*energies*

MDPI

*Article*

# Prediction of TOC in Lishui–Jiaojiang Sag Using Geochemical Analysis, Well Logs, and Machine Learning

**Xu Han [1,2], Dujie Hou [1,2,*], Xiong Cheng [1,2], Yan Li [1,2], Congkai Niu [1,2] and Shuosi Chen [2]**

[1] School of Energy Resources, China University of Geosciences, Beijing 100083, China
[2] Key Laboratory of Marine Reservoir Evolution and Hydrocarbon Accumulation Mechanism, Ministry of Education, China University of Geosciences, Beijing 100083, China
* Correspondence: houdj313@163.com

**Abstract:** Total organic carbon (TOC) is important geochemical data for evaluating the hydrocarbon generation potential of source rocks. TOC is commonly measured experimentally using cutting and core samples. The coring process and experimentation are always expensive and time-consuming. In this study, we evaluated the use of three machine learning (ML) models and two multiple regression models to predict TOC based on well logs. The well logs involved gamma rays (GR), deep resistivity (RT), density (DEN), acoustic waves (AC), and neutrons (CN). The ML models were developed based on random forest (RF), extreme learning machine (ELM), and back propagation neural network (BPNN). The source rock of Paleocene Yueguifeng Formation in Lishui–Jiaojiang Sag was taken as a case study. The number of TOC measurements used for training and testing were 50 and 27. All well logs and selected well logs (including AC, CN, and DEN) were used as inputs, respectively, for comparison. The performance of each model has been evaluated using different factors, including $R^2$, MAE, MSE, and RMSE. The results suggest that using all well logs as input improved the TOC prediction accuracy, and the error was reduced by more than 30%. The accuracy comparison of ML and multiple regression models indicated the BPNN was the best, followed by RF and then multiple regression. The worst performance was observed in the ELM models. Considering the running time, the BPNN model has higher prediction accuracy but longer running time in small-sample regression prediction. The RF model can run faster while ensuring a certain prediction accuracy. This study confirmed the ability of ML models for estimating TOC using well logs data in the study area.

**Keywords:** machine learning; geochemical analysis; well log data; source rock; Lishui–Jiaojiang Sag

check for updates

**Citation:** Han, X.; Hou, D.; Cheng, X.; Li, Y.; Niu, C.; Chen, S. Prediction of TOC in Lishui–Jiaojiang Sag Using Geochemical Analysis, Well Logs, and Machine Learning. *Energies* **2022**, *15*, 9480. https://doi.org/10.3390/en15249480

Academic Editor: Mohammad Sarmadivaleh

Received: 8 October 2022
Accepted: 12 December 2022
Published: 14 December 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

Resource potential is a basic parameter for evaluating whether a hydrocarbon-bearing block has commercial exploitation value [1]. Evaluation of resource potential requires geochemical test data, which can directly reflect the hydrocarbon generation potential [2,3]. Total organic carbon (TOC) is an important parameter of geochemical data to evaluate resource potential. To obtain accurate TOC, research mainly relies on direct measurement of cutting and core samples. Due to the deep depth of the target development layer, the formation pressure is high, indicating that the difficulty and cost of cutting and core samples collection are generally high [4]. Meanwhile, direct measurement of core samples in the laboratory is expensive. As a result, the difficulty of direct TOC measurement affects exploration and development of oil and gas in deep layers.

Well logs can reflect the physical properties of the rocks in the target interval. Due to the different physical properties, such as gamma radioactivity, resistivity, and density between organic-rich rocks and other surrounding rocks, logging information can effectively distinguish source rocks. TOC prediction methods based on logging parameters are increasingly used. A variety of different TOC prediction methods and empirical formulas have been established by many researchers through fitting actual measured TOC with

logging information. Beers [5] first established the relationship between gamma logging and TOC and calculated the TOC of the target interval. Schmoker [6] proposed a TOC prediction technology using density logging, and the technology has been applied to TOC prediction of shale until now. Mendelson and Toksoz [7] first used the multiple regression method to establish the relationship between multiple logging parameters and TOC with a correlation coefficient. Autric and Dumesnil [8] established a prediction relationship between acoustic transit time and TOC and found that physical properties of source rocks affected the prediction accuracy. The accuracy of empirical formulas to predict TOC cannot be guaranteed when applied in other study areas according to their study. Passey et al. [9] proposed the ΔlogR technique, which takes into account the changes in resistivity and porosity of source rocks before and after hydrocarbon generation and realizes effective unification of geological processes and TOC for the first time. However, the ΔlogR technique cannot be applied to TOC prediction of highly mature organic-rich rocks. Kamali and Mirshady [10] extended the ΔlogR technique by using the neuro-fuzzy method so that the technique can also be applied to TOC prediction of gas-producing source rocks. Passey et al. [11] extended the ΔlogR technique to include the TOC of highly mature organic-rich rocks in the prediction range. Hu et al. [12], Wang et al. [13], and Zhao et al. [14] calculated TOC using superposition between resistivity and neutron porosity logs. However, the huge differences in the maturity of source rocks and the background value of TOC content in different regions have a significant effect on the prediction results [13,14]. Since no linear or non-linear relationship occurs between TOC and various logging parameters, there are differences in petrol-physical properties in different study areas [15,16]. Many empirical formulas have different prediction performances in different research areas.

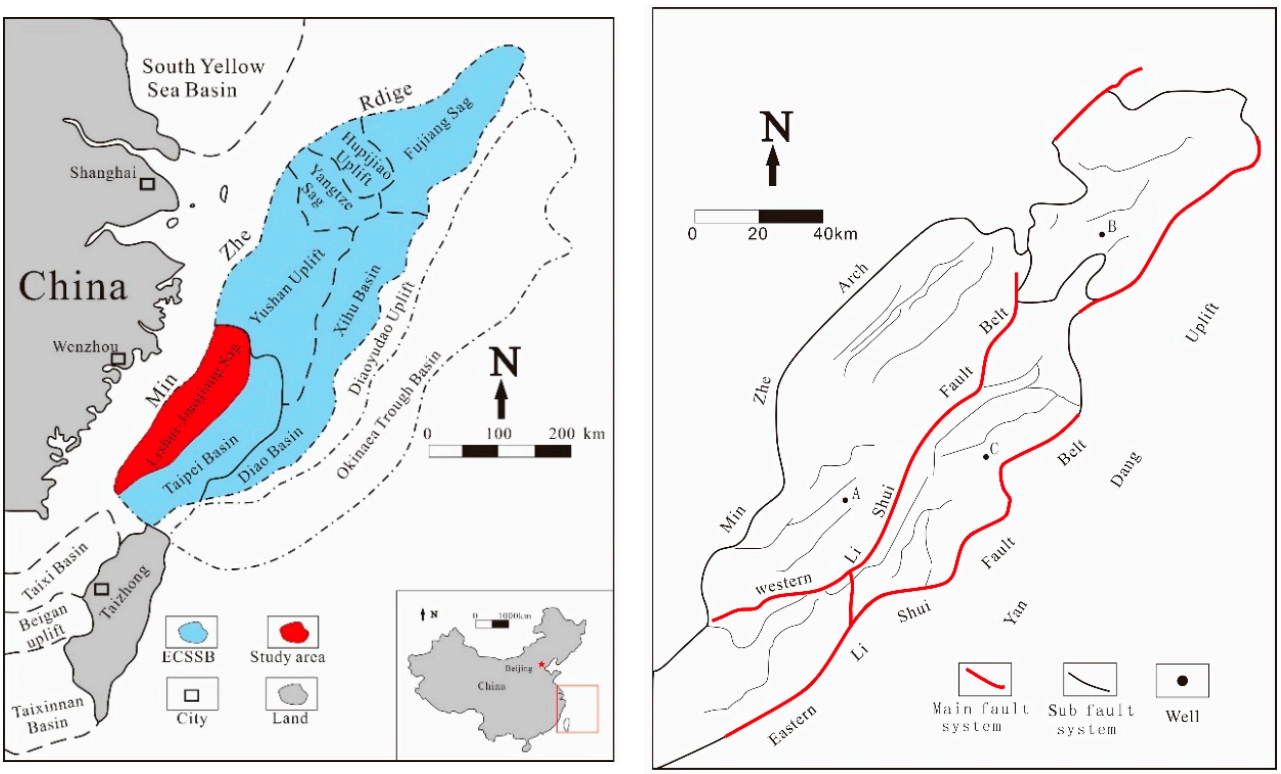
ML is a specific branch of artificial intelligence and is a method of empowering ML to do things that cannot be completed by direct programming [17]. ML is applied with great success in various industries, such as engineering [18–23], medicine [24–29], economy [30–32], and environmental and geospatial modeling [33–35]. ML has been applied to macroscopic features of target rocks, such as seismic facies classification [36–40] and logging lithofacies classification [41–43]. ML has demonstrated outstanding performance in these areas. An important aspect observed in recent research is that ML can learn and adapt to the dynamics of reservoir conditions, such as formation and depositional environment [44], while making use of geophysical data for lithology identification [45,46], porosity, and permeability [47–53]. With the development of ML from shallow learning to deep learning, deep learning has also been successfully applied to image-based, geoscience-related issues [54,55], such as seismic facies classification [40], lithology classification [56,57], mineral recognition [58], and carbonate rock recognition [59,60]. Current deep learning is a data-hungry technology. There are many problems in the real world, including in the medical field and oil and gas exploration, which are associated with an insufficient amount of labeled data [61]. The performance of deep learning in the above fields is not good [61]. In recent years, ML has been applied to TOC prediction, and the prediction results are ideal. TOC prediction mainly uses neural networks [16,44,62–65], extreme learning machine (ELM) [62,66,67], support vector machines [4,68], and decision trees [15,69,70]. Different machine learning algorithms have different operations and kernel functions. Therefore, the results obtained from different algorithms are quite different [63,71,72]. As a result, the choice of algorithm directly affects the accuracy of TOC prediction using ML.

The Lishui–Jiaojiang Sag is an important petroliferous basin in the East China Sea Basin [73,74]. Previous studies have shown that the Lishui–Jiaojiang Sag has excellent migration and storage capacity [75–78]. Basic research on source rocks is seriously lacking because the cost of core sampling and laboratory analysis of TOC is generally high. The ambiguity of the organic matter abundance in the source rocks severely restricts the oil and gas exploration process in the Lishui Sag. TOC prediction based on well logs is very important for Lishui–Jiaojiang Sag. In this paper, the source rock of Paleocene Yueguifeng Formation in the Lishui–Jiaojiang Sag is taken as the main research object. Three ML algorithms (random forest, extreme learning machine, and back propagation neural network)

are used for predicting TOC (target variable). Two well combinations of logging parameters are used as input. This study aims to: (1) establish the multiple regression and ML models of TOC prediction; (2) compare the performance of ML and multiple regression models of TOC prediction; (3) compare the performance of models of TOC prediction using different logging parameters; and (4) determine the best TOC prediction method for the study area.

## 2. Geological Setting and Stratigraphy

The Lishui–Jiaojiang Sag is located in the western Taipei Depression (see Figure 1a), which is located in the East China Sea Shelf Basin; the sag faces the Yushan Uplift to the north, the Fuzhou Sag and the Minjiang Sag to the east, which is separated by the Yandang Uplift, and the Minjiang Uplift to the west and south (see Figure 1b) [73,75]. Tectonically, the sag lies at the convergence of the Eurasian, Pacific, and Philippine plates. NE-trending segmental faults control the morphology of the basin and have the characteristics of north–south blocks, east–west zoning, and an overall northeast–southwest distribution. The Lishui–Jiaojiang Sag is divided into different structural units (such as the Lishui Western Subsag, the Lingfeng Ridge, and the Lishui Eastern Subsag) by the Lingfeng low bulge [78,79].




(**a**) (**b**)

**Figure 1.** (**a**) Tectonic location of the East China Sea Shelf Basin (modified after (modified after [51]). (**b**) Structural units of the Lishui–Jiaojiang Sag (modified after [51]).

### *2.1. Tectonic*

The Lishui–Jiaojiang Sag is a Cenozoic single-fault sag superimposed on a residual Mesozoic basin [78,80]. From the end of the late Cretaceous to the early Tertiary, the Lishui–Jiaojiang Sag experienced four tectonic evolution phases, including a rift stage (late Cretaceous to Paleocene), a post-rift depression stage (early Eocene to late Eocene), an uplift stage (late Eocene to late Miocene), and a regional subsidence stage (late Miocene to Quaternary), which were affected by the changes in the subduction direction of the western Pacific plate and rifting [75,76].

### 2.2. Stratigraphy

The stratigraphic column of the Lishui–Jiaojiang Sag is divided into late Cretaceous, Paleogene, Neogene, and Quaternary strata from bottom to top. Oil and gas in Lishui Sag have been mainly discovered in Paleocene strata (see Figure 2) [75,78]. The main source rock is the lower Paleocene Yueguifeng Formation (E1y). The main object of this study is the lower Paleocene Yueguifeng (E1y) Formation [73,80]. The lithology of the Paleocene Yueguifeng Formation (E1y) is dominated by gray and dark gray mudstone and silty mudstone interspersed with thin layers of light gray calcium-bearing siltstone, fine sandstone, and a small number of thin layers of fine calcareous sandstone. The Lauguifeng Formation includes two sets of coarse–fine–coarse composite sedimentary cycles. The descending cycle is dominated by dark brown and dark brown mudstone mixed with light gray and off-white fine-medium-grained sandstone. The upper cycle consists of light gray, gray, dark gray, dark gray mudstone, and light gray fine-medium-grained sandstone nearly equal-thickness interbedded, sandwiched by thin layers of light gray, gray siltstone, and two layers of black coal. The overall deposition thickness is 125–400 m.

