



## OPEN XGBoost algorithm optimized by simulated annealing genetic algorithm for permeability prediction modeling of carbonate reservoirs

Changbing Huang<sup>1✉</sup>, Xinyu Zhu<sup>1</sup>, Mingyu Lu<sup>2</sup>, Yuling Zhang<sup>1</sup> & Shengbo Yang<sup>1</sup>

Carbonate reservoir has strong heterogeneity, complex pore structure and poor correlation between porosity and permeability, so the traditional permeability model can not meet the needs of logging interpretation. Taking the carbonate reservoir of Longwangmiao Formation in Moxi block of central Sichuan as an example, this paper proposes to establish a permeability prediction model by using the XGBoost algorithm of simulated annealing genetic algorithm (SA-GA) hybrid optimization. Combined with core data, five permeability sensitive logging curves (CNL, DEN, DT,  $\ln RT$  and GR) are optimized by calculating correlation coefficients, and the permeability prediction model is established based on XGBoost algorithm, and the XGBoost hyperparameters are optimized by using SA-GA. The method is applied to the evaluation of logging permeability in the study area. The results show that the prediction results of SA-GA-XGBoost algorithm are more consistent with the core data. The adjusted  $R^2$  is 0.876, and the root mean square error (RMSE) is only 0.142. The prediction accuracy is better than the conventional permeability model and BP neural network model, which meets the industrial requirements of logging evaluation and provides a new idea for oil and gas exploration in carbonate reservoirs.

**Keywords** Carbonate rock, Permeability, Simulated annealing, Genetic algorithm, XGBoost

Permeability is a key parameter for reservoir evaluation, and the accuracy of this parameter calculation will directly affect the development effect of the reservoir. Due to the strong heterogeneity and complex pore structure of carbonate reservoirs, the traditional permeability interpretation model established by logging data has been unable to meet the needs. Therefore, how to use logging data to accurately predict the permeability of carbonate reservoirs is also a major difficulty in current research<sup>1–4</sup>.

At present, the traditional carbonate reservoir permeability calculation model mainly adopts the idea of clastic reservoir evaluation, using the correlation between core porosity and permeability to derive the permeability calculation model, or using the empirical formula of permeability, but the two methods are applied to the evaluation of carbonate reservoirs, and the results varied widely. In recent years, with the rapid development of artificial intelligence, more and more experts and scholars have been applying machine learning algorithms to geophysical logging permeability interpretation<sup>5–10</sup>. Yang Jian et al<sup>11</sup> used an improved BP neural network model to predict the logging permeability, and achieved certain application results; Wang Xinglong et al<sup>12</sup> used C5.0 Decision-tree algorithm to interpret the permeability of carbonate reservoir; Gu Yufeng et al<sup>13</sup> used a particle swarm optimization algorithm for parameter optimization of the gradient Decision-tree, and established a model for the permeability of clastic reservoir; Zhang Yanan et al<sup>14</sup> proposed to construct a logging interpretation model for the permeability of complex carbonate reservoir by using hybrid simulated annealing-genetic algorithm-random forest algorithm; Han Rubing et al<sup>15</sup> phase-control carbonate reservoir layers, and then use BP neural network for permeability prediction. Although the above method has played a certain application effect in permeability logging interpretation, due to the slow convergence speed of the model, easy over-fitting and high complexity of the model, the above algorithm cannot be widely used in logging interpretation. Compared with BP neural network, traditional Decision-tree, Random Forest and other algorithms, XGBoost algorithm adds

<sup>1</sup>School of Environment and Resources, Southwest University of Science and Technology, Mianyang 621010, China.

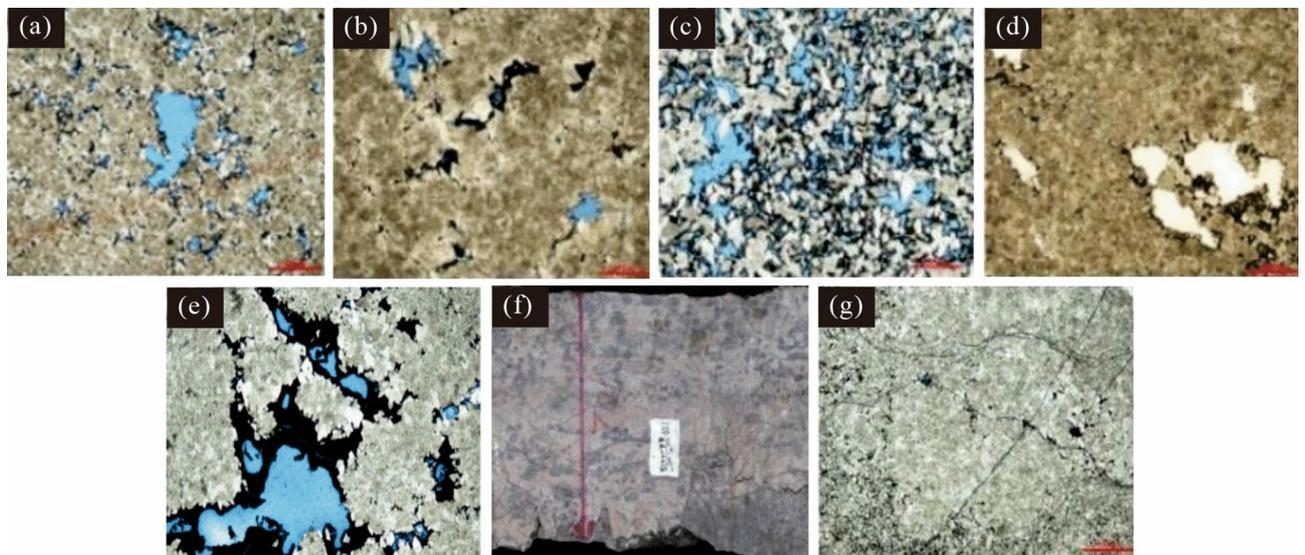
<sup>2</sup>China Petroleum Group Logging Co., Ltd. Tarim Branch, Korla 841000, China. ✉email: 1832873542@qq.com

