Journal of Geoscience and Environment Protection, 2025, 13(3), 68-86
https://www.scirp.org/journal/gep
ISSN Online: 2327-4344
ISSN Print: 2327-4336

Application of Machine Learning Methods in Parameter Prediction of Deep Lacustrine Oil Shale Reservoirs—Taking the Triassic Chang 7 in the Longdong Area as an Example

Yuxin Shen[1,2*], Zhichao Liu[1,2#]

[1]School of Earth Science and Engineering, Xi'an Shiyou University, Xi'an, China
[2]Key Laboratory of Petroleum Geology and Reservoir, Xi'an Shiyou University, Xi'an, China
Email: [#]1363874156@qq.com

How to cite this paper: Shen, Y. X., & Liu, Z. C. (2025). Application of Machine Learning Methods in Parameter Prediction of Deep Lacustrine Oil Shale Reservoirs—Taking the Triassic Chang 7 in the Longdong Area as an Example. *Journal of Geoscience and Environment Protection, 13,* 68-86.
https://doi.org/10.4236/gep.2025.133004

Received: February 9, 2025
Accepted: March 10, 2025
Published: March 13, 2025

Abstract

In the course of oil and gas exploration, understanding the petrophysical parameters such as reservoir porosity and permeability is crucial for evaluating oil and gas reserves and mining effectiveness. However, due to the complexity of well logging data and the interrelation between data, traditional analysis methods often have certain limitations and deficiencies. In order to overthrow the constraints of traditional methods, machine learning algorithms are used to build prediction models for porosity and permeability. In this study, Chang 7 reservoir in Longdong area was taken as the research object, and 320 core samples were selected for detailed measurement of porosity and permeability. In order to establish a generative prediction model, four different machine learning methods were used, including random forest (RF), K-nearest neighbor (KNN), extreme gradient boosting tree (XGBoost) and support vector machine (SVM). These methods were used to build an accurate prediction model for porosity and permeability combined with core and well logging data. In the experimental process, data preprocessing and feature selection were carried out, and four distinct machine learning methods were utilized to train and verify the model, and the optimal algorithm was selected according to the accuracy and stability of the model. Through single well analysis, the effectiveness of machine learning-based methods in predicting porosity and permeability was verified. Machine learning methods can build efficient and accurate prediction models by deeply mining information in data, providing researchers with a more detailed and comprehensive understanding of reservoir characteristics.

*First author.
#Corresponding author.

This technological progress not only optimizes the decision-making process of reservoir development, but also improves the effectiveness of resource application, which is of great value to the advancement of the petroleum industry.

Keywords

Machine Learning, Ordos Basin, Porosity, Permeability, Underground Reservoir Prediction

1. Introduction

In oilfield exploration and development, porosity and permeability are key parameters for evaluating geological reservoir and oil and gas resource potential. Porosity is an important index to measure the proportion of pore space in the total volume of rock or soil, which plays a key role in the evaluation and development of underground reservoirs (Shan, 2014). Permeability describes the ability of rock or soil to pass through fluid (liquid or gas), which is a key index for evaluating water permeability (Wang et al., 2024).

At present, many methods have been developed to calculate the porosity and permeability of rocks, each of which has its own characteristics and limitations. 1) Crossplot method: this method uses the crossplot of logging parameters such as AC and CNL, CNL and DEN to interpret porosity (Sun, 2016). However, the challenge of this method is that the determination of clay skeleton parameters is subjective, and the parameter selection may be different under different operators or different well conditions, resulting in increased difficulty in application and limited interpretation accuracy (Guo & Gong, 2018; Sima et al., 2008). 2) Lee formula method: through Willie formula, combined with acoustic Time difference Δt, Fluid acoustic moveout Δtf and Skeleton acoustic time difference Δtma to calculate porosity Φ (Wang, 2016; Zhang et al., 2014). However, when the porosity is very low (5% - 15%) or very high (>30%), the performance is poor, especially when calculating the porosity of tight reservoirs (porosity is often less than 15%), the accuracy of the formula is significantly reduced, resulting in large calculation errors. 3) Multiple regression analysis: this method is based on porosities, porosities POR and AC, DEN, CNL, Regression model of parameters such as shale content VSH and GR (Hou et al., 2019; She et al., 2019). However, the premise of accurately calculating porosity is to accurately determine the shale content VSH, and the high gamma sandstone in complex reservoirs makes the accurate calculation of VSH particularly difficult. The error of VSH directly affects the interpretation accuracy of porosity. Although some researchers directly use logging parameters such as AC to simplify the calculation, this method may lead to high porosity interpretation and deviation from the actual situation when the shale content fluctuates. 4) Special logging method: calculate porosity by using special logging tech-

niques such as NMR, imaging logging or ECS logging, which show high accuracy in complex reservoirs (Liu et al., 2023; Ren et al., 2017). However, due to the high cost of data acquisition and the relatively small amount of data, the wide application of these technologies is limited. 5) Spontaneous potential method: in the absence of porosity logging data, spontaneous potential (SP) is used to calculate reservoir porosity (Liang et al., 2017). However, the application of this method in tight reservoirs is limited because the SP anomaly amplitude is low, which may lead to low porosity interpretation. In addition, SP is also affected by many factors, such as argillaceous content, salinity of formation water and so on. When the porosity of high gamma reservoir is calculated solely by SP, its accuracy is often not high (Qu et al., 2019). In the calculation of permeability, the product of permeability coefficient and porosity is usually used. If the porosity calculation is wrong, the value of permeability will also be affected (Shi et al., 2019; Foalem et al., 2024).

As an important branch of artificial intelligence, machine learning has made significant progress in various fields in recent years (Wang et al., 2015). Taking oilfield exploration and development as an example, researchers have used machine learning methods to improve the accuracy of reservoir identification (Luo et al., 2022). The new geochemical indicators were excavated and the rapid prediction of mercury saturation was realized. However, it is still facing the challenge of rapid prediction of core porosity and permeability. The solution of this problem will help to promote the process of digital oilfield construction. Therefore, this paper aims to use machine learning algorithm to analyze and train the actual core data, so as to realize the rapid prediction of core porosity and permeability parameters, and promote the development of the oil industry in the direction of intelligence and efficiency.

2. Overview of Regional Geology

Ordos Basin is located in the central and western regions of China. It belongs to the Loess Plateau region, and its administrative regions span five provinces and regions of Shaanxi, Gansu, Ningxia, Inner Mongolia and Shanxi (Dai et al., 2018). The basin is bordered by Yinshan and Daqingshan in the north, Longshan, Huanglong and Qiaoshan in the south, Helan and Liupan Mountains in the west, and Luliang and Taihang Mountains in the East. With a total area of 370,000 square kilometers, it is the second largest sedimentary basin in China and is often referred to as the Shaanxi Gansu Ningxia basin (Wen et al., 2022; Zhan et al., 2021). Ordos Basin is famous for its unique tectonic evolution, which has experienced a complex process of overall rise and fall and depression migration, showing the typical characteristics of a large multicycle cratonic basin. Its basement is composed of Archean and Lower Proterozoic metamorphic rock series, while the sedimentary cover is extremely rich. From the great wall system to the quaternary system, it spans multiple geological ages, with a total thickness of 5000 to 10,000 meters (Zhang, 2019). In terms of structural morphology, the Ordos Basin presents a gentle trend of declining in the West and rising in the East, high in the East

and low in the West. The terrain changes very gently, and the slope per kilometer is even less than 1˚. This unique tectonic background creates favorable conditions for the accumulation of oil and gas, which makes the distribution of oil and gas resources in the basin show the distinct characteristics of half basin oil and full basin gas, that is, the south is dominated by oil, while the north is rich in natural gas, and the oil and gas resources also show the distribution pattern of upper oil and lower gas in the vertical direction, with a wide area and wide distribution. The multi-layer system is compound and connected, which is of great exploration and development value (Han et al., 2016).