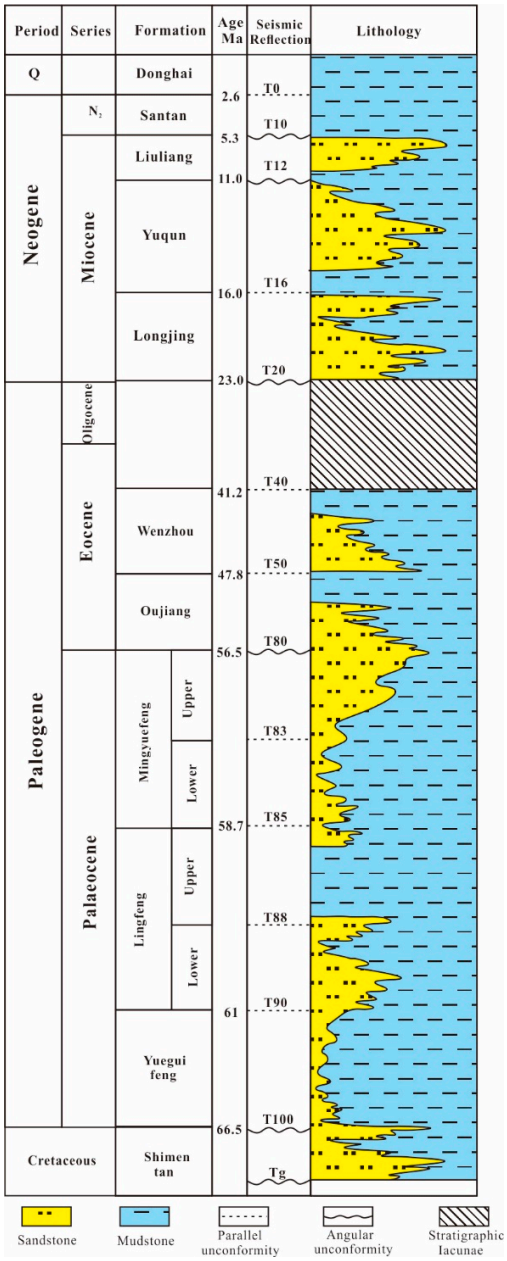



**Figure 2.** Stratigraphic units of the Paleocene in the Lishui Sag (modified after [51]).

## 3. Materials & Methods

In order to test the accuracy of the log multivariate fitting model and the ML models in TOC prediction, only 77 source rock samples were collected in the Lishui–Jiaojiang Sag. These 77 samples were derived from three wells (well A, well B, and well C) in different areas of the Lishui–Jiaojiang Sag. All samples were derived from the lower Paleocene Yueguifeng Formation.

### *3.1. Geochemical Analysis*

Each source rock sample was crushed to 80 meshes and acidified with dilute hydrochloric acid (HCl) to remove inorganic carbon. The TOC measurement was conducted by using the C/S analyzer CS-580A. The pretreatment of samples for pyrolysis analysis is the same as the TOC analysis. Rock-Eval pyrolysis analysis was conducted using a Rock-Eval-II. Before the Rock-Eval pyrolysis analysis, the Rock-Eval-II was calibrated with standard samples from the geochemical laboratory; samples were put into the Rock-Eval analyzer for testing. The measured parameters include TOC, $S_1$, $S_2$, and maximum pyrolysis yield temperature ($T_{max}$). Hydrogen index (HI) was calculated by TOC and $S_2$.

### *3.2. Log Series Selection and Well Log Models*

ML is the process of inferring correlations by continuously learning from input data, which can influence the accuracy of the model. In general, the more feature parameters are supplied, the more thoroughly the objectives are characterized and more accurately the predictions are obtained. If the selected feature parameters are weakly correlated with the target, the system may learn the wrong functional relationship. The prediction may have a larger deviation. Therefore, before performing ML prediction model, the feature parameters should be screened to ensure that these parameters can more accurately describe the characteristics of the target.

Numerous studies have been conducted to clarify the connection between logging settings and TOC. Studies have shown that organic matter has varying degrees of linear correlation with gamma rays (GR), deep resistivity (RT), density (DEN), acoustic waves (AC), and neutrons (CN) [5,6,8,9].

GR logging could record gamma radioactivity during rock formations. The higher the reflectivity of the rock formation, the stronger the GR logging response [5]. Strongly radioactive organic-rich mudstone is abundant in radioactive thorium, potassium, and uranium elements, meaning that the organic-rich mudstones generally have high GR values.

Density logs could reflect the volume density of the formation based on the assumed formation density and drilling fluid density [6,8]. The density of sand and mudstone is generally 2~2.7 g/cm$^3$, and the density of organic-rich stratum is usually lower than that of surrounding rocks. The density of solid organic matter is close to that of water (around 1.0 g/cm$^3$). Organic-rich stratum will produce organic matter pores after hydrocarbon generation and expulsion, and the density will be further reduced. Therefore, density logs can be used to indirectly estimate TOC concentrations.

Neutron logs may identify the degree of scattering of exposed neutrons in the formation, which, afterwards, reflects the formation's porosity. Organic matter is directly related to the hydrogen atoms and porosity of the rock [14]. Therefore, when the TOC of stratum increases, the values of CN increase. AC logs could measure the response of the rock mass near the borehole to the artificial elastic wave field to achieve the purpose of detecting the properties of the target formation (porosity, longitudinal and transverse wave velocity, etc.).

AC logging is mainly a logging tool for determining lithology, porosity, and fluid type [10,12]. Previous studies have shown that the acoustic transmission time of organic-rich source rocks is significantly shorter. Resistivity logs record the fluid resistivity in the formation and then calculate the true resistivity of the formation [9,11]. The response of TOC is not sensitive to resistivity logs. However, after the mature source rocks generation and expelling hydrocarbons, the resistivity will increase significantly due to the presence of hydrocarbon compounds. Therefore, the TOC content of the source rock section can

be indirectly reflected according to the change in the relative resistivity of the source rock section.

Different study areas have different geological conditions, and the relationship between logging parameters and TOC is also different. Figure 3 shows the visual relationship between the logging parameters and the measured TOC. It shows that, as AC, CN, GE, and RT fluctuate positively, TOC also fluctuates in a corresponding positive direction. When DEN fluctuates negatively, TOC also fluctuates positively. This is consistent with the physical principles of logging. However, the visualization law cannot show a quantitative correlation between the individual logging parameters and the TOC. A simple linear regression technique was used to determine the sensitivity of TOC to log data. During this process, a key metric (correction factor) labeled $R^2$ was used to investigate the effect of different logs on the true TOC value. $R^2$ represents the proportion of population variance explained by the model.

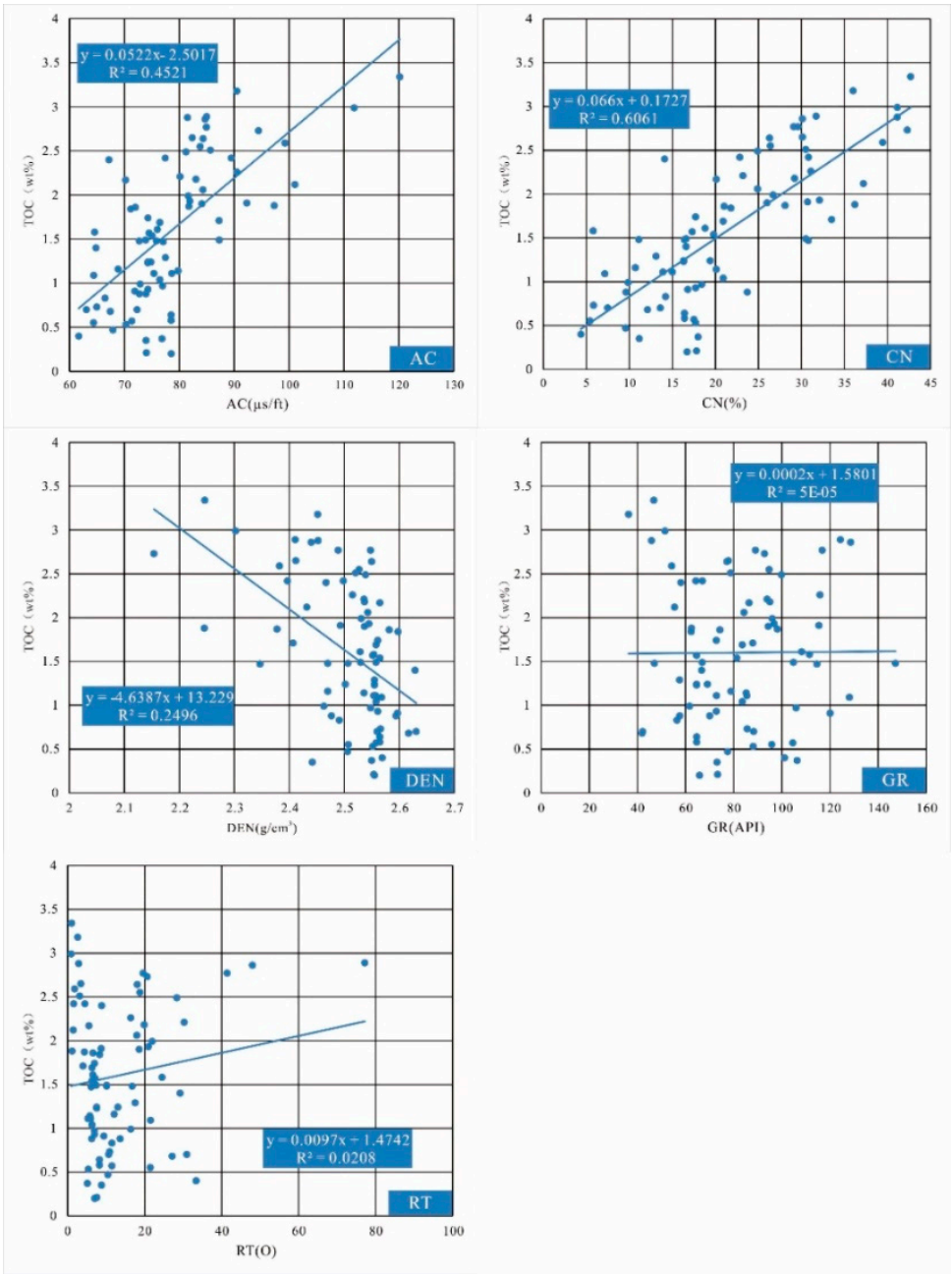



**Figure 3.** Cross plots of TOC laboratory measurements and logging data.

Figure 3 shows that the TOC of the Yueguifeng Formation source rock has a better response in DT and CN logging. With the increase in TOC, the logging response increases, showing an obvious positive correlation, with $R^2$ of 0.45 and 0.60, respectively. DEN logging was negatively correlated with TOC to a certain extent, with an $R^2$ of 0.24. However, there was no obvious correlation between TOC and GR and RT. In addition, the Pearson correlation coefficient was used to analyze the sensitivity of TOC to logging data. Table 1 shows that the measured TOC has a good correlation with AC, CN, and DEN, and the R values of the measured TOC value and AC, CN, and DEN are 0.67, 0.77, and −0.49, respectively. There was no obvious correlation between the measured TOC and GR and RT. The linear regression analysis results are basically consistent with the Pearson correlation coefficient analysis results. The sensitivity of the corresponding log and the comprehensiveness of the feature description will affect the performance of the model. Therefore, two feature parameter sets, HXC (including all well logs) and HX (including AC, CN, and DEN), were selected. HXC has high comprehensiveness feature description of TOC. HX and TOC have a high correlation. The accuracy of prediction results with different feature parameter sets has been compared to obtain the optimal combination as input.

**Table 1.** Data of Pearson's correlation coefficients.

| R | AC | CN | DEN | GR | RT | TOC |
|---|---|---|---|---|---|---|
| AC | 1 | 0.85133 | −0.68255 | −0.09409 | −0.16529 | 0.67236 |
| CN | | 1 | −0.61136 | −0.06937 | −0.08279 | 0.77855 |
| DEN | | | 1 | 0.04474 | 0.1027 | −0.49961 |
| GR | | | | 1 | 0.42611 | 0.00713 |
| RT | | | | | 1 | 0.1442 |
| TOC | | | | | | 1 |

It has been pointed out that, if there are two or more uncorrelated independent variables that have a good correlation with the dependent variable, the results of multiple regression are usually better than the results of university analysis, and the multiple regression equation can be based on the correlation of each dataset [7]. Matrix calculation is obtained. The supposed multivariate prediction model is as follows:

$$\mathrm{TOC} = \frac{\mathrm{A} \times \mathrm{AC} + \mathrm{B} \times \mathrm{CN} + \mathrm{C} \times \mathrm{GR} + \mathrm{D} \times \mathrm{RT} + \mathrm{E}}{\mathrm{DEN}} + \mathrm{F} \quad (1)$$

A, B, C, D, E, F, are all constants. Entering the combination of characteristic parameters and TOC into ORIGIN and using regression statistics to obtain two multivariate fitting equations as follows:

$$\mathrm{TOC} = \frac{0.004 \times \mathrm{AC} + 15.703 \times \mathrm{CN} - 4.13113}{\mathrm{DEN}} + 1.7546 \quad (2)$$

$$\mathrm{TOC} = \frac{0.03826 \times \mathrm{AC} + 12.48936 \times \mathrm{CNL} + 0.00327 \times \mathrm{GR} + 0.0473 \times \mathrm{LLD} - 7.18502}{\mathrm{DEN}} + 1.85181 \quad (3)$$

### *3.3. ML Methods*

In this study, three ML algorithms were used to predict the TOC of source rocks in the Lishui–Jiaojiang Sag, and 77 log parameter combinations and corresponding measured TOC values were used as input dataset and test set to build and test the prediction model. Three ML algorithms include extreme learning machine (ELM), back propagation neural network (BPNN), random forest (RF).

#### 3.3.1. Method of BPNN

BPNN is the most basic neural network, and its output results are propagated forward, and the error is propagated by back propagation [81]. BPNN models can have one or more

hidden layers [82,83]. Taking two hidden layers as an example, the forward propagation operation process is roughly as follows: the input value is $x_n = i_1 = O_1$, and the output value is y. After linear combination with the weight matrix in the middle, the input $i_2$ of the hidden layer 1 is obtained, and, in the hidden layer 1, the input $i_2$ is activated with the activation function to obtain the output $O_2$, and then it can be known that $i_2 = W_1 \cdot O_1$. Let f(x) is the activation function, $O_2 = f(i_2)$. Then, similarly, the weight from hidden layer 1 to 2 is $W_2$, the input of hidden layer 2 is $i_3 = W_2 \cdot O_2$, and the output is $O_3 = f(i_3)$. The output $y = W_3 \cdot O_3$ obtained from the hidden layer 2 to the output layer can also be written as $y = f(w_3 f(w_2 f(w_1 o_1)))$ [10,81,83–85]. At this time, the loss function will be used to further process y. Generally, our loss function takes $O_4 = 1/2(L-y)^2$, where L is the label of the training set [86]. If the desired output cannot be obtained in the output layer, then turn to back propagation. The error signal is returned along the original connection path, and the weight of each neuron is modified to minimize the error signal [10,44,84]. The weight update of each neuron can be calculated by the formula shown in Equation (5).

$$\Delta W_1 = (y - L)\left(\left(W_2^T\left[W_2^T \cdot f'(i_3)\right]\right) \cdot f'(i_2)\right)O_1^T \tag{4}$$

$$\Delta W_2 = (y - L) \cdot f'(i_4) \cdot W_3^T \cdot f'(i_3) \cdot O_2^T \tag{5}$$

$$\Delta W_3 = (y - L) \cdot f'(i_4) \cdot O_3^T \tag{6}$$

Finally, the update of the weight w can be completed by bringing the sample into the continuous forward. BPNN is widely used in pattern recognition, classification, data mining, and other disciplines [10,26,87,88]. The study found that BPNN has excellent performance in nonlinear mapping, and the flexible network structure ensures the accuracy of model prediction [86,89]. However, BPNN has a slow learning speed and is prone to fall into local minima [62,85].