regularization terms, the fitting model refinement, having the advantages of high efficiency, high scalability, and high prediction accuracy, and being widely used in reservoir evaluation<sup>16,17</sup>. Therefore, this paper takes the strong heterogeneous carbonate reservoir of Longwangmiao Formation in Moxi Block of Anyue Gas Field as an example. Based on the optimization of permeability sensitive logging curves, the XGBoost algorithm based on simulated annealing genetic algorithm (SA-GA) optimization is proposed to establish the logging permeability prediction model of carbonate reservoir. The optimization model can obtain the global optimal solution and the convergence speed is faster than the single optimization algorithm such as particle swarm optimization (PSO) and genetic algorithm (GA), which can give full play to the performance of XGBoost.

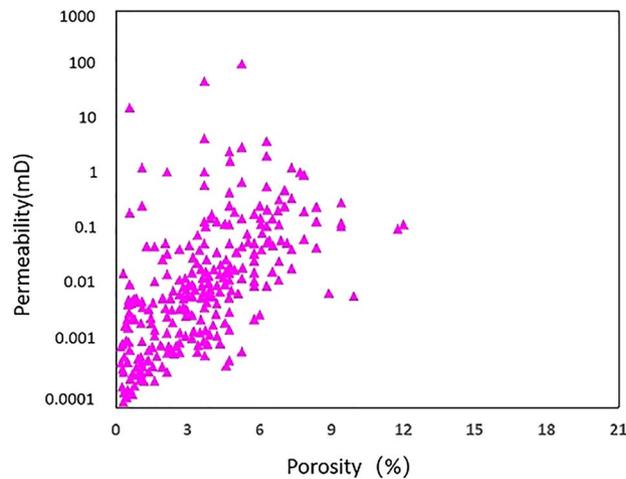
### Overview of the study area

In 2013, the high-yield Anyue super-large gas field was discovered in the Longwangmiao Formation of the Sichuan Basin. This gas field is the oldest and largest single super-large marine gas reservoir discovered in China at this stage. Moxi block of Anyue gas field is located in Weiyuan-Longnvisi structural group in the gentle structural area of the middle slope of the ancient uplift in central Sichuan. The main production layer Longwangmiao Formation is widely developed carbonate reservoirs, which have the characteristics of diverse lithology, complex pore structure and strong heterogeneity. The horizontal and vertical spatial distribution of this kind of carbonate rocks is well connected. The reservoir lithology is a large set of dolomite, and some core thin sections are shown in Fig. 1. The reservoir lithology is mainly medium-fine-grained dolomite, with residual granular (sandy and oolitic) structures (Fig. 1a, c, e). The main types of reservoir space are intergranular and intercrystalline solution pores (Fig. 1b, e), followed by intercrystalline pores, and some well sections develop cavities and fractures (Fig. 1d-f, g), with strong heterogeneity of the reservoir.

According to the analysis data of the core of the Longwangmiao Formation, the porosity of the small rock samples with porosity greater than 2%, ranges from 2.02% to 12.11%, with an average of 4.16%. Among them, the porosity of more than 4% accounted for 44.23%, more than 6% accounted for 13.74%. The permeability ranges from 0.0001 mD to 108.1 mD, with an average of 0.87 mD. The ones with permeability greater than 0.01 mD accounted for 54.84%, those with permeability greater than 0.1 mD accounted for 23.4% and those with permeability greater than 1 mD accounted for 9.14%. It can be seen that the reservoir of Longwangmiao Formation is characterized by medium porosity and medium permeability. Excluding the influence of fractures, the porosity-permeability relationship of the reservoirs in the study area was established, as shown in Fig. 2, it is not difficult to see that the pore structure of the reservoir has the characteristics of partial high porosity and low permeability and low porosity and high permeability. It can be found that the correlation between porosity and permeability of carbonate reservoirs in the study area is not obvious.