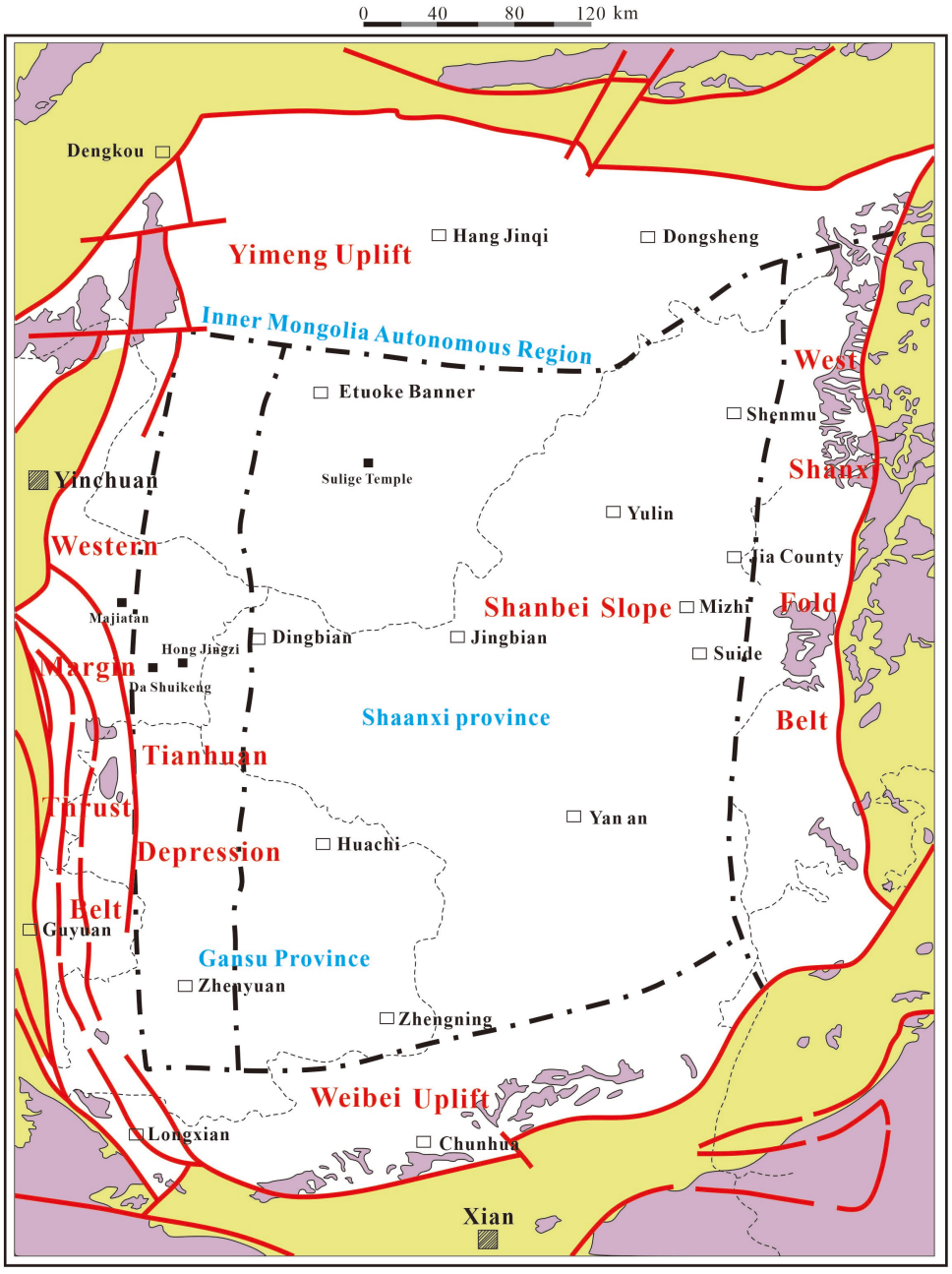

Figure 1. Structural zoning map of Ordos Basin.

According to the sequence analysis of the study area, the Triassic is mainly divided into five third-order sequences (As shown in **Figure 1**) (Gao et al., 2012). The rock properties of Chang 7 formation are mainly mudstone and shale. Especially in Longdong area, the deep lacustrine oil shale of Chang 7 formation is cleverly mixed with sandy turbidite and rich in oil and gas resources. This unique set of strata is the key oil generating strata bred in the most prosperous stage of Yanchang Formation Lake Basin, which is widely distributed in the whole lake basin. In the process of downhole logging, the formation shows typical "three high and one low" curve characteristics, i.e. high resistivity, high natural gamma value, high acoustic time difference and relatively low natural potential. These characteristics become important signs to identify the formation (Ma, 2010). Therefore, the research object of this paper is deep lacustrine oil shale (Pan et al., 2019).

3. Main Machine Learning Methods

3.1. Random Forest Algorithm

Random forest algorithm is an algorithm based on the principle of ensemble learning, and its basic building unit is the decision tree algorithm (Gao, Niu, & Sun, 2023). Decision tree is a tree structure, in which each internal node represents a feature or attribute, each branch represents a value of the feature or attribute, and each leaf node represents a classification or regression result (Zhao, Wang, & Rong, 2024; Hou, Wang, & Zai, 2022). Through the decision tree, the data set can be divided into multiple subsets, and each subset contains data with the same characteristics or attributes (Cui, Yang, & Wang, 2023; Huang, 2019). Then we can analyze each subset and classify or regress it. Objective function of random forest:

$$G(x) = \arg\max_y \sum_{i=1}^{k} I\left(g_i(x) = Y\right) \tag{1}$$

where: X is the feature set; $G(x)$ represents the final prediction result of the model;

The process of building the model with random forest algorithm is as follows (Guo, Zhong, & Li, 2019):

1) randomly select n samples from the original data set to form a new training data subset.

2) randomly select m features, and select the best feature from the M features for splitting.

3) split according to the selected feature to get a child node.

4) repeat steps 1 - 3 until the decision tree has grown.

5) repeat steps 1 - 4 to generate multiple decision trees.

During prediction, the test data set is run on each decision tree. The random forest algorithm will collect the prediction results of each decision tree, take the average value of these prediction results, and finally get the final prediction results. This can reduce the noise and bias that may be generated by a single decision tree, and make the overall prediction more stable and accurate.

