processes

MDPI

Article

Prediction of Capillary Pressure Curves Based on Particle Size Using Machine Learning

Xinghua Qi [1], Yuxuan Wei [2], Shimao Wang [2], Zhuwen Wang [1,*] and Mingyu Zhou [2]

[1] College of Geo-Exploration Science and Technology, Jilin University, Changchun 130026, China; qixh20@mails.jlu.edu.cn
[2] School of Mining Engineering and Geology, Xinjiang Institute of Engineering, Ürümqi 830023, China
* Correspondence: wangzw@jlu.edu.cn

Abstract: Capillary pressure curves are usually obtained through mercury injection experiments, which are mainly used to characterize pore structures. However, mercury injection experiments have many limitations, such as operation danger, a long experiment period, and great damage to the sample. Therefore, researchers have tried to predict capillary pressure data based on NMR data, but NMR data are expensive and unstable to obtain. This study aims to accurately predict capillary pressure curves. Based on rock particle size data, various machine learning methods, such as traditional machine learning and artificial neural networks, are used to build prediction models and predict different types of capillary pressure curves, aiming at studying the best prediction algorithm. In addition, through adjusting the amount of particle size characteristic data, the best amount of particle size characteristic data is explored. The results show that three correlation coefficients of the four optimal algorithms can reach more than 0.92, and the best performance is obtained using the Levenberg–Marquardt method. The prediction performance of this algorithm is excellent, with the three correlation coefficients being all higher than 0.96 and the root mean square error being only 5.866. When partial particle size characteristics are selected, the training performance is gradually improved with an increase in the amount of feature data, but it is far less than the performance of using all the features. When the interpolation increases the particle size characteristics, the best performance is achieved when the feature data volume is 50 groups and the root mean square error is the smallest, but the Kendall correlation coefficient decreases. This study provides a new way to obtain capillary pressure data accurately.

Keywords: capillary pressure curve; rock particle size; artificial neural network; machine learning

check for updates

Citation: Qi, X.; Wei, Y.; Wang, S.; Wang, Z.; Zhou, M. Prediction of Capillary Pressure Curves Based on Particle Size Using Machine Learning. *Processes* **2024**, *12*, 2306. https://doi.org/10.3390/pr12102306

Academic Editors: Liang Xiao, Xin Nie, Mehdi Ostadhassan and Hongyan Yu

Received: 19 September 2024
Revised: 3 October 2024
Accepted: 10 October 2024
Published: 21 October 2024

1. Introduction

Capillary pressure curves are widely used to characterize reservoir pore structures. Previous studies have used conventional mercury injection experiments to obtain pore volume distribution ranges of different types of reservoirs [1]. For complex reservoirs such as tight sandstone, high-pressure mercury injection experiments have been applied [2]. Mercury injection experiments are also often combined with other data to complete reservoir quality assessments. Some scholars combine mercury injection experiments with gas adsorption to improve the accuracy of the assessments [3,4]. Subsequently, techniques such as scanning electron microscopy, nuclear magnetic resonance, and a quenched solid density function have also been applied to jointly characterize the pore characteristics of unconventional reservoirs such as tight sandstone [5,6]. In addition, some scholars also use a variety of technical means, including mercury injection experiments, to conduct fractal dimension analyses on the pore distribution of tight sandstone reservoirs to further evaluate the pore structure of the reservoirs [7,8].

Capillary pressure curves can be used to characterize the reservoir quality in an intuitive way, which can only be obtained through mercury injection experiments. However, mercury injection experiments have many limitations, including high safety risks, high costs, and the potential for contaminating the core and limiting the performance of other experiments. For this reason, some scholars have proposed a method for constructing capillary pressure curves using a T_2 distribution curve via nuclear magnetic resonance [9]. Subsequently, some scholars proposed a method of combining NMR logging data with conventional logging data to predict pseudo-capillary pressure curves based on the estimated binary porosity [10]. However, this method can accurately construct capillary pressure curves only at low mercury saturation. In order to solve this problem, researchers have analyzed the possible causes of this situation in detail and made corresponding improvements [11]. Among them, one scholar proposed a segmented power function method, which has been widely used to predict pseudo-capillary pressure curves using the inverse accumulation curve via nuclear magnetic resonance [12,13]. In addition, in order to improve the prediction accuracy of complex reservoirs, researchers have also introduced a segmented multi-parameter power function method [14]. In short, previous studies have used the actual physical correlation between the two characteristics and used mathematical methods to predict capillary pressure curves. In recent years, artificial intelligence methods have also been applied to the prediction of capillary pressure curves [15].

