Hindawi
Geofluids
Volume 2021, Article ID 6155663, 18 pages
https://doi.org/10.1155/2021/6155663

Research Article

# Total Organic Carbon Content Prediction in Lacustrine Shale Using Extreme Gradient Boosting Machine Learning Based on Bayesian Optimization

**Xingzhou Liu, Zhi Tian, and Chang Chen**

*Research Institute of Exploration and Development, Liaohe Oilfield Company, Petrochina, Panjin 124010, China*

Correspondence should be addressed to Zhi Tian; 478131932@qq.com

Received 4 June 2021; Revised 14 August 2021; Accepted 16 August 2021; Published 7 October 2021

Academic Editor: Mehdi Ostadhassan

The total organic carbon (TOC) content is a critical parameter for estimating shale oil resources. However, common TOC prediction methods rely on empirical formulas, and their applicability varies widely from region to region. In this study, a novel data-driven Bayesian optimization extreme gradient boosting (XGBoost) model was proposed to predict the TOC content using wireline log data. The lacustrine shale in the Damintun Sag, Bohai Bay Basin, China, was used as a case study. Firstly, correlation analysis was used to analyze the relationship between the well logs and the core-measured TOC data. Based on the degree of correlation, six logging curves reflecting TOC content were selected to construct training dataset for machine learning. Then, the performance of the XGBoost model was tested using $K$-fold cross-validation, and the hyperparameters of the model were determined using a Bayesian optimization method to improve the search efficiency and reduce the uncertainty caused by the rule of thumb. Next, through the analysis of prediction errors, the coefficient of determination ($R^2$) of the TOC content predicted by the XGBoost model and the core-measured TOC content reached 0.9135. The root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) were 0.63, 0.77, and 12.55%, respectively. In addition, five commonly used methods, namely, $\Delta \log R$ method, random forest, support vector machine, $K$-nearest neighbors, and multiple linear regression, were used to predict the TOC content to confirm that the XGBoost model has higher prediction accuracy and better robustness. Finally, the proposed approach was applied to predict the TOC curves of 20 exploration wells in the Damintun Sag. We obtained quantitative contour maps of the TOC content of this block for the first time. The results of this study facilitate the rapid detection of the sweet spots of the lacustrine shale oil.

## 1. Introduction

Recently, unconventional shale oil and gas reservoirs have profoundly revolutionized the energy industry in North America and China [1, 2]. Unlike marine shales, the Bohai Bay Basin in northeast China mainly develops lacustrine shale plays, with frequent changes in the sedimentary environment and strong reservoir inhomogeneity. The accurate and efficient identification of the sweet spots in thin shale plays is a hot research issue. Studies found that the exploration potential of shale oil is primarily associated with three factors: the hydrocarbon generation potential, reservoir capacity, and recoverability [3]. Organic matter is an essential material for determining the hydrocarbon generation potential and hydrocarbon enrichment [4]. The total organic carbon (TOC) content is a key indicator to evaluate organic matter abundance [5]. An accurate TOC value is typically obtained from Rock-Eval pyrolysis of the rock core sample; however, drilling to obtain core samples is time-consuming and expensive, resulting in the discontinuous and nonuniform distribution of core-measured TOC data. Moreover, the thickness of organic-rich lacustrine shale plays is usually very small; thus, it is unreasonable to use discrete core-measured TOC data points to evaluate the hydrocarbon generation potential. Well logs have high resolution and provide continuous data. The variation of the organic matter content affects the petrophysical properties of the formation, such as radioactivity, resistivity, and density (DEN), forming a unique logging response; therefore, the TOC curve can be predicted using well logs [6].

At present, methods using well logs to predict the TOC content include statistical correlation, overlapping methods, multiple regression, and machine learning. Beers first proposed using the natural gamma radioactivity intensity to evaluate the TOC content [7]. Subsequently, many scholars established empirical relationship equations between the natural gamma (GR) logs and TOC in different areas [8, 9]. Swanson found that the radioactivity of organic matter was mainly related to the adsorption amount of uranium (U) in the formation. Thus, researchers predicted the TOC content using the GR ray spectrum logs [10, 11], such as establishing a linear correlation between the TOC content and the U log [12] or establishing a multivariate statistical relationship between the TOC content and the U log combined with the thorium (Th)/U ratio log [13]. Schmoker found that the main reason for a decrease in the DEN of an organic-rich formation was the increase in the organic matter content; therefore, a regression relationship was established between the DEN log and TOC content [14]. Herron proposed a method to determine the TOC content using the carbon-oxygen ratio log [15]. Passey et al. proposed the $\triangle\log R$ method [16], which overlaps the porosity logs with the deep resistivity (RD) log and uses the nonsource rock zone as the baseline to establish an empirical relationship formula between the TOC content and the well logs. Subsequently, many scholars proposed improved methods based on the $\triangle\log R$ method [17–20]. In recent years, the emergence of special logging methods has provided many approaches to predict TOC content. Examples include calculating the TOC content using element capture spectroscopy logs [21] or combining the nuclear magnetic resonance logs and the DEN log to estimate TOC content [22]. All the above methods are developed based on the rock physical model (RPM) and rely extensively on empirical formulas. Due to the third artificial intelligence boom, machine learning has been widely used for lithology identification [23–25] and reservoir evaluation [26, 27]. Machine learning methods for TOC content prediction include support vector machine (SVM) [28, 29], Gaussian process regression (GPR) [30, 31], extreme learning machine (ELM) [32, 33], neural network [34, 35], fuzzy clustering [36], and random forest (RF) [37]. Machine learning is data-driven, which improves the accuracy and efficiency of TOC prediction compared to conventional methods.

Practically, most of the core samples that can be used as machine learning samples are concentrated in key reservoir zones. However, few labeled data points exist in nonreservoir zones, leading to a significant imbalance in training samples. When individual models are used to optimize the objective function, it is easy to fall into local minima, and these models have poor generalization ability. Ensemble learning can effectively solve this problem by training multiple models and taking advantage of the composite output. The individual models are used to create an optimal predictive model, which provides higher prediction accuracy than an individual model. A popular example of an ensemble model is RF, which has been used for seismic reservoir prediction [38], lithology identification [39], and hydrocarbon source rock prediction [40]. However, RF is based on the bagging technique and is sensitive to noise and prone to overfitting when performing regression prediction. In contrast, the gradient boosting decision tree (GBDT) is based on the boosting technique and generally performs better for regression problems. Chen et al. first proposed the extreme gradient boosting (XGBoost) method based on GBDT [41]. Unlike the GBDT algorithm which utilizes first-order derivative information, XGBoost carries out a second-order Taylor expansion on the loss function and contains a regular term in the objective function to find the optimal solution to avoid overfitting, making the method highly efficient, flexible, and portable. Yan et al. applied XGBoost to well logging interpretation of tight sandstone and found that it performed better for fluid identification than the SVM and RF models [27]. Nguyen et al. used XGBoost for predicting compressional and shear waves in micritic limestones and achieved higher accuracy than an artificial neural network (ANN) and SVM [42]. Gu et al. used a particle swarm optimization (PSO) algorithm to determine the hyperparameters of the XGBoost algorithm and applied XGBoost to predict the permeability of tight sandstone [43]. To date, the XGBoost model has not been applied to the TOC prediction of reservoirs. Therefore, in this study, a workflow consisting of XGBoost machine learning based on Bayesian optimization for TOC prediction is proposed and applied to lacustrine shale oil in the Bohai Bay Basin. The prediction results are compared with the $\triangle\log R$ method and other typical machine learning methods to demonstrate the accuracy and reliability of the proposed method.

## 2. Theory of Machine Learning

*2.1. Theory of the XGBoost Model.* XGBoost is an ensemble boosting algorithm that consists of multiple decision tree iterations. It is an improvement of the GBDT. Multiple classification and regression tree (CART) models are first constructed to make predictions using the dataset; these trees are then combined into a new tree model. The models are continuously and iteratively enhanced, with each iteration generating a new tree model that fits the residuals of the previous tree. As more trees are added, the complexity of the ensemble model becomes progressively higher until it approaches the complexity of the data itself; thus, training achieves optimal results [41]. If there are $K$ regression trees, the expression of the prediction function is defined as

$$y_i\wedge = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F, \tag{1}$$

where $f_k$ is the $k^{\text{th}}$ regression tree, $F$ represents the set of CARTs, and $y_i\wedge$ is the predicted value of the $i^{\text{th}}$ sample.

The loss function $L$ is represented by the predicted value $y_i\wedge$ and the true value $y_i$:

$$L = \sum_{i=1}^{n} l(y_i, y_i\wedge), \tag{2}$$

where $n$ is the number of samples.

The prediction accuracy of the model is jointly determined by the deviation and the variance. The loss function represents the deviation of the model, and the variance is determined by the regular term $\Omega$ that suppresses the complexity of the model. Therefore, the objective function Obj can be defined as

$$\text{Obj} = \sum_{i=1}^{n} l(y_i, y_i\wedge) + \sum_{k=1}^{K} \Omega(f_k), \tag{3}$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2, \tag{4}$$

where $T$ represents the number of leaf nodes, $\omega$ is the leaf weight value, $\gamma$ is the penalty factor of the leaf tree, and $\lambda$ is the leaf weight penalty factor.

XGBoost uses a gradient boosting strategy where the newly generated regression tree needs to fit the residuals of the last prediction. The objective function at the $t^{\text{th}}$ iteration can be rewritten as

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, y\wedge_i^{t-1} + f_i(x_i)\right) + \Omega(f_t) + \mathscr{C}. \tag{5}$$

A Taylor expansion is performed on the objective function to obtain

$$L^{(t)} \cong \sum_{I=1}^{n} \left[ 1\left(y_i, y\wedge_i^{t-1}\right) + g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i) \right] + \Omega(f_t), \tag{6}$$

where $g_i = \partial_{y\wedge_i^{t-1}} l(y_i, y\wedge_i^{t-1})$ is the first-order derivative of the loss function and $h_i = \partial^2_{y\wedge_i^{t-1}} l(y_i, y\wedge_i^{t-1})$ is the second-order derivative of the loss function.

Therefore, it is only necessary to calculate the $g_i$ and $h_i$ values of the loss function for each step and optimize the objective function to obtain $f(x)$ for each step. Finally, an optimal ensemble model is obtained based on the additive method.

*2.2. Bayesian Optimization of the Hyperparameters.* When a machine learning model is established, the hyperparameters need to be determined in advance. The selection of the hyperparameters has a significant impact on prediction accuracy. Therefore, it is important to obtain the optimal combination of hyperparameters. The optimization of the hyperparameters is a typical black-box optimization problem. Commonly used optimization methods include grid search (GS), random search (RS), genetic algorithm (GA), PSO, and Bayesian optimization [44]. The GA and PSO algorithms require a sufficient number of initial sample points and are not very efficient for optimization. At present, GS, RS, and Bayesian optimization are the most common methods. The GS method needs to traverse all possible parameter combinations, which is very time-consuming for a large data volume and many hyperparameter dimensions. In contrast, the RS randomly samples the hyperparameters in a certain range and selects them by comparing the performance of different combinations, which does not guarantee that the optimal combination will be obtained. Moreover, the GS and RS are computed independently for each hyperparameter combination. The current computation does not use the result of the searched points, but this information guides the search process and can improve the quality of the results and the search speed. In contrast, Bayesian optimization selects the most promising hyperparameters by evaluating the past results, enabling the selection of the appropriate hyperparameters with fewer iterations than the RS method [45, 46]. Theoretically, Bayesian optimization solves the global optimal solution of the objective function:

$$x^* = \underset{x\in X}{\operatorname{argmin}} f(x), \tag{7}$$

where $x$ denotes the hyperparameters to be optimized, $X$ is the set of hyperparameters to be optimized, $f(x)$ is the objective function, and $x^*$ is the optimal combination of hyperparameters. The core of the Bayesian optimization algorithm consists of two parts: first, the posterior probability distribution is calculated based on past results using GPR to obtain the expected mean and variance of the hyperparameters at each sampling point. Second, an acquisition function is constructed to determine the next sampling point based on the posterior distribution.

