Hindawi
Geofluids
Volume 2021, Article ID 6794213, 13 pages
https://doi.org/10.1155/2021/6794213

Research Article

Machine Learning Method for TOC Prediction: Taking Wufeng and Longmaxi Shales in the Sichuan Basin, Southwest China as an Example

Jia Rong,[1,2,3] **Zongyuan Zheng**,[1,2,3] **Xiaorong Luo**,[1,2,3] **Chao Li**,[1,2] **Yuping Li**,[4] **Xiangfeng Wei**,[4] **Quanchao Wei**,[4] **Guangchun Yu**,[4] **Likuan Zhang**,[1,2] **and Yuhong Lei**[1,2]

[1]*Key Laboratory of Petroleum Resources Research, Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing 100029, China*
[2]*Innovation Academy for Earth Science, Chinese Academy of Sciences, Beijing 100029, China*
[3]*University of Chinese Academy of Sciences, Beijing 100049, China*
[4]*Exploration Branch Company of SINOPEC, Chengdu 610041, China*

Correspondence should be addressed to Jia Rong; rongjia@mail.iggcas.ac.cn

Received 18 June 2021; Revised 13 August 2021; Accepted 20 August 2021; Published 26 September 2021

Academic Editor: Liu Bo

The total organic carbon content (TOC) is a core indicator for shale gas reservoir evaluations. Machine learning-based models can quickly and accurately predict TOC, which is of great significance for the production of shale gas. Based on conventional logs, the measured TOC values, and other data of 9 typical wells in the Jiaoshiba area of the Sichuan Basin, this paper performed a Bayesian linear regression and applied a random forest machine learning model to predict TOC values of the shale from the Wufeng Formation and the lower part of the Longmaxi Formation. The results showed that the TOC value prediction accuracy was improved by more than 50% by using the well-trained machine learning models compared with the traditional $\Delta \mathrm{Log}R$ method in an overmature and tight shale. Using the halving random search cross-validation method to optimize hyperparameters can greatly improve the speed of building the model. Furthermore, excluding the factors that affect the log value other than the TOC and taking the corrected data as input data for training could improve the prediction accuracy of the random forest model by approximately 5%. Data can be easily updated with machine learning models, which is of primary importance for improving the efficiency of shale gas exploration and development.

1. Introduction

Shale gas is a very important unconventional energy. Shale gas production in the United States constitutes a major part of its energy structure, and China has also made breakthroughs in the shale gas field in recent years [1]. Quickly identifying sweet spots where oil and gas are enriched in shale formation has an important impact on guiding the economic and effective exploitation of shale oil and gas resources [2–4]. The total organic matter content (TOC) is an important index for evaluating the enrichment of oil and gas resources, which can effectively indicate the organic matter enrichment intervals in shale formation [5]. The TOC values are often obtained through laboratory testing of cores. However, shale formation has strong heterogeneity, and it is limited by sedimentary space, material source supply, and other factors [6]. Moreover, in the early stage of shale oil and gas resource exploration, the continuity and integrity of core data cannot be guaranteed. As a result, the use of discrete measured TOC values from the core test may lead to a misunderstanding of organic matter-rich intervals. In contrast, the geophysical log data are complete and continuous. Continuous vertical TOC values can be obtained by using log data, and then, the distribution of organic enrichment layers can be predicted [7–10].

In the 1980s, Schmoker first discovered the relationship between log data and organic matter abundance, and the density log value was used to calculate the organic carbon

content. With the continuous development of technology, many methods using log information to predict TOC values have been found, such as the log curve superposition evaluation method (ΔLog$R$ method and its modification method) [11–13], multiple linear regression evaluation methods [14], machine learning, and other mathematical analysis evaluation methods [15–17]. However, different methods have different scopes. The ΔLog$R$ method has a comparably wide range of applications among these methods because it is driven by the physical model. Nevertheless, neither the traditional ΔLog$R$ method nor the improved method can fully cover a variety of different formation conditions (such as abnormal fluid pressure, overmaturity, and tight reservoirs). In addition, most TOC evaluation methods based on ΔLog$R$ need to manually determine the baseline value of the porosity curve and the resistivity curve, which is a relatively cumbersome process.

In recent years, machine learning methods have become a useful tool for building prediction models, which can reveal hidden patterns and unknown correlations between independent variables and dependent variables [18, 19]. In the machine learning model, TOC prediction is a multiple regression problem. The machine learning algorithm can automatically determine the comprehensive relationship between the TOC values and the corresponding log values through the learning of samples. Machine learning methods are driven by data and thus are not subject to changes in geological conditions. A large amount of stratum information can be better used to comprehensively predict TOC values, so the accuracy will not be greatly reduced due to the distortion of a certain curve [20, 21]. The disadvantages of machine learning-based models are that they may have multiple solutions and overfit with a limited number of samples. In recent years, some machine learning methods have shown good application effects and prospects in TOC prediction of source rocks. Zhao et al. [22] used Bayesian methods to predict TOC values and achieved good results. Handhal et al. [23] used integrated learning methods which not only guaranteed the accuracy of the model but also solved the problems of overlearning and improved the generalization ability of the model.

This paper takes the shale of the Wufeng Formation and the lower part of the Longmaxi Formation in the Jiaoshiba area of the Sichuan Basin as the main research object. Based on a large amount of measured TOC values from drilling cuttings and cores in this area, Bayesian linear regression and the relatively stable random forest algorithm are selected to predict the TOC values. Comparing the results with the traditional ΔLog$R$ method, this paper discusses which method is more suitable for TOC prediction in this area and how to improve the calculation speed and accuracy based on existing methods.

2. Data and Methods

2.1. Geological Background and Data Source. The shale of the Upper Ordovician Wufeng Formation and the Lower Silurian Longmaxi Formation in the southeastern Sichuan Basin is a shelf deposit that was maintained over a long time in a deep-water anoxic environment. Black shale and silty shale with stable thickness and wide distribution are deposited, and they have high siliceous contents and are rich in graptolite. The organic matter content is mainly derived from high-productivity marine organisms and has a high degree of thermal evolution. Its Ro value is approximately 2.0% to 3.5% [24]. This basin is currently the most important shale gas reservoir in China [25, 26]. The first shale gas field in China was built in the Jiaoshiba area (Figure 1), and the main production layer of this shale gas field is the black shale section of the Wufeng Formation and the lower part of Longmaxi Formation. The geothermal field is relatively stable. The roof and floor plates constitute excellent sealing units, and the damage of faults is limited. Overall, it has good preservation conditions, which are conducive to the enrichment and storage of shale gas [26–28]. However, the black shale in the Wufeng Formation and the lower part of the Longmaxi Formation was once buried to a depth of 6000 m. The diagenesis of the shale was relatively thorough, which led to changes in the porosity log values under the effect of many factors. The application effect of the TOC prediction method is not ideal [29]. In this paper, the Jiaoshiba area is the research object used to investigate prediction effect of the TOC values with different machine learning methods based on log data and the measured TOC values from 9 typical wells (JY1, JY2, JY3, JY4, JY5, JY6, JY7, JY8, and JY9) in this area. Among them, the data from wells JY1, JY3, JY4, JY5, JY8, and JY9 are used for model establishment and verification. The data from wells JY2, JY6, and JY7 are used to verify the universality of the model.

2.2. Methods

2.2.1. Traditional ΔLogR Method. The traditional ΔLog$R$ method was proposed by Passey et al. [11]. This method uses the porosity curve, deep lateral resistivity curve, and maturity parameters to predict the TOC values. Normally, the porosity curve (usually the acoustic transit time log curve) and the resistivity curve overlap in the fine-grained organic-poor texture layer but show an amplitude difference (defined as ΔLog$R$) in the organic-rich texture layer. This amplitude difference has a linear relationship with the TOC values and is a function of maturity, which can be used to calculate TOC values.

