International Science and Technology Journal
المجلة الدولية للعلوم والتقنية
Volume 38 العدد
Part 2 المجلد
المجلة الدولية للعلوم والتقنية
International Science and Technology Journal
ISTJ

Machine Learning-Based Ensemble Models for Prediction of Reservoir Porosity

http://www.doi.org/10.62341/istj-vol38-2- irego42

Received	**2026/06/01**	**تم استلام الورقة العلمية في**
Accepted	**2026/06/19**	**تم قبول الورقة العلمية في**
Published	**2026/06/20**	**تم نشر الورقة العلمية في**

Machine Learning-Based Ensemble Models for Prediction of Reservoir Porosity

Reema O. Khalleefah[1]

1.Al-Furat University, Ajdabiya, Libya
Email: remaomar17@gmail.com

Abstract

Reservoir porosity determines how much fluid formation can store, making its accurate estimation a key element of petrophysical evaluation, reservoir characterization and geological studies. Core analysis provides reliable readings, but covers limited intervals and time-consuming, expensive. Log-based equations derived from well-logging interpretation introduce uncertainties that affect prediction reliability. Four ensemble machine learning models were used here to enhance porosity prediction, Decision Trees (DT), Random Forests (RF), Gradient Boosting (GB) and Extreme Gradient Boosting (XGB). Input dataset used 734 data point from seven wells in a Libyan field while a separate well with 202 data point kept as a blind well to evaluate model generalization. Bulk density (RHOB), gamma ray (GR), compressional travel time (DT), and neutron porosity (CNL) chosen as input features because each shows established relationship with porosity. Statistical metrics including correlation coefficients (R^2) and root-mean squared error (RMSE) were used to assess the model. The results indicate that ensemble models provide a robust and efficient alternative to conventional porosity estimation methods, offering improved predictive accuracy and reliability. This study highlights the potential of machine learning in reservoir characterization, contributing to more data-driven decision-making in petroleum engineering applications.

International Science and Technology Journal
المجلة الدولية للعلوم والتقنية
Volume 38 العدد
Part 2 المجلد
المجلة الدولية للعلوم والتقنية
International Science and Technology Journal
ISTJ

Machine Learning-Based Ensemble Models for Prediction of Reservoir Porosity

http://www.doi.org/10.62341/istj-vol38-2- irego42

This scientific manuscript was presented at the sessions of the International Renewable Energy, Gas, Oil and Climate Change Conference "iREGO" in the period of April 25-27, 2026. Tripoli - Libya

Keywords: Porosity Prediction, Ensemble models, Machine Learning, Well Logging Data.

نماذج التعلم الآلي التجميعية للتنبؤ بمسامية المكامن النفطية

ريما عمر خليفة
1. جامعة الفرات الأهلية
البريد الإلكتروني: remaomar17@gmail.com

ملخص

تُحدد المسامية قدرة المكمن على احتواء السوائل، وهو ما يجعل تقديرها بدقة ركيزة أساسية في التقييم البتروفيزيائي ووصف المكمن والدراسات الجيولوجية. يوفر تحليل العينات اللبية قراءات موثوقة ، إلا أنه يقتصر على فترات محددة من المكمن، إضافة إلى ارتفاع تكلفتها وطول الزمن اللازم لتنفيذها. أما المعادلات المستخدمة في تسجيلات الأبار تُدخل قدراً من عدم اليقين يحد من دقة التنبؤات. استُخدمت في هذه الدراسة أربعة نماذج من نماذج التعلم الآلي الجماعي وهي: أشجار القرار (DT)، والغابات العشوائية (RF)، والتعزيز التدريجي (GB) والتعزيز التدريجي المتطرف (XGB). اعتمدت الدراسة على مجموعة بيانات مكونة من 734 نقطة من سبعة آبار في حقل ليبي، في حين خصصت بئر مستقل يحتوي على 202 نقطة كبئر تحقيق (Blind well) لتقييم قدرة النموذج على التعميم. تم اعتماد أربعة تسجيلات كمدخلات الكثافة الكلية (RHOB)، أشعة غاما (GR)، زمن السفر الانضغاطي (DT)، والمسامية بالنيوترونات (CNL) كونها تظهر علاقة فيزيائية بين كل منها وقيم المسامية. تم استخدام مقياسين احصائيين رئيسيين هما معامل الإرتباط (R^2) والجذر التربيعي لمتوسط مربع الخطأ (RMSE) لتقييم أداء النموذج. تشير النتائج إلى أن نماذج التجميع توفر بديلاً قوياً وفعالاً مقارنة بأساليب تقدير المسامية التقليدية من حيث الدقة والموثوقية. وتسلط هذه الدراسة إمكانيات التعلم الآلي

Machine Learning-Based Ensemble Models for Prediction of Reservoir Porosity

http://www.doi.org/10.62341/istj-vol38-2- irego42

في توصيف المكامن، بما يُعزز اتخاذ قرارات أكثر اعتماداً على البيانات في تطبيقات هندسة البترول.