**Figure 1.** Core electron microscope and cast photos of carbonate reservoir in the study area. (a) Moxi 17, Longwangmiao Formation, 4612.50–4612.61, fine-silt crystal sand debris dissolved pore dolomite,  $\times 20$ , single polarization; (b) Moxi 17, 4628.07–4628.34, medium-coarse grained sandstone lithic dissolved pore dolomite, intergranular dissolved pore development, partially filled asphalt,  $\times 20$ , single polarization; (c) Moxi 13, 4640.15–4640.30, fine-medium crystalline dissolved pore dolomite, the dissolved pores see the asphalt ring edge,  $\times 20$ ; (d) Moxi 13, 4606.28, fine-grained residual sand debris dissolved pore dolomite, cave development, single polarization,  $\times 20$ ; (e) Moxi 12, 4622.05–4622.21, medium-fine-grained residual sandy dolomite, intergranular solution pore development, casting body, single polarization,  $\times 20$ ; (f) Moxi 12, Longwangmiao Formation, 4636.82–4637.10, fine-grained sand-clastic dolomite, see mesh seam development; (g) Moxi 12, 4628.32–4628.50, Longwangmiao Formation, fine-powder crystal residual sandy dolomite, fracture, casting body, single polarization,  $\times 20$ .



**Figure 2.** Image of reservoir porosity and permeability relationship in the study area.

## Principle of SA-GA-XGBoost method

### Principle of XGBoost

Friedman first proposed the Gradient Boosting Decision Tree algorithm (GBDT) in 2001, which is an algorithm that combines, Decision-tree and gradient boosting, the algorithm constructs a weak classifier function by making the loss function obtained from each round of iteration descend along the gradient direction, and the results of multiple weak classifiers are combined with certain weights to form a strong classifier as the final prediction output<sup>18</sup>. XGBoost is an optimized algorithm of GBDT, which is characterized by the fact that the model can automatically utilize the CPU to perform multi-threaded parallel computation, which is faster, and the second-order expansion of the loss function by the Taylor's formula makes the prediction accuracy higher, and the addition of the regular term can constrain the descent of the loss function and the overall complexity of the model. The XGBoost objective function expression is as follows:

$$Obj = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum \Omega(f_k) \quad (1)$$

Where:  $l(y_i, \hat{y}_i)$  denotes the loss error,  $\Omega(f_k)$  denotes the regular term,  $\hat{y}_i$  is the model prediction value,  $y_i$  is the true value of the  $i$ th sample,  $k$  is the number of trees, and  $f_k$  is the model of the  $k$ th tree.

The expression for the regular term is specified as follows:

$$\Omega(f_k) = \gamma J + 1/2\lambda \|\omega\|^2 \quad (2)$$

Where:  $J$  is the number of leaf nodes of each classification regression tree;  $\omega$  is the sum of the weights of the leaf nodes of that tree;  $\gamma$  and  $\lambda$  are penalty coefficients, which can be adjusted in specific applications.

When the  $t$ th iteration is performed, the objective function is second-order Taylor expanded and the infinitesimal term is removed to obtain:

$$Obj = \sum_{i=1}^n [l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega f_t + C \quad (3)$$

Where:  $C$  is a constant term;  $g_i = \partial \hat{y}_i^{(t-1)} l(y_i, \hat{y}_i^{(t-1)})$ , denotes the first-order derivative of  $g_i = \partial \hat{y}_i^{(t-1)} l(y_i, \hat{y}_i^{(t-1)})$  with respect to  $\hat{y}_i^{(t-1)}$ ;  $h_i = \partial^2 \hat{y}_i^{(t-1)} l(y_i, \hat{y}_i^{(t-1)})$ , denotes the second-order derivative of  $l(\hat{y}_i^{(t-1)})$  with respect to  $\hat{y}_i^{(t-1)}$ .

Let  $G_j = \sum_{i \in I_j} g_j$ ,  $H_j = \sum_{i \in I_j} h_j$ , be brought into Eq. 3 and simplify the objective function to finally obtain:

$$Obj = -\frac{1}{2} \sum_{j=1}^T \frac{G_j^2}{H_j + \lambda} + \lambda J \quad (4)$$

### SA-GA parameter optimization

Since XGBoost has to set many empirical parameters in modeling, such as the number of leaf nodes in the regression tree, the learning rate and the regularization coefficient, etc., optimization is necessary to ensure the modeling quality. The simulated annealing algorithm (SA) can jump out of the local values to find the global optimal solution; the genetic algorithm (GA) has the characteristics of fast convergence and is not easy to overfitting. The SA-GA algorithm combines the two algorithms to make up for the shortcomings of the GA, which has the disadvantage of local self-optimization in the rapid convergence, and the SA, which has the disadvantage of slow convergence and prone to oscillation. It has the characteristics of stronger search

ability, faster convergence speed and better adaptability to complex problems, so that it can give full play to the advantages of the respective algorithms<sup>19–21</sup>. The specific steps are as follows:

- (1) Initialize the population. Randomly generate  $n$  individuals as the XGBoost regression tree volume, learning rate, maximum regression depth, regular term coefficient, the sum of the minimum sample weights of the leaves, and the minimum splitting gradient descent value, and set the maximum number of iterations as  $M_{max}$ , and set the annealing initial and cutoff temperatures.
- (2) Calculate the fitness function. The adaptation function in the algorithm is the error function, and its expression is:

$$f = \frac{1}{n} \sum_{i=1}^n (y_k - y)^2 \quad (5)$$

- (3) Selection, crossover, and mutation. Evaluate the population and select the excellent individuals based on the individual fitness, the expression is:

$$p_i = \frac{f_i}{\sum_i^n f_i} \quad (6)$$

- (4) Local update using Metropolis criterion. Simulated annealing operation is performed on the new individuals after crossover and mutation, and the expression is:

$$P(\Delta E, T) = e^{\frac{\Delta E}{K_b T}} \quad (7)$$

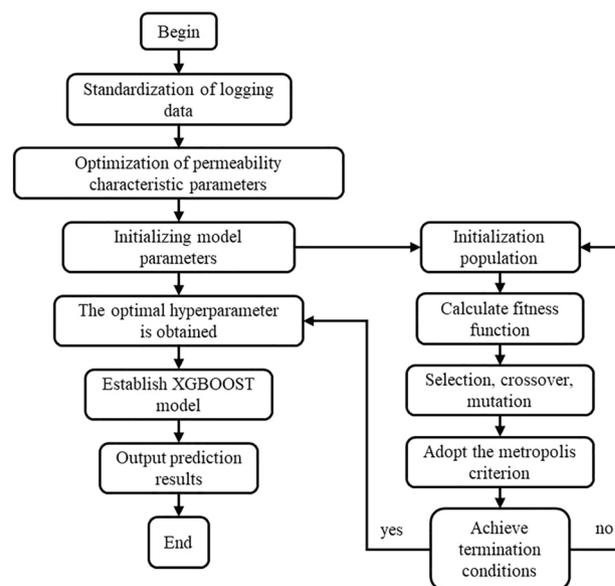
- (5) Iteration ends. The optimization search ends when the maximum number of iterations is reached, otherwise (1)-(4) are repeated.

The flowchart of the SA-GA-XGBoost method is shown in Fig.3.

## Analysis of application examples

### Standardization of logging data

Due to the different dimensions of different logging curves, the order of magnitude gap between the data is too large, which affects the calculation speed and accuracy of the model, so it is also necessary to preprocess the data before learning and training. In this paper, the standard deviation normalization formula is used to standardize



**Figure 3.** Flowchart of the SA-GA-XGBoost algorithm. Standardization of logging data: select the core analysis data as the training set. Optimization of permeability characteristic parameters: by calculating the correlation coefficient between logging curve and permeability, the characteristic parameters are optimized. Initialize model parameters: set the initial parameters of the model, and optimize the SA-GA algorithm to obtain the optimal parameters. Establish XGBoost model: according to the optimal parameters, establish the algorithm model, predict the test set, and obtain the prediction results.

all data samples (Eq. 8). Since the resistivity curve is a nonlinear logarithmic characteristic curve, it needs to be logarithmically transformed and then normalized.

$$X = \frac{x - \bar{x}}{\delta} \quad (8)$$

Where:  $X$  is the normalized data,  $x$  is the sample data,  $\bar{x}$  is the sample mean, and  $\delta$  is the sample data standard deviation.

### Permeability characterization parameter optimization

Since different logging sequences have certain responses to the logging permeability of carbonate reservoirs, such as natural gamma, neutron porosity, density, acoustic wave, resistivity, etc., and the interpretation conclusions by using single logging response has serious multi-resolution, the selection of logging characteristic parameters has a direct impact on the efficiency of the prediction model as well as the reliability of the results when establishing the permeability prediction model. In this paper, three correlation coefficients (Eq. 9, 10, 11), Pearson's coefficient, Kendall's coefficient, and Spearman's coefficient, which are commonly used in statistics, are prioritized to calculate the correlation between various types of logging curves and permeability<sup>22</sup>.

$$r_p = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \quad (9)$$

$$r_k = \frac{c - d}{\frac{1}{2}n(n - 1)} \quad (10)$$

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (11)$$

Where:  $r_p$ ,  $r_k$ ,  $r_s$  denote Pearson's coefficient, Kendall's coefficient, and Spearman's coefficient, respectively;  $x_i$ ,  $y_i$  are sample data, and  $\bar{x}$ ,  $\bar{y}$  are sample data averages;  $n$  is the number of sample data,  $c$  is the number of consistent values in the subjective and objective assessment values, and  $d$  is the number of values in which the subjective assessment values are not the same as the objective assessment values; and  $d_i$  is the number of rank differences in the paired variables.

Using three kinds of correlation coefficients to calculate the input parameters, the correlation coefficients  $|r|$  of each logging response characteristics and permeability were obtained, and the results are shown in Table 1, the correlation coefficients  $|r|$  are in the range of 0–1, and the larger value indicates the higher degree of correlation.

The correlation results between logging response and permeability calculated by the three correlation coefficients are basically the same, according to the principle that the correlation coefficient is greater than 0.5, this paper selects five permeability-sensitive logging curves of CNL, DEN, DT, In RT and GR as inputs, and integrally selects 10 core wells in carbonate rocks of the Longwangmiao Formation of Muxi block in Anyue Gas Field, with a total of 200 core data, of which 80% of the total 160 samples constitute the model training set, and 20% of the total 40 samples constitute the test set to establish the permeability logging prediction model.