The formula for calculating the amplitude difference from the acoustic transit time and resistivity is as follows:

$$\Delta\mathrm{Log}R = \lg\left(\frac{R}{R_\mathrm{b}}\right) + 0.02 * (\Delta t - \Delta t_\mathrm{b}), \tag{1}$$

where ΔLog$R$ is the curve amplitude difference measured in logarithmic resistivity units; R is the deep lateral resistivity, and the unit is Ω·m; Δt is the acoustic transit time, and the unit is μs/ft; R_b and Δt_b are the baseline values of the resistivity curve and the acoustic transit time curve, respectively, which correspond to the overlapping section of the two in the fine-grained organic-poor texture layer; and 0.02 is the calibration coefficient.

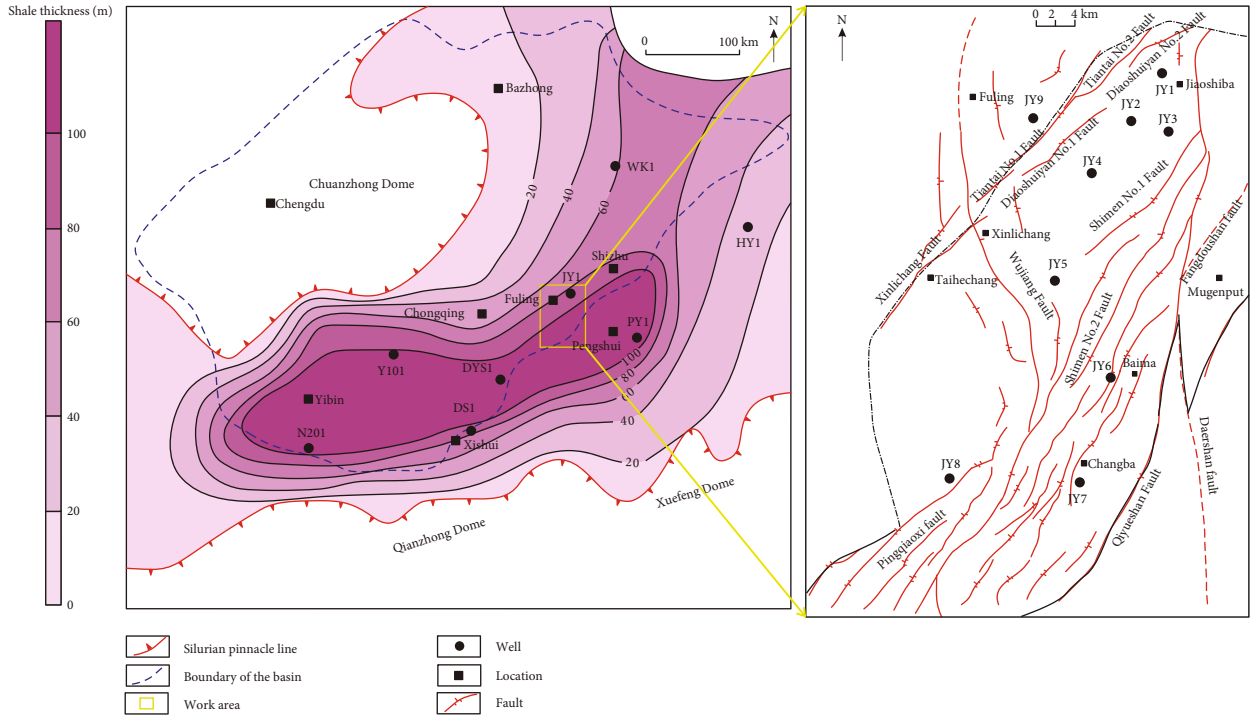

Figure 1: Geological background and distribution of study wells in the Jiaoshiba area (revised according to Wang et al. [30]).

The TOC value is obtained by the following empirical relationship:

$$\mathrm{TOC} = \Delta \mathrm{Log}R * 10^{(2.297-0.1688*\mathrm{LOM})}, \quad (2)$$

where TOC is the total organic carbon content (%) and LOM is the maturity parameter, which can be replaced with the Ro value.

2.2.2. Bayesian Linear Regression Method. Bayesian linear regression model is based on Bayesian inference in statistics [31, 32]. It regards the parameters of the linear model as random variables and finds the posterior by the prior of the model parameters (weight coefficients). This model has the basic properties of a Bayesian statistical model and can obtain the probability density function of the weight coefficient. In addition, it can carry out online learning and model hypothesis testing based on Bayesian factors [32, 33].

The purpose of Bayesian linear regression is not to find the single best value of the model parameters but to determine the posterior distribution of the model parameters. The response variables as well as the model parameters are from the probability distribution. The posterior distribution of the model parameters is based on the input and output of the training data [34]:

$$P(\beta \mid y, X) = \frac{P(y \mid \beta, X) * P(\beta \mid X)}{P(y \mid X)}, \quad (3)$$

where $P(\beta \mid y, X)$ is the posterior probability distribution of the model parameters based on the input and output, $P(y \mid \beta, X)$ is the likelihood probability of the output, $P(\beta \mid X)$ is the prior probability of the parameter β based on the input, and $P(y \mid X)$ is the normalization constant. This formula is a simple expression of the Bayesian theorem, which is the basis of Bayesian inference.

The probability density function of the parameter posterior distribution is as follows:

$$\ln p(w \mid y) = -\frac{\alpha}{2}\sum_{n=1}^{N}\left\{y_n - w^T\phi(x_n)\right\}^2 - \frac{\lambda}{2}w^T w + \text{const}. \quad (4)$$

In the formula, w is the weight coefficient of each vector of the model, α is the noise variance, λ is an individual hyperparameter that can measure the accuracy of w, y is the real target value, x_n is the vector value of each log, $w^T\phi(x_n)$ is the model prediction value, and const is the constant [35, 36].

For the TOC prediction problem, accurate prior information is not available for the weight proportion of each log vector; thus, the noninformation prior must be introduced. That is, the probability distribution of the prior parameter w is obtained by using the spherical Gaussian distribution. After that, the specific process is as follows. (1) Use the training data to build the model. In this process, parameters α and λ can be obtained by maximum likelihood estimation. Another method is to artificially specify an initial value and then update them continuously until the maximum log marginal likelihood is obtained. At this time, the model is most consistent with the actual situation. (2) Use the validation data to verify the accuracy of the model. In this process, the grid search method is used to optimize the hyperparameters of the gamma distribution that α and

λ obey to, so as to obtain the optimal model. (3) Use the optimal model to predict the test data. If the minimum accuracy is met, then the model parameters are returned and the whole dataset is trained to fit the final model. (4) Use the final model to predict the TOC values of other sections or wells.

Bayesian linear regression can solve the problem of overfitting in maximum likelihood estimation because the parameters are regarded as unknown fixed values in the maximum likelihood estimation linear regression, while they are regarded as random variables in the Bayesian linear regression, which is widely used in the field of machine learning. The utilization rate of the data samples is 100% by the Bayesian linear regression method, and the complexity of the model can be effectively and accurately determined by using training samples only. This method is suitable for processing small datasets like log values [37]. It has been applied in lithology recognition, fluid classification, etc., which has achieved good results [38, 39].

2.2.3. Random Forest Method. Random forest method is an integrated learning method for classification, regression, and other tasks, and it uses the prediction results of multiple decision trees to determine the final classification results and regression values. It is essentially a bagging method that uses limited data to obtain many new samples through repeated sampling, constructs multiple independent estimators, and takes the average results for overall prediction. When determining the final output, multiple decision trees are combined. Although a single decision tree has a large variance, the variance of the final comprehensive result can be very low since each decision tree is perfectly trained for a specific sample.

The corresponding basic steps of the algorithm are as follows: (1) Bootstrap sampling with return is carried out from the training data to generate several datasets. Each dataset generates a decision tree through training. (2) When the decision tree is divided into nodes, it is necessary to randomly select several features from all log vectors and make the branches of the optimal feature grow fully until they cannot regenerate. Pruning is not performed in this process. (3) Use out-of-bag data (unselected data) to test the effect and generalization ability of the model, determine the optimal number of decision trees, and rebuild the model. (4) Use the determined model to predict the new data [20, 40].