#### 3.3.2. Method of ELM

Extreme learning machine (ELM) was first proposed by Huang Guangbin in 2004 [90]. Extreme learning machine is to improve the back propagation algorithm (backward propagation, BP) to improve learning efficiency and simplify the setting of learning parameters [91,92]. The back propagation algorithm uses the gradient algorithm (back propagation). ELM adopts random selection of input layer weights and hidden layer deviations, and output layer (see Figure 4) weights are calculated and analyzed according to Moore–Penrose generalized inverse matrix theory by minimizing the loss function of the training error term and the output layer weight norm [93] as follows:

$$y = \sum_{i=1}^{L} \beta_i G(a_i x + b_i) \tag{7}$$

X is the input data, β is the output weight, G(x) is the feature map or activation function (activation function) [91,92,94]. The main operation steps of ELM are divided into three steps: (1) randomly assign node parameters; (2) calculate the output matrix of the hidden layer; (3) solve for output weights. The core of the ELM algorithm is to use the least squares method to solve the output weights to minimize the error function [90,92,95]. The weight value can be obtained by the following system of equations.

$$\min_{\beta} \|H\beta - T\| \tag{8}$$

By solving the least squares problem, the hidden layers' output weights are implemented as follows:

$$\hat{\beta} = \left(H^T T\right)^{-1} H^T T \tag{9}$$

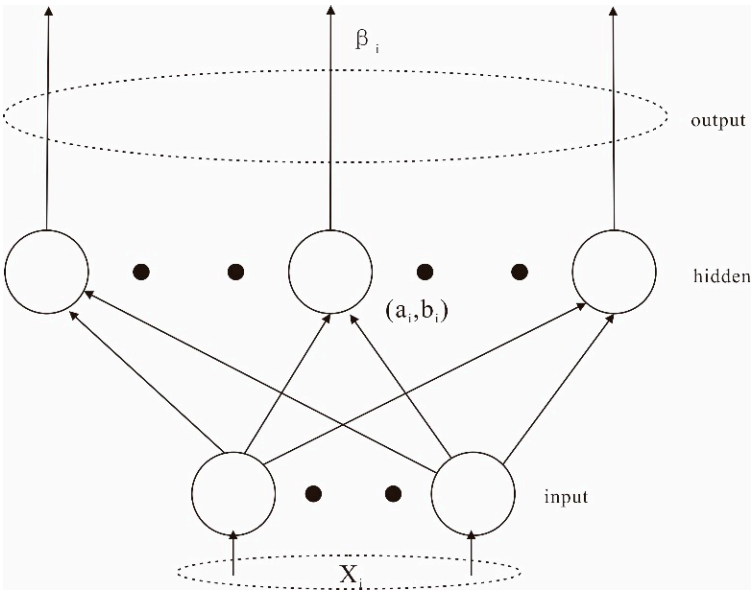



**Figure 4.** Schematic diagram of ELM.

The model can be established by using the weight value and the hidden layer deviation and make a prediction of the target value. The study found that ELM can make predictions on small sample data with fewer feature parameters. It can maintain high prediction accuracy set generalization [65].

3.3.3. Method of RF

RF is an ensemble learning algorithm [96,97]. The understanding of random forest can be divided into "random" and "forest": "random" refers to random sampling of data and random sampling of features, mainly reflected in Bagging; its full name is Bootstrap aggregation [96,98]. It is a kind of sampling method for replacement [99]. The "forest" refers to the use of multiple freely growing CART decision trees to form a "forest" [100]. The core of the RF is the regression tree (ntree: the default value is 500 trees) and the number of branches (mtry: the default value is 1/3 of the total number of features) [101]. The principle of RF operation (see Figure 5) is (1) based on the Bagging ensemble framework, sampling n training samples randomly and with replacement each time from the original samples, performing T rounds of sampling to obtain T training sets, (2) using the T training sets, train T CART decision trees independently and combine T CART decision trees into a strong learner. The core of a single CART decision tree lies in the recursive growth process. Assuming the input dataset Xc, according to the segmentation method of the p-th eigenvalue V as Sp, v, the input formula can be expressed as:

$$S_{p,v}{:}X_L = \{x_p{>}V|x \in X_c\}; X_R = \{x_p \leq V|x \in X_c\} \quad (10)$$

where XL and XR are two subsets of the dichotomy [73,74]. During the learning process, the CART decision tree obtains a nested IFELESE logic chain through stepwise recursive bisection segmentation [73,75]. This logical chain is stored and displayed using a binary tree structure [78]. Then, for the prediction problem, the prediction variance is used to optimize the segmentation method [72,73,79]. The prediction variance (variance) can be written as:

$$\mathrm{vari}(X) = \frac{\sum_{\hat{i}X}(y(x) - E(y|x \in X))^2}{\|X\|} \quad (11)$$

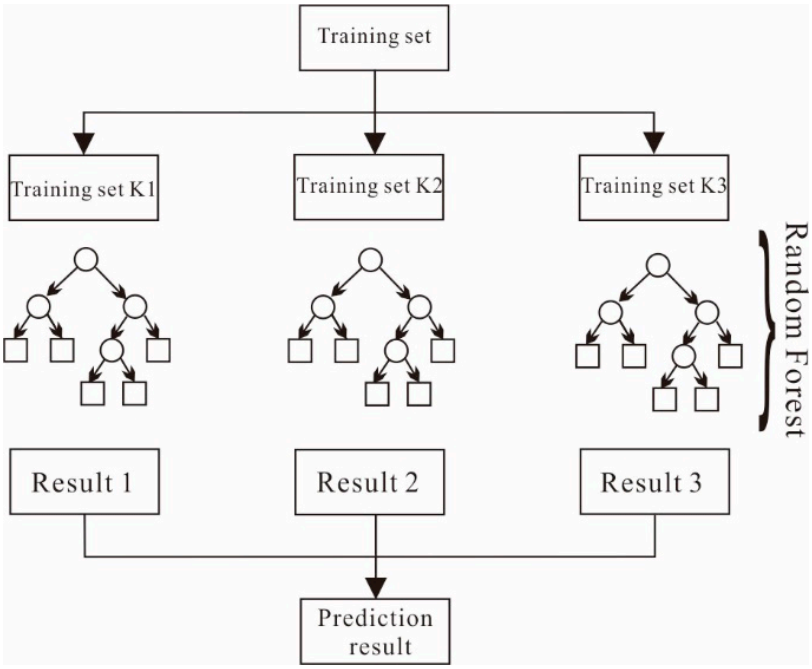



**Figure 5.** Schematic diagram of RF.

After fully recursive optimal growth, the entire input space is divided into L mutually exclusive subspaces. Inputting the sample x′ into the tree structure, its predicted value can be written as:

$$E(y|x = x') = \frac{\sum_{i=1}^{n} y_i \times I(x_i \in X_L)}{\sum_{i=1}^{n} I(x_i \in X_L)} \quad (12)$$

(3) In the end, Bagging will assemble all T CART decision tree prediction results in an equal weight manner to establish an integrated prediction result [96,102,103]. The resulting function can be written as:

$$H(x) = \frac{1}{T}\sum_{T=1}^{T} E_m(x) \quad (13)$$

To be consistent with the logging prediction model, two kinds of characteristic parameter combinations, HXC and HX, were selected. The data combination of 50 randomly selected well logging data and measured geochemical data were used as a training set for the ML algorithm to build a prediction model. The remaining 27 data combinations were used as the test set to test the performance of the three prediction models. The ML algorithm modeling and TOC prediction in this study were conducted using MATLAB2019 software.

### *3.4. Evaluation Criteria*

In this study, $R^2$, MAE, MSE, and RMSE have been selected to evaluate the performance of prediction models. The mean absolute error (MAE) reflects the average of the absolute errors between the predicted and observed values [69]. Mean squared error (MSE) is the most commonly used regression loss function and is calculated by summing the squares of the distances between the predicted and true values [15]. The root mean square error (RMSE) represents the sample standard deviation of the difference between the predicted and observed values (called residuals) [4]. The RMSE is used to describe the degree of dispersion of the samples.

## 4. Results and Discussion

### *4.1. TOC and Source Rock Study from Geochemical Analyses*

The TOC concentrations of the study area showed high values, ranging from 0.2% to 3.34%, with an average value of 1.6% (Table 2). There are some differences among the wells. The TOC concentration of well B (Jiaojiang Depression) is the highest, ranging from 1.88% to 3.34%, with an average of 2.63%. The TOC concentration in well C (the Lishui Eastern Subsag) is lower, ranging from 0.2% to 2.89%, with an average of 1.57%. The TOC of well A (the Lishui Western Subsag) is the lowest, ranging from 0.35% to 1.58%, with an average of 0.74%. The TOC of well A (the Lishui Western Subsag) is the lowest, ranging from 0.35% to 1.58%, with an average of 0.74%. The hydrogen index of Yueguifeng showed high values, ranging from 7.14 mg HC/g TOC to 390.64 mg HC/g TOC, with an average of 100.94 mg HC/g TOC. The HI of well B (Jiaojiang Depression) is the highest, ranging from 158.51 mg HC/g TOC to 388.68 mg HC/g TOC, with an average of 323.67 mg HC/g TOC. The HI of well A (the Lishui Western Subsag) is lower, ranging from 34 to 247.5 mg HC/g TOC, with an average of 102.82 mg HC/g TOC. The HI of well C (the Lishui Eastern Subsag) is the lowest, ranging from 17.14 to 100 mg HC/g TOC, with an average of 56.01 mg HC/g TOC. The total distribution of hydrocarbon generation potential ($S_1 + S_2$) in Yueguifeng Formation is ranging from 0.05 mg HC/g TOC to 12.51 mg HC/g TOC. The S1 + S2 of well A (the Lishui Western Subsag) is low, ranging from 3.2 mg HC/g TOC to 12.51 mg HC/g TOC, with an average of 8.96 mg HC/g TOC. The $S_1 + S_2$ of well B (Jiaojiang Sag) is the highest, ranging from mg HC/g TOC to 3.34 mg HC/g TOC, with an average of 8.96 mg HC/g TOC. The $S_1 + S_2$ of well C (the Lishui Eastern Subsag) is next, ranging from 0.05 mg HC/g TOC to 3.98 mg HC/g TOC, with an average of 1.45 mg HC/g TOC (see Table 2).

**Table 2.** Results from geochemical Rock-Eval/TOC analysis with calculated parameters from Yueguifeng Formation.

| Well | Depth (m) | TOC | S1 (mg/g TOC) | S2 (mg/g TOC) | Pg $S_1 + S_2$ (mg/g TOC) | HI $S_2$/TOC | $T_{max}$ °C |
|---|---|---|---|---|---|---|---|
| A | 2701 | 0.55 | 0.12 | 0.88 | 1 | 159.42 | 438 |
| | 2707 | 1.09 | 0.22 | 1.19 | 1.41 | 109.17 | 444 |
| | 2720 | 1.58 | 0.27 | 2.11 | 2.38 | 133.54 | 434 |
| | 2721 | 0.4 | 0.12 | 0.99 | 1.11 | 247.5 | 442 |
| | 2916.6 | 0.99 | 0.02 | 0.48 | 0.5 | 48 | 454 |
| | 2917 | 0.88 | 0.02 | 0.3 | 0.32 | 34 | 452 |
| | 2940 | 0.73 | 0.89 | 0.77 | 1.66 | 105.48 | 430 |
| | 2960 | 0.7 | 0.23 | 0.96 | 1.19 | 137.14 | 428 |
| | 3079 | 0.35 | 0.02 | 0.18 | 0.2 | 34 | 451 |
| | 3092 | 0.47 | 0.02 | 0.31 | 0.33 | 65.96 | 451 |
| | 3099 | 0.37 | 0.02 | 0.21 | 0.23 | 56.76 | 453 |
| B | 2356.5 | 3.34 | 0.13 | 11.72 | 11.85 | 350.9 | 438 |
| | 2362.5 | 2.99 | 0.33 | 11.68 | 12.01 | 390.64 | 441 |
| | 2395.5 | 1.88 | 0.22 | 2.98 | 3.2 | 158.51 | 438 |
| | 2401.5 | 2.42 | 0.23 | 7.21 | 7.44 | 297.93 | 440 |
| | 2422.5 | 2.59 | 0.42 | 8.57 | 8.99 | 330.89 | 443 |
| | 2425.5 | 2.12 | 0.37 | 7.21 | 7.58 | 340.09 | 440 |
| | 2434.5 | 3.18 | 0.15 | 12.36 | 12.51 | 388.68 | 440 |
| | 2488.5 | 2.88 | 0.14 | 7.48 | 7.62 | 259.72 | 440 |
| | 2495 | 2.42 | 0.22 | 9.2 | 9.42 | 380.17 | 438 |
| | 2506.5 | 2.65 | 0.34 | 9.12 | 9.46 | 344.15 | 441 |
| | 2518.5 | 2.51 | 0.51 | 8 | 8.51 | 318.73 | 440 |