Artificial intelligence methods, including a variety of machine learning models and deep learning models, have been widely used in the fields of lithology identification and data prediction. Based on logging data, previous studies have used a variety of machine learning methods, such as support vector machines (SVMs) and decision trees (DTs), to effectively identify igneous rocks [16,17]. Subsequently, for a variety of lithology applications, researchers have adopted a variety of neural networks and other methods to conduct research [18,19]. In the field of data prediction, previous studies have used a variety of machine learning and neural network methods to effectively predict shear waves [20,21]. In the prediction of pore structure, many scholars have adopted more complex deep learning methods, including deep and shallow neural networks and time series [22,23].

It is worth noting that there are many previous methods for constructing capillary pressure curves, including segmented power functions, segmented multi-parameter power functions, and artificial intelligence techniques, but they are all based on NMR data obtained from core analysis. However, NMR itself is costly and susceptible to external interference, and there may be insufficient data in the actual production process, which hinders the goal of using NMR data to construct capillary pressure curves. Therefore, the research focus of this paper turns to using rock particle size data.

The particle size data of rocks are often used to quantitatively describe the physical properties of rocks. In order to further study the influence of particle size on the energy evolution of coal mine waste rock, researchers designed a special compaction device [24]. Subsequently, they developed a bidirectional loading experimental system to systematically explore the relationship between particle size and porosity [25]. In the field of rock mechanics, many scholars have used the discrete element method to conduct numerical simulation research on particles [26–28]. To solve the problem of crack expansion caused by particle size, researchers introduced CT scanning and microscopic analysis [29]. Recently, some scholars found that there is a correlation between the particle size distribution curve and T_2 distribution via nuclear magnetic resonance, and the conversion coefficient between them was obtained using a nonlinear fitting method [30].

The above studies give a clear indication that there is a strong correlation between particle size data and NMR data, and NMR data can be applied to the construction of capillary pressure curves. This further reflects a certain correlation between particle size data and capillary pressure curves (i.e., mercury injection data). Therefore, this paper proposes a new method for capillary pressure curve prediction based on rock particle size data and using a variety of artificial intelligence techniques, including machine learning.

2. Overview of the Study Area

The particle size and mercury injection data in this study were selected from the Zhenwu–Caozhuang area of the Paleogene Danan Formation in the Gaoyou Depression (Figure 1). The Gaoyou Depression, located in the central region of the Dongtai Depression within the Subei Basin, covers an area of 2130 km^2. It exhibits a roughly rhomboidal shape, with its major axis oriented northeast. The depression forms a loop, characterized by steep slopes in the south part and gentler slopes in the north.