The main advantage of the random forest algorithm is that each decision tree only uses part of the samples and only extracts some of the attributes for modelling, which enhances the diversity of learners, corrects the habit of the decision tree for overadapting to its training set, and improves the generalization of the model [41]. Especially for the high-dimensional regression problem like TOC prediction, the stability and generalization of the model are more important than the small deviation to some extent. It has been successfully applied in many aspects, such as lithology identification [42], source rock prediction [43], and seismic reservoir prediction [44], and it has the advantages of simplicity and interpretability.

3. Results and Analysis

3.1. Traditional ΔLogR Method. According to the results of previous studies [45], the TOC values predicted by the Δ Log$R$ method have a poor correlation with the measured TOC values. The predicted results cannot objectively reflect the actual total organic carbon content (Figure 2).

3.2. Machine Learning Methods. The machine learning models used for TOC prediction include four steps: data preprocessing, log series selection, hyperparameter selection and model establishment, and model verification and application [46–49]. The specific workflow is as follows: find enough data points and preprocess the data, including deep homing, data cleaning, and data resampling; divide the data into the training set, validation set, and test set; use the training set to optimize the hyperparameters before the learning process because these parameters will affect the performance of the model and cannot be learned by machine learning algorithm; use the optimized hyperparameters to build the model; use the well-trained model to evaluate test set; and extrapolate and apply the evaluated model.

3.2.1. Data Preprocessing. There is often a deviation between the core and log depth because of the low core recovery rate and inaccurate estimation of the core depth, which leads to inconsistencies between the geological characteristics recorded by the core and the log records, which affects the accuracy of geological feature recognition by log data. Under actual geological conditions, the depth of the log records is more accurate than that of the cores. Therefore, the core depth needs to be corrected so that the TOC test sampling points can be calibrated to the log depth.

It is currently believed that the minimum resolution of the log data is 0.1 m, and it is impossible to distinguish two TOC test sampling points that are less than 0.1 m apart. In addition, some data not meeting the statistical significance often have an impact on the establishment of the model; thus, it is necessary to screen the TOC data. This study uses the DataFrames function of the Pandas tripartite library in Python to screen the TOC test points whose depth difference is less than 0.1 m and the invalid values beyond the mean value plus or minus 3 times the variance.

The log data are sampled uniformly and densely throughout the well section, which can be approximately regarded as a continuous variable, while the measured TOC values are discrete data with a fixed depth. There is a certain mismatch between the two in the sampling depth. A resampling operation is required for two kinds of data with different sampling intervals. At present, the commonly used resampling methods in the field of log technology include fast Fourier transform, Gaussian convolution, window data shift, linear transformation, and linear antialiasing [50, 51]. Considering that the log data generally present continuous linear transformation on the well section, this paper chose the linear transformation reprocessing method in the implementation process and set the maximum interval to 0.25 m (2 log intervals), which avoids the subjectiveness of manually selecting data.

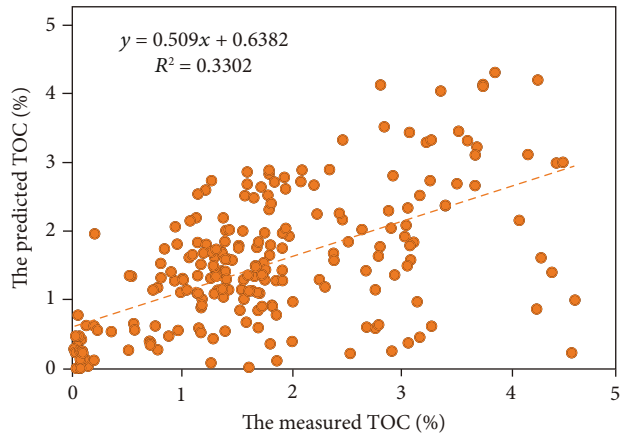

Figure 2: Correlation between the measured TOC and the predicted TOC by the traditional $\Delta \mathrm{Log}R$ method [45].

After preprocessing, 386 groups of modelling data from JY1, JY3, JY4, JY5, JY8, and JY9 and 242 groups of prediction test data from JY2, JY6, and JY7 are finally obtained. Each group of data includes RT, GR, AC, CAL, CNL, DEN, and SP log values and corresponding measured TOC value. Their statistical information is shown in Table 1 and Table 2. Statistical analyses show that most measured TOC values are less than 6%, which is generally low. Individual values exceeding this range are regarded as outliers and removed.

3.2.2. Log Series Selection. The accuracy of the machine learning methods used to predict the TOC values largely depends on the input data. If the correlation between the log and TOC values is weak or too complicated, then it is easy for the algorithm to learn the wrong function relationship with small numbers of samples, which may result in oversimulation. Therefore, it is necessary to analyze the correlation between the log data and the TOC values before building the machine learning model. Generally, more selected features correspond to more log series, more information that can be covered, and a more accurate model. However, redundant features will also affect the accuracy of the calculation and the generalization of the model.

Considering the difficulty of acquiring log data, this paper mainly uses the commonly available conventional log parameters (GR, SP, CAL, DEN, CNL, AC, and RT) to predict the TOC. Before modelling, it is necessary to perform a preliminary correlation analysis on the selected log series and the measured TOC values. This process can avoid overfitting caused by weak correlation or complex relationships between the log and TOC values. In statistical analyses, the Pearson product-moment correlation coefficient (Pearson's r) is widely used to measure the degree of linear correlation between two variables, and its value is between -1 and 1. A positive number indicates a positive correlation, and a negative number indicates a negative correlation. The closer the absolute value is to 1, the higher the correlation between the two variables [52]. The Pearson matrix can be used to analyze the correlation between different log curves and the measured TOC (Figure 3). In Figure 3, the number in each block is the Pearson's r of the two variables corresponding to the row and column. The Pearson's r values of the measured TOC value and GR, DEN, AC, and CNL are 0.65, -0.9, 0.48, and -0.67, respectively, which have relatively good correlations. For the regression prediction model, redundant information with low correlation needs to be excluded. According to the actual condition in the work area, the prediction of TOC is carried out by using log series with correlation coefficients greater than or equal to 0.2 as the input data, including GR, SP, DEN, AC, and CNL.

3.2.3. Hyperparameter Selection and Model Establishment. To enhance the generalization ability of the machine learning models, it is necessary to use cross-validation methods to optimize the hyperparameters. Then, the selected optimal hyperparameters are used to overcome overlearning and improve the prediction performance [53]. In this paper, the modelling data are divided into a training dataset, verification dataset, and test dataset at a ratio of 6 : 3 : 1. The TOC distribution of different datasets is similar to ensure that the results obtained from cross-validation are meaningful. The training data are the initial learning data for building the model. The verification data are used to test the accuracy of the model with different hyperparameters and screen the best hyperparameter. The test data will not participate in the establishment of the model or the selection of the model, although they will be used to test the accuracy of the final model. The accuracy of the model obtained by the test datasets can reflect the extrapolation ability of the model to a certain extent, which increases the credibility of the model.

The loss functions commonly used in cross-validation are the mean square error (MSE), mean absolute error (MAE), explained variance score (EVS), and coefficient of determination (R^2) [54]. The calculation needs to be repeated many times in the cross-validation. To save calculation costs and avoid losing accuracy, the mean absolute error (MAE) is used as the loss function in this paper. The formula is as follows [55]:

$$\mathrm{MAE}(y, \hat{y}) = \frac{1}{n_{\mathrm{samples}}} \sum_{i=0}^{n_{\mathrm{samples}}-1} |y_i - \hat{y}_i|, \tag{5}$$

where y is the real target value, $\hat{y}$ is the estimated target value, n_{samples} is the number of samples, y_i is the real target value of the i-th sample, and $\hat{y}_i$ is the estimated target value of the i-th sample.