*2.2.1. Gaussian Process.* The Gaussian process (GP) is a generalization of the multivariate Gaussian probability distribution defined by the mean function $m(x)$ and the covariance function $k(x, x')$.

$$m(x) = E[f(x)], \tag{8}$$

$$k\left(x, x'\right) = E\left[(f(x) - m(x))\left(f\left(x'\right) - m\left(x'\right)\right)\right]. \tag{9}$$

The GP can be expressed as

$$f(x) \backsim \left(\text{GP}\left(m(x), k\left(x, x'\right)\right)\right). \tag{10}$$

For convenience in practical applications, let the prior mean function be 0. There exists a Gaussian distribution satisfying

$$p(f \mid (\mathbf{X}, \theta) = N(0, \mathbf{K}(\mathbf{X}, \mathbf{X})). \tag{11}$$

The covariance matrix $\mathbf{K}(\mathbf{X}, \mathbf{X})$ can be expressed as

$$\mathbf{K}(\mathbf{X}, \mathbf{X}) = \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix}. \tag{12}$$

The corresponding covariance function can be expressed as

$$k\left(x, x'\right) = \exp\left(-\frac{\left|x - x'\right|^2}{2}\right). \tag{13}$$

According to the nature of the GP, after adding the sample $X*$ to be predicted, the new Gaussian distribution can be expressed as

$$\begin{pmatrix} f \\ f_* \end{pmatrix} \backsim N\left(0, \begin{bmatrix} \mathbf{K}(\mathbf{X}, \mathbf{X}) & \mathbf{K}(\mathbf{X}, \mathbf{X}_*) \\ \mathbf{K}(\mathbf{X}_*, \mathbf{X}) & \mathbf{K}(\mathbf{X}_*, \mathbf{X}_*) \end{bmatrix}\right). \tag{14}$$

Then, the joint posterior distribution of $f_*$ is

$$p(f_* \mid (\mathbf{X}, f, \mathbf{X}_*) = N(\langle f_* \rangle, \operatorname{cov}(f_*)). \tag{15}$$

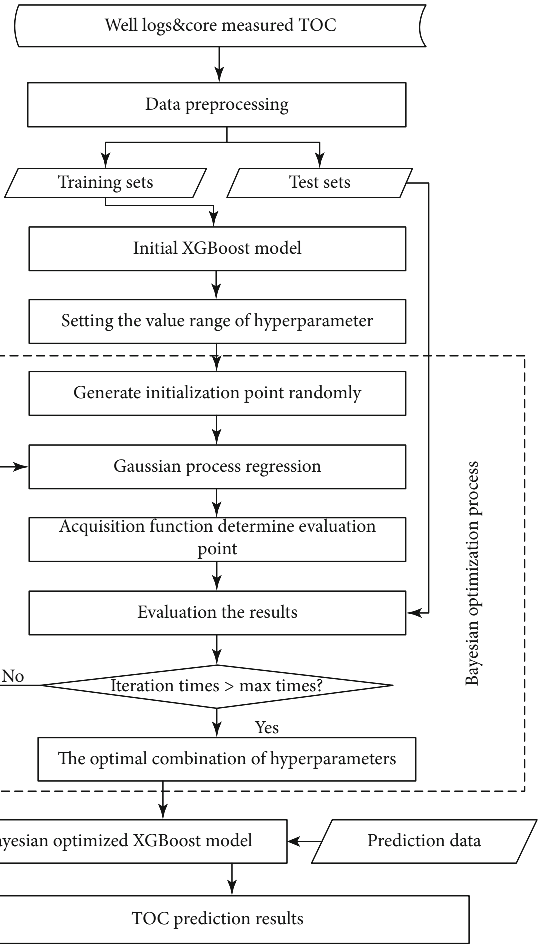
By evaluating the mean and covariance matrices, $f_*$ can be sampled from the joint posterior distribution.

*2.2.2. Acquisition Functions.* The acquisition function determines the next sample point based on the posterior results of the probabilistic agent model. Usually, the selection of sample points for the acquisition function requires both exploring new areas in the objective space and exploiting areas that are already known. The exploitation refers to searching for the global optimal solution based on the current optimal solution to improve the mean value of the objective function. The exploration refers to detecting the unevaluated sample points to reduce the uncertainty of the objective function. When the GP is used as the probabilistic agent model, the four commonly used acquisition functions include probability of improvement (PI), entropy search (ES), upper confidence bound (UCB), and expected improvement (EI) [45]. In this paper, the EI is chosen as the acquisition function; its mathematical expression is

$$a_{\mathrm{EI}}(x) = E\left[\max\left(0, f' - f(x) \mid x, \mathscr{D}\right], \tag{16}$$

where $\mathscr{D} = (\mathbf{X}, \theta)$ represent the observations and $f'$ is the minimum value of the current observation of $f$.

*2.3. TOC Prediction Process.* The flowchart of the TOC prediction based on the Bayesian optimization XGBoost model is shown in Figure 1. It contains three parts, namely, data preprocessing, model building, and model application, which are described as follows.



Well logs&core measured TOC
Data preprocessing
Training sets
Test sets
Initial XGBoost model
Setting the value range of hyperparameter
Generate initialization point randomly
Gaussian process regression
Acquisition function determine evaluation point
Evaluation the results
No
Iteration times > max times?
Yes
The optimal combination of hyperparameters
Bayesian optimization process
Bayesian optimized XGBoost model
Prediction data
TOC prediction results

FIGURE 1: Flowchart of TOC prediction using the Bayesian optimization XGBoost model.

(1) Data preprocessing: we first collect the core-measured TOC data and the corresponding well log data. The data are depth-corrected, outlier-processed, and normalized, and then, the well logs relevant to the TOC prediction are selected as machine learning input features using linear regression cross-plots and Pearson correlation coefficient techniques. Finally, the processed data are randomly divided into a training set and a test set using an appropriate rule

(2) Model building: we establish the initial XGBoost model and then optimize the hyperparameters of the model using the Bayesian optimization algorithm

(3) Model application: the optimal XGBoost model is applied to the unused well logs to predict the TOC content

## 3. Geology Settings and Data Analysis

*3.1. Study Area.* The Damintun Sag is located in the northern part of Liaohe Depression in the Bohai Bay Basin in northeast China, covering a region of about $800\,\mathrm{km}^2$ (Figure 2(a)). It is a Mesozoic-Cenozoic continental sedimentary sag developed in the basement of the Archean metamorphic rock and Proterozoic carbonate rock. Structurally, it has an irregular triangular shape that is wider in

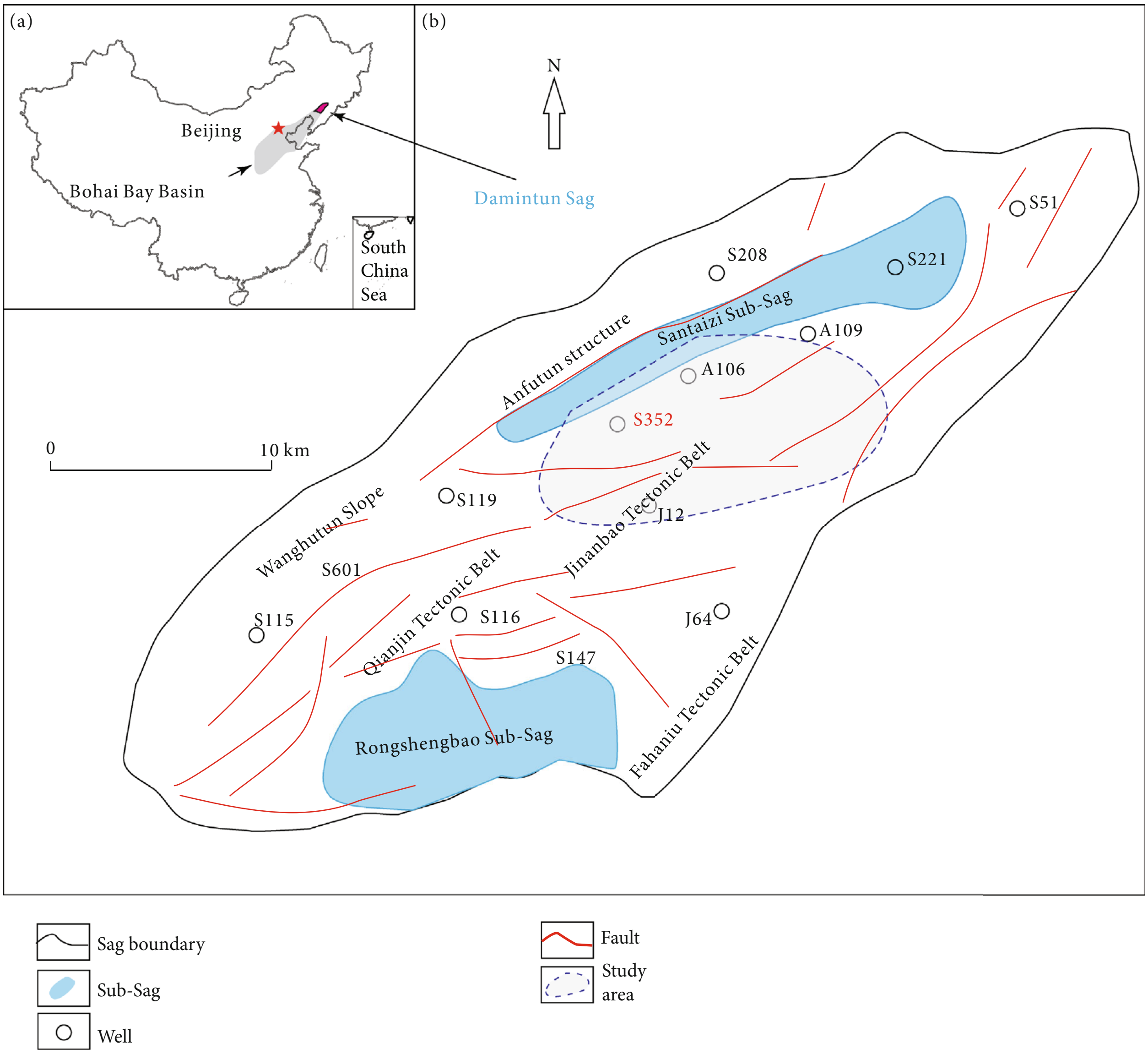


Figure 2: Location of the study area.

the south and narrower in the north and bounded by three major faults. The main source rocks are the oil shale plays of the fourth member and the dark mudstone plays of the third member in the Paleogene Shahejie Formation ($E_2S_4$, $E_2S_3$). The target formation of this study is the lower submember of $E_2S_4$ ($E_2s_4^L$), located in the central area of the Damintun Sag (Figure 2(b)), covering an area of about 200 km$^2$. Among 137 wells drilled in the $E_2s_4^L$ formation, favorable oil and gas conditions were observed in 53 wells, and 4 wells provide industrial oil production.