3.2. *k*-Nearest Neighbor Algorithm

KNN algorithm is based on case-based learning. Unlike other machine learning algorithms, it attempts to build a clear model or function to map input to output, but directly depends on the instance in the training data set to predict (Luo, 2023). In KNN algorithm, there is a labeled training data set, which contains the feature vector and its corresponding label (category label in classification problem or continuous value in regression problem) (Hubei Seismological Bureau, 2023). When a new and unlabeled data point needs to be classified or predicted, KNN algorithm will calculate the distance between the new data point and all points in the training data set (usually Euclidean distance, but it can also be other distance measures). Then, select the k training data points closest to the new data points, and predict based on the labels of these k points (Wang, 2023). The objective function is as follows:

$$D = \left\{ \left(x_1, y_1 \right), \left(x_2, y_2 \right), \cdots, \left(x_m, y_m \right) \right\} \tag{2}$$

$$x_i = \left(x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)} \right) \tag{3}$$

$$L_P\left(x_i, x_j \right) = \left(\sum_{l=1}^{N} \left| x_i^{(l)} - x_j^{(l)} \right|^p \right)^{\frac{1}{p}} \tag{4}$$

where, D is the training data set, m is the number of samples, y is the corresponding category of different samples, and there are n features in different areas, is the feature vector of the sample, p is the index, and the influence of the weight on the points closer is greater than that on the points farther away (Huang, Zhao, & Bai, 2023). The greater the p index, the greater the influence on the points closer.

3.3. XGBoost Algorithm

Xgboost is an efficient implementation of Boosting algorithm. It belongs to the boosting algorithm family. Its core idea is to train multiple weak learners (usually decision trees) iteratively, and combine them in a serial manner to gradually improve the performance of the prediction model, and finally form a powerful integrated classifier or regressor. The basic component of Xgboost is the decision tree. These decision trees are "weak learners", which together form Xgboost, and these decision trees that make up Xgboost are in order: the generation of the latter decision tree will consider the prediction results of the previous decision tree, that is, the deviation of the previous decision tree will be taken into account, so that the samples with wrong prediction of the previous decision tree will be adjusted later, and then the next decision tree will be trained on the basis of the adjusted sample distribution. The objective function is as follows:

$$\mathscr{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{t-1} + f_t\left(x_i \right) \right) + \Omega\left(f_t \right)$$

$$\text{where} \quad \Omega\left(f \right) = \gamma T + \frac{1}{2}\lambda \left\| \omega \right\|^2$$

where, $\Omega(f_t)$ represents the regular term of the t-th tree (Li, 2013), it consists of two parts:

1) constrain the number of leaf nodes. This step is mostly a simple model

2) L2 normal form constraint is used to ensure the stability and generalization ability of the model when predicting the score of each leaf node.

3.4. SVM Algorithm

SVM is an effective classification algorithm, which can effectively carry out linear or nonlinear classification in high-dimensional data sets. Its advantage lies in the separation of space, that is, reducing the dimension of some data characterized by spatial points to form a feature space that can be used for classification. The core of SVM is to find a hyperplane that can maximize the classification interval. At the same time, some optimization techniques are used to improve the generalization ability of the model (Guo & Chen, 2018; Ali & Ang, 2023; Wang, Zhang, & Chai, 2004) and make them easier to classify. Then, support vector is used to divide the space and establish a classifier.

The goal of SVM is to divide the samples into two categories and establish the decision boundary by maximizing the boundary space. We use a hidden Markov model to describe the support vector machine, Where the eigenvector = (,, …,) represents the sample, represents the label of the sample, and is taken as −1 or 1, indicating that they belong to two categories respectively. The decision boundary of support vector machine is to find a hyperplane $w^{*}x - b = 0$ of the crossing point = (,, …,), so that the positive and negative samples are on both sides respectively. The hyperplane can be determined by the normal vector and the decision offset b. the parameter in the feature space is = (,, …,). The decision offset b is determined by the nearest support vector of the hyperplane, which is the positive and negative sample point closest to the decision boundary. If the point satisfies (* + b) > 1, it is a support vector (Liu, Tian, Liu et al., 2023). In order to make the hyperplane fully supported by support vector machine, the hyperplane is solved as the maximum interval classification by support vector machine. The learning process of support vector machine is to find the maximum interval hyperplane in the training data set and make it a support vector.

4. Reservoir Parameter Prediction Based on Machine Learning and Logging Data

4.1. Logging Curve Preprocessing

The flow chart of this experiment is shown in **Figure 2**. The experimental data set is from eight coring wells in an oilfield in Gansu Province, China. A total of 320 core samples were obtained from these eight wells, and the porosity and permeability of these samples were measured in the laboratory. After screening, a data set containing a variety of logging curve data is finally constructed. The data set consists of nine key parameters: acoustic transit time (AC), borehole diameter (CAL), compensated neutron (CNL), compensated density (DEN), natural

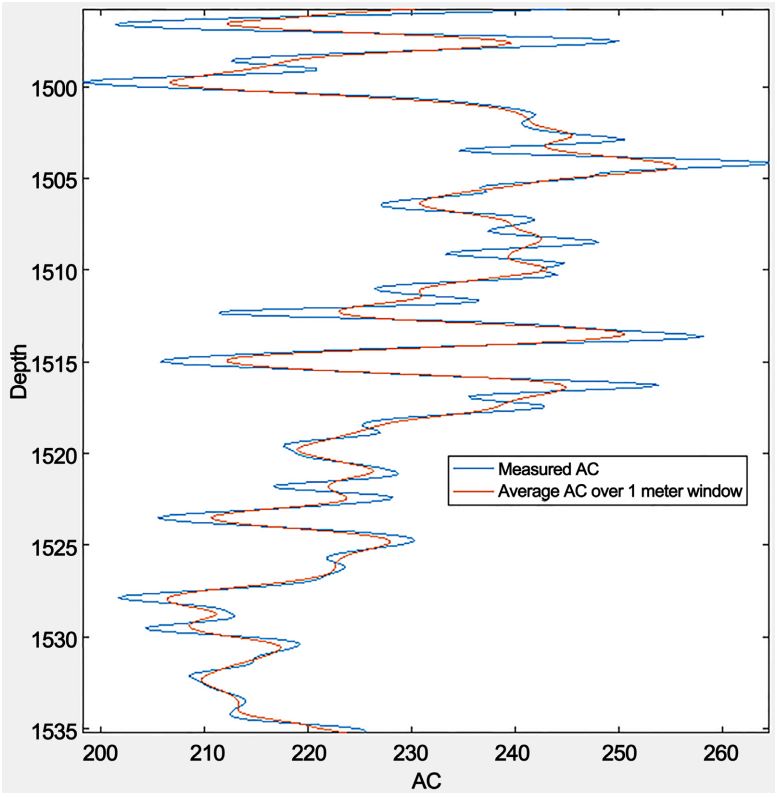

Figure 2. Porosity and permeability prediction flow chart based on machine learning and core and well logging data.