### Constructing the SA-GA-XGBoost prediction model

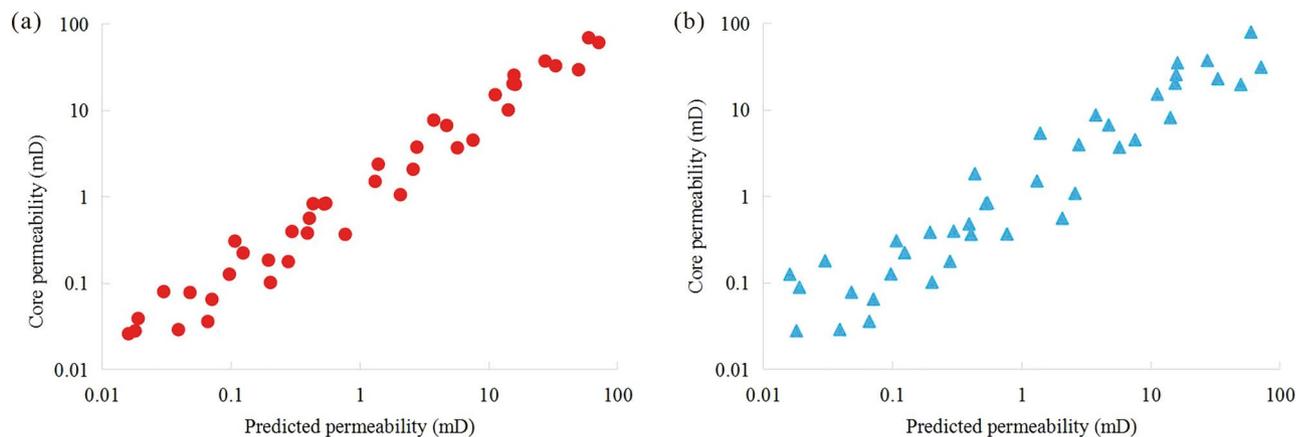
In the XGBoost algorithm model, the setting of various parameters will affect the prediction accuracy and convergence effect of the model. In this paper, the initial parameters of the model are set as follows, the regression tree volume  $k=100$ , the learning rate  $=0.01$ , the maximum regression depth  $\text{max-depth}=3$ , the regular term coefficient  $=0.1$ , the sum of the minimum sample weight of the leaf  $\text{min-chile-weight}=0.1$ , and the minimum splitting gradient descent. When using SA-GA to optimize the XGBoost parameters, the number of iterations is set to 200, the number of populations is set to 50, and the temperature reduction parameter is set to 0.98. When the number of iterations is greater than 120, the optimization results are also close to the global optimal solution, and the parameter optimization takes 0.5 h. The optimal parameters of the XGBoost are finally obtained as shown in Table 2.

Logging response	Pearson's coefficient	Kendall's coefficient	Spearman's coefficient	Sorting
CNL	0.8326	0.8207	0.8521	1
DEN	0.8065	0.7946	0.8173	2
DT	0.7534	0.7345	0.7832	3
InRT	0.6858	0.6927	0.7178	4
GR	0.6119	0.6592	0.6233	5
InRXO	0.4902	0.4571	0.4818	6
SP	0.4269	0.3865	0.4056	7
CAL	0.3671	0.3747	0.3544	8

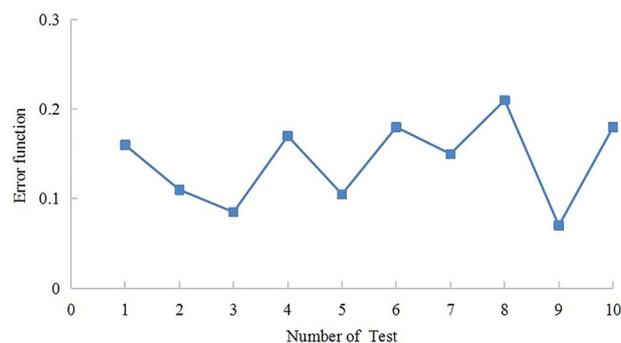
**Table 1.** Results of different correlation coefficient analysis.

Parameter name	Regression tree volume $k$	Learning rate $\eta$	Maximum regression depth $max\_depth$	Regular term coefficient (math.) $\lambda$	Sum of minimum sample weights of leaves $min\_chile\_weight$	Minimum splitting gradient value $\delta$
Optimum value	500	0.1	5	1.1	1.5	0