During the sedimentation of $E_2s_4^L$, the lake level oscillation caused by tectonic movement has led to cyclic changes in the sedimentary environments, and the lithology of the formation shows "sandwich" characteristics (Figure 3). The upper part is the $E_2s_4^L$-I group, characterized by dark oil shales, and thin-bedded sandstone is locally observed. The middle part is the $E_2s_4^L$-II group, which is composed of siltstone and argillaceous dolomite. The lower part is the $E_2s_4^L$-III group, characterized by intercalated oil shales, marl, and dolomite. The total thickness of $E_2s_4^L$ ranges from 20 m to 220 m, the TOC content ranges from 2% to 12.8%, $R_0$ ranges from 0.4 to 0.6%, and the organic matter is mainly type I, with some types $II_1$ and $II_2$. The hydrocarbon generation intensity is about $4200 \times 10^4$ t/km$^2$. The latest prediction showed that the $E_2s_4^L$ formation has $20.9 \times 10^8$ t hydrocarbon resources, demonstrating significant potential for shale oil exploration [47].

*3.2. Data Analysis.* The data used in the study is from the key exploration well S352 and consists of well log data and core-measured TOC data. Well S352 was drilled from 3150 to 3352 m to encounter $E_2s_4^L$ formation, and 145.92 m of sealed coring was completed at depths of 3169-3348.97 m, obtaining a core length of 122.47 m, with a core recovery rate of 83.9%. A total of 107 experimental core samples were obtained at nonequal intervals in this core section (3169-3348.97 m). A Leco carbon and sulfur analyzer was used to

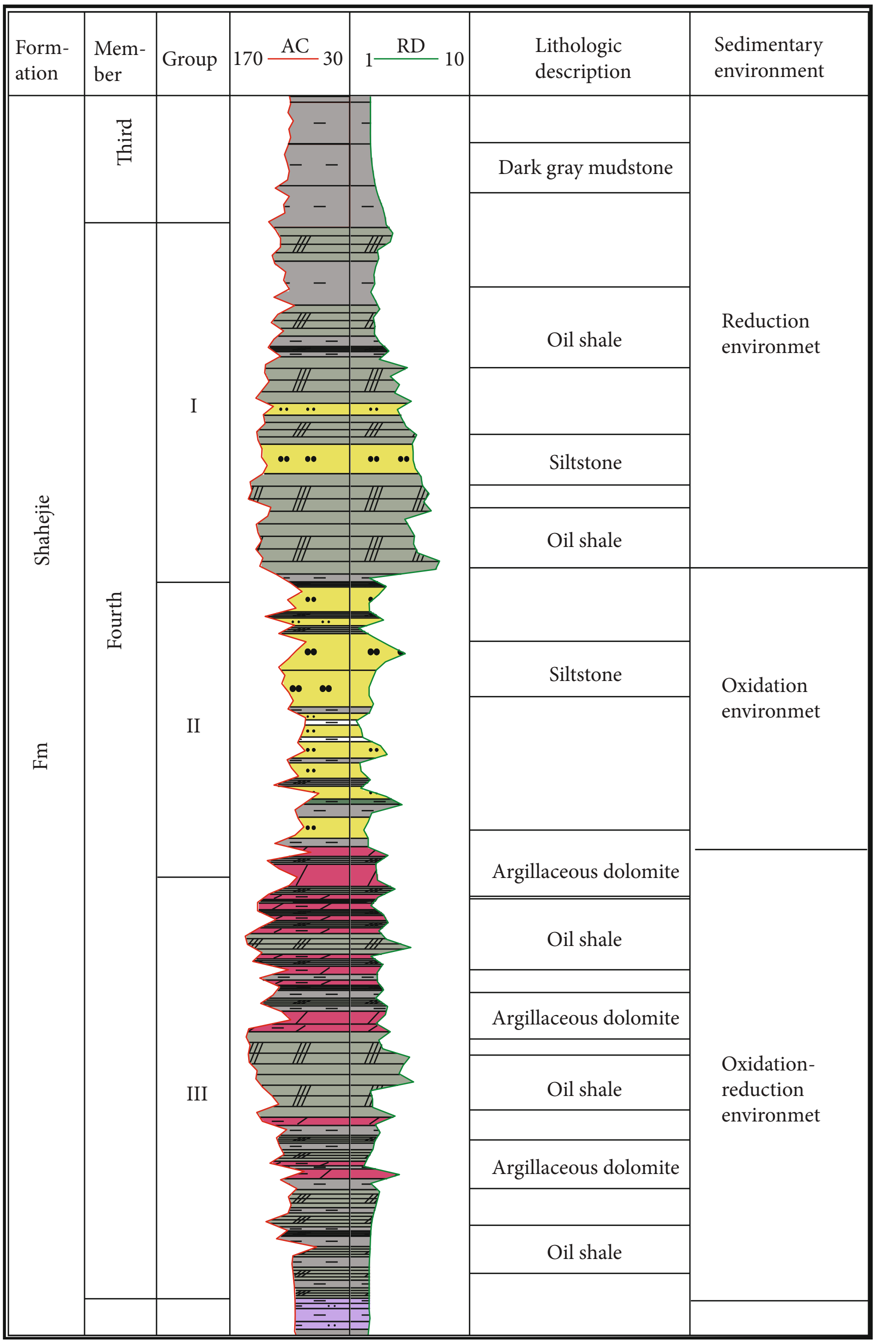


FIGURE 3: Geological section of $E_2s_4^L$ submember of the study area.

measure the TOC content according to Chinese standard GB/T 191452003, and 104 valid TOC data points were obtained. The available conventional well logs include GR, natural potential (SP), well diameter (CAL), neutron (CNL), DEN, transit time (AC), RD, and natural gamma energy spectrum (U, TH, K). Before using data, depth correction and outlier filtering were performed to ensure that the core-measured TOC data and the well log data had a one-to-one correspondence. Table 1 shows the distribution characteristics of the preprocessed well logs, including the mean, maximum and minimum values, standard deviation, skewness, and kurtosis. It can be seen that most of the log curves satisfy a Gaussian distribution, except for the RD, which has a large deviation. Thus, we applied a logarithmic transformation of the RD data before use.

Crossplots were created to analyze the correlation between the core-measured TOC content and the well logs, and linear regression was used to fit the data. The coefficient of determination ($R^2$) was calculated to evaluate the goodness of fit of the linear model. It is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - y\wedge_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}. \tag{17}$$

Table 1: Results of the statistical analysis of well S352 well logs.

| | GR | RD | AC | DEN | CNL | U | TH | K | TOC |
|---|---|---|---|---|---|---|---|---|---|
| | API | ohm·m | $\mu$s/ft | g/cm$^3$ | % | ppm | ppm | % | wt.% |
| Mean | 53.34 | 18.23 | 91.56 | 2.27 | 36.84 | 2.14 | 6.99 | 1.52 | 4.11 |
| Std | 7.61 | 41.11 | 14.19 | 0.17 | 7.41 | 0.71 | 1.78 | 0.56 | 2.37 |
| Max | 70.73 | 1.76 | 117.95 | 2.64 | 49.23 | 3.97 | 10.27 | 3.50 | 10.18 |
| Min | 29.68 | 296.08 | 62.80 | 1.96 | 19.89 | 0.63 | 2.45 | 0.52 | 0.29 |
| Skewness | -0.52 | 5.80 | -0.42 | 0.47 | -0.57 | 0.26 | -0.39 | 1.33 | 0.39 |
| Kurtosis | 1.64 | 35.71 | -0.81 | -0.55 | -0.61 | -0.25 | -0.12 | 2.07 | -0.49 |

The crossplots are shown in Figure 4. It is observed that AC, CNL, RD, GR, TH, and U have a positive linear relationship with the TOC content. $R^2$ of AC is the highest (0.3431), followed by CNL (0.2984). The linear relationship between RD, GR, TH, and the TOC content is weaker, with $R^2$ values of 0.0408, 0.0112, and 0.0957, respectively. The DEN and potassium (K) have a negative linear relationship with the TOC content, with higher $R^2$ for DEN (0.2805) and lower $R^2$ for K (0.1002).

For the multisource data, Pearson's correlation coefficient was calculated to measure the degree of linear correlation between the well log data and the TOC content. It is calculated using

$$p_{x,y} = \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y}, \tag{18}$$

where $p_{x,y}$ reflects the degree of linear correlation between variables $x$ and $y$, $\operatorname{cov}(x, y)$ is the covariance of variables $x$ and $y$, $\sigma_x$ is the standard deviation of $x$, and $\sigma_y$ is the standard deviation of $y$.

The correlation between the variables can be evaluated by creating a heat map of the Pearson correlation coefficient. As shown in Figure 5, the values represent the correlation coefficient $p_{x,y}$. A negative number represents negative correlation, a positive number represents positive correlation, 0 means no correlation, and a value close to 1 or -1 indicates a strong correlation. It can be seen that the highest correlation occurs between AC and TOC (0.59), followed by CNL (0.55) and DEN (-0.53), respectively. The correlation between GR, U, and TOC is relatively poor (0.02 and 0.07, respectively).

In summary, none of the well logs were significantly correlated with the TOC content. However, the results provide a ranking of the well log data according to their association with the core-measured TOC content. Thus, we can identify and remove irrelevant and redundant features from the training dataset, reduce the complexity of the model by reducing the dimensionality of the input data, and improve the efficiency of the model [37]. Therefore, based on the results, we selected six logs (AC, DEN, CNL, K, TH, and RD) as input training features.

## 4. Evaluation Method of Model Performance

*4.1. K-Fold Cross-Validation (CV).* In machine learning, the data are typically randomly divided into three parts: training set, test set, and validation set. However, we had very few labeled data points, resulting in strong uncertainty when using a small validation dataset to evaluate the model performance and robustness. The optimum method to avoid this problem is $K$-fold CV. The dataset is split into $K$ parts, and for each iteration, $K-1$ parts are used as the training set, and the remaining part is used as the test set, obtaining $K$ models. The $K$-fold CV makes use of all data, substantially improves the learning ability of the model, and increases the model's robustness. In this paper, following the suggestions of Zhang et al. [48] and Wong [49], the folding number $K$ was set to 5 and is related to the trade-off between computation time and bias (Figure 6).