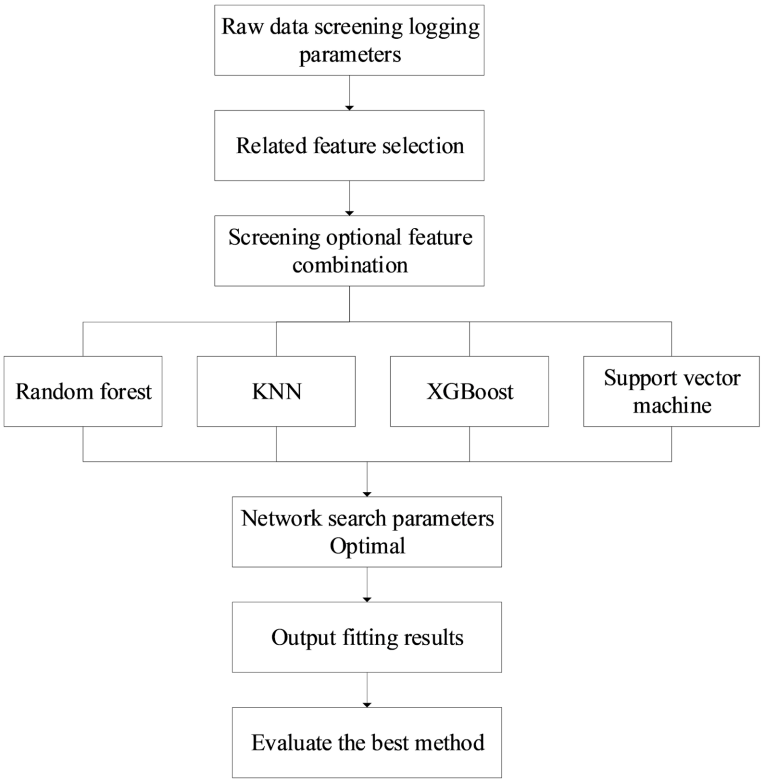

Figure 3. AC curve filter processing diagram of well Ning 191 in the study area.

gamma (GR), true formation resistivity (RT), spontaneous potential (SP), permeability (perm) and porosity (POR) (Beijing University of Posts and Telecommunications, 2020). In order to better analyze and process the logging curve data and reduce the error in data processing, firstly, the MATLAB software is used to filter and reduce the noise of some original logging curves of 8 wells in the study area, and the (moving average) is used to filter and reduce the noise of the acoustic time difference (AC) curve. The length of the sliding window selected here is 9, that is, 9 data points, corresponding to $0.125 \times (9 - 1) = 1$ m (Zhang, 2023) in the logging data. **Figure 3** shows the AC curve filtering and noise reduction treatment of well Ning 11 in the study area. It can be clearly seen that after filtering, the burr of acoustic time difference (AC) curve is effectively removed, showing a smoother curve shape (Zhang, Shi, Zhang et al., 2019), which is more convenient for the statistics of logging curve data below.

4.2. Dataset Data Cleaning

After collecting the logging curve data of each coring well, the data set needs to be cleaned (Liu & Zhang, 2024). The core task of data cleaning is to screen out the data that does not meet the established standards or requirements, and present the cleaned results to users to ensure the accuracy and reliability of the data. The rules of data cleaning mainly include: Inspection and processing of null values, detection and processing of illegal values, detection and processing of inconsistent data, and detection and processing of duplicate records. There are 320 pieces of data in this experiment. After data cleaning, there are 313 pieces of data left. 247 pieces of data in the experimental data set are used for the training set, and the remaining 66 pieces of data are used for the test set.

4.3. Standardization

Before data analysis, in order to ensure the unity of comparison standards and the reliability of analysis results, it is necessary to standardize the original data. Standardization processing is to convert the original data into a small specific interval in proportion by specific mathematical transformation means, so as to eliminate these differences, so that each variable can be on the same scale in data analysis or model training, so as to avoid some variables with large numerical range or variability occupying a dominant position in the algorithm and affecting the accuracy and reliability of the results. Taking the data of plate 66 as an example, the data is standardized (As shown in **Table 1**). This transformation process makes all index values unified to the same numerical level (Kumar, David, & Vikram, 2023). After standardization, indicators of different units or orders of magnitude can also be comprehensively analyzed and compared. This process has laid a solid foundation for further data analysis. At present, there are many methods of data standardization. Min max, Z-score standardization and decimal scaling are commonly used. In this experiment, Z-score standardization is used. Z-score is a common data preprocessing method. Z-score standardization is based on the mean and stand-

ard deviation of the original data. The calculation method of normalizing the original value V of attribute a to V'using Z-score is as follows:

$$\text{New data} = \frac{\text{Original data} - \text{Mean}}{\text{Standard deviation}}$$

Table 1. The data table obtained by standardizing the data using z-score.

Well Name	Original data	New data
Ban 66	70.24	1.77
Ban 66	69.84	1.74
Ban 66	69.41	1.72
Ban 66	68.98	1.69
Ban 66	68.53	1.66
Ban 66	68.04	1.63
Ban 66	67.53	1.60
Ban 66	67.04	1.57
Ban 66	66.62	1.54
Ban 66	66.26	1.52

Note: Only partial data of parameter SP is shown.

4.4. Feature Parameter Selection

The purpose of feature parameter selection is to delete features that are not related to annotation, so as to ensure the accuracy and efficiency of the model (Abdelhamid, Ouafi, Rabah et al., 2023). In this experiment, the labels are PERM and POR, and the characteristics are AC, CAL, CNL, DEN, GR, RT, SP. A simple model is constructed by using the idea of wrapping. The weak features are removed according to the coefficient, and the rest of the features repeat the process until the evaluation index drops greatly or is lower than the threshold value. According to the correlation degree between the annotation and the feature, the correlation thermodynamic diagram is drawn. Correlation thermograph (also known as correlation coefficient graph) is a commonly used visualization method, which is used to display the data distribution of the correlation coefficient matrix or contingency table of a group of variables (Liu et al., 2022). This graph can intuitively display the difference between the given values, and help researchers quickly identify the correlation between variables (Anifowose, Abdulraheem, & Al-Shuhail, 2019).

In the correlation thermodynamic diagram, each cell carries the correlation information between different attribute pairs in the data. The correlation coefficient values of these attribute pairs are cleverly transformed into a series of unique color tones, which are visually displayed by filling the cells. The change of color depth or hue can indicate the size of correlation coefficient, thus reflecting the degree of correlation between variables (Liu et al., 2022). According to the thermodynamic