**Table 2.** *Cont.*

| Well | Depth (m) | TOC | S1 (mg/g TOC) | S2 (mg/g TOC) | Pg $S_1 + S_2$ (mg/g TOC) | HI $S_2$/TOC | $T_{max}$ °C |
|---|---|---|---|---|---|---|---|
| | 3576 | 0.97 | 0.21 | 0.45 | 0.66 | 46.39 | 451 |
| | 3577.5 | 1.61 | 0.55 | 1.28 | 1.83 | 79.5 | 458 |
| | 3586.5 | 1.24 | 0.33 | 0.66 | 0.99 | 53.23 | 452 |
| | 3595.5 | 1.71 | 0.36 | 0.98 | 1.34 | 57.31 | 461 |
| | 3604.5 | 1.86 | 0.21 | 1.16 | 1.37 | 62.37 | 458 |
| | 3607.5 | 1.87 | 0.38 | 1.24 | 1.62 | 66.31 | 459 |
| | 3630.5 | 1.48 | 0.17 | 0.61 | 0.78 | 41.22 | 460 |
| | 3640 | 0.68 | 0.02 | 0.19 | 0.21 | 27.94 | 470 |
| | 3640.05 | 0.7 | 0.01 | 0.12 | 0.13 | 17.14 | 479 |
| | 3640.8 | 1.29 | 0.13 | 0.83 | 0.96 | 64.34 | 466 |
| | 3641.6 | 2.4 | 1.08 | 1.98 | 3.06 | 82.5 | 470 |
| | 3641.7 | 0.83 | 0.05 | 0.21 | 0.26 | 25.3 | 478 |
| | 3642.7 | 2.17 | 0.11 | 1.08 | 1.19 | 49.77 | 466 |
| | 3643 | 0.53 | 0.03 | 0.21 | 0.24 | 39.55 | 472 |
| | 3643.6 | 0.2 | 0.01 | 0.04 | 0.05 | 20 | 471 |
| | 3643.7 | 0.64 | 0.03 | 0.16 | 0.19 | 25 | 476 |
| | 3643.79 | 0.58 | 0.03 | 0.18 | 0.21 | 31.03 | 452 |
| | 3645.8 | 1.14 | 0.17 | 0.6 | 0.77 | 52.63 | 465 |
| | 3646 | 1.11 | 0.1 | 0.62 | 0.72 | 55.86 | 469 |
| | 3646.4 | 1.11 | 0.11 | 0.6 | 0.71 | 54.05 | 466 |
| | 3646.95 | 1.69 | 0.14 | 0.55 | 0.69 | 32.54 | 463 |
| | 3647 | 1.04 | 0.13 | 0.6 | 0.73 | 57.69 | 462 |
| | 3647.09 | 1.54 | 0.38 | 0.83 | 1.21 | 53.9 | 461 |
| | 3647.2 | 0.93 | 0.1 | 0.36 | 0.46 | 38.71 | 471 |
| | 3647.25 | 1.74 | 0.37 | 0.71 | 1.08 | 40.8 | 470 |
| | 3647.42 | 1.49 | 0.12 | 0.6 | 0.72 | 40.27 | 468 |
| | 3647.5 | 1.24 | 0.1 | 0.62 | 0.72 | 50 | 462 |
| C | 3647.5 | 1.23 | 0.11 | 0.74 | 0.85 | 60.16 | 463 |
| | 3647.81 | 1.57 | 0.16 | 0.67 | 0.83 | 42.68 | 465 |
| | 3648.16 | 0.21 | 0.03 | 0.21 | 0.24 | 100 | 455 |
| | 3688.5 | 1.84 | 0.29 | 1.54 | 1.83 | 83.7 | 448 |
| | 3702 | 0.88 | 0.16 | 0.32 | 0.48 | 36.36 | 458 |
| | 3712.5 | 1.16 | 0.15 | 0.6 | 0.75 | 51.72 | 471 |
| | 3724.5 | 0.57 | 0.07 | 0.26 | 0.33 | 45.61 | 461 |
| | 3739.5 | 1.91 | 0.88 | 1.47 | 2.35 | 76.96 | 455 |
| | 3745 | 1.49 | 0.52 | 1.21 | 1.73 | 81.21 | 456 |
| | 3751.5 | 2.26 | 1.27 | 1.87 | 3.14 | 82.74 | 457 |
| | 3755 | 2.77 | 1.32 | 2.01 | 3.33 | 72.56 | 457 |
| | 3769.5 | 1.99 | 1.3 | 1.3 | 2.6 | 65.33 | 455 |
| | 3772.5 | 2.49 | 1.38 | 2.43 | 3.81 | 97.59 | 454 |
| | 3775.5 | 2.21 | 2.05 | 1.93 | 3.98 | 87.33 | 453 |
| | 3784.5 | 2.77 | 1.08 | 1.68 | 2.76 | 60.65 | 446 |
| | 3793.5 | 2.06 | 1.25 | 1.25 | 2.5 | 60.68 | 457 |
| | 3793.5 | 2.64 | 1.03 | 1.61 | 2.64 | 60.98 | 451 |
| | 3795 | 1.93 | 0.99 | 1.22 | 2.21 | 63.21 | 453 |
| | 3795 | 2.18 | 1.01 | 1.25 | 2.26 | 57.34 | 455 |
| | 3796.5 | 1.9 | 0.94 | 1.15 | 2.09 | 60.53 | 459 |
| | 3796.5 | 2.55 | 1.21 | 2.4 | 3.61 | 94.12 | 458 |
| | 3813 | 1.4 | 0.2 | 0.68 | 0.88 | 48.57 | 457 |
| | 3852 | 2.73 | 0.89 | 1.61 | 2.5 | 58.97 | 464 |
| | 3883 | 2.86 | 1.27 | 1.83 | 3.1 | 63.99 | 447 |
| | 3888 | 2.89 | 1.52 | 2.15 | 3.67 | 74.39 | 456 |
| | 3903 | 1.47 | 0.13 | 0.58 | 0.71 | 39.46 | 455 |
| | 3910.5 | 1.48 | 0.37 | 0.61 | 0.98 | 41.22 | 453 |
| | 3913.5 | 0.91 | 0.28 | 0.43 | 0.71 | 47.25 | 455 |

$S_1$: volatile hydrocarbon (HC) content, mg HC/g rock. $S_2$: remaining HC generative potential, mg HC/g rock. HI: hydrogen index = $S_2 \times 100$/TOC, mg HC/g TOC. Pg: potential for hydrocarbon generation = S1 + S2 (mg/g).

The type of organic matter has been classified according to the van Krevelen plot of whole rock hydrogen index (S2/TOC × 100) and TMAX shown in Figure 6. The intersection diagram shows that the source rock kerogen of the Yueguifeng Formation is dominated by type II, with a small amount of type I and type III kerogen. Type II kerogen is dominated by type $II_2$, and type $II_2$ kerogen accounts for only about 15%. The type of kerogen in well A is dominated by type $II_2$, with a small amount of type I and type III kerogen. The type of kerogen in well B is dominated by type $II_1$, with a small amount of type $II_2$ kerogen. The kerogen type of well C is dominated by type $II_2$, with a small amount of type I and type III kerogen. At the same time, the high degree of thermal evolution of the source rock in well C results in a low hydrogen index. The distribution of each well reflects the heterogeneity of kerogen. Based on the above geochemical evaluation, the source rocks of the Yueguifeng Formation have objective hydrocarbon generation potential. The source rocks mainly generate oil with a small amount of gas.

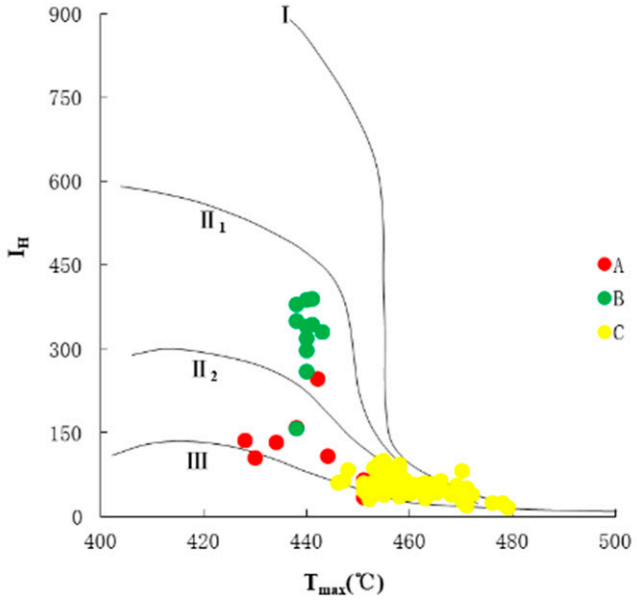



**Figure 6.** Cross plot of hydrogen index (HI) and $T_{max}$.

### *4.2. TOC Quantification from Multiple Regression Models*

After using two logging multiple regression models to predict the TOC of the Yueguifeng Formation source rock in the Lishui–Jiaojiang Sag (see Table 3), $R^2$, MAE, MSE, and RMSE have been calculated to evaluate the performance of the two models (see Table 4). The calculation results show that the $R^2$ of the multiple regression model using the HXC is 0.63 (see Figure 7). The values of MAE, MSE, and RMSE are 0.39, 0.24, and 0.49, respectively. The $R^2$ of the multiple regression model using the HX is 0.60. The values of MAE, MSE, and RMSE are 0.41, 0.26, and 0.51, respectively. The $R^2$ of the multiple regression model using the HX is 0.60 (see Figure 7). The values of MAE, MSE, and RMSE are 0.41, 0.26, and 0.51, respectively. The $R^2$ of the multiple regression model using the HXC is higher than the $R^2$ of the multiple regression model using the HX. Error values calculated by MAE, MSE, and RMSE also reflect the lower prediction error of the multiple regression model using the HXC. The analysis shows that the accuracy of the multiple regression model using the HXC is significantly higher than that of the multiple regression model using the HX, but the performance is limited, which is also consistent with previous analysis. The relationship between TOC and various logging parameters is not a simple linear or nonlinear relationship, and the internal functional relationship is complex. Simple linear correlation analysis does not confirm whether low correlation parameters are helpful for establishment of predictive models.

**Table 3.** Results of TOC prediction from multiple regression.

| Well | Depth (m) | TOC % | Prediction Multiple Regression | |
|---|---|---|---|---|
| | | | HXC | HX |
| A | 2701 | 0.55 | 0.77 | 0.55 |
| | 2707 | 1.09 | 0.92 | 0.68 |
| | 2720 | 1.58 | 0.89 | 0.59 |
| | 2721 | 0.4 | 0.93 | 0.51 |
| | 2916.6 | 0.99 | 0.96 | 0.82 |
| | 2917 | 0.88 | 0.91 | 0.81 |
| | 2940 | 0.73 | 0.61 | 0.6 |
| | 2960 | 0.7 | 0.8 | 0.71 |
| | 3079 | 0.35 | 0.91 | 0.9 |
| | 3092 | 0.47 | 0.8 | 0.81 |
| | 3099 | 0.37 | 1.3 | 1.36 |
| B | 2356.5 | 3.34 | 3.16 | 3.11 |
| | 2362.5 | 2.99 | 2.91 | 2.96 |
| | 2395.5 | 1.88 | 2.44 | 2.62 |
| | 2401.5 | 2.42 | 1.59 | 1.68 |
| | 2422.5 | 2.59 | 2.61 | 2.79 |
| | 2425.5 | 2.12 | 2.5 | 2.62 |
| | 2434.5 | 3.18 | 2.27 | 2.52 |
| | 2488.5 | 2.88 | 2.4 | 2.84 |
| | 2495 | 2.42 | 1.87 | 2.16 |
| | 2506.5 | 2.65 | 1.91 | 2.14 |
| | 2518.5 | 2.51 | 1.97 | 2.15 |
| C | 3576 | 0.97 | 1.35 | 1.39 |
| | 3577.5 | 1.61 | 1.35 | 1.41 |
| | 3586.5 | 1.24 | 1.43 | 1.44 |
| | 3595.5 | 1.71 | 2.19 | 2.37 |
| | 3604.5 | 1.86 | 1.37 | 1.54 |
| | 3607.5 | 1.87 | 1.84 | 2.01 |
| | 3630.5 | 1.48 | 1.16 | 1.25 |
| | 3640 | 0.68 | 1.21 | 1.01 |
| | 3640.05 | 0.7 | 1.29 | 1.09 |
| | 3640.8 | 1.29 | 1.24 | 1.06 |
| | 3641.6 | 2.4 | 0.94 | 1.09 |
| | 3641.7 | 0.83 | 0.99 | 1.1 |
| | 3642.7 | 2.17 | 1.29 | 1.48 |
| | 3643 | 0.53 | 1.17 | 1.34 |
| | 3643.6 | 0.2 | 1.25 | 1.29 |
| | 3643.7 | 0.64 | 1.26 | 1.27 |
| | 3643.79 | 0.58 | 1.26 | 1.27 |
| | 3645.8 | 1.14 | 1.43 | 1.5 |
| | 3646 | 1.11 | 1.16 | 1.18 |
| | 3646.4 | 1.11 | 1.05 | 1.11 |
| | 3646.95 | 1.69 | 1.43 | 1.54 |
| | 3647 | 1.04 | 1.43 | 1.54 |
| | 3647.09 | 1.54 | 1.35 | 1.47 |
| | 3647.2 | 0.93 | 1.24 | 1.34 |
| | 3647.25 | 1.74 | 1.24 | 1.34 |
| | 3647.42 | 1.49 | 1.18 | 1.27 |
| | 3647.5 | 1.24 | 1.17 | 1.26 |
| | 3647.5 | 1.23 | 1.17 | 1.26 |
| | 3647.81 | 1.57 | 1.21 | 1.32 |
| | 3648.16 | 0.21 | 1.25 | 1.35 |
| | 3688.5 | 1.84 | 1.41 | 1.59 |
| | 3702 | 0.88 | 1.5 | 1.71 |

**Table 3.** *Cont.*

| Well | Depth (m) | TOC % | Prediction Multiple Regression | |
|---|---|---|---|---|
| | | | HXC | HX |
| C | 3712.5 | 1.16 | 0.89 | 0.87 |
| | 3724.5 | 0.57 | 1.31 | 1.33 |
| | 3739.5 | 1.91 | 2.24 | 2.18 |
| | 3745 | 1.49 | 2.16 | 2.15 |
| | 3751.5 | 2.26 | 2.37 | 2.2 |
| | 3755 | 2.77 | 2.7 | 2.1 |
| | 3769.5 | 1.99 | 2.1 | 1.91 |
| | 3772.5 | 2.49 | 2.13 | 1.8 |
| | 3775.5 | 2.21 | 2.06 | 1.69 |
| | 3784.5 | 2.77 | 2.21 | 2.06 |
| | 3793.5 | 2.06 | 1.96 | 1.8 |
| | 3793.5 | 2.64 | 2.02 | 1.89 |
| | 3795 | 1.93 | 2.35 | 2.24 |
| | 3795 | 2.18 | 2.2 | 2.06 |
| | 3796.5 | 1.9 | 2.04 | 1.87 |
| | 3796.5 | 2.55 | 2.06 | 1.89 |
| | 3813 | 1.4 | 1.46 | 1.27 |
| | 3852 | 2.73 | 3.24 | 3.1 |
| | 3883 | 2.86 | 2.88 | 2.14 |
| | 3888 | 2.89 | 3.55 | 2.25 |
| | 3903 | 1.47 | 1.97 | 2.19 |
| | 3910.5 | 1.48 | 1.19 | 0.91 |
| | 3913.5 | 0.91 | 1.27 | 1.29 |