وقد تم عرض هذه الورقة العلمية في جلسات المؤتمر الدولي للطاقة المتجددة والنفط والغاز وتغير المناخ "أيريقو" في الفترة 25–27 ابريل 2026م. طرابلس – ليبيا

الكلمات المفتاحية :التنبؤ بالمسامية، النماذج التجميعية، التعلم الآلي، بيانات تسجيلات الابار.

1. Introduction

Porosity is a fundamental parameter in reservoir engineering, critical for evaluating reservoir capacity, estimating hydrocarbon reserves, and serving as a key input in reservoir simulation studies. Accurate estimation of porosity is crucial for determining the volume of hydrocarbons in place. Generally, it is defined as the storage capacity of a reservoir (Dandekar, 2013). Mathematically, porosity is the measure of total pore space volume (pore volume) divided by the total bulk volume of the rock (Alyafei, 2021).

Typically, this property cannot be directly available from petrophysical logs. Traditionally, porosity is determined through laboratory-based techniques, such as Routine Core Analysis (RCA), which provide direct and accurate measurements of reservoir properties but are costly and time-consuming (Erofeev et al., 2019). Although RCA gives useful information, it is not always possible to core every well in the petroleum field. That is because RCA is limited by excessive cost and time intensive (Alatefi et al., 2023).

As a result, coring is usually reserved for rare situations, thereby leaving a significant number of wells without direct measurements. Also, it can be determined from well logging such as density, neutron, sonic, and NMR logs. While logging-based methods are more cost-effective and time-efficient, they are often affected by uncertainties arising from environmental and tool-specific factors, making porosity estimation less reliable without advanced analytical techniques. To address this challenge, the standard oil fields used the traditional method of well logging tools to analyze petrophysical parameters. These methods are economical because

http://www.doi.org/10.62341/istj-vol38-2- irego42

the log data is collected during drilling and completion operations (Alatefi et al., 2023) . Consequently, there is a growing demand for cost-effective, time-efficient, and accurate methods to estimate reservoir porosity from well logs.

Machine learning (ML) has recently emerged as a powerful data-driven approach for modelling complex relationships in petrophysical datasets. ML, as defined by Tom Mitchell, is enhanced by computer programs automatically from experience by studying computer algorithms (Bhattacharya, 2021). Researchers have increasingly focused on utilizing machine learning algorithms such as Artificial Neural Networks (ANN) (Abdullah Al-Qahtani, n.d.; Bhatt & Helle, n.d.; Nyein & Ali Hamada, 2023), Decision Tree (DT)(Al-Fakih et al., 2023), Random Forest (RF) (Anifowose et al., 2023; Kumar et al., 2024; Zou et al., 2021), Gradient Boosting (GB), Extreme Gradient Boosting (XGB) (He et al., 2025), and Support Vector Machine (SVM) (Elkatatny et al., 2018) to model the complex, non-linear relationships between well log data and porosity values. For instance, Nourani et al. (2022) used machine learning techniques, including the Random Forest (RF), Multilayer Perceptron (MLP), and their optimized versions using Genetic Algorithms (GA-RF and GA-MLP) to predict porosity in chalk formation. The author found that this approach highlights the potential of combining advanced ML models with rapid elemental analysis techniques for efficient porosity estimation. Similarly, Sun et al. (2024) a CNN-Transformer model was proposed to enhance porosity prediction from well logs by capturing both spatial and sequential features. CNN effectively extracts local correlations, while the Transformer handles complex depth-based dependencies. evaluated against traditional ML approaches, demonstrating superior accuracy and generalization. This approach offers a novel and efficient tool for improving porosity estimation in complex geological settings. In another study, Mulashani et al. (2022) applied machine learning models included ANN, DT, and RF to predict porosity from advanced mud gas (AMG) data across multiple wells. Their results demonstrated the feasibility of using AMG data for real-time porosity estimation before wireline logging.

http://www.doi.org/10.62341/istj-vol38-2- irego42

This study highlights the potential of ML to extend AMG utility beyond traditional fluid typing and petrophysical correlations. These studies collectively demonstrate the effectiveness of ML techniques in predicting porosity, especially in the absence of core data. By leveraging ML algorithms, it becomes feasible to improve both the accuracy and reliability of porosity estimation, enabling more informed reservoir characterization in data-limited environments.

2. Methodology

During this research, four tree-based models were used, including Decision Trees (DT), Random Forests (RF), Gradient Boosting (GB), and Extreme Gradient Boosting (XGB).