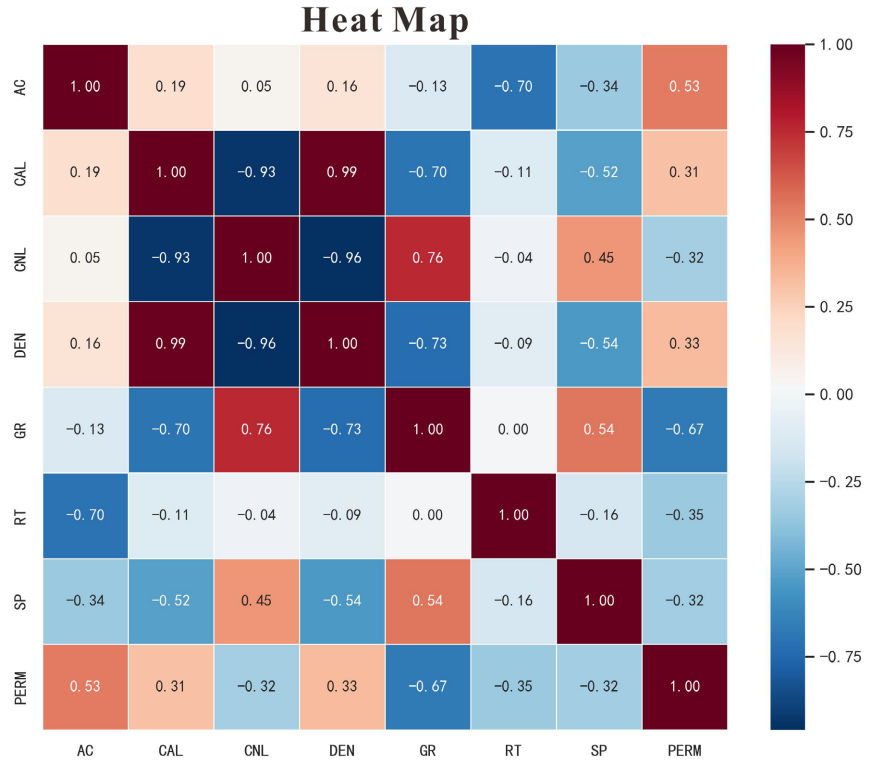

Figure 4. One-Hotmaps of correlation coefficient between permeability and logging variables in the study area.

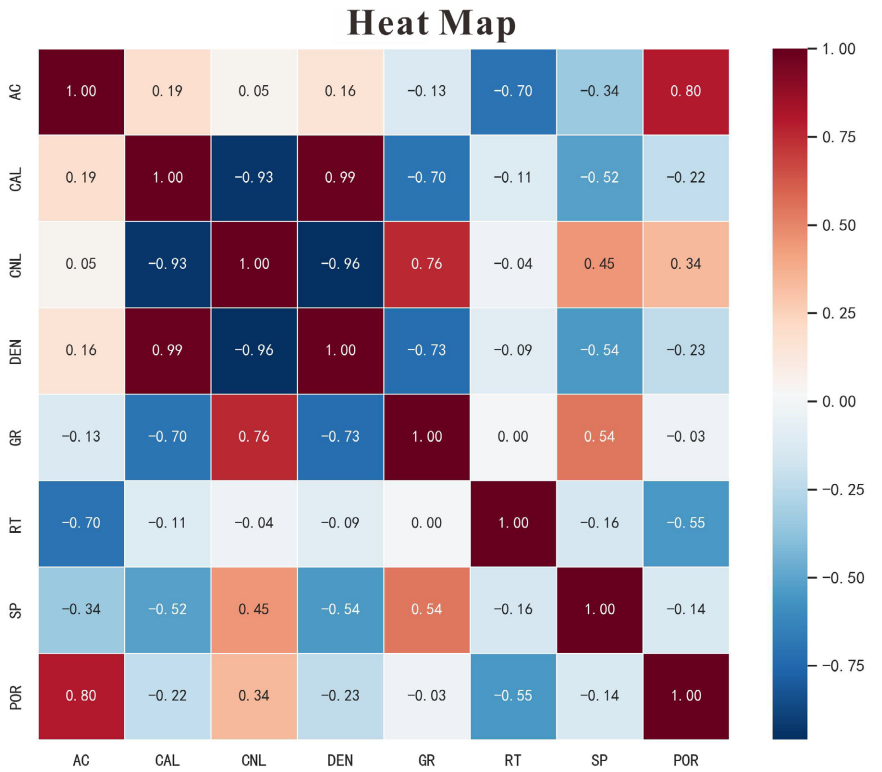

Figure 5. One-Hot map of correlation coefficient between porosity and logging variables in the study area.

diagram of the correlation coefficient between permeability and logging variables in the study area (As shown in **Figure 4**) (Anifowose, Abdulraheem, & Al-Shuhail, 2019), acoustic time difference (AC), caliper logging (CAL), density logging (DEN) and permeability are positively correlated, and the correlation coefficients from high to low are 0.53, 0.31 and 0.33, respectively. Other logging parameters are negatively correlated with permeability. Therefore, AC, cal and den are selected as the parameters of the model.

According to the thermodynamic diagram of correlation coefficient between porosity and logging variables in the study area (As shown in **Figure 5**), acoustic time difference (AC) and compensated neutron (CNL) are positively correlated with porosity, with correlation coefficients of 0.80 and 0.34, and other logging parameters are negatively correlated with porosity. Therefore, AC and CNL are selected as the parameters of the model.

5. Conclusion

In this experiment, the parameter adjustment method of grid search is selected, and the optimal parameter combination is found by traversing the given parameter combination. Compared with manual parameter adjustment, grid search does not require multiple tests and comparisons (Martin, Samuel, Max et al., 2024). The results of manual parameter adjustment are affected by human factors, such as personal experience, prejudice, intuition, etc., which are subjective and uncertain.

Random forest is a very representative Bagging ensemble algorithm. All of its base evaluators are decision trees. The n-estimators parameter is used to set the number of trees in the forest. The larger the n-estimators, the better the effect of the model. However, any model has an upper limit. When the n-estimators reach a certain level, the accuracy of the random forest will not rise or start to fluctuate with the rise of n-estimators. Set the n-estimators parameters to 500, 800, and 1000. The max_depth parameter is the maximum tree depth, and its setting depends on the complexity of the dataset (Zhang, 2019). If the data set is simple, you can set a smaller max_depth to prevent over fitting. If the data set is complex, you can set a larger max_depth to improve the accuracy of the model. Set max_depth to 8, 9, 10.

Support vector machine is a supervised learning technology, which can be used not only for classification, but also for regression. Kernel function is the general name of many algorithms in the field of pattern analysis. These algorithms are good at using linear classifiers to deal with nonlinear problems. Set the kernel functions to Linear, Poly, and Rbf. Linear is a linear kernel that deals with linear problems. Poly is a polynomial kernel that deals with partial linearity. RBF is a radial basis kernel, which deals with partial nonlinearity.

KNN is a machine learning algorithm for classification and regression. In KNN, the n_neighbors parameter represents the number of nearest neighbors used for prediction. When n_neighbors is set to 1, the algorithm will find the nearest single

neighbor to predict. When n_neighbors is set to a larger value, the algorithm will consider more neighbors to predict. Set n_neighbors to 2, 3, 4, and 5. Xgboost is a lifting tree model, so it integrates many tree models to form a strong classifier. Parameter learning rate refers to the extent of model parameter update in each iteration. Smaller learning rate can make the model more stable, but it needs more iterations. The higher the learning rate, the faster the convergence speed of the model, but it may lead to the instability of the model. Generally, the initial learning rate is set to 0.1, and then adjusted according to the performance of the model. Set the learning rate to 0.1, 0.001 and 0.0001. According to the grid search algorithm, the optimal parameter combination is obtained, and the mean square error (MSE) and mean absolute error (MAE) are output.