*4.2. Comparison of Models.* We compared the performance of the XGBoost model with other machine learning algorithms. Four methods were selected, i.e., RF, SVM, $K$ -nearest neighbor (KNN), and multiple linear regression (MLR). The detailed description of these algorithms can be found in the book of Mohri et al. [50]. The hyperparameters of each machine learning algorithm were determined using a Bayesian optimization method to ensure fairness. Additionally, we included the most widely used $\Delta\log R$ method for comparison. This method overlays the RD logs in logarithmic coordinates and the porosity logs in arithmetic coordinates to calculate the TOC content in organic-rich shales, where the two logs are separated. The difference between the two logs, $\Delta\log R$, is then derived empirically using

$$\Delta\log R = \log 10\left(\frac{R}{R_{\text{baseline}}} + 0.02(\Delta t - \Delta t_{\text{baseline}})\right), \tag{19}$$

where $R$ is the resistivity ($\Omega$·m), $\Delta t$ is the measured transit time ($\mu$s/ft), and $R_{\text{baseline}}$ and $\Delta t_{\text{baseline}}$ are the resistivity and transit time values, respectively, where the two logs overlap in the baseline of the organic-deficient zone.

The $\Delta\log R$ and the organic maturity are used to determine the TOC content in the organic-rich zones, as shown in

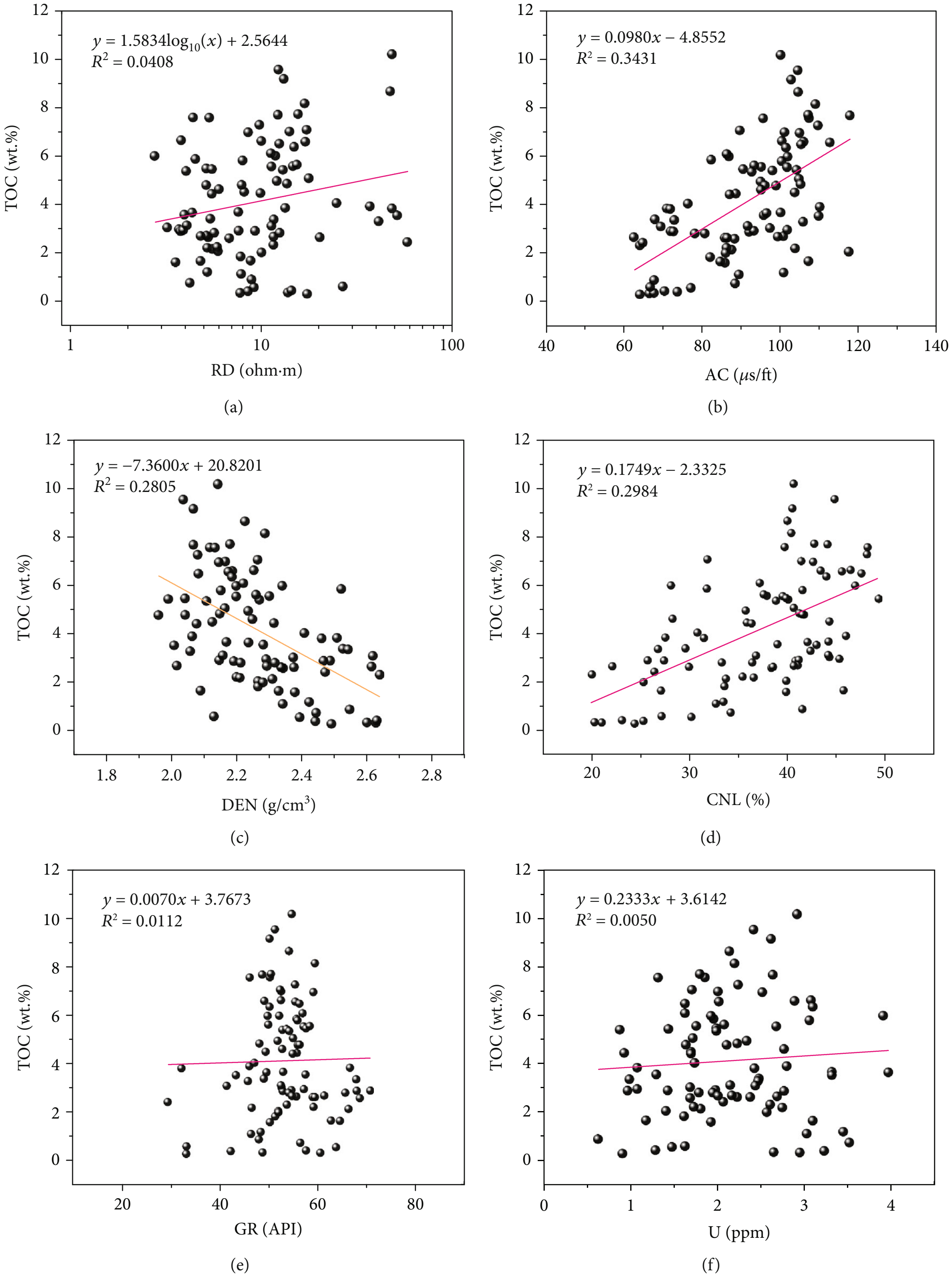


Figure 4: Continued.

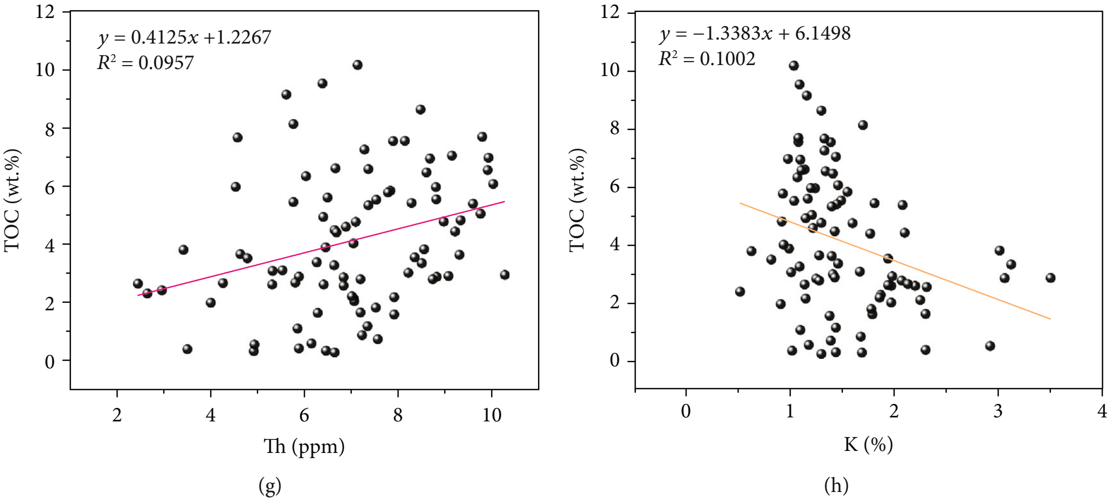
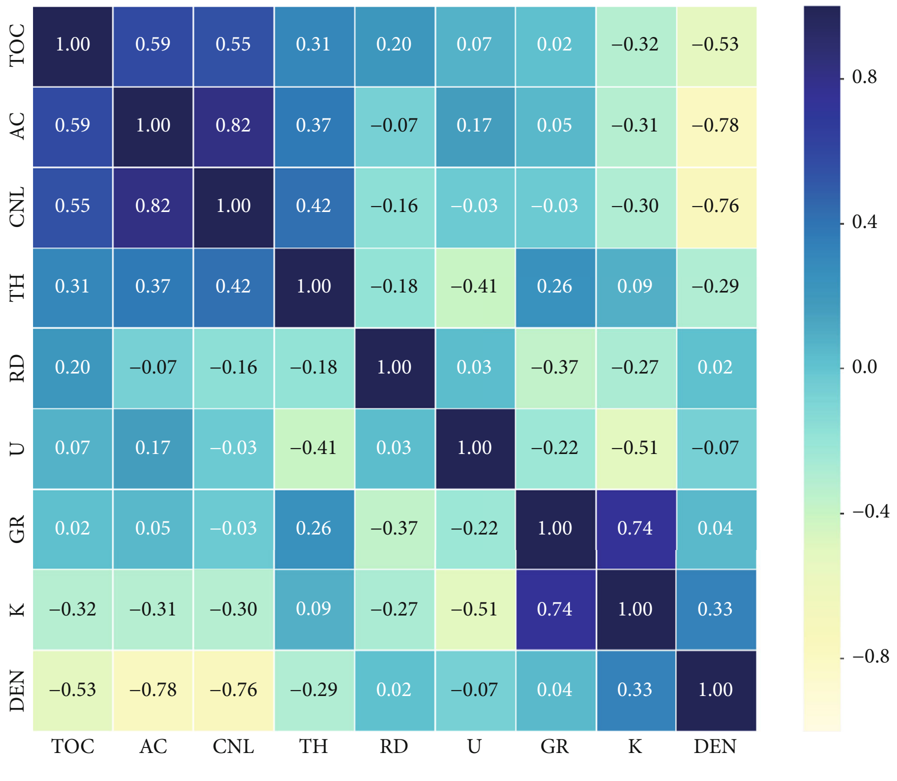


Figure 4: Crossplots between core-measured TOC and well logs: (a) RD-TOC, (b) AC-TOC, (c) DEN-TOC, (d) CNL-TOC, (e) GR-TOC, (f) U-TOC, (g) TH-TOC, and (h) K-TOC.



|  | TOC | AC | CNL | TH | RD | U | GR | K | DEN |
|---|---|---|---|---|---|---|---|---|---|
| TOC | 1.00 | 0.59 | 0.55 | 0.31 | 0.20 | 0.07 | 0.02 | −0.32 | −0.53 |
| AC | 0.59 | 1.00 | 0.82 | 0.37 | −0.07 | 0.17 | 0.05 | −0.31 | −0.78 |
| CNL | 0.55 | 0.82 | 1.00 | 0.42 | −0.16 | −0.03 | −0.03 | −0.30 | −0.76 |
| TH | 0.31 | 0.37 | 0.42 | 1.00 | −0.18 | −0.41 | 0.26 | 0.09 | −0.29 |
| RD | 0.20 | −0.07 | −0.16 | −0.18 | 1.00 | 0.03 | −0.37 | −0.27 | 0.02 |
| U | 0.07 | 0.17 | −0.03 | −0.41 | 0.03 | 1.00 | −0.22 | −0.51 | −0.07 |
| GR | 0.02 | 0.05 | −0.03 | 0.26 | −0.37 | −0.22 | 1.00 | 0.74 | 0.04 |
| K | −0.32 | −0.31 | −0.30 | 0.09 | −0.27 | −0.51 | 0.74 | 1.00 | 0.33 |
| DEN | −0.53 | −0.78 | −0.76 | −0.29 | 0.02 | −0.07 | 0.04 | 0.33 | 1.00 |

Figure 5: Heat map of Pearson's correlation coefficient.

$$\mathrm{TOC} = \Delta\log R \times 10(2.297 - 0.1688\mathrm{LOM}) + \Delta\mathrm{TOC}, \tag{20}$$

where LOM is the level of organic maturity. $\Delta$TOC is the TOC content background level in organic-rich shale.