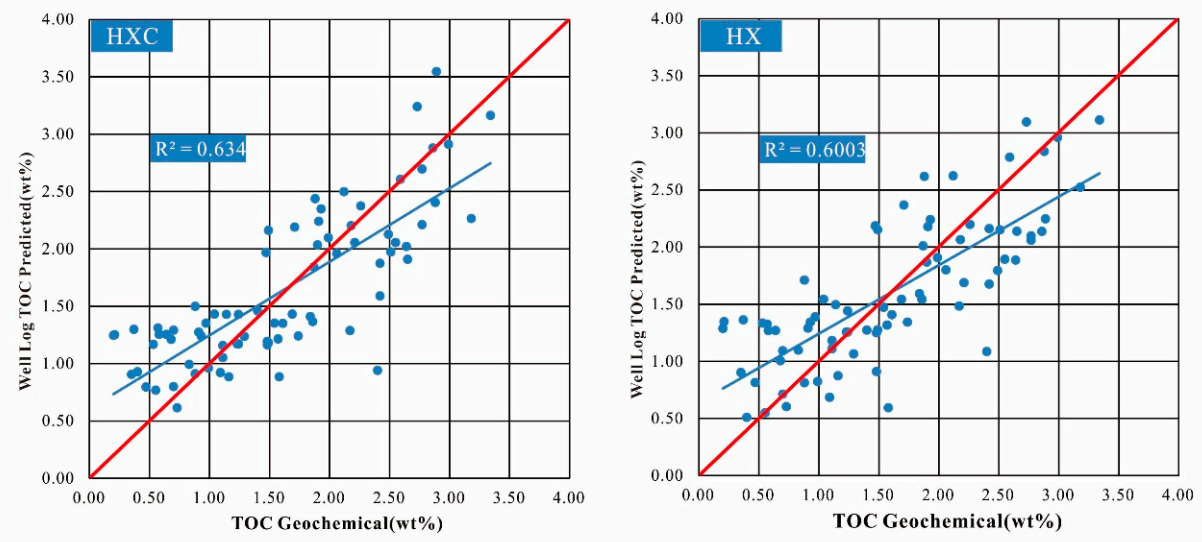



**Figure 7.** Cross plots between TOC values from multiple regression models and TOC values from core data.

**Table 4.** Mean error values of different methods.

| Method | RF | | | | BPNN | | | | ELM | | | | Multiple Regression | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Parameter Type | HXC Training | HX Training | HXC Tested | HX Tested | HXC Training | HX Training | HXC Tested | HX Tested | HXC Training | HX Training | HXC Tested | HX Tested | HXC Tested | HX Tested |
| $R^2$ | 0.97 | 0.97 | 0.65 | 0.49 | 0.91 | 0.84 | 0.70 | 0.53 | 0.88 | 0.82 | 0.66 | 0.57 | 0.63 | 0.60 |
| MAE | 0.03 | 0.03 | 0.39 | 0.56 | 0.22 | 0.26 | 0.37 | 0.47 | 0.23 | 0.28 | 0.42 | 0.45 | 0.39 | 0.41 |
| MSE | 0.02 | 0.02 | 0.26 | 0.45 | 0.07 | 0.26 | 0.23 | 0.43 | 0.10 | 0.12 | 0.27 | 0.33 | 0.24 | 0.26 |
| RMSE | 0.15 | 0.15 | 0.51 | 0.67 | 0.26 | 0.51 | 0.48 | 0.66 | 0.31 | 0.35 | 0.52 | 0.57 | 0.49 | 0.51 |

### *4.3. TOC Quantification Using ML Methods*

After using three ML test models to predict the TOC of the Yueguifeng Formation source rock in the Lishui–Jiaojiang Sag (see Table 5). $R^2$, MAE, MSE, and RMSE were calculated to evaluate the performance of the models(see Table 4). In terms of the RF model using the HXC, the $R^2$ of the training result is 0.97 (Figure 8). The MAE, MSE, and RMSE of the training results are 0.03, 0.02, and 0.15 (Table 4), respectively. The $R^2$ of the test results is 0.65 (Figure 8). The MAE, MSE, and RMSE of the test results are 0.39, 0.26, and 0.61 (Table 5), respectively. In terms of the RF model using the HX (Figure 9), the $R^2$ of the training result is 0.97. The MAE, MSE, and RMSE of the training results are 0.03, 0.02, and 0.15 (Table 5), respectively. The $R^2$ of the test results is 0.49 (Figure 9). The MAE, MSE, and RMSE of the test results are 0.56, 0.45, and 0.67 (Table 5), respectively.

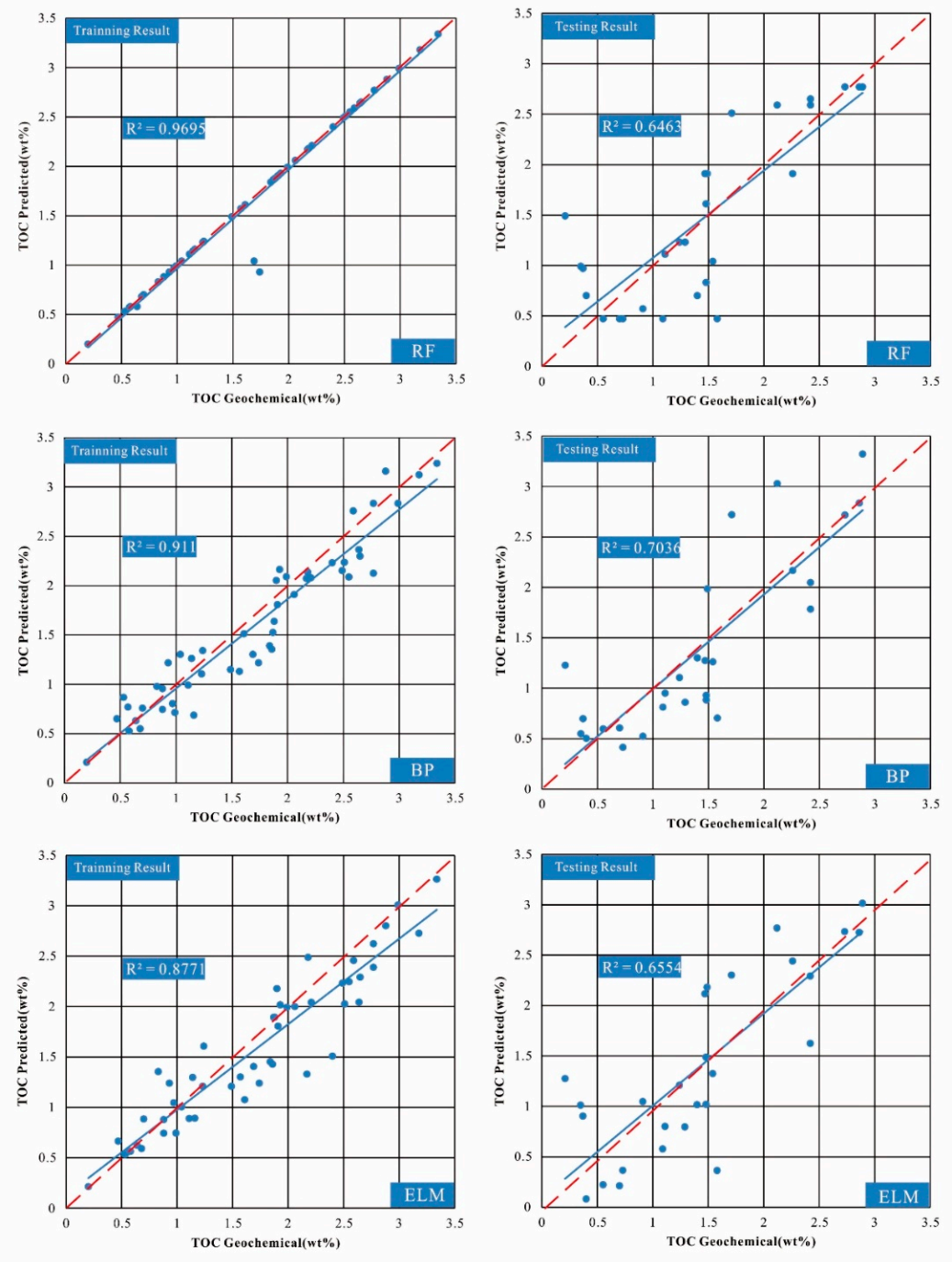



**Figure 8.** Cross plots between TOC values from ML models using HXC and TOC values from core data.

**Table 5.** Results of TOC prediction from ML.

| Core Data | Training Dataset Prediction | | | | | | Core Data | Tested Dataset Prediction | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RF | | BPNN | | ELM | | | RF | | BPNN | | ELM | |
| | HXC | HX | HXC | HX | HXC | HX | | HXC | HX | HXC | HX | HXC | HX |
| 0.99 | 0.99 | 0.99 | 0.71 | 1.11 | 0.75 | 0.82 | 1.40 | 0.70 | 0.70 | 1.30 | 1.22 | 1.02 | 1.07 |
| 0.88 | 0.88 | 0.88 | 0.75 | 0.92 | 0.74 | 0.78 | 2.73 | 2.77 | 1.88 | 2.72 | 2.77 | 2.73 | 2.81 |
| 0.47 | 0.47 | 0.47 | 0.65 | 0.92 | 0.66 | 0.42 | 2.86 | 2.77 | 2.77 | 2.84 | 2.82 | 2.73 | 2.42 |
| 3.34 | 3.34 | 3.34 | 3.24 | 3.22 | 3.26 | 2.79 | 2.89 | 2.77 | 2.65 | 3.32 | 2.86 | 3.02 | 2.56 |
| 2.99 | 2.99 | 2.99 | 2.83 | 2.96 | 3.01 | 2.90 | 1.47 | 1.91 | 1.87 | 1.27 | 2.94 | 2.12 | 2.53 |
| 1.88 | 1.88 | 1.88 | 1.64 | 2.50 | 1.90 | 2.68 | 1.48 | 1.61 | 1.16 | 0.93 | 1.00 | 1.02 | 0.92 |
| 2.59 | 2.59 | 2.59 | 2.76 | 3.07 | 2.46 | 3.01 | 0.91 | 0.57 | 1.84 | 0.52 | 1.48 | 1.05 | 1.13 |
| 3.18 | 3.18 | 3.18 | 3.12 | 3.13 | 2.73 | 2.84 | 0.73 | 0.47 | 0.47 | 0.41 | 0.72 | 0.36 | 0.44 |
| 2.88 | 2.88 | 2.88 | 3.16 | 2.87 | 2.80 | 3.15 | 0.55 | 0.47 | 0.47 | 0.60 | 0.75 | 0.22 | 0.44 |
| 2.65 | 2.65 | 2.65 | 2.30 | 2.72 | 2.29 | 2.44 | 1.09 | 0.47 | 0.47 | 0.81 | 0.79 | 0.58 | 0.52 |
| 2.51 | 2.51 | 2.51 | 2.24 | 2.58 | 2.03 | 2.36 | 1.58 | 0.47 | 0.47 | 0.70 | 0.71 | 0.36 | 0.44 |
| 0.97 | 0.97 | 0.97 | 0.80 | 1.11 | 1.05 | 1.34 | 0.35 | 0.99 | 0.99 | 0.55 | 1.47 | 1.01 | 0.96 |
| 1.61 | 1.61 | 1.61 | 1.51 | 1.51 | 1.08 | 1.39 | 0.37 | 0.97 | 0.97 | 0.70 | 1.48 | 0.90 | 1.30 |
| 1.24 | 1.24 | 1.24 | 1.34 | 1.64 | 1.61 | 1.47 | 2.12 | 2.59 | 2.59 | 3.03 | 3.04 | 2.77 | 2.92 |
| 1.86 | 1.86 | 1.86 | 1.36 | 1.41 | 1.43 | 1.44 | 2.42 | 2.65 | 1.24 | 2.05 | 2.50 | 2.29 | 2.35 |
| 1.87 | 1.87 | 1.87 | 1.53 | 2.42 | 1.89 | 2.33 | 2.42 | 2.59 | 3.18 | 1.78 | 2.11 | 1.62 | 1.99 |
| 0.68 | 0.68 | 0.68 | 0.55 | 0.68 | 0.59 | 0.81 | 1.71 | 2.51 | 3.18 | 2.72 | 2.97 | 2.30 | 2.68 |
| 0.7 | 0.7 | 0.70 | 0.76 | 1.20 | 0.88 | 0.89 | 1.48 | 0.83 | 0.88 | 0.88 | 1.27 | 1.49 | 1.22 |
| 2.4 | 2.4 | 2.40 | 2.23 | 1.74 | 1.51 | 2.33 | 1.29 | 1.23 | 1.11 | 0.86 | 1.10 | 0.80 | 0.95 |
| 0.83 | 0.83 | 0.83 | 0.98 | 1.40 | 1.36 | 1.05 | 0.40 | 0.70 | 0.47 | 0.50 | 0.76 | 0.08 | 0.36 |
| 2.17 | 2.17 | 2.17 | 2.07 | 2.31 | 1.33 | 1.39 | 0.70 | 0.47 | 0.47 | 0.61 | 0.69 | 0.21 | 0.54 |
| 0.531 | 0.531 | 0.53 | 0.87 | 0.53 | 0.53 | 0.53 | 1.49 | 1.91 | 2.51 | 1.99 | 2.58 | 2.18 | 2.36 |
| 0.2 | 0.2 | 0.20 | 0.21 | 0.20 | 0.21 | 0.22 | 2.26 | 1.91 | 1.91 | 2.17 | 2.82 | 2.44 | 2.45 |
| 0.64 | 0.58 | 0.64 | 0.63 | 0.63 | 0.62 | 1.18 | 1.11 | 1.11 | 0.20 | 0.95 | 1.31 | 0.80 | 1.10 |
| 0.58 | 0.58 | 0.64 | 0.53 | 0.53 | 0.56 | 1.18 | 1.54 | 1.04 | 2.17 | 1.26 | 1.55 | 1.32 | 1.40 |
| 1.14 | 1.14 | 1.14 | 1.26 | 1.77 | 1.30 | 1.51 | 1.24 | 1.23 | 1.23 | 1.11 | 1.27 | 1.21 | 1.16 |
| 1.11 | 1.11 | 1.11 | 0.99 | 1.12 | 0.89 | 0.99 | 0.21 | 1.49 | 0.53 | 1.23 | 1.36 | 1.28 | 1.27 |
| 1.69 | 1.04 | 1.04 | 1.30 | 1.67 | 1.41 | 1.50 | | | | | | | |
| 1.04 | 1.04 | 1.04 | 1.30 | 1.67 | 1.01 | 1.50 | | | | | | | |
| 0.93 | 0.93 | 0.93 | 1.22 | 1.39 | 1.24 | 0.91 | | | | | | | |
| 1.74 | 0.93 | 0.93 | 1.22 | 1.39 | 1.24 | 1.25 | | | | | | | |
| 1.49 | 1.49 | 1.49 | 1.15 | 1.29 | 1.21 | 1.17 | | | | | | | |
| 1.23 | 1.23 | 1.23 | 1.11 | 1.27 | 1.21 | 1.16 | | | | | | | |
| 1.57 | 1.57 | 1.57 | 1.13 | 1.34 | 1.30 | 1.24 | | | | | | | |
| 1.84 | 1.84 | 1.84 | 1.39 | 1.31 | 1.45 | 1.47 | | | | | | | |
| 0.88 | 0.88 | 0.88 | 0.96 | 1.40 | 0.88 | 0.90 | | | | | | | |
| 1.16 | 1.16 | 1.16 | 0.69 | 1.36 | 0.89 | 0.85 | | | | | | | |
| 0.57 | 0.57 | 0.57 | 0.77 | 1.24 | 0.56 | 0.58 | | | | | | | |
| 1.91 | 1.91 | 1.91 | 1.81 | 1.95 | 1.81 | 1.90 | | | | | | | |
| 2.77 | 2.77 | 2.77 | 2.83 | 2.77 | 2.62 | 2.33 | | | | | | | |
| 1.99 | 1.99 | 1.99 | 2.09 | 2.32 | 1.99 | 2.03 | | | | | | | |
| 2.49 | 2.49 | 2.49 | 2.15 | 2.17 | 2.23 | 1.87 | | | | | | | |
| 2.21 | 2.21 | 2.21 | 2.08 | 2.05 | 2.04 | 1.74 | | | | | | | |
| 2.77 | 2.77 | 2.77 | 2.13 | 2.32 | 2.39 | 2.21 | | | | | | | |
| 2.06 | 2.06 | 2.06 | 1.91 | 2.29 | 2.00 | 1.90 | | | | | | | |
| 2.64 | 2.64 | 2.64 | 2.36 | 2.29 | 2.04 | 2.46 | | | | | | | |
| 1.93 | 1.93 | 1.93 | 2.16 | 1.98 | 2.02 | 2.41 | | | | | | | |
| 2.18 | 2.18 | 2.18 | 2.14 | 2.32 | 2.49 | 2.22 | | | | | | | |
| 1.9 | 1.9 | 1.90 | 2.05 | 2.36 | 2.18 | 1.99 | | | | | | | |
| 2.55 | 2.55 | 2.55 | 2.09 | 2.44 | 2.25 | 2.03 | | | | | | | |