Moreover, the coefficient of determination (R^2) is used as the standard for evaluation in the test set. A value closer to 1 corresponds to a better final regression prediction result, and a value closer to 0 corresponds to a worse regression prediction result. The formula is as follows [56]:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n_{\mathrm{samples}}-1} (y_i - y\wedge_i)^2}{\sum_{i=0}^{n_{\mathrm{samples}}-1} (y_i - \bar{y})^2}, \tag{6}$$

Table 1: Statistical information of the input data for building the model.

Statistical information	RT (Ω·m)	GR (API)	AC (μs/ft)	CAL (inch)	CNL (%)	DEN (g/cm^3)	SP (mV)	TOC (%)
Maximum value	137.11	247.88	88.29	13.39	25.06	2.74	137.21	6.02
Minimum value	5.78	100.52	60.53	8.34	7.70	2.47	5.43	0.01
Average value	41.07	162.66	73.57	9.78	16.92	2.64	63.92	1.92
Standard deviation	21.13	25.72	4.39	1.84	3.18	0.06	36.17	1.21

Table 2: Statistical information of the input data for verifying the universality of the model.

Statistical information	RT (Ω·m)	GR (API)	AC (μs/ft)	CAL (inch)	CNL (%)	DEN (g/cm^3)	SP (mV)	TOC (%)
Maximum value	351.04	253.64	89.72	12.20	23.87	2.74	96.68	5.43
Minimum value	6.33	110.41	62.57	8.63	9.87	2.43	18.51	0.26
Average value	74.44	169.31	72.40	8.89	16.27	2.65	47.36	2.11
Standard deviation	54.00	18.66	6.27	0.46	3.13	0.07	21.92	1.12

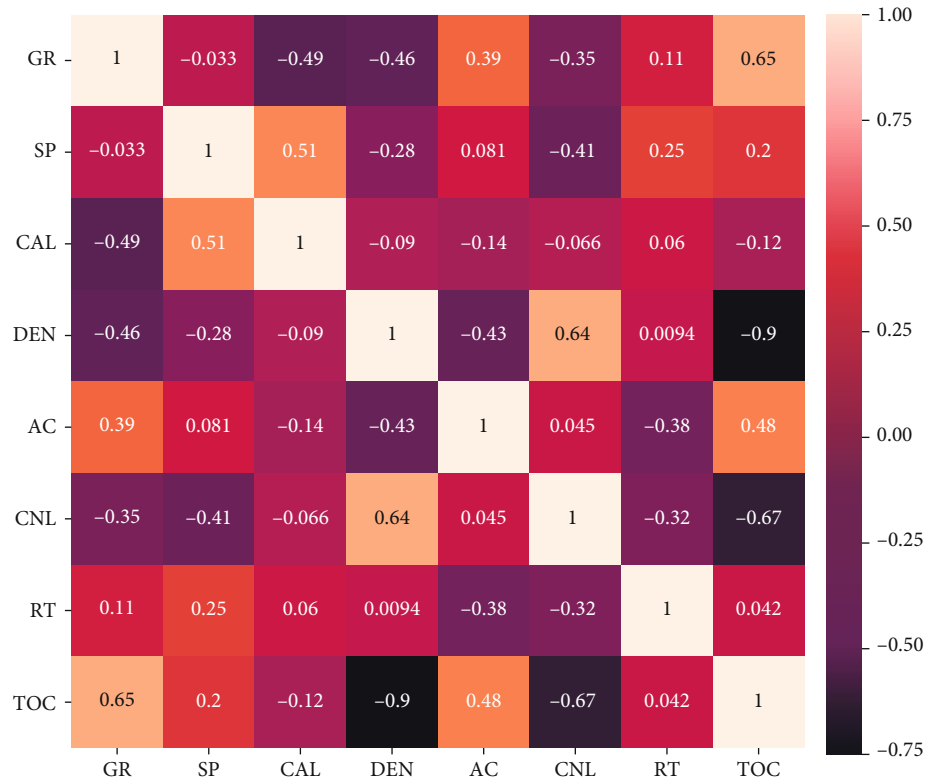

Figure 3: Heat map of Pearson's correlation coefficients.

where y is the true target value, $\hat{y}$ is the estimated target value, n_{samples} is the number of samples, y_i is the true target value of the i-th sample, $\hat{y}_i$ is the estimated target value of the i-th sample, and $\bar{y}$ is the average value of the true target value.

In general, it is important to choose the initial value of the regularization parameter (α, λ) when fitting a curve to a polynomial by Bayesian linear regression method because the regularization parameter is determined by an iterative process that depends on the initial value [57]. Whether the regularization parameters are the default values ($\alpha\_\text{init} = 0.74$, $\lambda\_\text{init} = 1.00$) or the relative extreme values ($\alpha\_\text{init} = 100.00$, $\lambda\_\text{init} = 0.001$), the training results are good. The sample data can be considered relatively consistent with the Gaussian prior; therefore, the results do not depend on the initial values. Moreover, a comparison between the test set and the training set (Figure 4) shows that the generalization of the model is relatively good, with an R^2 score of 0.8997.

The random forest regression model also uses the cross-validation to optimize the hyperparameters. However, the difference is that the random forest method has too many

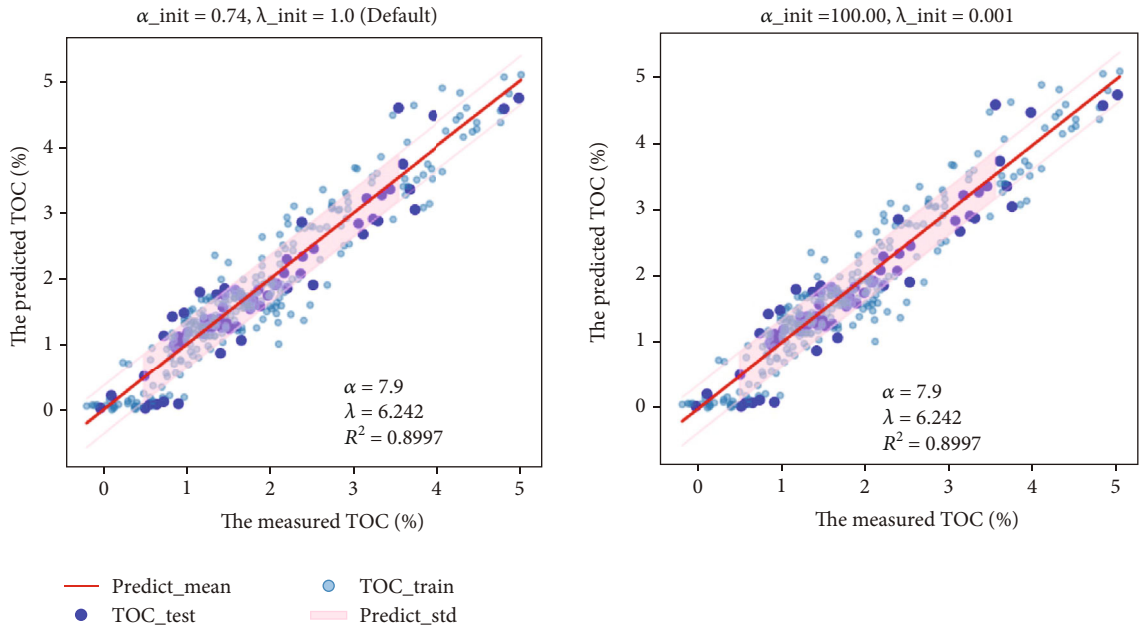

Figure 4: Training and testing results of the Bayesian linear regression model with different hyperparameters.

hyperparameters, such as the maximum depth of the tree (max_depth), the minimum number of samples required to split the internal nodes (min_sample_split), the maximum number of features to be considered when looking for the best split (max_features), and the number of samples for training each basic estimator (max_samples) [33]. Conventional search methods, such as grid search algorithms, can exhaust all parameter combinations. However, the efficiency is too low, which will cause a waste of computing power. Therefore, it is necessary to use a randomized search algorithm for optimization. Randomized search cross-validation samples a fixed number of hyperparameters from a given distribution. Because not all the parameters are sampled, it can improve the speed of operation. However, the speed of searching for a good combination of parameters is still not ideal, which takes approximately 20 minutes.