*4.3. Evaluation Criteria.* In addition to $R^2$, we chose the root mean square error (RMSE), the mean absolute error (MAE), and the mean absolute percentage error (MAPE) to evaluate the model performance. These criteria are defined as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} |y_i - y\wedge_i|^2}, \tag{21}$$

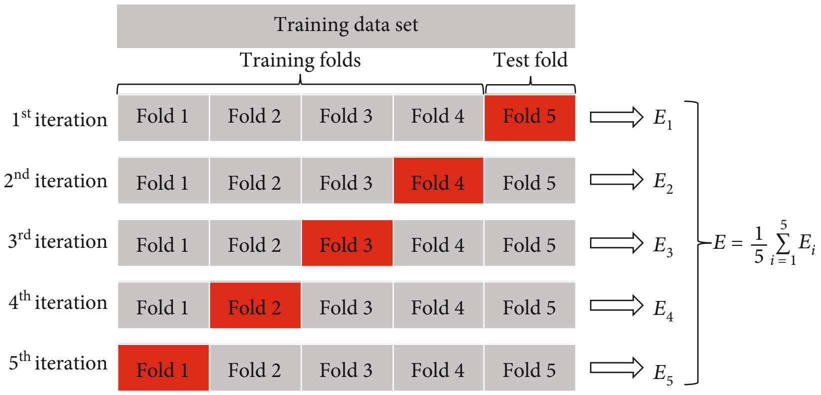

Training data set
Training folds
Test fold
1st iteration
Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
$E_1$
2nd iteration
Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
$E_2$
3rd iteration
Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
$E_3$
4th iteration
Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
$E_4$
5th iteration
Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
$E_5$
$E = \frac{1}{5}\sum_{i=1}^{5} E_i$

Figure 6: Fivefold cross-validation.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - y\wedge_i|, \tag{22}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n} \frac{|y_i - y\wedge_i|}{\max\left(\varepsilon, |y_i|\right)}, \tag{23}$$

where $y_i$ is the true value, $y_i\wedge$ is the predicted value, $\epsilon$ is a positive minimum value, and $n$ is the number of samples.

## 5. Results and Discussion

In this study, a 5-fold CV was adopted to test the model performance and robustness. The code was implemented on a microcomputer with an Intel Core i7-7700 CPU with 32 GB RAM and a Windows 10 system. The programming language was Python. The SVM, KNN, MLR, and RF models were implemented in the open-source Scikit-learn machine learning package. We used the open-source XGBoost toolkit to run the XGBoost algorithm, and the $\triangle\log R$ method code was written by the authors.

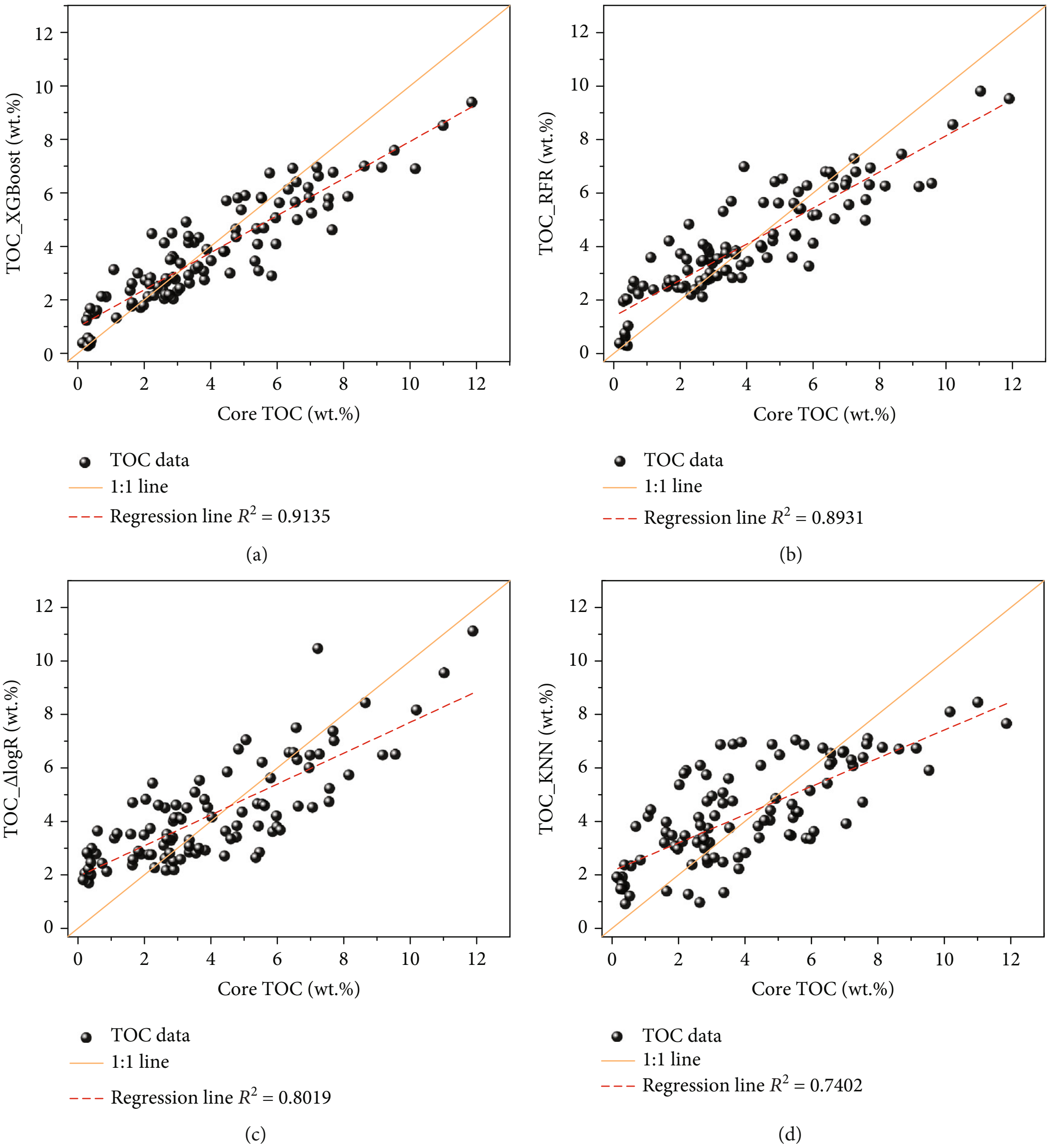
*5.1. Comparison of Model Performance.* The dataset was randomly divided into a training set and a test set for the 5-fold CV. All data were normalized to eliminate the effect of unit and scale differences between different well logging parameters. The crossplots of the predicted and core-measured TOC content are shown in Figure 7; the solid line is the 1 : 1 line, and the dashed line is the linear regression line. It should be noted that the $\Delta \log R$ method used all available data for analysis, and no 5-fold CV was used. The results showed that the XGBoost model has the best prediction performance, with $R^2$ of 0.9135, followed by the RMF model with an $R^2$ value of 0.8931 and the $\Delta \log R$ method with an $R^2$ value of 0.8345. In contrast, the other three methods have mediocre prediction performances, with $R^2$ values around 0.74.

Furthermore, we compared the RMSE and MAPE of the different methods using 5-fold CV. Figure 8 shows the RMSEs of the test set, which indicates that the RMSEs of XGBoost and RF are substantially lower than those of the other methods. Moreover, it can be inferred that the XGBoost model is the most reliable because its RMSE value is the lowest in all cases, except when $k$ is 1. Figure 9 shows the MAPEs of the test set. In terms of the relative error performance, the XGBoost model outperforms the other models, with a maximum MAPE value of 16.14% for $k = 4$, a minimum of 9.77% for $k = 1$, and a mean MAPE value of 12.55%. The second-best model is RF, with a maximum MAPE value of 17.18% for $k = 4$, a minimum of 9.05% for $k = 1$, and a mean MAPE value of 12.97%. The MAPE of SVM fluctuates considerably; the maximum value is 22.86%, and the minimum value is 11.06%. The mean MAPE value of KNN is 16.49%. The MLR had the lowest performance, with MAPE values exceeding 20% in each test.

Table 2 lists the mean values of the MAE, RMSE, and MAPE of the different methods for 5-fold CV. The mean values of the MAE, RMSE, and MAPE of the XGBoost model are 0.63, 0.77, and 12.55%, respectively, and each is the lowest value compared with the other methods. The error analysis results indicate that the XGBoost method has the highest accuracy, providing a significant advantage over other machine learning methods, as well as the $\Delta \log R$ method, for TOC prediction.

*5.2. Model Validation.* We selected well S352 to validate the prediction results of the TOC content of different methods. The well logs, core-measured TOC data, and TOC curves predicted by different methods are shown in Figures 10 and 11. The first track represents the mud logging lithology, the second track shows the lithology indicator logs, the third track is the resistivity logs, the fourth track shows the porosity logs, the fifth track is the GR ray spectrum logs, and the 6th-10th tracks are the TOC curves predicted by the $\triangle\log R$, MLR, KNN, SVM, RF, and XGBoost methods; the red dots represent the core-measured TOC data.

Figure 10 shows the predicted results of the $E_2s_4^{\text{L}}$-I group. Between 3150 and 3200 m, the lithology is oil shale, and the fluctuations of the well logs are small, indicating good formation homogeneity. The TOC curves predicted by all methods are highly correlated with the core-



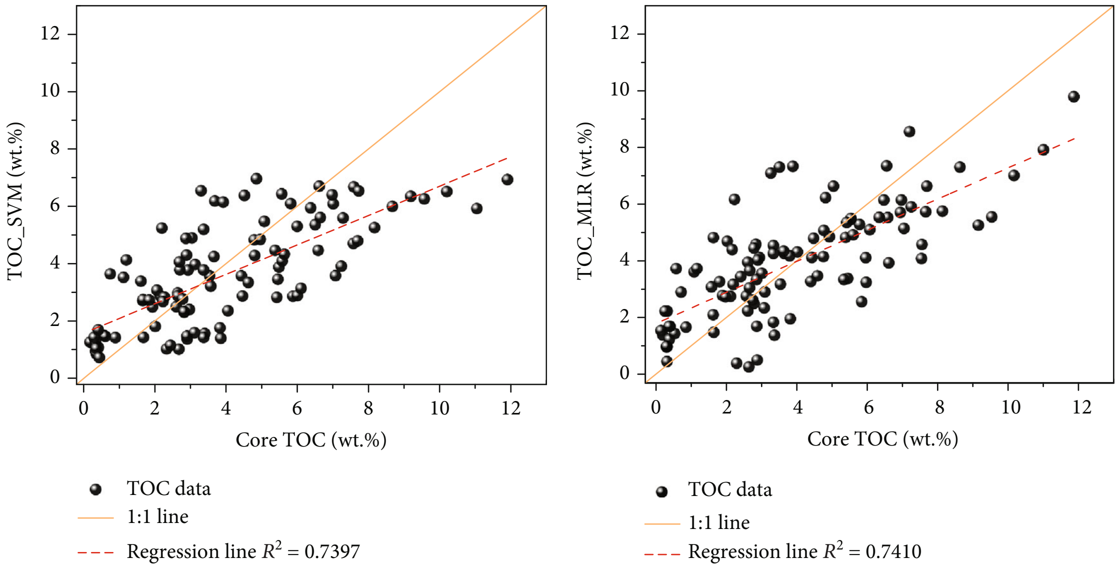
(a)

(b)

(c)

(d)

FIGURE 7: Continued.