2.1 Machine Learning Models

2.1.1 Decision Trees (DT)

A Decision Tree (DT) is a machine learning model constructed by a series of decisions based on variable values to choose one path or another (Rebala et al., 2019). It is used for classification and regression (Bhattacharya, 2021). A classification tree provides a categorial output (class, text, colour, etc.), whereas a regression is a decision tree that provides a numerical value (Erofeev et al., 2019). The algorithms create a tree structure, like a flowchart. Figure 1,illustrates the tree components, which are divided into three parts: decision node, branches, and leaf node. The DT begins with the root node, which is at the top of the tree and ends with multi-leaf nodes. It uses the specific input feature condition to partition the data into recursive segments. The primary goal of a Decision Tree (DT) is to break down a complex problem into more manageable components. This is achieved through a series of steps: splitting, stopping, and pruning. The process begins at the root node, where the training data is initially divided. Splitting continues through the internal nodes until certain stopping conditions are satisfied. To prevent overfitting and streamline the model, pruning techniques are applied. These techniques enhance the DT by eliminating redundant branches,

Machine Learning-Based Ensemble Models for Prediction of Reservoir Porosity

http://www.doi.org/10.62341/istj-vol38-2- irego42

resulting in a clearer and more accurate representation of the solution (Larestani et al., 2022).

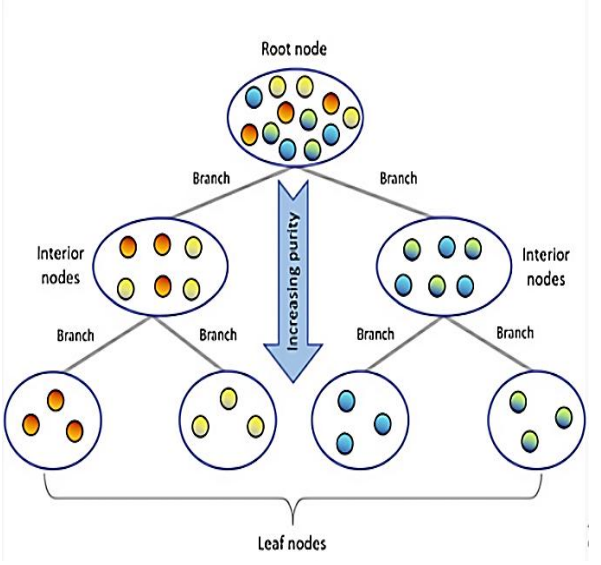

Figure1 :The concept of a decision tree. *(Bhattacharya, 2021)*

2.1.2 Random Forests (RF)

Random Forest (RF) is a supervised machine learning algorithm developed based on an ensemble of decision trees. This algorithm can be used for classification and regression problems (Bhattacharya, 2021; Pandey et al., 2020). Generally, RF algorithms create multiple decision trees using randomly selected subsets from the dataset, and the final prediction for the classification task is computed by majority voting, whereas for the regression task, the result is the average results from each decision tree (Awad Mariette & Khanna Rahul, 2015).

RF algorithms differ from DT algorithms, which start with creating multiple parallel decision trees to train and predict the dataset with bagging or bootstrap aggregation. The term beginning refers to the short form of bootstrap aggregation. In machine learning algorithms with poor performance or overfitting issues, it is useful to use bagging, which trains the dataset in parallel (Bhattacharya, 2021). During this process, the prediction for regression is provided

http://www.doi.org/10.62341/istj-vol38-2- irego42

through averaging while voting through a classification. This procedure minimizes the chances of overfitting and yields a more stable model. However, bootstrap aimed to reduce underfitting by training the model previously trained model with low performance. This approach provides the best model performance (Pandey et al., 2020). Figure 2, illustrates the basic concept of the Random Forest algorithm.

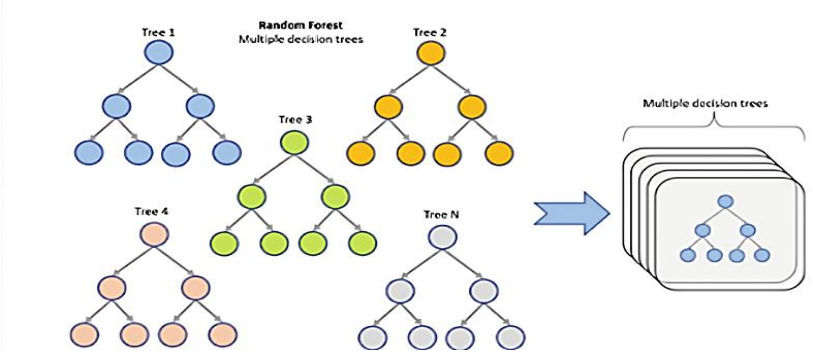

Figure 2: Basic concept of the Random Forest model (Bhattacharya,2021)

2.1.3 Gradient Boosting (GB)

This approach relies on the "boosting" technique, which builds an ensemble of weak learners typically decision trees. The boosting algorithm constructs trees in sequence, where each new tree focuses on correcting the errors made by the previous ones. This iterative process involves training each subsequent tree on the residuals (pseudo-errors) of the preceding model. Like other machine learning algorithms, Gradient Boosting minimizes a loss function, using gradient descent to reduce the error introduced by each new estimator. The final prediction is generated by combining the initial model with all following estimators, each weighted appropriately (Erofeev et al., 2019).