To solve this problem, previous studies proposed the halving random search cross-validation [58, 59], which is an iterative selection process in which all parameter combinations (replaced with candidates in the following part) use a small amount of resources for the evaluation in the first iteration and only some candidates are selected in the next iteration; therefore, more resources will be allocated. In other words, the search strategy begins to use a small amount of resources to evaluate all candidates and uses an increasing number of resources to iteratively select the best candidate. The resource usually refers to the number of training samples and can also be any numeral parameters, such as the number of basic estimators in the random forest algorithm.

As shown in Figure 5, in the first iteration, a small amount of resources (the number of samples) were used to evaluate all candidates. In the second iteration, only the better half of the previous candidates were evaluated, while the number of resources allocated doubled. This process was repeated until the last iteration, in which only 2 candidates remained. With the iteration and the increase of input samples, different candidates were eliminated according to the score of the verification set, and then, the hyperparameters were optimized. The line segments with different colors in Figure 5 represent different candidates (parameter combinations), thus reflecting the score changes in their verification sets during the iterative process. As the number of iterations (abscissa) increases, candidates with low scores (ordinate) are eliminated, and candidates with high scores continue to participate in the next iteration. Only one candidate will remain until the end of the iteration process. The best parameter candidate is the one with the highest score in the last iteration (the black line), resulting in the best hyperparameter combination: {'bootstrap': False, 'criterion': 'mae', 'max_depth': None, 'max_features': 1, 'min_samples_split': 4}. In this case, R^2 of the verification set is 0.9464.

The performance of the halving random search cross-validation method in the test set was statistically analyzed. As shown in Figure 6, the best R^2 value is approximately 0.9 and the lowest MAE is approximately 0.3. The whole search was completed in only 19.2 seconds, and the speed increased by more than 60 times.

However, the prediction effect of the above random forest model is not ideal. Especially when the TOC value is less than 0.6%, the relative error between the predicted value and the measured value is large. After excluding certain factors, such as algorithms and parameters, they are considered related to the input data structure. In the process of the log data acquisition, noise will be inevitably generated due to interference from the environment and random factors, which brings errors to the calculation of geological parameters [60]. In addition to the TOC variation, the factors that affect the changes in log values include abnormal fluid pressure, hydrocarbons, tight reservoirs, overmature organic

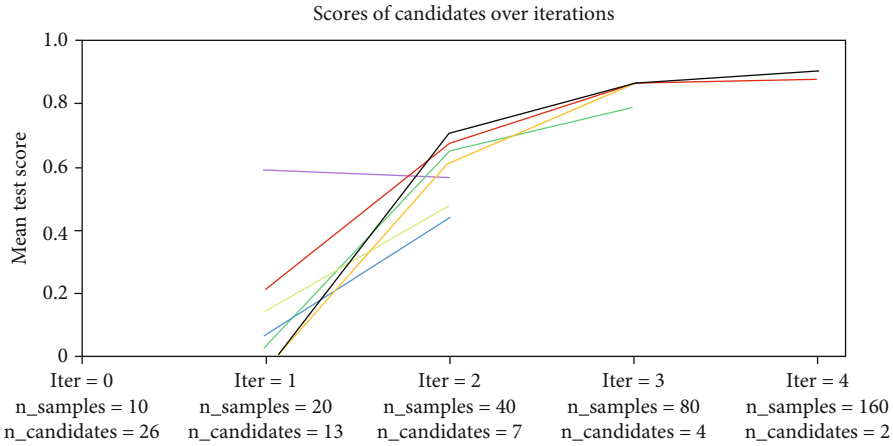

Figure 5: Schematic diagram of the iterative process of the halving random search cross-validation (The line segments with different colors in the figure represent some typical parameter combinations.).

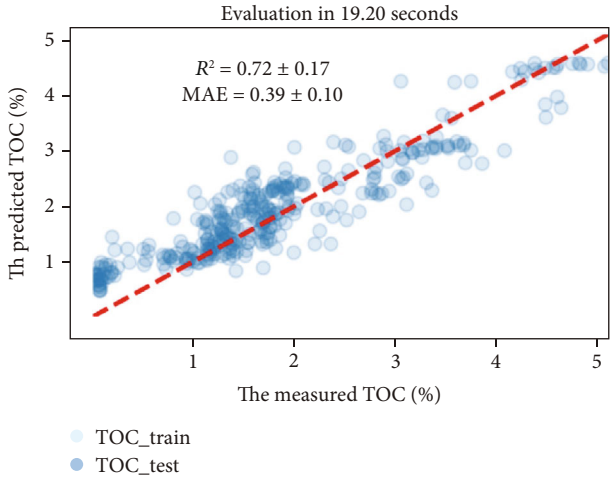

Figure 6: Results of the halving random search cross-validation in the test set.

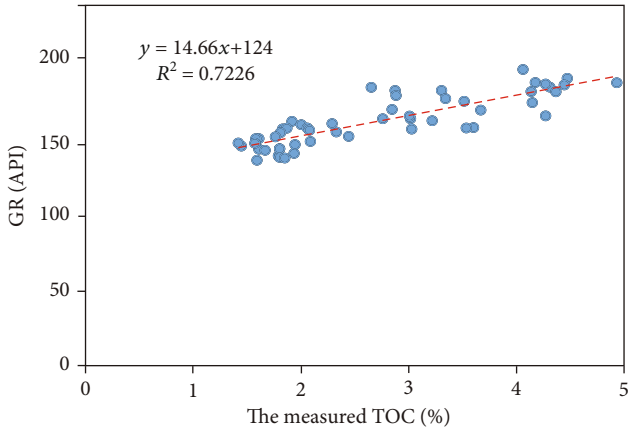

Figure 7: Relationship between GR and the measured TOC in well JY1.

matter, and other formation information [61]. The fluctuation of log values caused by these factors may cover the log variation caused by the TOC, especially when the TOC value is small. The contribution of other formation information to the variation in log values may be much greater than that of the TOC, resulting in inaccurate prediction results. Therefore, when using log data to predict the TOC values, it is better to exclude the interference of unrelated factors in advance. Referring to previous research work [45], the log base value corresponding to the TOC of 0% should be found based on the correlation between the log values and the measured TOC values, and then, the input data should be changed to the absolute value after the actual log value minus the log base value. For example, the relationship between the GR value and the measured TOC value of well JY1 is shown in Figure 7, which can be expressed as the following formula:

$$\mathrm{GR} = 14.66 \times \mathrm{TOC} + 124 \left(R^2 = 0.7226\right). \quad (7)$$

As a result, the base value GRb is 124 API with the TOC value of 0%. If the input value of this part of well JY1 is defined as GR′, then $\mathrm{GR}' = |\mathrm{GR} - \mathrm{GRb}| = |\mathrm{GR} - 124|$. If a measured TOC value is not available to confirm the base value where TOC is 0%, the average value of the predicted shale interval with a relatively small GR, AC, and RT, relatively large DEN and CNL, and no obvious changes in log value is taken as the base value. According to the above method, the modelling process is repeated after reprocessing the input data. The result is shown in Figure 8. R^2 increased by approximately 5%, and the large error when the TOC value is less than 0.6% has been reduced.