(e)

(f)

Figure 7: Crossplots of the predicted TOC and core-measured TOC content: (a) XGBoost model, (b) RF model, (c) △log *R* method, (d) KNN model, (e) SVM model, and (f) MLR model.

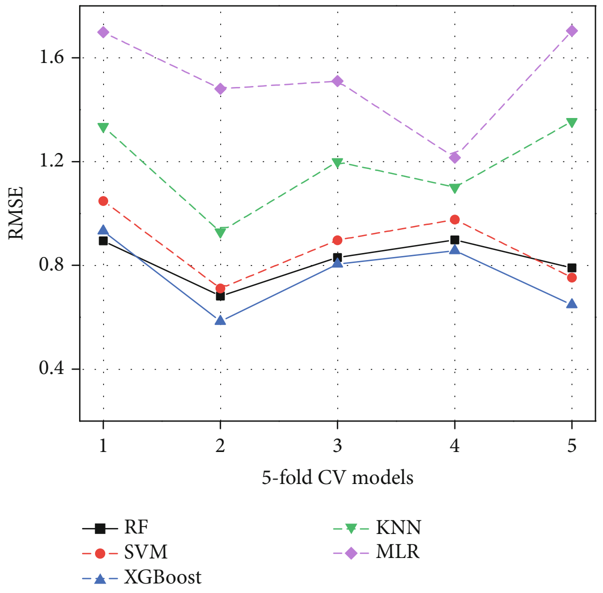


Figure 8: RMSE results of the 5-fold cross-validation.

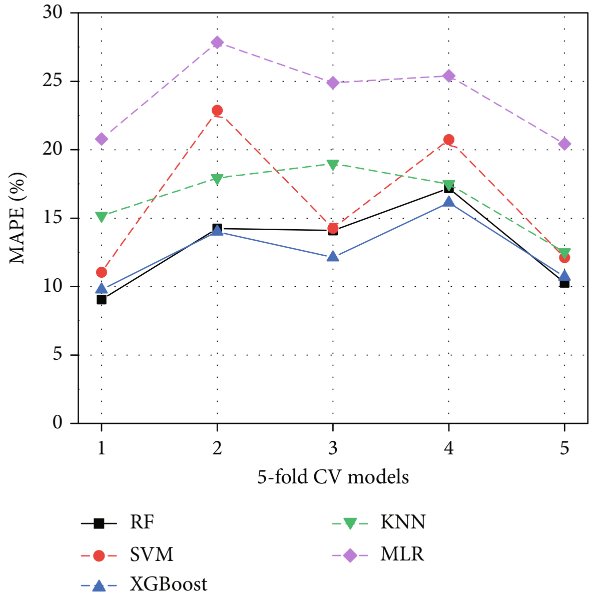


Figure 9: MAPE results of the 5-fold cross-validation.

measured TOC content, and the trends are similar. However, between 3200 and 3236 m, the lithology starts to change. The resistivity logs show high resistance characteristics, and the core-measured TOC content increases significantly. The prediction results of the XGBoost, RF, and △log *R* methods are in good agreement with the core-measured TOC content. In contrast, the predicted values of the MLR, SVM, and KNN methods are considerably smaller than the core-measured TOC content.

Figure 11 shows the prediction results of the $E_2s_4{}^L$-II and $E_2s_4{}^L$-III groups. The depth of the $E_2s_4{}^L$-II group is

Table 2: Mean error values of different methods for 5-fold CV.

| Model | MAE | RMSE | MAPE (%) |
|---|---|---|---|
| XGBoost | 0.63 | 0.77 | 12.55 |
| RF | 0.71 | 0.82 | 12.97 |
| SVM | 0.71 | 0.88 | 16.21 |
| KNN | 0.82 | 1.18 | 16.42 |
| Linear | 1.26 | 1.52 | 23.87 |
| *ΔlogR* | 0.79 | 1.07 | 14.80 |

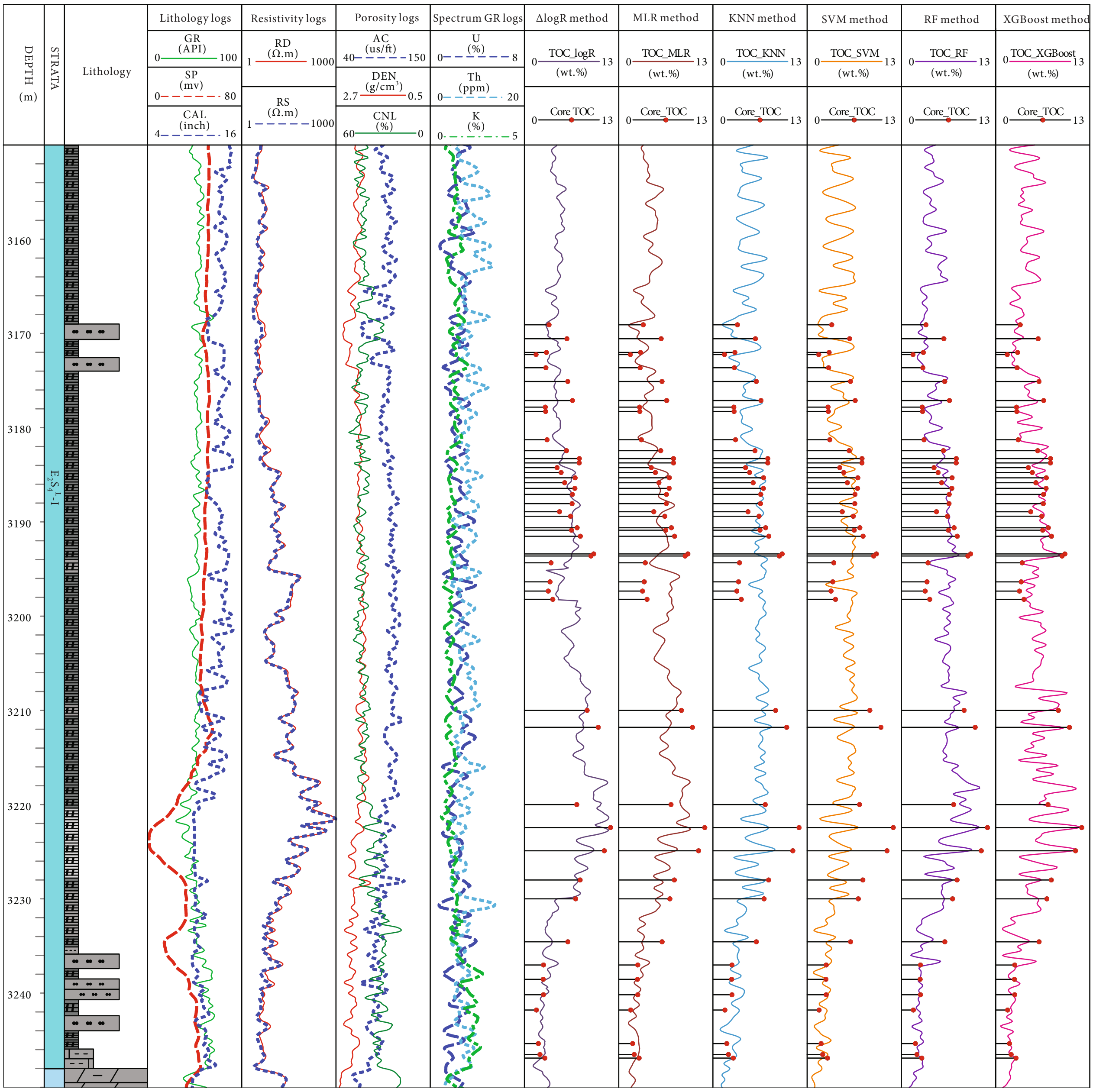


Figure 10: Comparison of prediction results by different methods ($E_2s_4^L$-I group).

3248-3278 m, and the lithology is argillaceous dolomite interbedded with a small amount of oil shale. The core-measured TOC content ranges from 0.17% to 3.84%, indicating a weak hydrocarbon generation potential. The accuracy of the predictions of the different methods is highly variable. The XGBoost and RF methods show higher prediction accuracy than the other methods. The TOC values predicted by the $\triangle \log R$ method are significantly larger than the core-measured TOC values. The likely reason is that the mineral composition of this play differs greatly from that at the baseline formation; thus, the AC and RD logs are substantially affected by the lithology and do not reflect the changes in organic matter content. The depth of the $E_2s_4^L$-III group is 3278-3350 m. This group shows strong heterogeneity. Oil shales and argillaceous dolomite are frequently interbedded, and the thickness of each layer is less than 3 m. The well logs show fluctuations, and the TOC trend is unclear. The core-measured TOC content ranges from 0.29% to 9.77%. The XGBoost method provides the highest agreement with the core-measured TOC data, followed by the RF. The predicted values obtained from the $\triangle \log R$, MLR, KNN, and SVM methods are substantially lower than the core-measured TOC values.

Overall, the prediction results from well S352 show that the Bayesian optimization XGBoost method performed most

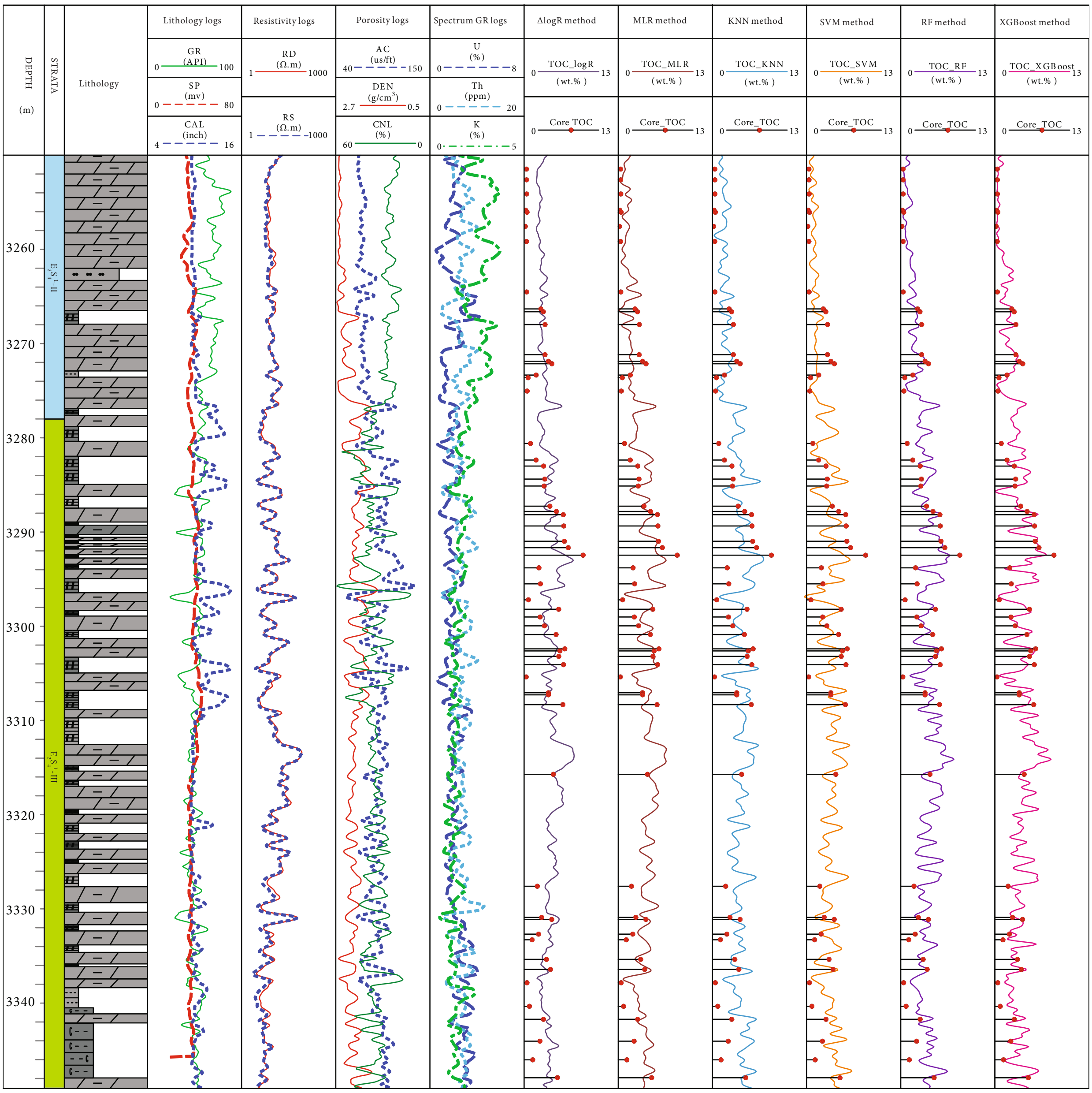


Figure 11: Comparison of prediction results by different methods ($E_2s_4^L$-II group and $E_2s_4^L$-III group).