2.1.4 Extreme Gradient Boosting (XGB)

XGBoost, or Extreme Gradient Boosting, is an ensemble learning method that can be applied to both classification and regression

http://www.doi.org/10.62341/istj-vol38-2- irego42

tasks. In regression problems, it works by continuously building new regression trees, where each new tree is trained to fit the residual errors of the previous model. The final prediction is obtained by summing the outputs of all the individual trees (Kumar et al., 2024). XGBoost is an improved and scalable version of the gradient boosting method. It builds a strong model by gradually combining several weak models. First, it fits the model to the training data, then fits another model to the errors (residuals) of the first one to improve the results. This process of correcting errors continues until a stopping rule is reached. The final prediction is the sum of all the models' outputs. To reduce the risk of overfitting, XGBoost includes a regularization term in its objective function. Unlike standard gradient boosting, which uses loss values to guide tree splitting, XGBoost follows a depth-first method and removes unnecessary branches from the trees using a set maximum depth. It also speeds up tree building by using parallel computing: the inner part of the algorithm calculates tree details, and the outer part explores possible leaf nodes, making the process more efficient (Youcefi et al., 2024).

2.2 Methodology Using Machine Learning

The research methodology utilized Python programming to develop machine learning algorithms, leveraging several Python libraries, including Scikit-Learn (a machine learning library), NumPy (a numerical library), Pandas (a Data Frame library), and Matplotlib (a graphical and visual library), among others.

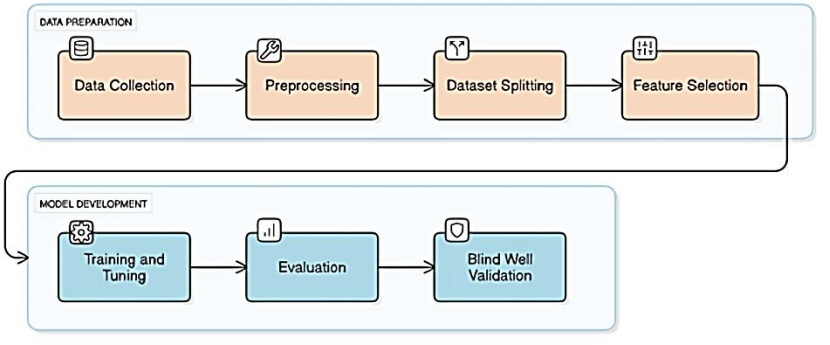

Figure 3:Research workflow (Prepared by the author).

http://www.doi.org/10.62341/istj-vol38-2- irego42

2.2.1 Dataset Description

This research focused on utilizing eight wells located in a mature oil field in the Sirte Basin. This oil field has enough RCA as well as open-hole logging data, which will facilitate achieving the research objectives. Hence, seven wells were used in this research as a training dataset, while one well (A-05) was used as a blind well to test the model accuracy. The research workflow is illustrated in Figure 3 started with data collection, which involved various well logging data that were available from the sandstone reservoir, as mentioned in Table 1 which includes caliper (CAL), deep resistivity (ILD), bulk density (RHOB), gamma ray (GR), medium resistivity (ILM), micro resistivity (MSFL), compensated neutron (CNL), sonic (DT), spontaneous potential (SP) and correlated core data (phi). However, due to limitations in core data for the selected areas, the label selection used was correlated to core data; these data were used as labels for machine learning algorithms.

2.2.2 Exploratory Data Analysis (EDA)

The data available in the real world is not available in a perfect format and is not prepared for use in machine learning (ML) models, as well as petrophysical data. Thus, the critical step before starting training ML algorithms is Exploratory Data Analysis (EDA), which helps in understanding and preparing data for ML. EDA involves understanding the relationship between variables, identifying inconsistencies, and verifying assumptions(Pandey et al., 2020). Therefore, during this step, the Pandas and NumPy libraries were used to summarize statistics and for visualization techniques. Python libraries such as matplotlib and Seaborn were used and detect the nature of the dataset and answer the questions about the available data. Table 1 illustrates the statistical summary of the dataset. The total number of data points for training data is about 734 points; all of these points were in the reservoir zone. The number of data points in the blind well is about 202 points, as shown in Table 2. Data visualization is a critical method in EDA that is used to understand the data distribution and identify the outliers. An outlier is a parameter that has a range outside a reasonable range. The boxplot was used to visualize the dataset, which lead to detecting the outliers using Eq (1), (2) and (3). Figure 4 shows the

http://www.doi.org/10.62341/istj-vol38-2- irego42

outlier for the training dataset. However, all outlier in the dataset is related to the geology. Hence, the removal of any of them affects the model's reality.

$$IQR = Q3 - Q1 \quad \text{Eq (1)}$$

$$Rmin = Q1 - 1.5 * IQR \quad \text{Eq (2)}$$

$$Rmax = Q3 + 1.5 * IQR \quad \text{Eq (3)}$$

Where: IQR = the interquartile range. Q3 = 75th percentile. Q1 = 25th percentile. Rmin = minimum outliers. Rmax = maximum outliers.