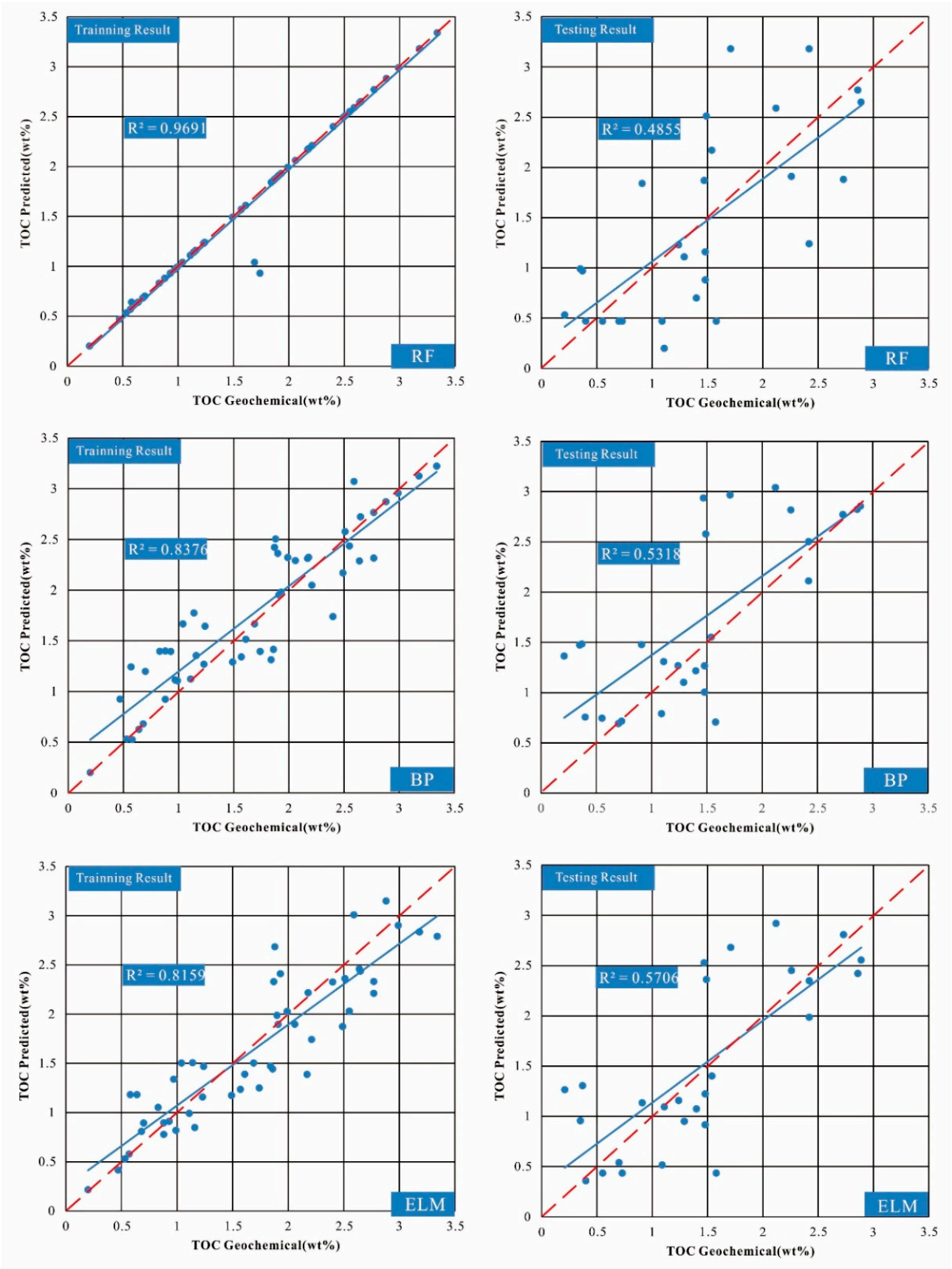



**Figure 9.** Cross plots between TOC values from ML models using HX and TOC values from core data.

In terms of the BP model using the HXC, the $R^2$ of the training result is 0.91 (Figure 8). The MAE, MSE, and RMSE of the training results are 0.22, 0.07, and 0.26 (Table 5), respectively. The $R^2$ of the test results is 0.70 (Figure 8). The MAE, MSE, and RMSE of the test results are 0.37, 0.24, and 0.48 (Table 5), respectively. In terms of BP model using the HX, the $R^2$ of the training result is 0.84 (Figure 9). The MAE, MSE, and RMSE of the training results are 0.26, 0.26, and 0.51 (Table 5), respectively. The $R^2$ of the test results is 0.53 (Figure 9). The MAE, MSE, and RMSE of the test results are 0.47, 0.43, and 0.66 (Table 5), respectively.

In terms of the ELM model using the HXC, the $R^2$ of the training result is 0.88 (Figure 8). The MAE, MSE, and RMSE of the training results are 0.23, 0.10, and 0.31 (Table 5), respectively. The $R^2$ of the test results is 0.65 (Figure 8). The MAE, MSE, and RMSE of the test results are 0.42, 0.27, and 0.52 (Table 5), respectively. In terms of the ELM model using the HX, the $R^2$ of the training result is 0.82 (Figure 9). The MAE, MSE, and RMSE of the training results are 0.28, 0.12, and 0.35 (Table 5), respectively. The $R^2$ of the test results is 0.57 (Figure 9). The MAE, MSE, and RMSE of the test results are 0.45, 0.33, and 0.57 (Table 5), respectively.

The error comparison analysis of the training data and test data obtained using three different ML methods has been carried out. It shows that the $R^2$ of the training results of all ML methods is greater than 0.8, regardless of which combination was used. The values of MAE, MSE, and RMSE are ideal. It indicates that the models using the three ML methods have strong stability and are very accurate for re-prediction of input data. However, the three ML methods performed differently in terms of the accuracy of the test results. The prediction accuracy of the ELM model using the HX is the highest, followed by the BPNN model and is the lowest in the RF model. This may be because ELM is different from the back propagation algorithm of the neural network, and the output layer weights are calculated by generalized inverse matrix theory. This allows ELM to ensure a certain prediction accuracy with fewer training parameters. The core of the RF algorithm is random sampling with replacement. The RF model has a poor learning effect on the target because of the lesser feature input using the HX. It leads to the lowest prediction accuracy of the RF model. The BPNN model using the HXC has the highest prediction accuracy, and the ELM and RF perform about the same. This verifies that the difference in the core algorithm affects the accuracy of the prediction again. At the same time, the dimension and nature of the input data and the matching degree of the algorithm will also affect the prediction accuracy. Compared with ELM and RF, BPNN has a simpler structure and simpler operation. This advantage makes BPNN potentially more accurate when dealing with low-latitude data. It also shows that advanced ML algorithms do not always perform better.

The accuracy of ML models using two different parameter combinations has been compared. The average $R^2$ of the ML model using the HXC is 0.67. The average MAE, MSE, and RMSE are 0.4, 0.25, and 0.5, respectively. The average R2 of the ML models using the HX is 0.53. The average MAE, MSE, and RMSE are 0.49, 0.40, and 0.63, respectively. It shows that the accuracy of different ML models using HXC is significantly higher than those using HX, which is related to the completeness of feature information. The HX has the highest correlation with the expected value but provides no sufficient feature information, resulting in a weakening of the information transfer process of the conditional entropy and a decrease in the prediction accuracy. Although the two additional logging parameters are not highly correlated with the expected value, they supplement more characteristic information of the expected value and improve the prediction performance. Therefore, more parameter inputs that can describe the characteristics of the target value are beneficial to the prediction of the target value.

In terms of running time, the RF model has the shortest running time, followed by the ELM model, and the BPNN model has the longest running time (see Figure 10). The running time of the RF model using HXC parameters is 0.24 s. The running time of the RF model using HX parameters is 0.288 s. The running time of the ELM model using HXC parameters is 1.031 s. The running time of the ELM model using HX parameters is 1.022 s.

The running time of the BP model using HXC parameters is 2.601 s. The running time of the BP model using the HX parameter is 2.505 s. The ELM model and BP model running time display fewer feature parameters that have faster running times. The less feature parameter input, the less model computation time, but the predictions will be even worse, while the RF model run time shows different characteristics: the more feature parameters, the faster the run time. This may be due to the fact that random sampling of samples is more complicated because there are fewer feature parameters. In the final integration of the prediction results, the RF model using the HX parameter requires more computation time. Considering the prediction accuracy and running time, the BPNN model has higher prediction accuracy but longer running time in small-sample regression prediction. The RF model can run faster while ensuring a certain prediction accuracy. According to the run time comparison, the RF model is much faster than the BPNN model, which indicates that RF should be chosen as the better option if processing speed is important.

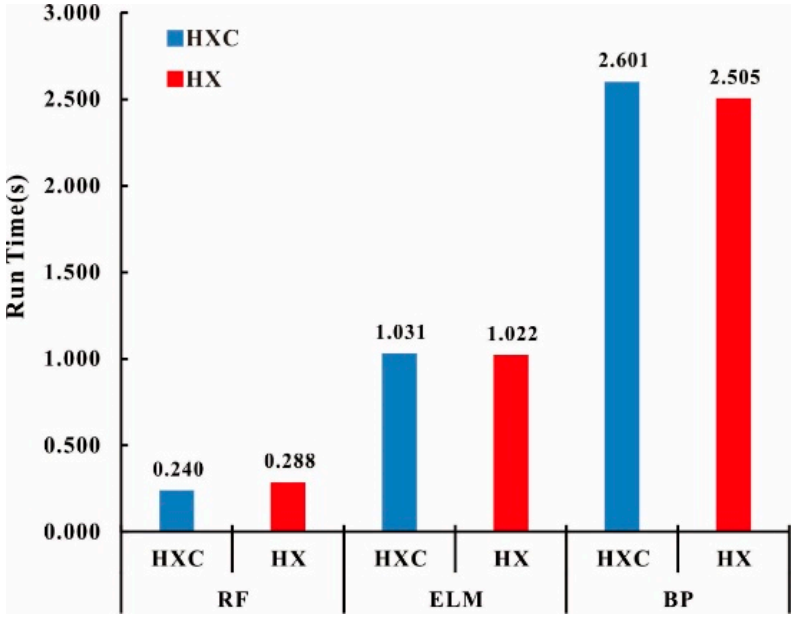



**Figure 10.** Run time of ML models using HXC and HX.

This study also found that the prediction accuracy of the training set was significantly higher than that of the test set. All models exhibit varying degrees of overfitting. Possible reasons are: (1) the number of training samples is insufficient, resulting in a gap in the feature extraction and distribution simulation of the expected value by each ML algorithm; (2) since the purpose of this research is to test the universality of different regions, there are a great deal of well logging data. There must be uncontrollable human error, resulting in low prediction accuracy. However, in general, the correlation between the measured value and the predicted value of the BPNN test set still reaches 70%, and it still has a good application prospect. However, the generalized TOC prediction model in the study area seems to need more data supplementation and in-depth research. We believe that the prediction model will perform better with more data.

### *4.4. Compare between Multiple Regression and ML*

The predicted values of each model using the better parameters (HXC) were compared. Figure 11 shows that the ML prediction accuracy is significantly higher than the multiple regression model. The results suggest that the prediction accuracy of the BPNN model is 10% higher than that of the multiple regression model. It also indicates the importance of feature parameters. When there are enough feature parameters, the more comprehensive the description of the target value, the better the prediction performance of the ML model.

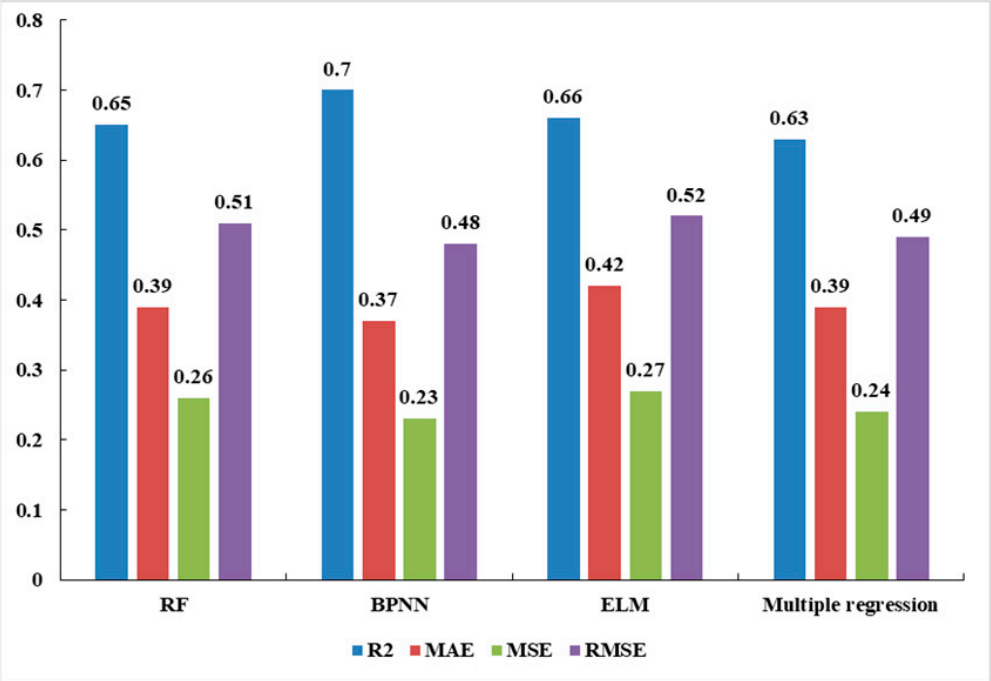



**Figure 11.** Error analysis of each model using HXC.

## 5. Conclusions

In this study, the authors established ML models and multiple regression models for TOC prediction with logging parameters and made a comparison of the prediction effect of each model. The datasets including all available well logs and TOC core measurements conducted in Lishui–Jiaojiang Sag served as necessary data for constructing the models. The ML models were developed based on random forest (RF), extreme learning machine (ELM), and back propagation neural network (BPNN). The performance of each model has been evaluated using different factors, including $R^2$, MAE, MSE, and RMSE. Based on the above results, the following conclusions can be drawn:

1. The TOC content of the source rocks in Yueguifeng Formation is relatively high, with an overall distribution of 0.2–3.34%. The $S_1 + S_2$ is generally distributed in the range of 0.5 mg HC/g TOC to 12.51 mg HC/g TOC. The type of kerogen is mainly type II. The source rocks of Yueguifeng Formation have good hydrocarbon generation potential.
2. The correlations between each logging parameter and TOC were evaluated through linear regression method and Pearson correlation coefficient analysis. The results indicate that the TOC of Yueguifeng Formation source rock has a better response in DEN, DT, and CN logging. The performance of each model using all well logs and selected well logs shows that each model with all well logs as input performed much better than the models with selected well logs.
3. In terms of accuracy, the results of error analysis show that each ML model with all well logs as input performed much better than the multiple regression models. In addition, it can be concluded that the BPNN model outperforms the other ML models. According to the run time comparison, the RF model is much faster than the BPNN model, which indicates that RF should be chosen as the better option if processing speed is important. This study confirmed the ability of ML models for building an efficient model for estimating TOC from readily available borehole logs data in the study area.