reliably in nonhomogeneous formations, providing the highest prediction accuracy and best generalization ability. Thus, this method is more suitable for TOC prediction of lacustrine shale oil than the other methods used in this study.

*5.3. Prediction of the TOC Distribution.* We selected 20 exploratory wells drilled in the $E_2s_4^L$ formation to predict the TOC content using the XGBoost model. The contour maps of the predicted TOC content of the $E_2s_4^L$-I, $E_2s_4^L$-II, and $E_2s_4^L$-III groups in the study area are shown in Figure 12. In the $E_2s_4^L$-I group, the TOC content is relatively higher on the west side of well A10-A49-A95 (>4%), and the highest value occurs near well S166 (>6%). The area with a TOC content exceeding 4% is 73 km$^2$ (Figure 12(a)). In the $E_2s_4^L$-II group, the TOC content is relatively low, ranging from 1.5% to 3.1%. The area with a TOC value greater than 2% covers 115 km$^2$ (Figure 12(b)). In the $E_2s_4^L$-III group, areas with a TOC content greater than 4% are located near wells S224, A49, A104, Sh25, and Sh17, with an area of 23 km$^2$. The TOC content of the other areas is below 4% (Figure 12(c)). Vertically, the $E_2s_4^L$-I group has the highest TOC content, followed by the $E_2s_4^L$-III group and the $E_2s_4^L$-II group. Horizontally, high-quality source rocks are mainly distributed on the west slope of the study area and sporadically in other regions.

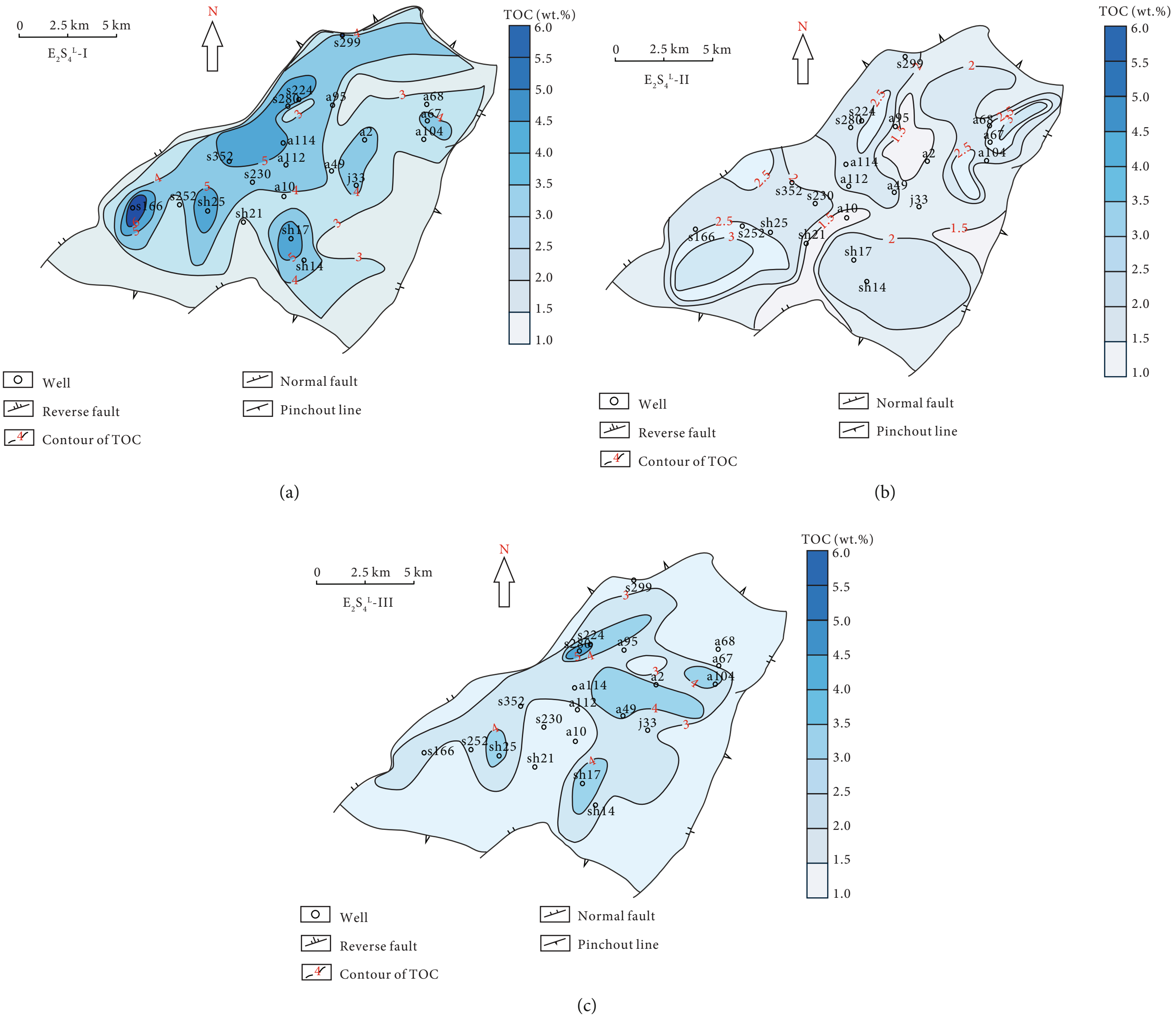


FIGURE 12: Contour maps of the predicted TOC content: (a) $E_2s_4^L$-I group, (b) $E_2s_4^L$-II group, and (c) $E_2s_4^L$-III group.

## 6. Conclusions

We proposed a robust data-driven Bayesian optimization XGBoost model to predict the TOC content using wireline log data. The data were obtained from the Damintun Sag, Bohai Bay Basin, China, consisted of well logs and core-measured TOC data. Linear regression crossplots were obtained, and Pearson's correlation coefficients were calculated to evaluate the relationship between the well logs and the core-measured TOC data. The results indicated that none of the well logs were significantly correlated with the TOC content. However, the correlation analysis enabled us to identify and remove irrelevant and redundant well logging features for the TOC prediction and reduce the model complexity by reducing the dimensionality of the input data. The model performance was evaluated using 5-fold CV. The quantitative error analysis of the four criteria showed that the proposed approach performs better compared to the traditional method ($\Delta\log R$), with $R^2$ increasing from 0.8345 to 0.9135 and MAE, RMSE, and MAPE decreasing from 0.79, 1.07, and 14.80% to 0.63, 0.77, and 12.55%, respectively. Also, the XGBoost model outperforms other popular machine learning algorithm (i.e., RF, SVM, KNN, and MLR) in terms of robustness, accuracy, and generalization in predicting TOC for strongly nonhomogeneous lacustrine shale plays. We used the proposed approach for the TOC prediction of 20 exploration wells in the Damintun Sag and obtained contour maps of the TOC content in the $E_2s_4^L$ formation. The maps enabled the identification of areas with high hydrocarbon generation potential, which is useful for finding sweet spots. Generally, machine learning relies extensively on the quality and quantity of the training data. As new exploration occurs, additional data should be added in real time to improve the reliability and generalization ability of the model. In the future, we plan to create a database for machine learning. In addition to predicting the TOC content, this database can be used to predict other petrophysical and geomechanical properties of reservoirs.

## Abbreviations

| | |
|---|---|
| AC: | Acoustic log |
| ANN: | Artificial neural network |
| CAL: | Well diameter log |
| CART: | Classification and regression tree |
| CNL: | Neutron log |
| DEN: | Density log |
| EI: | Expected improvement |
| ELM: | Extreme learning machine |
| ES: | Entropy search |
| GA: | Genetic algorithm |
| GBDT: | Gradient boosting decision tree |
| GP: | Gaussian process |
| GPR: | Gaussian process regression |
| GR: | Gamma ray log |
| GS: | Grid search |
| K: | Potassium log |
| KNN: | $K$-nearest neighbor |
| MAE: | Mean absolute error |
| MAPE: | Mean absolute percentage error |
| MLR: | Multiple linear regression |
| PI: | Probability of improvement |
| PSO: | Particle swarm optimization |
| $R^2$: | Coefficient of determination |
| RD: | Deep resistivity log |
| RF: | Random forest |
| RMSE: | Root mean square error |
| RPM: | Rock physical model |
| RS: | Random search |
| SP: | Natural potential log |
| SVM: | Support vector machine |
| TH: | Thorium log |
| TOC: | Total organic carbon |
| UCB: | Upper confidence bound |
| XGBoost: | Extreme gradient boosting. |

## Data Availability

The data used to support the findings of this study are included within the article.

## Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

## Acknowledgments

This work was funded by the Major Science and Technology Project of China National Petroleum Corporation "Study on Key Technologies for Sustainable Production of 10 Million Tons of Crude Oil in Liaohe Oilfield" (No. 2017E-16).

## References

[1] G. X. Li, K. Luo, and D. Q. Shi, "Key technologies, engineering management and important suggestions of shale oil/gas development: case study of a Duvernay shale project in Western Canada Sedimentary Basin," *Petroleum Exploration and Development*, vol. 47, no. 4, pp. 791–802, 2020.

[2] W. Zhao, S. Hu, L. Hou et al., "Types and resource potential of continental shale oil in China and its boundary with tight oil," *Petroleum Exploration and Development*, vol. 47, no. 1, pp. 1–11, 2020.

[3] L. Zhou, X. Zhao, G. Chai et al., "Key exploration & development technologies and engineering practice of continental shale oil: a case study of Member 2 of Paleogene Kongdian Formation in Cangdong Sag, Bohai Bay Basin, East China," *Petroleum Exploration and Development*, vol. 47, no. 5, pp. 1138–1146, 2020.