Mean squared error (MSE) is a loss function commonly used in regression analysis. It is used to measure the difference between the predicted value and the real value of the model. The specific calculation method is to first calculate the square of the difference between the predicted value and the real value at each sample point, and then sum these square errors and average them. The value range is [0, +∞) and the formula is as follows:

$$\mathrm{MSE} = \frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{y}_i\right)^2$$

Mean absolute error (MAE) is also a loss function commonly used in regression analysis, which focuses on measuring the absolute value of the difference between the predicted value of the model and the target value (or the real value). Unlike the mean square error (MSE), MAE calculates the sum of the absolute values of the difference between the predicted value and the target value at all sample points, and then takes the average value of this sum. Regardless of direction, the value range is [0, +∞), and the formula is as follows:

$$\mathrm{MAE} = \frac{1}{m}\sum_{i=1}^{m}\left|\left(y_i - \hat{y}_i\right)\right|$$

Table 2. The error and determination coefficient of permeability and porosity predicted by 4 methods.

Algorithm type	Permeability		Porosity	
	MSE	MAE	MSE	MAE
RF	2.73	1.05	0.59	0.56
SVM	23.29	2.90	1.38	0.87
KNN	12.22	1.90	0.88	0.69
XGBoost	2.36	0.95	1.08	0.69

Note: MSE—mean square error of output; MAE—mean absolute error.

It can be seen from Table 2 that the best method for predicting permeability is xgboost, and the mean square error and mean absolute value error are 2.73 and 1.05 respectively. The best method to predict porosity is RF, and the mean square

error and mean absolute error are 0.59 and 0.56, respectively. The prediction effect of permeability is worse than that of porosity, indicating that permeability is more difficult to characterize than porosity. The XGBoost method is used to predict the intersection between the actual permeability value and the predicted permeability value, as shown in Figure 6. The RF method is used to predict the intersection between the actual porosity value and the predicted porosity value, as shown in Figure 7.

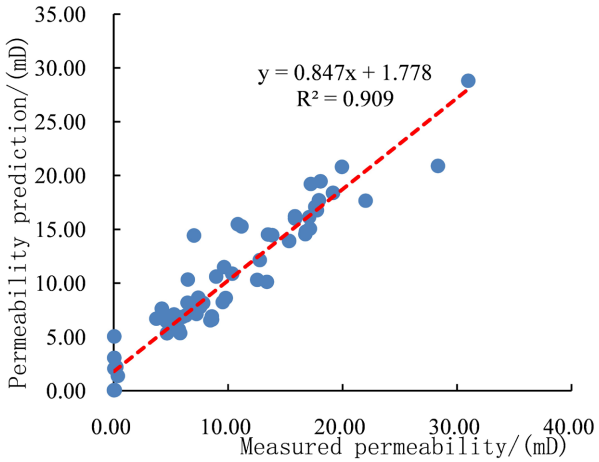

Figure 6. XGBoost method is used to predict the crossplot between the real and predicted permeability values.

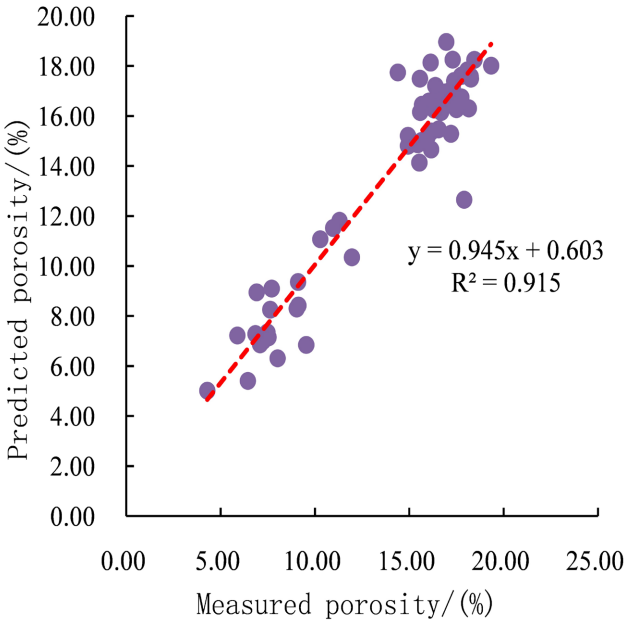

Figure 7. RF method is used to predict the crossplot between the real and predicted porosity values.

6. Single Well Analysis

The porosity and permeability of Zhuang 111 well with more cores are predicted.

The results show that the predicted porosity and permeability are basically consistent with the measured porosity and permeability (as shown in Figure 8). Especially in the well section of 1780 - 1810 m, the trend of predicted porosity and permeability is basically the same as that of measured porosity and permeability. However, there are also points with large errors. The reason for the large difference near the depth of about 1755 meters may be that the correlation between each logging parameter and porosity and permeability is low. The main reason is the significant influence of reservoir heterogeneity and secondary changes, which seriously affect the representativeness of the experimental parameters of this layer. In addition, the complexity of geological structure may also be another factor that reduces the accuracy of model interpretation. In view of this complexity, more in-depth research and data analysis are needed to establish an accurate prediction model to improve the accuracy and reliability of prediction. In addition, core homing is also one of the reasons affecting the accuracy of this model. It can be seen from 1795 - 1810 m in the figure that the porosity and permeability predicted by the model are relatively accurate, which are in good agreement with the measured porosity. In general, the reliability and accuracy of the model are high, which can provide the porosity and permeability of the study area, but the accuracy of the model still has room for improvement, which can be