2.2.3 Data Preprocessing

Data preprocessing plays a critical role in ensuring the quality and reliability of the input data for machine learning models. That can confirm better results from machine learning models. Thus, the scikit learn library was used during this research to perform data preprocessing and machine learning models.

The label selection is a part of data preprocessing, which involves the variables that the model is trying to predict. Porosity is selected as the label for this research. Feature selection is a method to select the needed features used to build the model and contribute to predicting the output. The main advantage of feature selection is to reduce overfitting. However, feature selection has another advantage increases accuracy, where modelling accuracy increases with fewer false data. As well as, Shortens Training Time (STT), which allows algorithms can learn more quickly with less data (Brownlee, 2016). In this research, RHOB, GR, CNL, and DT were utilized as input for predicting reservoir porosity (Phi).

Table 1: statistics summary for the training dataset.

	CAL	ILD	RHOB	GR	ILM	MSFL	CNL	phi	DT	SP
count	734	734	734	734	734	734	734	734	734	734
mean	8.24	11.1	2.34	22.92	10.64	15.81	0.19	0.17	74.6	3.19
std	0.17	13.26	0.05	15.9	11.24	7.26	0.03	0.03	3.57	44.38
min	7.89	1.11	2.2	9.59	1.39	2.71	0.14	0.01	65.5	-88.07
max	9.24	101.61	2.59	148.62	69.04	61.53	0.45	0.25	88.4	71.25

Machine Learning-Based Ensemble Models for Prediction of Reservoir Porosity

http://www.doi.org/10.62341/istj-vol38-2- irego42

Table 2: statistics summary of the blind well.

	CAL	ILD	RHOB	GR	ILM	MSFL	CNL	phi	DT	SP
count	202	202	202	202	202	202	202	202	202	202
mean	8.15	21.92	2.34	19.92	20.18	14.92	0.19	0.17	73.34	49.94
std	0.11	17.57	0.06	9.53	15.85	4.09	0.02	0.03	2.42	7.16
min	8.04	3.1	2.25	10.09	3.05	7.43	0.15	0.04	67	40.69
max	8.72	57.66	2.56	59.59	54.94	29.78	0.29	0.22	81.8	79.75

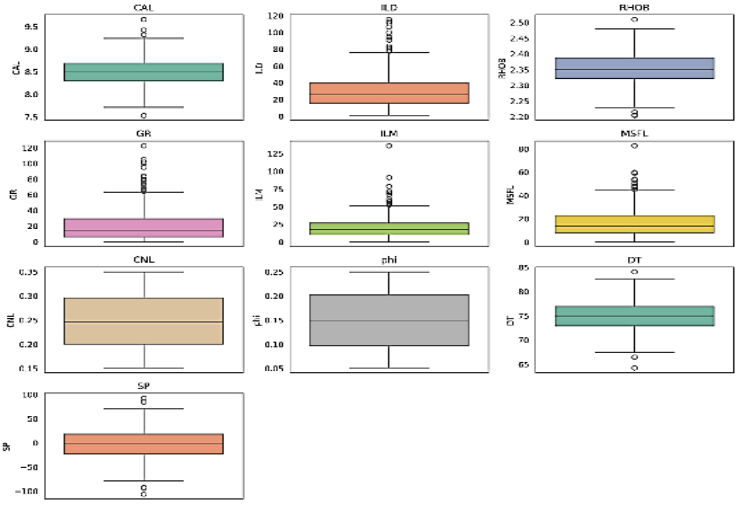

Figure 4:dataset boxplot (Prepared by the author).

2.2.4 Model Generation

In the present research, the algorithms were generated by using a Python script which used the Scikit-Learn library, which contained several ML models. For model selection until now, nobody has been able to select specific machine learning algorithms for specific problems. It can use different algorithms to compare the results and select the best model based on model accuracy. According to Grinsztajn et al. (2022), found that Tree-based models surpass deep learning methods in medium data size. Therefore, given the limited number of data points in this research, an ensemble learning approach was adopted, incorporating Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), and XGBoost (XGB) regression models to predict reservoir porosity and achieve the

http://www.doi.org/10.62341/istj-vol38-2- irego42

research objectives. Initially, the training dataset was split into two subsets: 75% used to train the model, while 25% for testing. This step was carried out using the Scikit-Learn library by importing the train-test-split function. Typically, the quality of the training and testing datasets can significantly influence model development, thereby affecting its performance. To ensure reliable validation, a blind well was employed in this research.

2.2.5 Evaluation Metric

Model evaluation is an important step in the machine learning workflow. Which uses quantitative metrics to evaluate the model performance to comprehend how well the model generalizes on an unseen dataset (Pandey et al. (2020). In this research, the model performance was evaluated using the correlation coefficient (R2) and root-mean squared error (RMSE), as shown in Eq (4) and (5). Typically, these evaluation metrics are commonly used to ensure the model's accuracy and provide insights about the model's predictive capabilities.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad \text{Eq (4)}$$

$$RSME = \sqrt{\frac{\sum_1^n (\hat{y}_i - y_i)^2}{n}} \quad \text{Eq (5)}$$

Where: SS_{res} is the sum of squares residual. SS_{tot} is the total sum of squares. $\hat{y}_i$: is the predicted value by the regression model, n: is the number of data points.