**Table 2.** Optimal XGBoost parameters after SA-GA optimization.



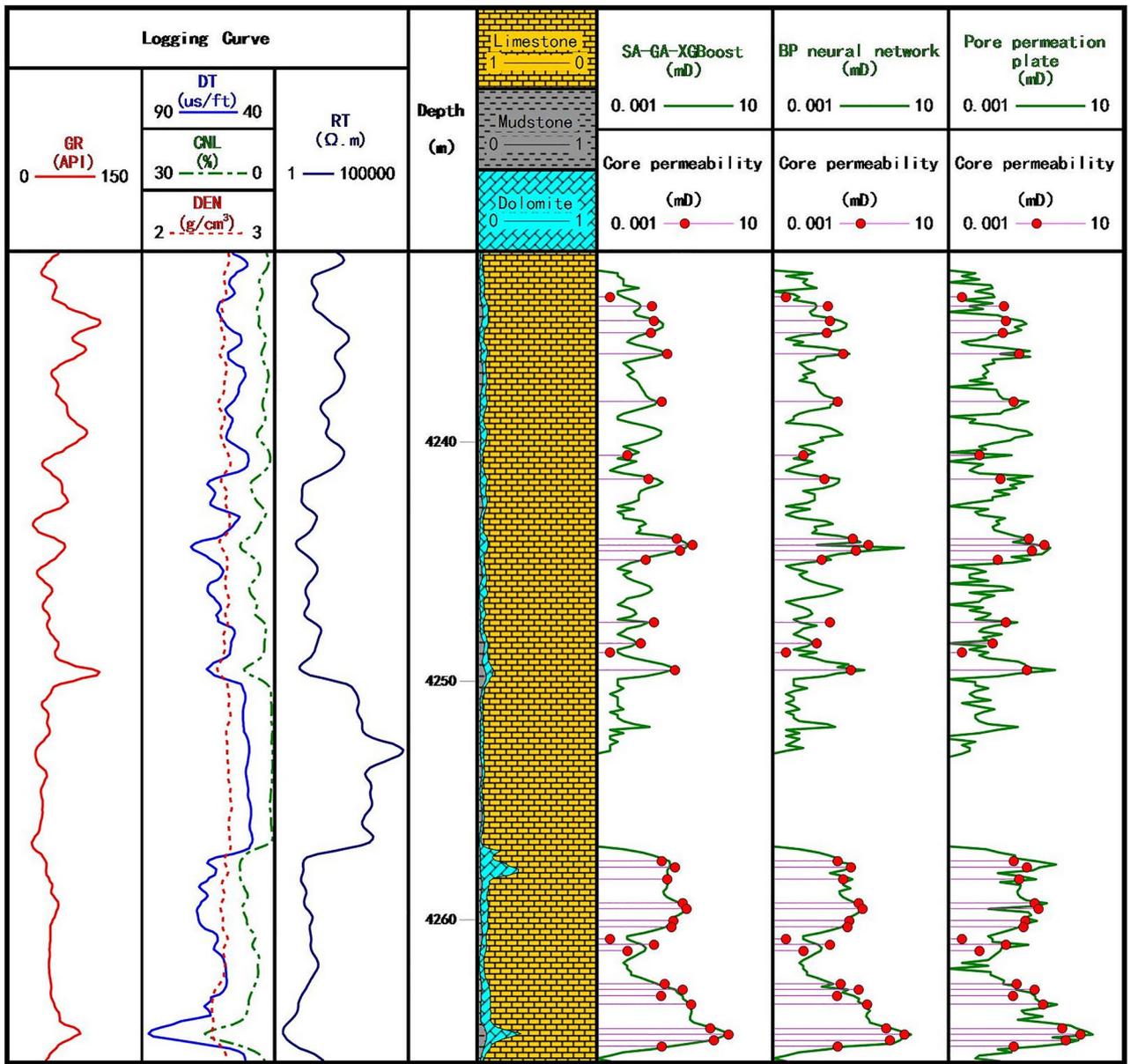
**Figure 4.** Prediction results of permeability evaluation model. (a) SA-GA-XGBoost model; (b) XGBoost model.



**Figure 5.** The result of 10-fold cross-validation.

### Model results validation analysis

Using the established SA-GA-XGBoost model, 40 samples of the test set are predicted, and the prediction results are shown in Fig. 4a, the correlation coefficient  $R^2$  between the core permeability and the predicted permeability reaches 0.94, and meanwhile, in order to validate the prediction effect of the model, The adjusted  $R^2$  and root mean square error (RMSE) commonly used in statistics are selected (Eq. 12, 13). The adjusted  $R^2$  is 0.887 and RMSE is 0.132. In order to compare the prediction effect of the model after parameter optimization, this paper also uses the conventional XGBoost algorithm and same data to model and predict, and the results are shown in Fig. 4b, with the correlation coefficient  $R^2$  of 0.88 and the adjusted  $R^2$  is 0.774 and RMSE is 0.215. It is not difficult to see that after entering the model after SA-GA parameter optimization, the adjusted  $R^2$  is increased by 0.113, RMSE is increased by 0.083, and the accuracy of prediction results is significantly improved. In order to verify the extrapolation ability of the SA-GA-XGBoost model, this paper uses the ten-fold cross-validation method to validate all the sample data, the data set was randomly divided into 10 copies, one of which was selected as the test set, and the remaining 9 were used as the training set for 10 training and testing<sup>23</sup>, and the results are shown in Fig. 5, the average value of the error mean of the SA-GA-XGBoost model is 0.15, indicating that the model has certain generalization ability and effectiveness. and the prediction accuracy meets the needs of logging interpretation.