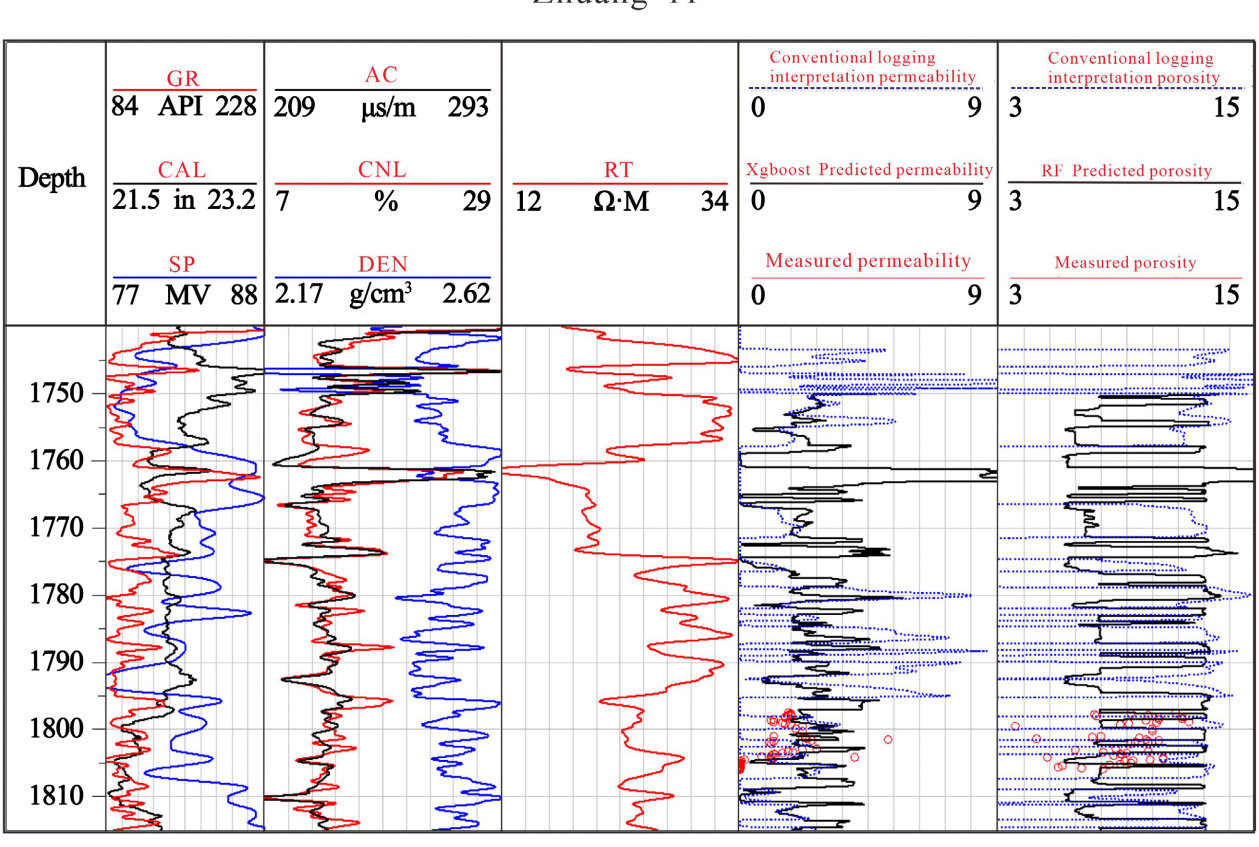

Figure 8. Predicted pore permeability column of Zhuang 111 well in the study area.

further optimized (Hofmann, Li, Taphorn et al., 2024).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Abdelhamid, O., Ouafi, A., Rabah, K. et al. (2023). Integrating Drilling Parameters and Machine Learning Tools to Improve Real-Time Porosity Prediction of Multi-Zone Reservoirs. Case Study: Rhourd Chegga Oilfield, Algeria. *Geoenergy Science and Engineering, 223,* Article 211511. https://doi.org/10.1016/j.geoen.2023.211511

Ali, F., & Ang, R. P. (2023). Many-Dimensional Model of Adolescent School Enjoyment: A Test Using Machine Learning from Behavioral and Social-Emotional Problems. *Education Sciences, 13,* Article 1103. https://doi.org/10.3390/educsci13111103

Anifowose, F., Abdulraheem, A., & Al-Shuhail, A. (2019). A Parametric Study of Machine Learning Techniques in Petroleum Reservoir Permeability Prediction by Integrating Seismic Attributes and Wireline Data. *Journal of Petroleum Science and Engineering, 176,* 762-774. https://doi.org/10.1016/j.petrol.2019.01.110

Beijing University of Posts and Telecommunications (2020). *A Method and Device for Information Reliability Evaluation Based on Knowledge Graph*. CN202010245428.8.

Cui, J. F., Yang, J. L., Wang, M. et al. (2023). Shale Porosity Prediction Based on Random Forest Algorithm. *Petroleum Geology and Recovery Efficiency, 30,* 13-21.

Dai, L. F., Chen, S. J., Wang, P. et al. (2018). The Influence of Differences in Physical Properties of Tight Sandstone Reservoirs in the Chang-7 Section of the Ordos Basin on Oil Content. *World Petroleum Industry,* 1-11.

Foalem, P. L., Khomh, F., & Li, H. (2024). Studying Logging Practice in Machine Learning-Based Applications. *Information and Software Technology, 170,* Article 107450. https://doi.org/10.1016/j.infsof.2024.107450

Gao, F. M., Niu, K., Sun, X. P. et al. (2023). Reservoir Pore Structure Characterization and Production Prediction Based on Machine Learning. *Progress in Geophysics,* 1-10.

Gao, W., Jiao, C. Y., Qu, Y. L. et al. (2012). Study on the Characteristics and Diagenesis of the Heshui Tarwan Chang 7 Reservoir in the Longdong Area of the Ordos Basin. *Natural Gas Exploration and Development, 35,* 22-27.

Guo, G. H., Zhong, S. H., Li, S. Z. et al. (2019). Application of Machine Learning and Trace Elements of Zircon to Construct an Identification Diagram of Granite Metallogenic Potential: A Case Study of Qimantage, East Kunlun. *Northwest Geology, 56,* 57-70.

Guo, J. X., & Chen, M. (2018). Based on Support Vector Machine (SVM) of Turbidite Fan Sedimentary Microfacies Automatically Identify. *Journal of Gansu Science, 30,* 25-31.

Guo, X. L., & Gong, H. J. (2018). Application of Entropy Weight Fuzzy Comprehensive Evaluation in Reservoir Evaluation. *Journal of Shaanxi University of Technology (Natural Science Edition), 34,* 21-27.

Han, W. X., Tao, S. Z., Yao, J. L. et al. (2016). Fine Characterization of Tight Reservoirs in the Longdong Area of the Ordos Basin. *Natural Gas Geosciences, 27,* 820-826.

Hofmann, J., Li, Z., Taphorn, K., Herzen, J., & Wudy, K. (2024). Porosity Prediction in Laser-Based Powder Bed Fusion of Polyamide 12 Using Infrared Thermography and Machine Learning. *Additive Manufacturing, 85,* Article 104176. https://doi.org/10.1016/j.addma.2024.104176

Hou, K. J., Wu, J. M., Ge, X. et al. (2019). Calculation Method of Reservoir Porosity in

Expanded Section Based on Two-Dimensional Nuclear Magnetic Data. In *Proceedings of the 31st National Natural Gas Academic Conference (Geological Exploration)* (p. 8). Sinopec Southwest Petroleum Engineering Ltd.

Hou, X. M., Wang, F. Y., Zai, Y. et al. (2022) Based on Machine Learning and Log Data of Carbonate Porosity and Permeability Prediction. *Journal of Jilin University (Earth Sciences), 52,* 644-653.

Huang, Q. L., Zhao, J. L., Bai, Q. et al. (2023). Automatic Identification of Sedimentary Microfacies Based on Adaboost Algorithm: A Case Study of Shanxi Formation, Q Area, Longdong Gas Field. *Geological Bulletin, 43,* 658-666.

Huang, X. J. (2019). *Simulation and Optimization of Ground Source Heat Pump System Based on TRNSYS*. Suzhou University of Science and Technology.

Hubei Seismological Bureau (Institute of Seismology, China Earthquake Administration) (2023). *A Method and System for Pre-Earthquake Gravity Disturbance Signal Recognition*. CN202310089610.2.