**Author Contributions:** X.H.: Investigation, Methodology, Writing—review and editing; D.H.: Conceptualization, Methodology, Supervision, Writing—review and editing; X.C. and C.N.: Investigation; Y.L. and S.C.: software. All authors have read and agreed to the published version of the manuscript.

**Funding:** The National Natural Science Foundation of China under contract No. 41872131.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** We would like to thank the CNOOC Shanghai Company for providing the core samples used in this work.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## References

1. Bechtel, A.; Movsumova, U.; Strobl, S.A.; Sachsenhofer, R.F.; Soliman, A.; Gratzer, R.; Püttmann, W. Organofacies and paleoenvironment of the Oligocene Maikop series of Angeharan (eastern Azerbaijan). *Org. Geochem.* **2013**, *56*, 51–67. [CrossRef]
2. Hakimi, M.H.; Abdullah, W.H.; Alqudah, M.; Makeen, Y.M.; Mustapha, K.A. Organic geochemical and petrographic characteristics of the oil shales in the Lajjun area, Central Jordan: Origin of organic matter input and preservation conditions. *Fuel* **2016**, *181*, 34–45. [CrossRef]
3. Hakimi, M.H.; Abdullah, W.H.; Makeen, Y.M.; Saeed, S.A.; Al-Hakame, H.; Al-Moliki, T.; Al-Sharabi, K.Q.; Hatem, B.A. Geochemical characterization of the Jurassic Amran deposits from Sharab area (SW Yemen): Origin of organic matter, paleoenvironmental and paleoclimate conditions during deposition. *J. Afr. Earth. Sci.* **2017**, *129*, 579–595. [CrossRef]
4. Tan, M.J.; Song, X.D.; Yang, X.; Wu, Q.Z. Support-vector-regression machine technology for total organic carbon content prediction from wireline logs in organic shale: A comparative study. *J. Nat. Gas Sci. Eng.* **2015**, *26*, 792–802. [CrossRef]
5. Beers, R.F. Radioactivity and Organic Content of Some Paleozoic Shales1. *AAPG Bull.* **1945**, *29*, 1–22.
6. Schmoker, J.W.; Hester, T.C. Organic Carbon in Bakken Formation, United States Portion of Williston Basin1. *AAPG Bull.* **1983**, *67*, 2165–2174.
7. Mendelzon, J.D.; Toksoz, M.N. Source Rock Characterization Using Multivariate Analysis of Log Data. In Proceedings of the SPWLA 26th Annual Logging Symposium, Dallas, TX, USA, 17 June 1985; p. SPWLA-1985-UU.
8. Autric, A. Resistivity, Radioactivity And Sonic Transit Time Logs To Evaluate The Organic Content Of Low Permeability Rocks. *Log Anal.* **1985**, *26*, SPWLA-1985-vXXVIn3a3.
9. Passey, Q.R.; Creaney, S.; Kulla, J.B.; Moretti, F.J.; Stroud, J.D. A Practical Model for Organic Richness from Porosity and Resistivity Logs1. *AAPG Bull.* **1990**, *74*, 1777–1794.
10. Kamali, M.R.; Allah Mirshady, A. Total organic carbon content determined from well logs using ΔLogR and Neuro Fuzzy techniques. *J. Pet. Sci. Eng.* **2004**, *45*, 141–148. [CrossRef]
11. Passey, Q.R.; Bohacs, K.M.; Esch, W.L.; Klimentidis, R.; Sinha, S. From Oil-Prone Source Rock to Gas-Producing Shale Reservoir—Geologic and Petrophysical Characterization of Unconventional Shale-Gas Reservoirs. In Proceedings of the International Oil and Gas Conference and Exhibition in China, Beijing, China, 8 June 2010; p. SPE-131350-MS.
12. Hu, H.T.; Lu, S.F.; Liu, C.; Wang, W.M.; Wang, M.; Li, J.J.; Shang, J.H. Models for Calculating Organic Carbon Content from Logging Information: Comparison and Analysis. *Acta Sedimentol. Sin.* **2011**, *29*, 1199–1205.
13. Wang, P.; Chen, Z.; Pang, X.; Hu, K.; Sun, M.; Chen, X. Revised models for determining TOC in shale play: Example from Devonian Duvernay Shale, Western Canada Sedimentary Basin. *Mar. Pet. Geol.* **2016**, *70*, 304–319. [CrossRef]
14. Zhao, P.; Ma, H.; Rasouli, V.; Liu, W.; Cai, J.; Huang, Z. An improved model for estimating the TOC in shale formations. *Mar. Pet. Geol.* **2017**, *83*, 174–183. [CrossRef]
15. Siddig, O.; Ibrahim, A.F.; Elkatatny, S. Application of Various Machine Learning Techniques in Predicting Total Organic Carbon from Well Logs. *Comput. Intell. Neurosci.* **2021**, *2021*, 7390055. [CrossRef] [PubMed]
16. Zheng, D.Y.; Wu, S.X.; Hou, M.C. Fully connected deep network: An improved method to predict TOC of shale reservoirs from well logs. *Mar. Pet. Geol.* **2021**, *132*, 105205. [CrossRef]
17. Filipič, B.; Junkar, M. Using inductive machine learning to support decision making in machining processes. *Comput. Ind.* **2000**, *43*, 31–41. [CrossRef]
18. Kim, D.H.; Kim, T.J.; Wang, X.; Kim, M.; Quan, Y.J.; Oh, J.W.; Min, S.H.; Kim, H.; Bhandari, B.; Yang, I.; et al. Smart Machining Process Using Machine Learning: A Review and Perspective on Machining Industry. *Int. J. Precis. Eng. Manuf.-Green Technol.* **2018**, *5*, 555–568. [CrossRef]
19. Rehman, T.U.; Mahmud, M.S.; Chang, Y.K.; Jin, J.; Shin, J. Current and future applications of statistical machine learning algorithms for agricultural machine vision systems. *Comput. Electron. Agric.* **2019**, *156*, 585–605. [CrossRef]
20. Shin, S.J.; Kim, Y.M.; Meilanitasari, P. A Holonic-Based Self-Learning Mechanism for Energy-Predictive Planning in Machining Processes. *Processes* **2019**, *7*, 739. [CrossRef]
21. Ucar, F.; Cordova, J.; Alcin, O.F.; Dandil, B.; Ata, F.; Arghandeh, R. Bundle Extreme Learning Machine for Power Quality Analysis in Transmission Networks. *Energies* **2019**, *12*, 1449. [CrossRef]

22. Xu, Y.; Zhou, Y.; Sekula, P.; Ding, L. Machine learning in construction: From shallow to deep learning. *Dev. Built Environ.* **2021**, *6*, 100045. [CrossRef]
23. Wong, L.J.; Michaels, A.J. Transfer Learning for Radio Frequency Machine Learning: A Taxonomy and Survey. *Sensors* **2022**, *22*, 1416. [CrossRef] [PubMed]
24. Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug design by machine learning: Support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26*, 5–14. [CrossRef] [PubMed]
25. Sample, P.A.; Goldbaum, M.H.; Chan, K.; Boden, C.; Lee, T.W.; Vasile, C.; Boehm, A.G.; Sejnowski, T.; Johnson, C.A.; Weinreb, R.N. Using machine learning classifiers to identify glaucomatous change earlier in standard visual fields. *Invest. Ophthalmol. Visual Sci.* **2002**, *43*, 2660–2665.
26. Bax, J.J.; van der Bijl, P.; Delgado, V. Machine Learning for Electrocardiographic Diagnosis of Left Ventricular Early Diastolic Dysfunction∗. *J. Am. Coll. Cardiol.* **2018**, *71*, 1661–1662. [CrossRef] [PubMed]
27. Singh, K.; Beam, A.L.; Nallamothu, B.K. Machine Learning in Clinical Journals Moving From Inscrutable to Informative. *Circ. Cardiovasc. Qual. Outcomes* **2020**, *13*, e007491. [CrossRef]
28. Mainali, S.; Darsie, M.E.; Smetana, K.S. Machine Learning in Action: Stroke Diagnosis and Outcome Prediction. *Front. Neurol.* **2021**, *12*, 2153. [CrossRef]
29. Shuhaiber, J.H.; Conte, J.V. Machine learning in heart valve surgery. *Eur. J. Cardio-Thorac. Surg.* **2021**, *60*, 1386–1387. [CrossRef]
30. Tong, S.; Koller, D. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.* **2002**, *2*, 45–66.
31. Chen, Z.Y.; Khoa LD, V.; Teoh, E.N.; Nazir, A.; Karuppiah, E.K.; Lam, K.S. Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: A review. *Knowl. Inf. Syst.* **2018**, *57*, 245–285. [CrossRef]
32. Andres, A.R.; Otero, A.; Amavilah, V.H. Using deep learning neural networks to predict the knowledge economy index for developing and emerging economies. *Expert Syst. Appl.* **2021**, *184*, 115514. [CrossRef]
33. Xie, X.; Sun, S. Multi-view Laplacian twin support vector machines. *Appl. Intell.* **2014**, *41*, 1059–1068. [CrossRef]
34. de Bruijne, M. Machine learning approaches in medical image analysis: From detection to diagnosis. *Med. Image Anal.* **2016**, *33*, 94–97. [CrossRef] [PubMed]
35. De Iaco, S.; Hristopulos, D.T.; Lin, G. Special Issue: Geostatistics and Machine Learning. *Math. Geosci.* **2022**, *54*, 459–465. [CrossRef]
36. de Matos, M.C.; Osorio PL, M.; Johann, P.R.S. Unsupervised seismic facies analysis using wavelet transform and self-organizing maps. *Geophysics* **2007**, *72*, P9–P21. [CrossRef]
37. de Matos, M.C.; Yenugu, M.; Angelo, S.M.; Marfurt, K.J. Integrated seismic texture segmentation and cluster analysis applied to channel delineation and chert reservoir characterization. *Geophysics* **2011**, *76*, P11–P21. [CrossRef]
38. Roy, A.; Romero-Pelaez, A.S.; Kwiatkowski, T.J.; Marfurt, K.J. Generative topographic mapping for seismic facies estimation of a carbonate wash, Veracruz Basin, southern Mexico. *Interpret.—A J. Subsurf. Charact.* **2014**, *2*, SA31–SA47. [CrossRef]
39. Qi, J.; Lin, T.F.; Zhao, T.; Li, F.Y.; Marfurt, K. Semisupervised multiattribute seismic facies analysis. *Interpret.—A J. Subsurf. Charact.* **2016**, *4*, SB91–SB106. [CrossRef]
40. Qian, F.; Yin, M.; Liu, X.Y.; Wang, Y.J.; Lu, C.; Hu, G.M. Unsupervised seismic facies analysis via deep convolutional autoencoders. *Geophysics* **2018**, *83*, A39–A43. [CrossRef]
41. Harris, J.R.; Grunsky, E.C. Predictive lithological mapping of Canada's North using Random Forest classification applied to geophysical and geochemical data. *Comput. Geosci.* **2015**, *80*, 9–25. [CrossRef]
42. Sebtosheikh, M.A.; Motafakkerfard, R.; Riahi, M.A.; Moradi, S.; Sabety, N. Support vector machine method, a new technique for lithology prediction in an Iranian heterogeneous carbonate reservoir using petrophysical well logs. *Carbonates Evaporites* **2015**, *30*, 59–68. [CrossRef]
43. Ai, X.; Wang, H.Y.; Sun, B.T. Automatic Identification of Sedimentary Facies Based on a Support Vector Machine in the Aryskum Graben, Kazakhstan. *Appl. Sci.* **2019**, *9*, 4489. [CrossRef]
44. Mulashani, A.K.; Shen, C.B.; Asante-Okyere, S.; Kerttu, P.N.; Abelly, E.N. Group Method of Data Handling (GMDH) Neural Network for Estimating Total Organic Carbon (TOC) and Hydrocarbon Potential Distribution (S-1, S-2) Using Well Logs. *Nat. Resour. Res.* **2021**, *30*, 3605–3622. [CrossRef]
45. Hossain, T.M.; Watada, J.; Aziz, I.A.; Hermana, M. Machine Learning in Electrofacies Classification and Subsurface Lithology Interpretation: A Rough Set Theory Approach. *Appl. Sci.* **2020**, *10*, 5940. [CrossRef]
46. Ashraf, U.; Zhang, H.; Anees, A.; Mangi, H.N.; Ali, M.; Zhang, X.; Imraz, M.; Abbasi, S.S.; Abbas, A.; Ullah, Z.; et al. A Core Logging, Machine Learning and Geostatistical Modeling Interactive Approach for Subsurface Imaging of Lenticular Geobodies in a Clastic Depositional System, SE Pakistan. *Nat. Resour. Res.* **2021**, *30*, 2807–2830. [CrossRef]
47. Shokir, E.M.E. A novel model for permeability prediction in uncored wells. *SPE Reserv. Eval. Eng.* **2006**, *9*, 266–273. [CrossRef]
48. Al-Anazi, A.; Gates, I.D. Support-Vector Regression for Permeability Prediction in a Heterogeneous Reservoir: A Comparative Study. *SPE Reserv. Eval. Eng.* **2010**, *13*, 485–495. [CrossRef]
49. Kaydani, H.; Mohebbi, A.; Eftekhari, M. Permeability estimation in heterogeneous oil reservoirs by multi-gene genetic programming algorithm. *J. Pet. Sci. Eng.* **2014**, *123*, 201–206. [CrossRef]
50. Zhang, G.Y.; Wang, Z.Z.; Li, H.J.; Sun, Y.A.; Zhang, Q.C.; Chen, W. Permeability prediction of isolated channel sands using machine learning. *J. Appl. Geophys.* **2018**, *159*, 605–615. [CrossRef]