**Figure 6.** Comparison of permeability model prediction in well M8.

$$AdjustedR^2 = 1 - \left[ \frac{(1 - R^2)(n - 1)}{n - k - 1} \right] \tag{12}$$

Where:  $R^2$  is the model coefficient of determination;  $n$  is the number of samples;  $k$  is the number of predictors.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \tag{13}$$

Where:  $y_i$  denotes the true value of permeability;  $\hat{y}_i$  denotes the predicted value of permeability;  $n$  denotes the total number of samples of the training set data.

**Comparison of model application**

The established SA-GA-XGBoost model is practically applied to the prediction of logging permeability of carbonate reservoir in M8 well in the study area, in order to compare the prediction effect, the traditional pore-permeability plate method and the BP neural network are used to predict the permeability of M8 well, the prediction results of the three methods are shown in Fig. 6. It can be seen that all the three prediction methods can reflect the change of reservoir permeability. However, the prediction accuracy is significantly

different. The prediction results of BP neural network and SA-GA-XGBoost models are in good agreement with the core permeability. There is a significant deviation between the low and high permeability values in the BP permeability evaluation model, while the SA-GA-XGBoost model has a good application effect in the whole well section, and the error is small, especially in the low permeability interval, the advantage is more obvious. The adjusted  $R^2$  and RMSE of the three models were calculated respectively. The adjusted  $R^2$  of the porosity and permeability plate method was 0.672, and the RMSE was 0.364. The adjusted  $R^2$  of the BP neural network was 0.738, and the RMSE was 0.253. The adjusted  $R^2$  of the SA-GA-XGBoost model was 0.876, and the RMSE was 0.142. Therefore, the SA-GA-XGBoost permeability evaluation model can fully reflect the difference of logging response characteristics. The model has strong stability and high calculation accuracy. It is more suitable for the permeability evaluation of carbonate reservoirs with complex pore structure and strong reservoir heterogeneity.

## Conclusion

For the evaluation of logging permeability of Cambrian Longwangmiao Formation carbonate reservoirs in the Moxi block in central Sichuan, an XGBoost carbonate reservoir permeability prediction model based on SA-GA optimization is proposed from the optimization of model parameters, and it has been applied well in the study area, and meanwhile, the following understandings have been achieved:

- (1) The correlation between different logging response characteristics and permeability is calculated by three correlation coefficients, and five permeability-sensitive logging curves of CNL, DEN, DT,  $\ln RT$ , and GR with higher correlation are identified as input parameters of the model, which simplifies the model calculations and improves the prediction accuracy at the same time.
- (2) Using simulated annealing genetic algorithm, the parameter of XGBoost model is optimized to improve the generalization ability of the prediction model and the reliability of the results, and the prediction effect meets the actual needs of the study area.
- (3) Compared with the traditional network model, the SA-GA-XGBoost model has the advantages of high precision and strong generalization ability. In the future, it can also be applied to carbonate fracture identification and other aspects to provide more accurate and efficient reservoir parameter prediction for oil and gas exploration.

## Data availability

All data generated or analysed during this study are included in this published article.