[4] M. Wen, Z. Jiang, K. Zhang et al., "Difference analysis of organic matter enrichment mechanisms in upper Ordovician-lower Silurian shale from the Yangtze region of southern China and its geological significance in shale gas exploration," *Geofluids*, vol. 2019, Article ID 9524507, 14 pages, 2019.

[5] B. Liu, H. Wang, X. Fu et al., "Lithofacies and depositional setting of a highly prospective lacustrine shale oil succession from the Upper Cretaceous Qingshankou Formation in the Gulong sag, northern Songliao Basin, northeast China," *AAPG Bulletin*, vol. 103, no. 2, pp. 405–432, 2019.

[6] X. L. Zhang, L. Z. Xiao, R. H. Xie, H. Wu, and Y. Gao, "Petrophysical workflow for shale gas evaluation," *Progress in Geophysics*, vol. 4, 2013.

[7] R. F. Beers, "Radioactivity and organic content of some Paleozoic shales," *AAPG Bulletin American Association of Petroleum Geologists*, vol. 29, pp. 1–22, 1945.

[8] I. R. Supernaw, M. Dan, and A. J. Link, "Method for in situ evaluation of the source rock potential of earth formations," US Patent 4-071755, 1978.

[9] J. W. Schmoker, "Determination of organic-matter content of Appalachian Devonian shales from gamma-ray logs," *AAPG Bulletin*, vol. 65, no. 7, pp. 1285–1298, 1981.

[10] V. E. Swanson, *Oil Yield and Uranium Content of Black Shales*, Center for Integrated Data Analytics Wisconsin Science Center, 1960.

[11] V. E. Swanson, *Geology and Geochemistry of Uranium in Marine Black Shales: A Review*, US Government Printing Office, Washington, DC, 1961.

[12] W. H. Fertl and G. V. Chilingar, "Total organic carbon content determined from well logs," *SPE Formation Evaluation*, vol. 3, no. 2, pp. 407–419, 1988.

[13] Z. H. Chen, M. Cha, and Q. Jin, "Application of natural gamma ray logging and natural gamma spectrometry logging to recovering paleoenvironment of sedimentary basin," *Chinese Journal of Geophysics*, vol. 47, no. 6, pp. 1145–1150, 2004.

[14] J. W. Schmoker, "Organic content of Devonian shale in western Appalachian Basin," *AAPG Bulletin*, vol. 64, 1980.

[15] S. L. Herron, "A total organic carbon log for source rock evaluation," *The Log Analyst*, vol. 28, 1987.

[16] Q. R. Passey, S. Creaney, J. B. Kulla, F. J. Moretti, and J. D. Stroud, "A practical model for organic richness from porosity and resistivity logs," *AAPG Bulletin*, vol. 74, no. 12, pp. 1777–1794, 1990.

[17] Q. R. Passey, K. M. Bohacs, W. L. Esch, R. Klimentidis, and S. Sinha, "From oil-prone source rock to gas-producing shale reservoir-geologic and petrophysical characterization of unconventional shale gas reservoirs," in *International Oil and Gas Conference and Exhibition in China*, Beijing, China, 2010.

[18] P. Wang, Z. Chen, X. Pang, K. Hu, M. Sun, and X. Chen, "Revised models for determining TOC in shale play: example from Devonian Duvernay Shale, Western Canada Sedimentary Basin," *Marine & Petroleum Geology*, vol. 70, pp. 304–319, 2016.

[19] P. Zhao, Z. Mao, Z. Huang, and C. Zhang, "A new method for estimating total organic carbon content from well logs," *AAPG Bulletin*, vol. 100, no. 8, pp. 1311–1327, 2016.

[20] L. Zhu, C. Zhang, Z. Zhang, X. Zhou, and W. Liu, "An improved method for evaluating the TOC content of a shale formation using the dual-difference ΔlogR method," *Marine and Petroleum Geology*, vol. 102, pp. 800–816, 2019.

[21] R. R. Pemper, X. Han, F. E. Mendez et al., "The direct measurement of carbon in wells containing oil and natural gas using a pulsed neutron mineralogy tool," in *SPE Annual Technical Conference and Exhibition*, New Orleans, Louisiana, 2009.

[22] M. M. Herron, J. A. Grau, S. L. Herron et al., "Total organic carbon and formation evaluation with wireline logs in the Green River oil shale," *Plasma Physics & Controlled Fusion*, vol. 46, no. 4, pp. 593–609, 2013.

[23] D. Mou, Z. W. Wang, Y. L. Huang, S. Xu, and D. P. Zhou, "Lithological identification of volcanic rocks from SVM well logging data: case study in the eastern depression of Liaohe Basin," *Chinese Journal of Geophysics-Chinese Edition*, vol. 58, no. 5, pp. 1785–1793, 2015.

[24] A. A. Silva, M. W. Tavares, A. Carrasquilla, R. Misságia, and M. Ceia, "Petrofacies classification using machine learning algorithms," *Geophysics*, vol. 85, no. 4, pp. WA101–WA113, 2020.

[25] Y. Xie, C. Zhu, R. Hu, and Z. Zhu, "A coarse-to-fine approach for intelligent logging lithology identification with extremely randomized trees," *Mathematical Geosciences*, vol. 53, no. 5, pp. 859–876, 2021.

[26] J. Cao, J. Yang, Y. Wang, D. Wang, and Y. Shi, "Extreme learning machine for reservoir parameter estimation in heterogeneous sandstone reservoir," *Mathematical Problems in Engineering*, vol. 2015, Article ID 287816, 10 pages, 2015.

[27] Y. A. Xingyu, G. U. Hanming, and X. I. Yifei, "XGBoost algorithm applied in the interpretation of tight-sand gas reservoir on well logging data," *Oil Geophysical Prospecting*, vol. 54, no. 2, pp. 447–455, 2019.

[28] M. Tan, Q. Liu, and S. Zhang, "A dynamic adaptive radial basis function approach for total organic carbon content prediction in organic shale," *Geophysics*, vol. 78, no. 6, pp. D445–D459, 2013.

[29] M. Tan, X. Song, X. Yang, and Q. Wu, "Support-vector-regression machine technology for total organic carbon content prediction from wireline logs in organic shale: a comparative study," *Journal of Natural Gas Science & Engineering*, vol. 26, pp. 792–802, 2015.

[30] H. Yu, R. Rezaee, Z. Wang et al., "A new method for TOC estimation in tight shale gas reservoirs," *International Journal of Coal Geology*, vol. 179, pp. 269–277, 2017.

[31] J. Rui, H. Zhang, Q. Ren, L. Yan, Q. Guo, and D. Zhang, "TOC content prediction based on a combined Gaussian process regression model," *Marine and Petroleum Geology*, vol. 118, p. 104429, 2020.

[32] X. Shi, J. Wang, G. Liu, L. Yang, X. Ge, and S. Jiang, "Application of extreme learning machine and neural networks in total organic carbon content prediction in organic shale with wire line logs," *Journal of Natural Gas Science and Engineering*, vol. 33, pp. 687–702, 2016.

[33] L. Zhu, C. Zhang, C. Zhang, X. Zhou, J. Wang, and X. Wang, "Application of multiboost-KELM algorithm to alleviate the collinearity of log curves for evaluating the abundance of organic matter in marine mud shale reservoirs: a case study in Sichuan Basin, China," *Acta Geophysica*, vol. 66, no. 5, pp. 983–1000, 2018.

[34] L. Zhu, C. Zhang, C. Zhang et al., "Prediction of total organic carbon content in shale reservoir based on a new integrated hybrid neural network and conventional well logging curves," *Journal of Geophysics & Engineering*, vol. 15, no. 3, pp. 1050–1061, 2018.

[35] A. A. A. Mahmoud, S. Elkatatny, M. Mahmoud, M. Abouelresh, A. Abdulraheem, and A. Ali, "Determination of the total organic carbon (TOC) based on conventional well logs using artificial neural network," *International Journal of Coal Geology*, vol. 179, pp. 72–80, 2017.

[36] Y. Bai and M. Tan, "Dynamic committee machine with fuzzy-c-means clustering for total organic carbon content prediction from wireline logs," *Computers & Geosciences*, vol. 146, p. 104626, 2021.

[37] A. Handhal, A. M. al-Abadi, H. E. Chafeet, and M. J. Ismail, "Prediction of total organic carbon at Rumaila oil field, Southern Iraq using conventional well logs and machine learning algorithms," *Marine and Petroleum Geology*, vol. 116, p. 104347, 2020.

[38] J. Song, Q. Gao, and L. Zhe, "Application of random forests for regression to seismic reservoir prediction," *Oil Geophysical Prospecting*, vol. 51, no. 6, pp. 1202–1211, 2016.

[39] M. J. Cracknell and A. M. Reading, "The upside of uncertainty: identification of lithology contact zones from airborne geophysics and satellite data using random forests and support vector machines," *Geophysics*, vol. 78, no. 3, pp. WB113–WB126, 2013.

[40] L. X. Zhao, J. S. Liu, Y. X. Yao et al., "Quantitative seismic characterization of source rocks in lacustrine depositional setting using the random forest method: an example from the Changjiang sag in East China Sea basin," *Chinese Journal of Geophysics*, vol. 64, no. 2, pp. 700–715, 2021.

[41] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *The 22nd ACM SIGKDD International Conference*, pp. 785–794, New York, NY, USA, 2016.

[42] T. Nguyen-Sy, M. N. Vu, A. D. Tran-Le, B. V. Tran, and T. T. Nguyen, "Studying petrophysical properties of micritic limestones using machine learning methods," *Journal of Applied Geophysics*, vol. 184, no. 4, 2021.

[43] Y. F. Gu, D. Y. Zhang, and Z. D. Bao, "Permeability prediction using PSO-XGBoost based on logging data," *Oil Geophysical Prospecting*, vol. 56, no. 1, pp. 26–37, 2021.

[44] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," *Advances in Neural Information Processing Systems*, vol. 24, pp. 2546–2554, 2011.

[45] A. Klein, S. Falkner, S. Bartels, P. Hennig, and F. Hutter, “Fast Bayesian optimization of machine learning hyperparameters on large datasets,” in *Artificial Intelligence and Statistics*, pp. 528–536, PMLR, 2016.

[46] J. Wu, M. Poloczek, A. G. Wilson, and P. I. Frazier, “Bayesian optimization with gradients,” 2017, https://arxiv.org/abs/1703.04389.

[47] L. Xiaoguang, L. Xingzhou, L. Jinpeng, and T. Zhi, “Comprehensive evaluation and exploration practice of Sha 4 lacustrine shale oil in Damintun sag, Liaohe depression,” *China Petroleum Exploration*, vol. 24, no. 5, pp. 636–648, 2019.

[48] W. Zhang, C. Wu, H. Zhong, Y. Li, and L. Wang, “Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization,” *Geoscience Frontiers*, vol. 12, no. 1, pp. 469–477, 2021.

[49] T. T. Wong, “Performance evaluation of classification algorithms by _k_ -fold and leave- one-out cross validation,” *Pattern Recognition*, vol. 48, no. 9, pp. 2839–2846, 2015.

[50] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, MIT Press, Cambridge, MA, USA, 2012.