51. Zhang, G.Y.; Wang, Z.Z.; Mohaghegh, S.; Lin, C.Y.; Sun, Y.N.; Pei, S.J. Pattern visualization and understanding of machine learning models for permeability prediction in tight sandstone reservoirs. *J. Pet. Sci. Eng.* **2021**, *200*, 108142. [CrossRef]
52. Liu, J.Z.; Liu, X.Y. Recognition and Classification for Inter-well Nonlinear Permeability Configuration in Low Permeability Reservoirs Utilizing Machine Learning Methods. *Front. Earth Sci.* **2022**, *10*, 218. [CrossRef]
53. Wang, Y.J.; Li, H.Y.; Xu, J.C.; Liu, S.Y.; Wang, X.P. Machine learning assisted relative permeability upscaling for uncertainty quantification. *Energy* **2022**, *245*, 123284. [CrossRef]
54. de Lima, R.P.; Surianam, F.; Marfurt, K.J.; Pranter, M.J. Convolutional neural networks as aid in core lithofacies classification. *Interpret.—A J. Subsurf. Charact.* **2019**, *7*, SF27–SF40.
55. Koeshidayatullah, A.; Morsilli, M.; Lehrmann, D.J.; Al-Ramadan, K.; Payne, J.L. Fully automated carbonate petrography using deep convolutional neural networks. *Mar. Pet. Geol.* **2020**, *122*, 104687. [CrossRef]
56. de Lima, R.P.; Duarte, D.; Nicholson, C.; Slatt, R.; Marfurt, K.J. Petrographic microfacies classification with deep convolutional neural networks. *Comput. Geosci.* **2020**, *142*, 104481. [CrossRef]
57. Baraboshkin, E.E.; Ismailova, L.S.; Orlov, D.M.; Zhukovskaya, E.A.; Kalmykov, G.A.; Khotylev, O.V.; Baraboshkin, E.Y.; Koroteev, D.A. Deep convolutions for in-depth automated rock typing. *Comput. Geosci.* **2020**, *135*, 104330. [CrossRef]
58. Izadi, H.; Sadri, J.; Bayati, M. An intelligent system for mineral identification in thin sections based on a cascade approach. *Comput. Geosci.* **2017**, *99*, 37–49. [CrossRef]
59. Saporetti, C.M.; da Fonseca, L.G.; Pereira, E.; de Oliveira, L.C. Machine learning approaches for petrographic classification of carbonate-siliciclastic rocks using well logs and textural information. *J. Appl. Geophys.* **2018**, *155*, 217–225. [CrossRef]
60. Silva, A.A.; Tavares, M.W.; Carrasquilla, A.; Missagia, R.; Ceia, M. Petrofacies classification using machine learning algorithms. *Geophysics* **2020**, *85*, WA101–WA113. [CrossRef]
61. Zhu, L.; Zhang, C.; Zhang, C.; Zhang, Z.; Nie, X.; Zhou, X.; Liu, W.; Wang, X. Forming a new small sample deep learning model to predict total organic carbon content by combining unsupervised learning with semisupervised learning. *Appl. Soft Comput.* **2019**, *83*, 105596. [CrossRef]
62. Shi, X.; Wang, J.; Liu, G.; Yang, L.; Ge, X.M.; Jiang, S. Application of extreme learning machine and neural networks in total organic carbon content prediction in organic shale with wire line logs. *J. Nat. Gas Sci. Eng.* **2016**, *33*, 687–702. [CrossRef]
63. Johnson, L.M.; Rezaee, R.; Kadkhodaie, A.; Smith, G.; Yu, H.Y. Geochemical property modelling of a potential shale reservoir in the Canning Basin (Western Australia), using Artificial Neural Networks and geostatistical tools. *Comput. Geosci.* **2018**, *120*, 73–81. [CrossRef]
64. Zhu, L.; Zhang, C.; Zhang, C.; Wei, Y.; Zhou, X.; Cheng, Y.; Huang, Y.; Zhang, L. Prediction of total organic carbon content in shale reservoir based on a new integrated hybrid neural network and conventional well logging curves. *J. Geophys. Eng.* **2018**, *15*, 1050–1061. [CrossRef]
65. Wang, H.J.; Wu, W.; Chen, T.; Dong, X.J.; Wang, G.X. An improved neural network for TOC, S-1 and S-2 estimation based on conventional well logs. *J. Pet. Sci. Eng.* **2019**, *176*, 664–678. [CrossRef]
66. Zhu, L.Q.; Zhang, C.; Zhang, C.M.; Zhou, X.Q.; Wang, J.; Wang, X. 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 Geophys* **2018**, *66*, 983–1000. [CrossRef]
67. Liu, X.Z.; Tian, Z.; Chen, C. Total Organic Carbon Content Prediction in Lacustrine Shale Using Extreme Gradient Boosting Machine Learning Based on Bayesian Optimization. *Geofluids* **2021**, *2021*, 6155663. [CrossRef]
68. Rui, J.W.; Zhang, H.B.; Zhang, D.L.; Han, F.L.; Guo, Q. Total organic carbon content prediction based on support-vector-regression machine with particle swarm optimization. *J. Pet. Sci. Eng.* **2019**, *180*, 699–706. [CrossRef]
69. Amosu, A.; Imsalem, M.; Sun, Y.F. Effective machine learning identification of TOC-rich zones in the Eagle Ford Shale. *J. Appl. Geophys.* **2021**, *188*, 104311. [CrossRef]
70. Rong, J.; Zheng, Z.; Luo, X.; Li, C.; Li, Y.; Wei, X.; Wei, Q.; Yu, G.; Zhang, L.; Lei, Y. Machine Learning Method for TOC Prediction: Taking Wufeng and Longmaxi Shales in the Sichuan Basin, Southwest China as an Example. *Geofluids* **2021**, *2021*, 6794213. [CrossRef]
71. Zhu, L.; Zhang, C.; Zhang, C.; Zhang, Z.; Zhou, X.; Liu, W.; Zhu, B. 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. *J. Pet. Sci. Eng.* **2020**, *188*, 106944. [CrossRef]
72. Handhal, A.M.; Al-Abadi, A.M.; Chafeet, H.E.; Ismail, M.J. Prediction of total organic carbon at Rumaila oil field, Southern Iraq using conventional well logs and machine learning algorithms. *Mar. Pet. Geol.* **2020**, *116*, 104347. [CrossRef]
73. Ao, S.U.; Honghan, C.H.E.N.; Laisheng, C.A.O.; Mingzhu, L.E.I.; Cunwu, W.A.N.G.; Yanhua, L.I.U.; Peijun, L.I. Genesis, source and charging of oil and gas in Lishui sag, East China Sea Basin. *Pet. Explor. Dev.* **2014**, *41*, 574–584.
74. Jiang, Z.; Li, Y.; Du, H.; Zhang, Y. The Cenozoic structural evolution and its influences on gas accumulation in the Lishui Sag, East China Sea Shelf Basin. *J. Nat. Gas Sci. Eng.* **2015**, *22*, 107–118. [CrossRef]
75. Zhang, M.; Zhang, J.; Xu, F.; Li, J.; Liu, J.; Hou, G.; Zhang, P. Paleocene sequence stratigraphy and depositional systems in the Lishui Sag, East China Sea Shelf Basin. *Mar. Pet. Geol.* **2015**, *59*, 390–405. [CrossRef]
76. Li, Y.; Zhang, J.; Liu, Y.; Shen, W.; Chang, X.; Sun, Z.; Xu, G. Organic geochemistry, distribution and hydrocarbon potential of source rocks in the Paleocene, Lishui Sag, East China Sea Shelf Basin. *Mar. Pet. Geol.* **2019**, *107*, 382–396. [CrossRef]

77. Liu, L.; Li, Y.; Dong, H.; Sun, Z. Diagenesis and reservoir quality of Paleocene tight sandstones, Lishui Sag, East China Sea Shelf Basin. *J. Pet. Sci. Eng.* **2020**, *195*, 107615. [CrossRef]
78. Lei, C.; Yin, S.; Ye, J.; Wu, J.; Wang, Z.; Gao, B. Characteristics and deposition models of the paleocene source rocks in the Lishui Sag, east China sea shelf basin: Evidences from organic and inorganic geochemistry. *J. Pet. Sci. Eng.* **2021**, *200*, 108342. [CrossRef]
79. Huang, Y.; Tarantola, A.; Wang, W.; Caumon, M.C.; Pironon, J.; Lu, W.; Yan, D.; Zhuang, X. Charge history of CO2 in Lishui sag, East China Sea basin: Evidence from quantitative Raman analysis of CO2-bearing fluid inclusions. *Mar. Pet. Geol.* **2018**, *98*, 50–65. [CrossRef]
80. Li, Y.; Zhang, J.; Xu, Y.; Chen, T.; Liu, J. Improved understanding of the origin and accumulation of hydrocarbons from multiple source rocks in the Lishui Sag: Insights from statistical methods, gold tube pyrolysis and basin modeling. *Mar. Pet. Geol.* **2021**, *134*, 105361. [CrossRef]
81. Nagasaka, Y.; Iwata, A. Performance evaluation of BP and PCA neural networks for ECG data compression. In Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan), Nagoya, Japan, 25–29 October 1993; Volume 1, pp. 1003–1006.
82. Fung, C.C.; Wong, K.W.; Eren, H.; Charlebois, R.; Crocker, H. Modular artificial neural network for prediction of petrophysical properties from well log data. In Proceedings of the Quality Measurement: The Indispensable Bridge between Theory and Reality (No Measurements? No Science! Joint Conference-1996: IEEE Instrumentation and Measurement Technology Conference and IMEKO Tec, Brussels, Belgium, 4–6 June 1996; Volume 2, pp. 1010–1014.
83. Huang, Y.; Wong, P.M.; Gedeon, T.D. An improved fuzzy neural network for permeability estimation from wireline logs in a petroleum reservoir. In Proceedings of the Proceedings of Digital Processing Applications (TENCON '96), Perth, WA, Australia, 29 November 1996; Volume 2, pp. 912–917.
84. Wong, P.M.; Gedeon, T.D.; Taggart, I.J. Fuzzy ARTMAP: A new tool for lithofacies recognition. *Ai Appl.* **1996**, *10*, 29–39.
85. Wong, P.M.; Henderson, D.J.; Brooks, L.J. Permeability determination using neural networks in the Ravva Field, offshore India. *SPE Reserv. Eval. Eng.* **1998**, *1*, 99–104. [CrossRef]
86. Sfidari, E.; Kadkhodaie-Ilkhchi, A.; Najjari, S. Comparison of intelligent and statistical clustering approaches to predicting total organic carbon using intelligent systems. *J. Pet. Sci. Eng.* **2012**, *86–87*, 190–205. [CrossRef]
87. Hansen, K.B. The virtue of simplicity: On machine learning models in algorithmic trading. *Big Data Soc.* **2020**, *7*, 2053951720926558. [CrossRef]
88. Goz, E.; Yuceer, M.; Karadurmus, E. Total Organic Carbon Prediction with Artificial Intelligence Techniques. In *Computer Aided Chemical Engineering*; Elsevier: Amsterdam, The Netherlands, 2019; pp. 889–894.
89. Amiri Bakhtiar, H.; Telmadarreie, A.; Shayesteh, M.; Heidari Fard, M.H.; Talebi, H.; Shirband, Z. Estimating Total Organic Carbon Content and Source Rock Evaluation, Applying ΔlogR and Neural Network Methods: Ahwaz and Marun Oilfields, SW of Iran. *Pet. Sci. Technol.* **2011**, *29*, 1691–1704. [CrossRef]
90. Huang, G.B.; Zhu, Q.Y.; Siew, C.K. Extreme learning machine: A new learning scheme of feedforward neural networks. In Proceedings of the 2004 IEEE international joint conference on neural networks, Budapest, Hungary, 25–29 July 2004; Volumes 1–4, pp. 985–990.
91. Huang, G.B.; Zhou, H.M.; Ding, X.J.; Zhang, R. Extreme Learning Machine for Regression and Multiclass Classification. *IEEE Trans. Syst. Man Cybern. Part B-Cybern.* **2012**, *42*, 513–529. [CrossRef] [PubMed]
92. Peng, X.L.; Lin, P.; Zhang, T.S.; Wang, J. Extreme Learning Machine-Based Classification of ADHD Using Brain Structural MRI Data. *PLoS ONE* **2013**, *8*, e79476. [CrossRef]
93. Ahmad, I.; Basheri, M.; Iqbal, M.J.; Rahim, A. Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection. *IEEE Access* **2018**, *6*, 33789–33795. [CrossRef]
94. Zong, W.W.; Huang, G.B.; Chen, Y.Q. Weighted extreme learning machine for imbalance learning. *Neurocomputing* **2013**, *101*, 229–242. [CrossRef]
95. Tang, J.X.; Deng, C.W.; Huang, G.B. Extreme Learning Machine for Multilayer Perceptron. *IEEE Trans. Neural Networks Learn. Syst.* **2016**, *27*, 809–821. [CrossRef]
96. Lariviere, B.; Van den Poel, D. Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Syst. Appl.* **2005**, *29*, 472–484. [CrossRef]
97. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
98. Breiman, L.; Last, M.; Rice, J. Random forests: Finding quasars. In *Statistical Challenges in Astronomy*; Springer: New York, NY, USA, 2003; pp. 243–254.
99. Ziegler, A.; Konig, I.R. Mining data with random forests: Current options for real-world applications. *Wiley Interdiscip. Rev.-Data Min. Knowl. Discov.* **2014**, *4*, 55–63.
100. Scornet, E. Random Forests and Kernel Methods. *IEEE Trans. Inf. Theory* **2016**, *62*, 1485–1500. [CrossRef]
101. Shalaby, M.R.; Malik, O.A.; Lai, D.; Jumat, N.; Islam, M.A. Thermal maturity and TOC prediction using machine learning techniques: Case study from the Cretaceous-Paleocene source rock, Taranaki Basin, New Zealand. *J. Pet. Explor. Prod. Technol.* **2020**, *10*, 2175–2193.
102. Biau, G. Analysis of a Random Forests Model. *J. Mach. Learn. Res.* **2012**, *13*, 1063–1095.
103. Mutanga, O.; Adam, E.; Cho, M.A. High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. *Int. J. Appl. Earth Obs. Geoinf.* **2012**, *18*, 399–406.