Received: 20 February 2025; Accepted: 22 April 2025

Published online: 28 April 2025

## References

1. Wang, X. & Fan, T. Progress of research on permeability of carbonate rocks. *Earth Sci. Front.* **20**(5), 166–174 (2013).
2. Li, X. et al. Dolomite reservoir characteristics and permeability evaluation methods: An example from the Paleogene, Mesopotamian basin, Iraq. *China Offshore Oil Gas* **36**(3), 81–94. <https://doi.org/10.11935/j.issn.1673-1506.2024.03.008> (2024).
3. Luo, D. et al. The application of the nmr logging in the reservoir evaluation in western sichuan gas field. *Well Logg. Technol.* **47**(05), 585–591. <https://doi.org/10.16489/j.issn.1004-1338.2023.05.008> (2023).
4. Jiang, W. et al. Permeability calculation method of deep tight sandy conglomerate reservoir. *Well Logg. Technol.* **46**(01), 102–108. <https://doi.org/10.16489/j.issn.1004-1338.2022.01.017> (2022).
5. Subasi, A., El-Amin, M. F., Darwich, T. & Dossary, M. Permeability prediction of petroleum reservoirs using stochastic gradient boosting regression. *Ambient Intell. Hum. Comput.* **13**, 3555–3564. <https://doi.org/10.1007/s12652-020-01986-0> (2022).
6. Shao, G. et al. The phase control modeling of porous carbonate reservoir by machine learning for the cretaceous mishrif formation reservoir of h oilfield in the middle east. *Mar. Geol. Front.* **39**(11), 76–85. <https://doi.org/10.16028/j.1009-2722.2022.221> (2023).
7. Xu, P. et al. Permeability prediction of carbonate reservoirs by conventional logging with adaptive model. *Oil Geophys. Prospect.* **57**(5), 1192–1203. <https://doi.org/10.13810/j.cnki.issn.1000-7210.2022.05.021> (2022).
8. Kamali, M. Z. et al. Permeability prediction of heterogeneous carbonate gas condensate reservoirs applying group method of data handling. *Mar. Pet. Geol.* <https://doi.org/10.1016/j.marpetgeo.2022.105597> (2022).
9. Anifowose, F., Abdurraheem, A. & Al-Shuhail, A. A parametric study of machine learning techniques in petroleum reservoir permeability prediction by integrating seismic attributes and wireline data. *J. Pet. Sci. Eng.* **176**, 762–774. <https://doi.org/10.1016/j.petrol.2019.01.110> (2019).
10. Hou, X., Wang, F., Zai, Y. & Lian, P. Prediction of carbonate porosity and permeability based on machine learning and logging data. *J. Jilin Univ. (Earth Sci. Ed.)* **52**(02), 644–653. <https://doi.org/10.13278/j.cnki.jjuese.20210151> (2022).
11. Yang, J., Yang, C., Zhang, Y., Cui, L. & Wang, L. Permeability prediction method based on improved bp neural network. *Lithol. Reserv.* **23**(1), 98–102. <https://doi.org/10.3969/j.issn.1673-8926.2011.01.017> (2011).
12. Wang, X. et al. Application of c5.0 decision-tree algorithm to interpret permeability of carbonate reservoir. *Well Logg. Technol.* **44**(3), 300–304. <https://doi.org/10.16489/j.issn.1004-1338.2020.03.016> (2020).
13. Gu, Y., Z. D. Y. & Bao, Z. Permeability prediction using pso-xgboost based on logging data. *Oil Geophys. Prospect.* **56**(1), 26–37. <https://doi.org/10.13810/j.cnki.issn.1000-7210.2021.01.003> (2021).
14. Zhang, Y., Zhang, C., Sun, K., Yang, W. & Wang, M. Logging interpretation model on complex carbonate reservoir permeability based on hybrid simulated annealinggenetic algorithm-random forest algorithm. *Pet. Geol. Recov. Effic.* **29**(01), 53–61. <https://doi.org/10.13673/j.cnki.cn37-1359/te.2022.01.007> (2022).
15. Han, R., Gao, Y. & Zhang, Y. Intelligent prediction method for permeability of layered phase controlled carbonate reservoirs based on BP neural network. *China Offshore Oil Gas* **36**(01), 100–108. <https://doi.org/10.11935/j.issn.1673-1506.2024.01.010> (2024).
16. Sun, Y. et al. Identification of complex carbonate lithology by logging based on XGboost algorithm. *Lithol. Reserv.* **32**(4), 98–106. <https://doi.org/10.12108/xyyqc.20200410> (2023).
17. Pan, S., Zheng, Z., Lei, J. & Wang, Y. Porosity prediction of sandstone reservoirs based on hybrid optimization XGboost algorithm. *Comput. Appl. Softw.* **40**(05), 103–109. <https://doi.org/10.3969/j.issn.1000-386x.2023.05.015> (2023).
18. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29**(5), 1189–1232 (2001).
19. Qu, Z., Zhang, X., Cao, Y., Liu, X. & Feng, X. Research on genetic algorithm based on adaptive mechanism. *Appl. Res. Comput.* **32**(11), 3222–3225. <https://doi.org/10.3969/j.issn.1001-3695.2015.11.004> (2015).

20. Li, Q. et al. A method for evaluating complete pore-throat structure of carbonate rocks based on digital cores: A case study of qixia formation in northwest sichuan. *Pet. Geol. Recov. Effic.* **28**(150(03)), 53–61. <https://doi.org/10.13673/j.cnki.cn37-1359/te.2021.03.006> (2021).
21. Sun, J., Liu, p, Li, H. & Chen, P. Solving mtsp with two-stage sa and ga based on spark. *J. Zhengzhou Univ. (Eng. Sci.)* **45**(4), 62–69. <https://doi.org/10.13705/j.issn.1671-6833.2024.01.019> (2024).
22. Chen, H. et al. Support vector machine-based initial productivity prediction for srv of horizontal wells in tight oil reservoirs. *China Offshore Oil Gas* **34**(01), 102–109. <https://doi.org/10.11935/j.issn.1673-1506.2022.01.012> (2022).
23. Al-Mudhafar, W. J. Incorporation of bootstrapping and cross-validation for efficient multivariate facies and petrophysical modeling. *SPE Low Perm. Sympos.* <https://doi.org/10.2118/180277-MS> (2016).

### Author contributions

H, C.B.: Conceptualization, Methodology, Software, Visualization, Data Curation, Formal analysis, Writing-Original draft preparation; Z, X.Y.: Investigation, Validation, Methodology, Supervision, Writing - Review & Editing; L, M.Y.: Conceptualization, Methodology, Formal analysis, Writing-Review and Editing, Supervision; Z, Y.L.: Formal analysis, Visualization, Validation; Y, S.B.: Investigation, Validation, Supervision. All authors reviewed the manuscript.

### Funding

This research was supported by the National Natural Science Foundation of China (Grant No. 41672206) and the National Science and Technology Major Project of the Ministry of Science and Technology of China (Grant No. 2024ZD1003305).

### Declarations

### Competing interests

The authors declare no competing interests.

### Additional information

**Correspondence** and requests for materials should be addressed to C.H.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025