3.2.4. Extrapolation and Application of the Model. Based on the models established by the two machine learning methods, TOC prediction of the other three wells JY2, JY6, and JY7 in the study area was carried out. The comprehensive results are shown in Figure 9. R^2 is above 0.85, the mean absolute error (MAE) is approximately 0.3, and the mean relative error (MRE) is approximately 0.2 (Table 3). The values with large relative error mainly occur in the part

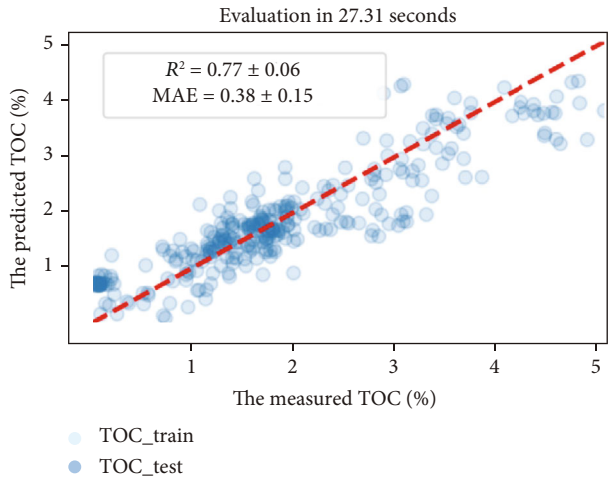

Figure 8: Modelling effect diagram of the random forest method after reprocessing the input data.

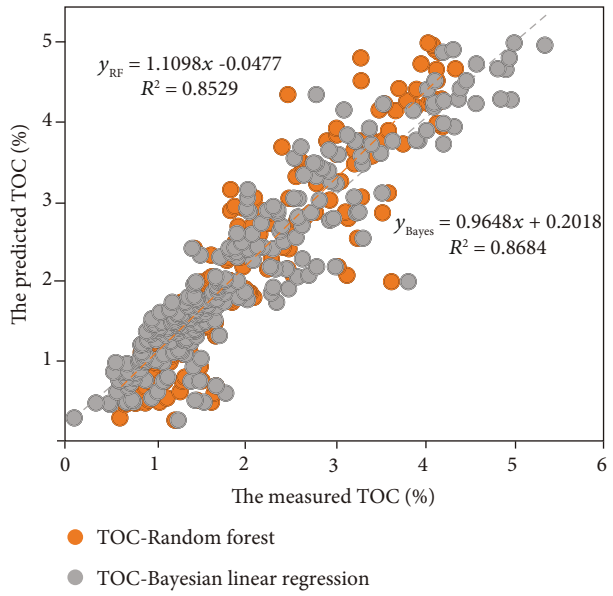

Figure 9: Comparison of the measured TOC value and predicted TOC value of the two machine learning methods.

Table 3: Evaluation table of the TOC prediction results of two machine learning methods.

Evaluation index	Random forest	Bayesian linear regression
MAE	0.3338	0.3241
MRE	0.2161	0.2064
R^2	0.8529	0.8684
Maximum relative error	1.4195	1.4181
Minimum relative error	0.0013	0.0006

where the TOC value is less than 0.6% because the log values of this part are greatly disturbed by factors other than the TOC. However, on the whole, the model has strong extrapolation ability and good generalization.

In addition, a comparison of the two machine learning methods with the traditional ΔLog$R$ method by extrapolation (Figure 10) showed that these two machine learning methods lead to great accuracy improvements in the results.

4. Discussion

The greatest advantage of the traditional ΔLog$R$ method is that it can eliminate the influence of porosity on the log response of organic carbon. However, it is not reasonable to use a fixed empirical coefficient with many limitations to predict TOC [62, 63]. In addition, the amplitude difference of only two log curves is used to calculate the organic matter content, and other important log information may be ignored, resulting in poor anti-interference ability of the model. Machine learning-based methods can synthesize a variety of log information to predict TOC. The results show that a large amount of geophysical information can reflect the changes in the composition of materials in formations from different physical quantities. The method integrated from a variety of information has a relatively better anti-interference ability.

In this study, using a variety of shale data from the Wufeng Formation and the lower part of the Longmaxi Formation in the Sichuan Basin, the accuracy of TOC prediction by Bayesian linear regression and the random forest method is more than 50% higher than that by the traditional ΔLog$R$ method. Of the two, the Bayesian linear regression model is more accurate. This method includes relevant domain knowledge and the guess of the model parameters. It assumes that not all the required parameter information will be provided by the available data, breaking through the limitations from data itself. If there are no prediction in advance, no prior information can be used for the parameters, which facilitates the construction of the model.

In the field of machine learning, the random forest method is more suitable for regression problems than other common algorithms, especially the problem of nonlinear or complex relationships between elements and labels similar to TOC prediction problem. It uses a set of irrelevant decision trees on the subsamples of the data and improves the prediction accuracy and reduces the variance by averaging. It is insensitive to the noise in the training set and thus is more conducive to obtaining a robust model to avoid overfitting. However, due to the need to connect a large number of decision trees together, general parameter optimization methods require considerable training time. Therefore, attention should be given to the method of parameter optimization in TOC prediction. In this paper, the halving random search cross-validation method was used to optimize the hyperparameters in the random forest model, which greatly improved the learning efficiency and increased the calculation speed by more than 60 times. In other words, the use of a well-trained machine learning model can quickly and easily predict the organic carbon content of shale.

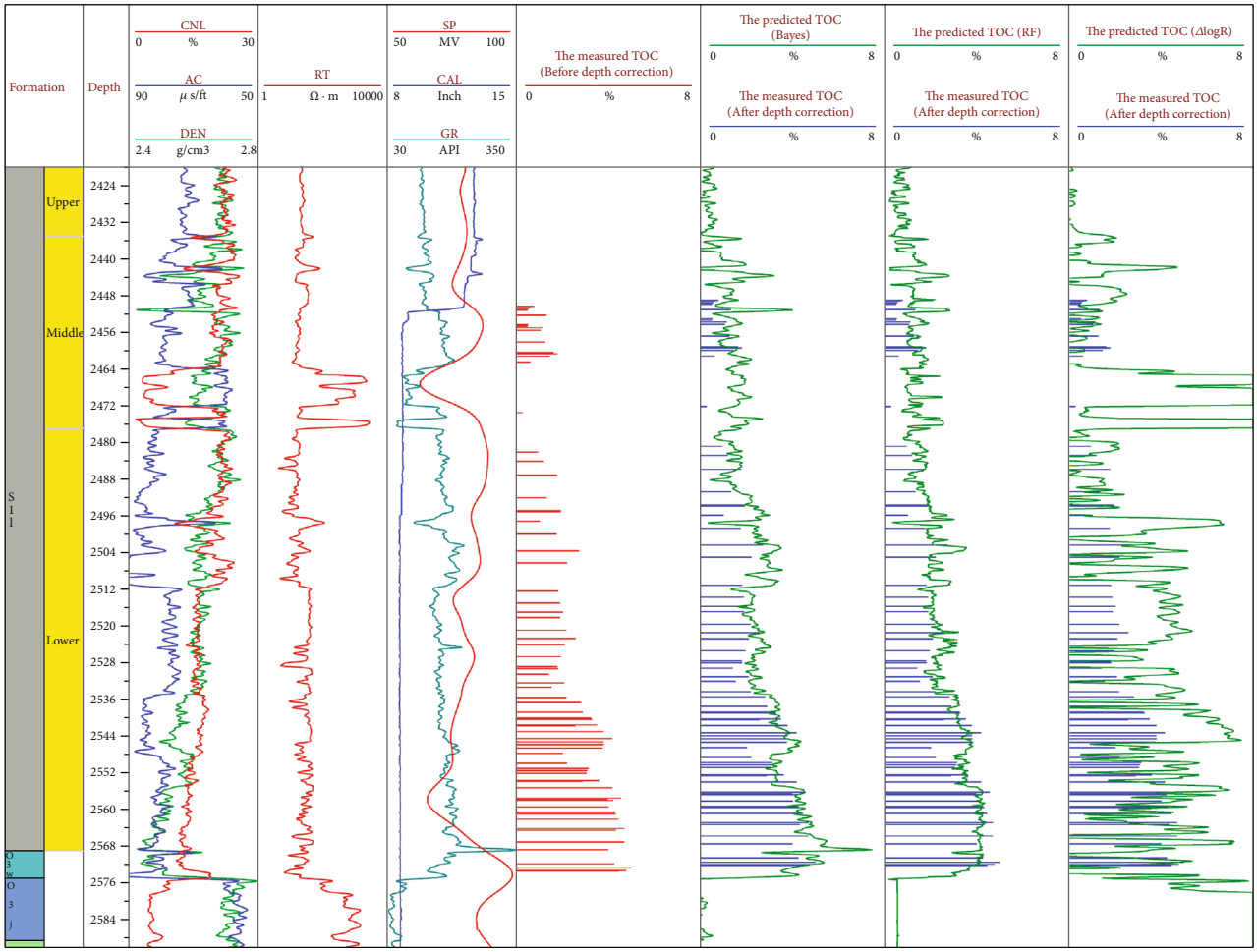

Figure 10: Comparison of the TOC prediction results of different methods in well JY2.