Kumar, P. S., David, L., & Vikram, V. (2023). A Robust Mechanistic Model for Pore Pressure Prediction from Petrophysical Logs Aided by Machine Learning in the Gas Hydrate-Bearing Sediments over the Offshore Krishna-Godavari Basin, India. *Natural Resources Research, 32,* 2727-2752.

Li, L. (2013). *Research on Micromagnetic Detection and Imaging Techniques of Crystalline Silicon Defects*. Nanchang Hangkong University.

Liang, X., Wang, S., Zhou, M. S. et al. (2017). Coal Reservoir Porosity Evaluation Based on Nuclear Magnetic Resonance and Resistivity Logging. *Coal Engineering, 49,* 130-133.

Liu, H., Xu, J. X., Chen, H. B. et al. (2023) Porosity Calculation Method for Shale Gas Reservoirs Based on Nuclear Magnetic Resonance Logging Correction. *China Offshore Oil and Gas, 35,* 89-95.

Liu, J., Tian, L., Liu, S. X. et al. (2023). Tight Gas Well Productivity Prediction Model Based on Composite Machine Algorithm: A Case Study of SM Block in Ordos Basin. *Petroleum Geology and Development in Daqing, 43,* 69-78.

Liu, K., Niri, M. F., Apachitei, G., Lain, M., Greenwood, D., & Marco, J. (2022). Interpretable Machine Learning for Battery Capacities Prediction and Coating Parameters Analysis. *Control Engineering Practice, 124,* Article 105202. https://doi.org/10.1016/j.conengprac.2022.105202

Liu, T., & Zhang, R. (2024). A Machine Learning-Based Hybrid Model for Fracture Parameterization and Distribution Prediction in Unconventional Reservoirs. *Computers and Geotechnics, 168,* Article 106146. https://doi.org/10.1016/j.compgeo.2024.106146

Luo, C. Q. (2023). *Research on PCB Assembling Scheme Optimization and Material Prediction Method under Multiple Constraints*. Chongqing University of Technology.

Luo, G., Xiao, L. Z., Shi, Y. Q. et al. (2022). Research on Fluid Identification Method of Tight Reservoir Based on Machine Learning. *Bulletin of Petroleum Science, 7,* 24-33.

Ma, D. B., Li, M., Cui, W. J. et al. (2010). Distribution Pattern of Turbidite Rocks in the Chang 6-Chang 7 Sections of the Yanchang Formation in Longdong Area. *Xinjiang Petroleum Geology, 31,* 33-36.

Martin, E., Samuel, P., Max, O. et al. (2024). Analysis of Data Generation and Preparation for Porosity Prediction in Cold Spray Using Machine Learning. *Journal of Thermal Spray Technology, 33,* 1270-1291.

Pan, S. W., Zheng, Z. C., Lei, J. Y. et al. (2019). Sandstone Reservoir Porosity Prediction Based on Hybrid Optimization XGBoost Algorithm. *Computer Applications and Software, 40,* 103-109+206.

Qu, H. J., Pu, R. H., Chen, S. et al. (2019). Facies and Potential Coupling Control of Mesozoic Oil Accumulation in Ordos Basin. *Oil & Gas Geology, 40,* 752-762+874.

Ren, J., Zhai, F. F., Li, F. L. et al. (2017). Based on The Natural Potential Method of Argillaceous Sandstone Reservoir Porosity Calculation. *Journal of Logging Technology, 9,* 292-295.

Shan, X. G. (2014). *Identification and Potential Evaluation of Chang 6 Thin Reservoir in Yanchang Formation of Qilicun Oilfield*. Xi'an Shiyou University.

She, G., Gui, P. F., Chen, Y. et al. (2019). Ping West Qaidam Basin Region Bedrock Reservoir Logging Evaluation. *Journal of Yangtze University (Natural Science Edition), 9,* 18-26.

Shi, P. Y., Wang, C. Y., Yan, W. L. et al. (2019). Metamorphic Mineral Content Was Calculated by the Logging Data Element Method. *Journal of Logging Technology, 6,* 597-600.

Sima, L. Q., Wang, P. C., Deng, X. H. et al. (2008). Discussion on Reservoir Porosity Calculation Method of Feixianguan Formation in Northeast Sichuan. *Journal of Southwest Petroleum University (Natural Science Edition), 2,* 1-4+183.

Sun, B. (2016). Research on the Method of Calculating Porosity of Tight Sandstone Formations Using Acoustic Logging Curves. *Petrochemical Technology, 23,* 163.

Wang, D. W., Li, J. Y., Liu, D. et al. (2024). Characterization of Capillary Pressure Curve and Pore Throat Distribution Based on Reservoir Physical Parameters. *Unconventional Oil and Gas, 11,* 1-9.

Wang, L. J. (2023). *Reservoir Porosity Prediction Based on Deep Learning*. Qingdao University of Technology.

Wang, X. G., Fang, Y., Fang, J. et al. (2015). Discussion on Calculation Method of Matrix Porosity in Limestone Reservoir. *Petroleum Geology and Engineering, 29,* 81-83.

Wang, Z. H., Zhang, C. G., Chai, C. Y. et al. (2004). Logging Identification Model of Low Permeability Reservoir Types. *Natural Gas Industry, 9,* 36-38.

Wang, Z. L. (2016). *Calculation and Application of Complex Lithologic Reservoir Parameters of Chang 6 Formation in Zhoujiawan area, Changqing Oilfield*. China University of Petroleum.

Wen, Z. G., Luo, Y. S., Liu, J. Y. et al. (2022). Pore Structure Characteristics and Genesis Mechanism of the Triassic Chang 7 Shale Oil Reservoir in Longdong Area. *Lithological Oil and Gas Reservoir, 34,* 47-59.

Zhan, Y. X., Chen, L., Lin, C. H. et al. (2021). Determination of Reservoir Formation Stages and Periods of Chang-7 Shale Oil in Longdong Area of Ordos Basin. *Yunnan Chemical Industry, 48,* 121-125.

Zhang, P., & Zhang, X.L. (2014). Research on Permeability Logging Interpretation Model of Low Porosity and Low Permeability Reservoir. *Groundwater, 36,* 74-76.

Zhang, Q. (2019). *Study on the Microstructural Characteristics of the Tight Sandstone Reservoir of Group 7 in the Triassic Yanchang Formation in the Longdong Area of the Ordos Basin*. Northwest University.

Zhang, Y. H., Shi, B. H., Zhang, Y. J. et al. (2019). Application of Machine Learning Method to Prediction of Shallow Beach and Bar Facies Thin Reservoir Porosity: A Case Study of Cretaceous in Chepaizi Area, Junggar Basin. *Acta Sedimentologica Sinica, 41,* 1559-1567.

Zhang, Z. Q. (2023). Research on the Identification Method of Fluid Properties of Pyroclastic Sandstone Conglomerate Reservoir Based on Machine Learning. Northeast Petroleum University.

Zhao, J., Wang, Q., Rong, W., Zeng, J., Ren, Y., & Chen, H. (2024). Permeability Prediction of Carbonate Reservoir Based on Nuclear Magnetic Resonance (NMR) Logging and Machine Learning. *Energies, 17,* Article 1458. https://doi.org/10.3390/en17061458