3. Results and Discussion

In this section, the performance of the base models was assessed. The prediction was generated using a validation dataset (a blind well) that the models had not seen during the training process. To evaluate the model's performance, two main statistical metrics were utilized on both the training and validation datasets: correlation coefficient (R^2) and root mean squared error (RMSE). Table 3 Highlights the performance of the applied machine learning models.

http://www.doi.org/10.62341/istj-vol38-2- irego42

On the training dataset, the DT model yields the lowest performance (R^2 = 0.93 and RMSE = 0.0082). In contrast, RF achieved an R^2 of 0.9768 with a corresponding RMSE of 0.0049 ,while both GB and XGB achieved the highest training accuracy (R^2 = 0.9834, RMSE = 0.0041). On the validation dataset (blind well), although RF achieved the lowest RMSE and the highest R^2, similar results were observed across all models, suggesting strong generalization capability. The DT model achieved (R^2= 0.9928 and RMSE = 0.0029), whereas RF recorded the best validation performance with (R^2 = 0.9974 and RMSE= 0.0017), GB and XGB followed closely with R^2 values of (0.9972 and 0.9969), respectively, and comparably low RMSE values (0.0018 and 0.0019), as illustrated in Figure 5 and Figure 6. Figure 5, present a comprehensive visualization of actual and predicted porosity, allowing for a thorough assessment of the performance and accuracy of the employed machine learning models. Based on this figure, can compare the machine learning result and the actual result. All models closely follow the trends of the actual porosity, with minor deviations observed at specific depths. DT predictions show slightly more variability compared to the other models (RF, GB, and XGB), which exhibit smoother and more consistent fits. Overall, the results confirm that ensemble learning techniques are highly suitable for porosity prediction from well log data, providing both accuracy and stability across unseen wells.

Machine Learning-Based Ensemble Models for Prediction of Reservoir Porosity

http://www.doi.org/10.62341/istj-vol38-2- irego42

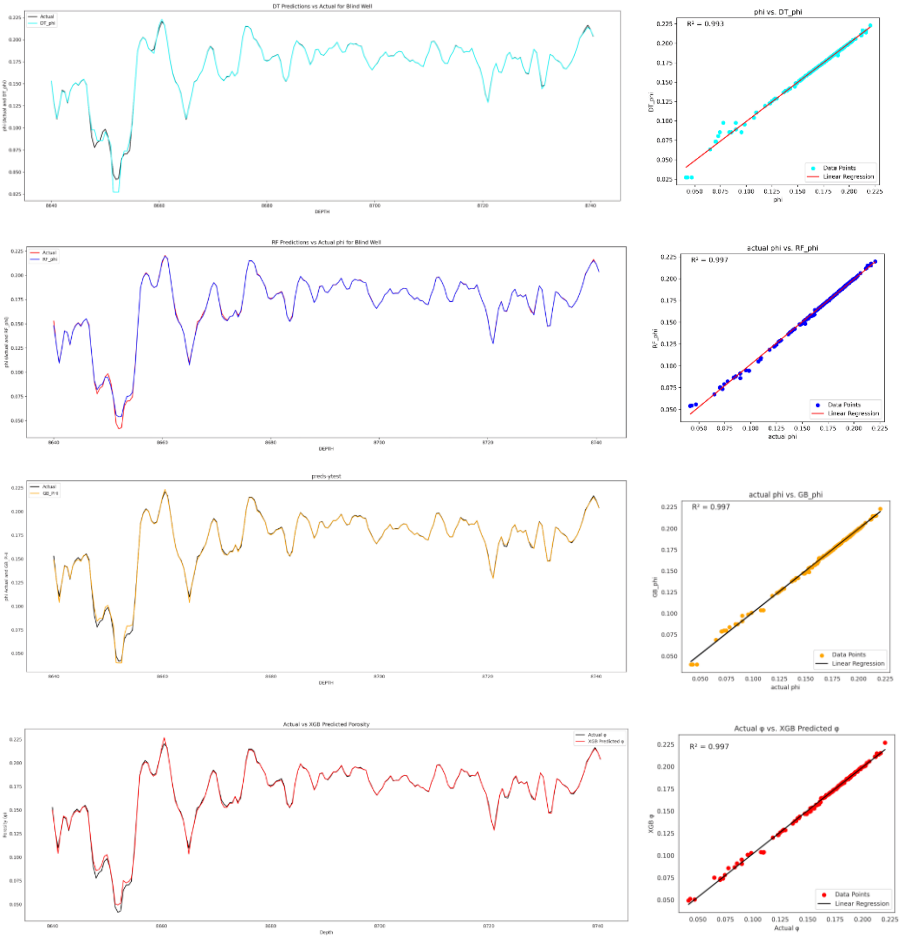

Figure 5: Comparison of actual versus predicted porosity (phi) for a blind well using four machine learning models: Decision Tree (DT, cyan), Random Forest (RF, blue), Gradient Boosting (GB, orange), and XGBoost (XGB, red). The left column shows scatter plots with linear regression lines and corresponding R^2 values, indicating the goodness of fit, while the right column presents line plots of predicted versus actual porosity along the well depth (Prepared by the author).