In addition, the machine learning model can be updated conveniently. If there is a new dataset, the machine learning model can be upgraded and provide broader applications. Therefore, compared with traditional methods, machine learning models are data-driven based, thereby avoiding a large number of theoretical assumptions and mathematical derivations. Moreover, it should be noted that the input data structure has a great impact on the building of the machine learning model, so the data preprocessing is very important before training. In response to TOC prediction, this paper provided a new data preprocessing strategy. It was to eliminate the log value changes caused by factors other than TOC before inputting, which improved the prediction accuracy by approximately 5%.

The machine learning models proposed in this paper can provide more accurate prediction results using both training data and test data with reasonable extrapolation. However, from the perspective of application, it has certain limitations. First, the data used in this paper are from the same research area with similar geological conditions. Thus, the reliability of the model needs to be further verified in other areas with large differences in geological conditions. Second, due to the frequent changes of sedimentary water properties in geological history, the heterogeneity of shale strata is strong, and the TOC values vary greatly. The limited TOC values may not fully reflect the relevant characteristics of the entire formation. The applicability of the model is unknown for strata that are not covered by the TOC test. In addition, the measured TOC values used in this study range from 0.01% to 6.02%. The applicability of the model in the case of higher TOC values is not discussed. For areas with a TOC value less than 0.6%, it is necessary to perform further research to improve the prediction accuracy. In this regard, it can be considered to collect the TOC test data and corresponding log data from different basins, sedimentary environments, and structural backgrounds, and a more comprehensive model should be built in a large numerical framework. Therefore, the general relationship between the TOC value and log value can be found by this method.

5. Conclusion

This paper uses Bayesian linear regression and random forest algorithms to predict TOC values. Compared with the ΔLog$R$ method, both machine learning methods have higher TOC prediction accuracy and better generalization in overmature and tight shale in the study area. When the random forest method is used for modelling, the halving random

search cross-validation can be applied to find the optimal hyperparameters and improve the training speed. The log data with the corresponding log base value removed can be taken as the input data for modelling. Thus, the factors other than TOC that affect the log values can be avoided to ensure the accuracy of the predicted results. In addition, if a new dataset is provided, the machine learning model can be updated more conveniently, which is of great significance for improving the efficiency of shale gas exploration and development.

Data Availability

The main data used to support the findings of this study are included within the article, and the others are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study is supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA14010202) and the National Science and Technology Major Project (2017ZX05008–004).

References

[1] C. N. Zou, S. Q. Pan, Z. H. Jing et al., "Shale oil and gas revolution and its impact," *Acta Petrolei Sinica*, vol. 41, no. 1, pp. 1–12, 2020.

[2] C. N. Zou, M. G. Zhai, G. Y. Zhang et al., "Formation, distribution, potential and prediction of global conventional and unconventional hydrocarbon resources," *Petroleum Exploration and Development*, vol. 42, no. 1, pp. 13–25, 2015.

[3] Z. J. Jin, Z. Q. Hu, B. Gao, and J. H. Zhao, "Controlling factors on the enrichment and high productivity of shale gas in the Wufeng-Longmaxi Formations, southeastern Sichuan Basin," *Earth Science Frontiers*, vol. 23, no. 1, pp. 001–010, 2016.

[4] 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.

[5] D. M. Jarvie, "Shale resource systems for oil and gas: part 2-shale-oil resource systems," in *Shale Reservoirs-Giant Resources for the 21st Century*, J. A. Breyer, Ed., Worldwide Geochemistry, 2012.

[6] B. Liu, Y. Song, K. Zhu, P. Su, X. Ye, and W. Zhao, "Mineralogy and element geochemistry of salinized lacustrine organic-rich shale in the Middle Permian Santanghu Basin: implications for paleoenvironment, provenance, tectonic setting and shale oil potential," *Marine and Petroleum Geology*, vol. 120, article 104569, 2020.

[7] R. F. Beers, "Radioactivity and organic content of some Paleozoic shales," *AAPG Bulletin*, vol. 29, no. 1, pp. 1–22, 1945.

[8] J. W. Schmoker, "Determination of organic-matter content of Appalachian Devonian shales from gamma-ray logs," *AAPG Bulletin*, vol. 65, pp. 1285–1298, 1981.

[9] J. W. Schmoker and T. C. Heste, "Organic carbon in Bakken Formation, United States portion of Williston Basin," *AAPG Bulletin*, vol. 67, pp. 2165–2174, 1983.

[10] B. L. Meyer and M. H. Nederlof, "Identification of source rocks on wireline logs by density/resistivity and sonic transit time/resistivity cross plots," *AAPG Bulletin*, vol. 68, pp. 121–129, 1984.

[11] Q. R. Passey, S. Creaney, and J. B. Kulla, "A practical model for organic richness from porosity and resistivity logs," *AAPG Bulletin*, vol. 74, no. 6, pp. 1777–1794, 1990.

[12] L. Q. Zhu, C. M. Zhang, S. Zhang, X. Q. Zhou, and W. N. Liu, "An improved method for evaluating the TOC content of a shale formation using the dual-difference ΔlogR method," *Marine and Petroleum Geology*, vol. 1025, pp. 800–816, 2019.

[13] C. Liu, W. Zhao, L. Sun, Y. Zhang, X. Chen, and J. Li, "An improved ΔlogR model for evaluating organic matter abundance," *Journal of Petroleum Science and Engineering*, vol. 206, article 109016, 2021.

[14] J. D. Mendelzon and M. N. Toksoz, "Source rock characterization using multivariate analysis of log data," in *SPWLA 26th Annual Log Symposium*, Dallas, Texas, USA, 1985.

[15] V. Bolandi, A. Kadkhodaie, and R. Farzi, "Analyzing organic richness of source rocks from well log data by using SVM and ANN classifiers: a case study from the Kazhdumi Formation, the Persian Gulf Basin, offshore Iran," *Journal of Petroleum Science & Engineering*, vol. 151, pp. 224–234, 2017.

[16] 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.

[17] M. R. Shalaby, N. Jumat, D. Lai, and O. Malik, "Integrated TOC prediction and source rock characterization using machine learning, well logs and geochemical analysis: case study from the Jurassic source rocks in Shams Field, NW Desert, Egypt," *Journal of Petroleum Science and Engineering*, vol. 176, pp. 369–380, 2019.

[18] A. Liaw and M. Wiener, "Classification and regression by random forest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[19] G. Huang, G. B. Huang, S. Song, and K. You, "Trends in extreme learning machines: a review," *Neural Networks*, vol. 61, pp. 32–48, 2015.

[20] M. G. Feng, W. Yan, X. M. Ge, and L. Q. Zhu, "Predicting total organic carbon content by random forest regression algorithm," *Bulletin of Mineralogy, Petrology and Geochemistry*, vol. 37, no. 3, pp. 475–481, 2018.

[21] Y. H. Sun, *Research on prediction method of total organic carbon in shale based on machine learning, [M.S. thesis]*, China University of Petroleum, Beijing, 2019.

[22] W. J. Zhao, H. Y. Gao, G. L. Yan, and T. C. Guo, "TOC prediction technology based on optimal estimation and Bayesian statistics," *Lithologic Reservoirs*, vol. 32, no. 1, pp. 86–93, 2020.