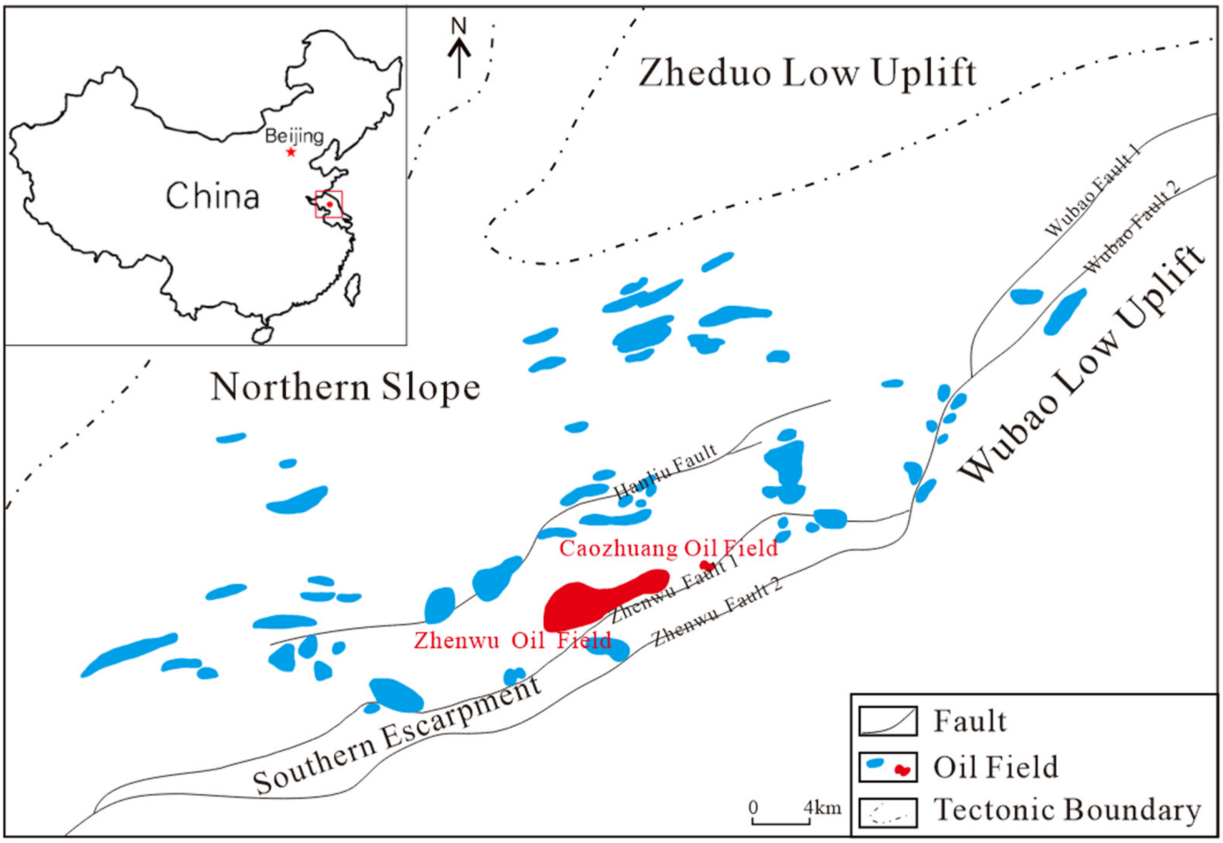


Figure 1. Structural zoning map of the Gaoyou Depression. The red point in the upper left figure represents the general position of the Gaoyou Depression, and the red area in the main figure is the study area.

The lithology primarily consists of siltstone, constituting 40–70% of the total, with pebbly siltstone and fine sandstone as the secondary components. The lithological section is characterized by a tectonic rhythm of interbedded dark gray and dark brown mudstones. The rock exhibits relatively good sorting, with a median particle size averaging 0.12 mm. The primary bedding types include graded bedding, massive bedding, and parallel bedding. The bases of some bedding sequences contain bands of gravel and mudstone, with intersecting surfaces exhibiting scour marks with a thickness in the 10–30 cm range. Some massive beds contain a small amount of fine gravel. Parallel bedding within siltstone shows variations in grain size and the orientation of mica sheets or carbon chips, with a typical thickness of about 10 cm. Core porosity ranges from 0.9% to 31.5%, with an average porosity of 18.47% and a concentrated distribution of 12–26.8%. The average air permeability is 844 × 10^{-3} μm^2, with a wide distribution range. Some cores in the study area are tight and characterized by low permeability and tightness, but the overall physical properties are relatively favorable, as shown in Figure 2. Typical cores were deposited by underwater distributary channels at the front of a fan delta (Figure 3).

