.105886

[7] L. khudhur Abbas, T. A. Mahdi, "Reservoir units of Mishrif Formation in Majnoon oil field, southern Iraq", *Iraqi Journal of Science*, vol. 60, no. 12, pp. 2656–2663, 2019. https://doi.org/10.24996/ijs.2019.60.12.15

[8] W. J. Al-Mudhafar, "Integrating machine learning and data analytics for geostatistical characterization of Clastic Reservoirs", *Journal of Petroleum Science and Engineering*, vol. 195, p. 107837, 2020. https://doi.org/10.1016/j.petrol.2020.107837

[9] M. A. Abbas, and E. M. Al Lawe, "Clustering Analysis and Flow Zone Indicator for Electrofacies Characterization in the Upper Shale Member in Luhais Oil Field, Southern Iraq", *Abu Dhabi International Petroleum Exhibition & Conference*, Abu Dhabi, UA, 2019. https://doi.org/10.2118/197906-MS

[10] M. Sarmad, "Robust data analysis for factorial experimental designs: Improved methods and software", *Durham theses, Durham University*, 2006.

[11] H. Belyadi, and A. Haghighat, "Machine Learning Guide for oil and gas using python",

Open Access
Vol. 15, No. 1, March 2025, pp. 43-61

P- ISSN: 2220-5381
E- ISSN: 2710-1096

2021. https://doi.org/10.1016/C2019-0-03617-5

[12] D. C. Howell, "Statistical Methods for Psychology", Seventh Edition, *Wadsworth Cengage Learning*, Belmont, 2010.

[13] T. Chen, T. and C. Guestrin, "XGBoost: A Scalable Tree Boosting System", *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. https://doi.org/10.1145/2939672.2939785

[14] W. J. Al-Mudhafar, "Incorporation of bootstrapping and cross-validation for efficient multivariate facies and Petrophysical Modeling", Paper presented at the *SPE Low Perm Symposium*, Denver, Colorado, USA, May 2016. https://doi.org/10.2118/180277-MS