Machine Learning-Based Ensemble Models for Prediction of Reservoir Porosity

http://www.doi.org/10.62341/istj-vol38-2- irego42

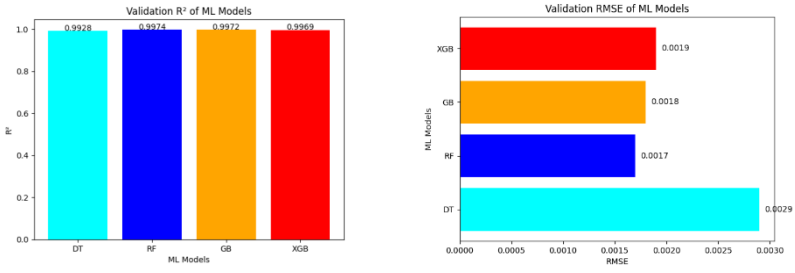

Figure 6:This figure shows a comprehensive comparison of machine learning models' overall performance, based on R^2 for the training and validation datasets (Prepared by the author).

Table 3: Performance analysis for the training dataset and the blind well.

Model	Training dataset		Validation dataset (blind well)	
	R^2	RMSE	R^2	RMSE
DT	0.9349	0.0082	0.9928	0.0029
RF	0.9768	0.0049	0.9974	0.0017
GB	0.9834	0.0041	0.9972	0.0018
XGB	0.9834	0.0041	0.9969	0.0019

4. Conclusions:

This research explored the performance of various ML models to predict reservoir porosity from well logging data, including Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), and Extreme Gradient Boosting (XGB). The analysis was conducted on 734 data points with correlated porosity (phi) as label output. The model was trained using well logging data, including GR, DT, CNL, and RHOB. The models were evaluated using unseen validation datasets (blind well) with two key statistical metrics: R^2 and RMSE. The results demonstrated that all models successfully captured the trends of the actual porosity, with minor deviations observed at specific depths. Among the evaluated models, ensemble learning techniques, particularly RF, achieved the highest accuracy and the

http://www.doi.org/10.62341/istj-vol38-2- irego42

lowest error values, indicating superior predictive capability and generalization performance compared to the single DT model.
GB and XGB also provided reliable and consistent predictions, closely following the performance of RF. Overall, the study confirms that ensemble machine learning approaches are highly effective and stable tools for porosity prediction from well log data. Their application can significantly enhance reservoir characterization by providing accurate and reliable estimates of porosity in both observed and previously unseen wells.

Although the study achieved strong result, this study has several limitations. The data size in low porosity range was relatively limited, which may cause higher prediction error observed in these intervals. Furthermore, the blind well used for validation originated from the same field and formation as the training dataset, which may restrict the assessment of the model's generalization to the other geological settings. Additionally, the use only four input's features including RHOB, DT, CNL, and GR may not fully capture more complex porosity types such as secondary or fracture porosity.

On this basis, the future study should:

- Increase the size and the number of data point especially in low porosity value to enhance the prediction in these intervals.
- Test the models on wells from different field's and lithologies to better assess true generalization capability.
- For more complex porosity types such as secondary or fracture porosity, it's important to incorporate additional features such as nuclear magnetic resonance (NMR).

5. References:

Abdullah Al-Qahtani, F. (n.d.). Porosity distribution prediction using artificial neural networks. Retrieved https://researchrepository.wvu.edu/etd/1010

Alatefi, S., Abdel Azim, R., Alkouh, A., & Hamada, G. (2023). Integration of Multiple Bayesian Optimized Machine Learning

http://www.doi.org/10.62341/istj-vol38-2- irego42

Techniques and Conventional Well Logs for Accurate Prediction of Porosity in Carbonate Reservoirs. Processes, 11(5). https://doi.org/10.3390/pr11051339

Al-Fakih, A., Kaka, S. I., & Koeshidayatullah, A. I. (2023). Reservoir Property Prediction in the North Sea Using Machine Learning. IEEE Access, 11, 140148–140160. https://doi.org/10.1109/ACCESS.2023.3336623

Alyafei, N. (2021). Fundamentals of reservoir rock properties (second). Hamad Bin Khalifa University Press. https://doi.org/https://doi.org/10.5339/Fundamentals_of_Reservoir_Rock_Properties_2ndEdition

Anifowose, F., Mezghani, M., Badawood, S., & Ismail, J. (2023). From Well to Field: Reservoir Rock Porosity Prediction from Advanced Mud Gas Data Using Machine Learning Methodology. SPE Middle East Oil and Gas Show and Conference, MEOS, Proceedings. https://doi.org/10.2118/213339-MS

Awad Mariette, & Khanna Rahul. (2015). Efficient Learning Machines Theories, Concepts, and Application for Engineers and System Designers. Springer nature.