[23] A. M. 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, aricle 104347, 2020.

[24] Y. F. Li, D. Y. Shao, H. G. Lv, Y. Zhang, X. L. Zhang, and T. W. Zhang, "A relationship between elemental geochemical characteristics and organic matter enrichment in marine shale of Wufeng Formation-Longmaxi Formation, Sichuan Basin," *Acta Petrolei Sinica*, vol. 36, no. 12, pp. 1470–1483, 2015.

[25] C. M. Zhang, W. S. Zhang, and Y. H. Guo, "Sedimentary environment and its effect on hydrocarbon source rocks of Longmaxi Formation in Southeast Sichuan and Northern Guizhou," *Earth Science Frontiers*, vol. 19, pp. 136–145, 2012.

[26] X. S. Guo, *Enrichment Mechanism and Exploration Technology of Jiaoshiba Area in Fuling Shale Gas Field*, Science Press, Beijing, 2014.

[27] S. G. Liu, W. X. Ma, J. Luba, W. M. Huang, X. L. Zeng, and C. J. Zhang, "Characteristics of the shale gas reservoir rocks in the Lower Silurian Longmaxi Formation, East Sichuan Basin, China," *Acta Petrologica Sinica*, vol. 27, no. 8, pp. 2239–2252, 2011.

[28] R. B. Liu, "Analyses of influences on shale reservoirs of Wufeng-Longmaxi Formation by overpressure in the southeastern part of Sichuan Basin," *Acta Sedimentologica Sinica*, vol. 33, no. 4, pp. 817–827, 2015.

[29] R. V. Tyson, "Sedimentation rate, dilution, preservation and total organic carbon: some results of a modelling study," *Organic Geochemistry*, vol. 32, no. 2, pp. 333–339, 2001.

[30] C. Wang, B. Q. Zhang, Y. C. Lu et al., "Lithofacies distribution characteristics and main development controlling factors of shale in Wufeng Formation-Member 1 of Longmaxi Formation in Jiaoshiba area," *Acta Petrolei Sinica*, vol. 39, no. 6, pp. 631–644, 2018.

[31] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes in Machine Learning*, MIT Press, 2006.

[32] J. Wakefield, *Bayesian and Frequentist Regression Methods*, Springer Science & Business Media, 2013.

[33] D. Polykovskiy and A. Novikov, *Bayesian Methods for Machine Learning*, Coursera and National Research University Higher Economics, 2017.

[34] K. W. Fornalski, "Applications of the robust Bayesian regression analysis," *International Journal of Society Systems Science*, vol. 7, no. 4, pp. 314–333, 2015.

[35] D. J. Mackay, "Bayesian interpolation," *Neural Computation*, vol. 4, no. 3, pp. 415–447, 1992.

[36] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

[37] R. M. Neal, *Bayesian Learning for Neural Networks, vol. 118*, Springer Science & Business Media, 2012.

[38] K. Rimstad and H. Omre, "Impact of rock-physics depth trends and Markov random fields on hierarchical Bayesian lithology/fluid prediction," *Geophysics*, vol. 75, no. 4, pp. R93–R108, 2010.

[39] Y. Xie, C. Zhu, W. Zhou, X. Liu, M. Tu, and M. Tu, "Evaluation of machine learning methods for formation lithology identification: a comparison of tuning processes and model performances," *Journal of Petroleum Science and Engineering*, vol. 160, pp. 182–193, 2018.

[40] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[41] T. K. Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, pp. 278–282, Montreal, QC, Canada, 1995.

[42] 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.

[43] L. Zhao, J. Liu, Y. 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 (in Chinese)*, vol. 64, no. 2, pp. 700–715, 2021.

[44] J. G. Song, Q. S. Gao, and Z. Li, "Application of random forests for regression to seismic reservoir prediction," *Oil Geophysical Prospectin*, vol. 51, no. 6, pp. 1202–1211, 2016.

[45] J. Rong, X. R. Luo, Y. P. Li, C. Li, Q. Wei, and G. Yu, "Improved TOC prediction method for shale reservoir of Wufeng Formation and the lower part of Longmaxi Formation in Jiaoshiba area, Sichuan Basin," *IOP Conference Series: Earth and Environmental Science*, vol. 600, no. 1, pp. 012–023, 2020.

[46] S. Elkatatny, "A self-adaptive artificial neural network technique to predict total organic carbon (TOC) based on well logs," *Arabian Journal for Science and Engineering*, vol. 44, no. 6, pp. 6127–6137, 2019.

[47] L. Zhu, C. Zhang, C. Zhang et al., "A new and reliable dual model- and data-driven TOC prediction concept: a TOC logging evaluation method using multiple overlapping methods integrated with semi-supervised deep learning," *Journal of Petroleum Science and Engineering*, vol. 188, article 106944, 2020.

[48] D. Ahangari, R. Daneshfar, M. Zakeri, S. Ashoori, and B. S. Soulgani, *On the Prediction of Geochemical Parameters (TOC, S1 and S2) by Considering Well Log Parameters Using ANFIS and LSSVM Strategies*, Petroleum, 2021.

[49] M. A. Meng, R. Zhong, and Z. Wei, "Prediction of methane adsorption in shale: classical models and machine learning based models," *Fuel*, vol. 278, no. 15, article 118358, 2020.

[50] S. R. Cai, "An Akima interpolation method for borehole data resampling," *Well Log Technology*, vol. 28, no. 2, pp. 112–114, 2004.

[51] A. N. Peng and D. P. Cao, "Research and application of log lithology identification based on deep learning," *Progress in Geophysics (in Chinese)*, vol. 33, no. 3, pp. 1029–1034, 2018.

[52] M. T. Puth, M. Neuhäuser, and G. D. Ruxton, "Effective use of Pearson's product-moment correlation coefficient," *Animal Behaviour*, vol. 93, pp. 183–189, 2014.

[53] M. W. Browne, "Cross-validation methods," *Journal of Mathematical Psychology*, vol. 44, no. 1, pp. 108–132, 2000.

[54] P. Christoffersen and K. Jacobs, "The importance of the loss function in option valuation," *Journal of Financial Economics*, vol. 72, no. 2, pp. 291–318, 2004.

[55] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Research*, vol. 30, no. 1, pp. 79–82, 2005.

[56] N. R. Draper and H. Smith, *Applied Regression Analysis, vol. 326*, John Wiley & Sons, 1998.

[57] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[58] K. Jamieson and A. Talwalkar, "Non-stochastic best arm identification and hyperparameter optimization," in *Proceedings of the 19th international conference on artificial intelligence and statistics*, pp. 240–248, 2016.

[59] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, "Hyperband: a novel bandit-based approach to hyperparameter optimization," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6765–6816, 2017.

[60] M. Jamshidian, M. Hadian, M. M. Zadeh, Z. Kazempoor, P. Bazargan, and H. Salehi, "Prediction of free flowing porosity

and permeability based on conventional well logging data using artificial neural networks optimized by imperialist competitive algorithm - a case study in the South Pars gas field," *Journal of Natural Gas Science and Engineering*, vol. 24, pp. 89–98, 2015.

[61] C. Li, L. K. Zhang, X. R. Luo et al., "Calibration of the mudrock compaction curve by eliminating the effect of organic matter in organic-rich shales: application to the southern Ordos Basin, China," *Marine and Petroleum Geology*, vol. 86, pp. 620–635, 2017.

[62] C. Liu, C. H. Yin, and S. F. Lu, "Predicting key parameters for variable-coefficient Δ lgR log technique and its application in source rocks evaluation," *Natural Gas Geoscience*, vol. 26, no. 10, pp. 1925–1931, 2015.

[63] C. Liu, L. D. Sun, J. Li et al., "Evaluation of the TOC of source rocks in lacustrine basins using the variable-coefficient ΔLg R technique-a case study of the Xujiaweizi Fault Depression in the Songliao Basin," *Interpretation*, vol. 7, no. 86, pp. 67–75, 2019.