Bhatt, A., & Helle, H. B. (n.d.). Committee neural networks for porosity and permeability prediction from well logs.

Bhattacharya, S. (2021). A Primer on Machine Learning in Subsurface Geosciences (Vol. 1). Springer. https://doi.org/https://doi.org/10.1007/978-3-030-71768-1

Brownlee, J. (2016). Machine Learning Mastery With Python. Machine Learning Mastery.

Dandekar, A. Y. (2013). Petroleum reservoir rock and fluid properties (second). CRC press.

Elkatatny, S., Tariq, Z., Mahmoud, M., & Abdulraheem, A. (2018). New insights into porosity determination using artificial intelligence techniques for carbonate reservoirs. Petroleum, 4(4), 408–418. https://doi.org/10.1016/j.petlm.2018.04.002

Erofeev, A., Orlov, D., Ryzhov, A., & Koroteev, D. (2019). Prediction of Porosity and Permeability Alteration based on Machine Learning Algorithms. http://arxiv.org/abs/1902.06525

http://www.doi.org/10.62341/istj-vol38-2- irego42

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data?

He, Y., Zhang, H., Wu, Z., Zhang, H., Zhang, X., Zhuo, X., Song, X., Dai, S., & Dang, W. (2025). Porosity prediction of tight reservoir rock using well logging data and machine learning. Scientific Reports, 15(1), 13124. https://doi.org/10.1038/S41598-025-95578-7

Kumar, J., Mukherjee, B., & Sain, K. (2024). Porosity prediction using ensemble machine learning approaches: A case study from Upper Assam basin. Journal of Earth System Science, 133(2). https://doi.org/10.1007/s12040-024-02310-6

Larestani, A., Hemmati-Sarapardeh, A., Samari, Z., & Ostadhassan, M. (2022). Compositional Modeling of the Oil Formation Volume Factor of Crude Oil Systems: Application of Intelligent Models and Equations of State. ACS Omega, 7(28), 24256–24273. https://doi.org/10.1021/ACSOMEGA.2C01466/ASSET/IMAGES/LARGE/AO2C01466_0017.JPEG

Mulashani, A. K., Shen, C., Nkurlu, B. M., Mkono, C. N., & Kawamala, M. (2022). Enhanced group method of data handling (GMDH) for permeability prediction based on the modified Levenberg Marquardt technique from well log data. Energy, 239, 121915. https://doi.org/10.1016/J.ENERGY.2021.121915

Nourani, M., Alali, N., Samadianfard, S., Band, S. S., Chau, K. wing, & Shu, C. M. (2022). Comparison of machine learning techniques for predicting porosity of chalk. Journal of Petroleum Science and Engineering, 209. https://doi.org/10.1016/j.petrol.2021.109853

Nyein, C. Y., & Ali Hamada, G. M. M. (2023, November 28). Artificial Neural Network (ANN) Prediction of Porosity and Water Saturation of Shaly Sandstone Reservoirs. https://doi.org/10.1306/51559nyein2019

Pandey, Y. N., Rastogi, A., Kainkaryam, S., Bhattacharya, S., & Saputelli, L. (2020). Machine Learning in the Oil and Gas Industry: Including Geosciences, Reservoir Engineering, and

http://www.doi.org/10.62341/istj-vol38-2- irego42

Production Engineering with Python. In Machine Learning in the Oil and Gas Industry: Including Geosciences, Reservoir Engineering, and Production Engineering with Python. Springer Science+Business Media. https://doi.org/10.1007/978-1-4842-6094-4

Rebala, G., Ravi, A., & Churiwala, S. (2019). An Introduction to Machine Learning. Springer International Publishing. https://doi.org/10.1007/978-3-030-15729-6

Sun, Y., Pang, S., Zhang, J., & Zhang, Y. (2024). Porosity prediction through well logging data: A combined approach of convolutional neural network and transformer model (CNN-transformer). Physics of Fluids, 36(2). https://doi.org/10.1063/5.0190078/3262471

Youcefi, M. R., Alshokri, A. I., Boussebci, W., Ghalem, K., & Hadjadj, A. (2024). Enhancing Porosity Prediction in Reservoir Characterization through Ensemble Learning: A Comparative Study between Stacking, Bayesian Model Optimization, Boosting, and Random Forest. Petroleum and Coal Pet Coal, 66(3), 1085–1098.

Zou, C., Zhao, L., Xu, M., Chen, Y., & Geng, J. (2021). Porosity Prediction With Uncertainty Quantification From Multiple Seismic Attributes Using Random Forest. Journal of Geophysical Research: Solid Earth, 126(7). https://doi.org/10.1029/2021JB021826