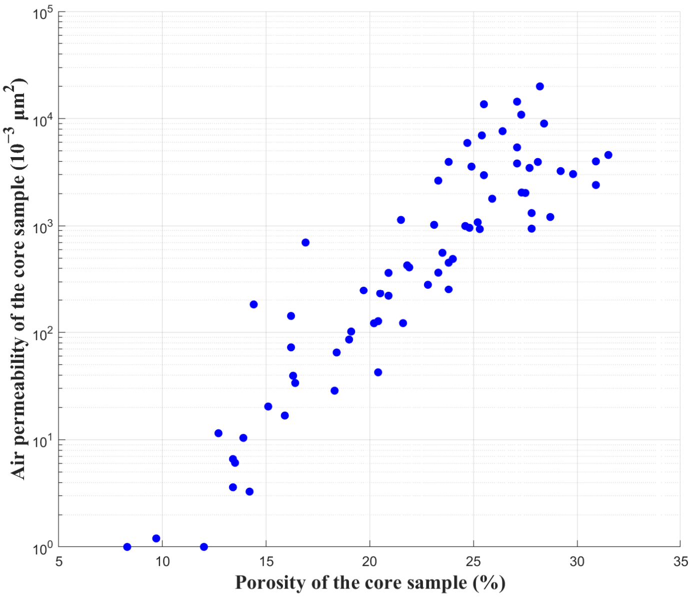
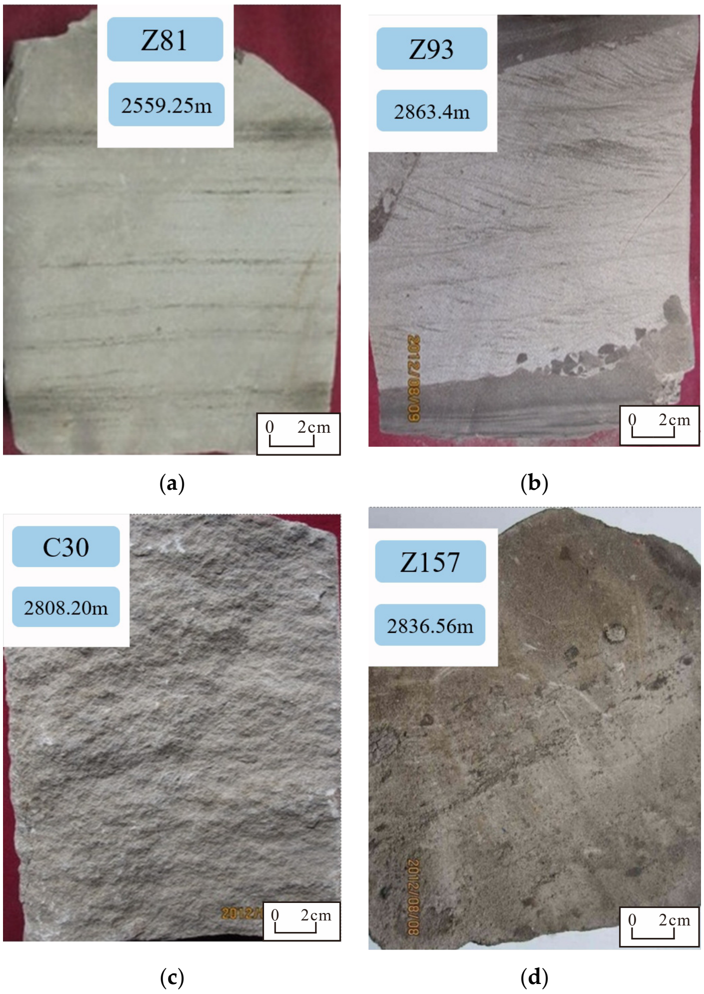


Figure 2. Scatter diagram showing the relationship between the porosity and air permeability of core samples.



(**a**) (**b**) (**c**) (**d**)

Figure 3. Typical cores deposited by underwater distributary channels in front of a fan delta: (**a**) parallel bedding; (**b**) cross-bedding; (**c**) channel scouring surface; (**d**) massive bedding.

3. Research Methods

3.1. Research Basis

Previous studies suggest a correlation between particle size and capillary pressure. This link is verified by the statistical analysis of actual data (Figure 4). This finding indicates that this relationship is appropriate for applying artificial intelligence methods for data fitting.

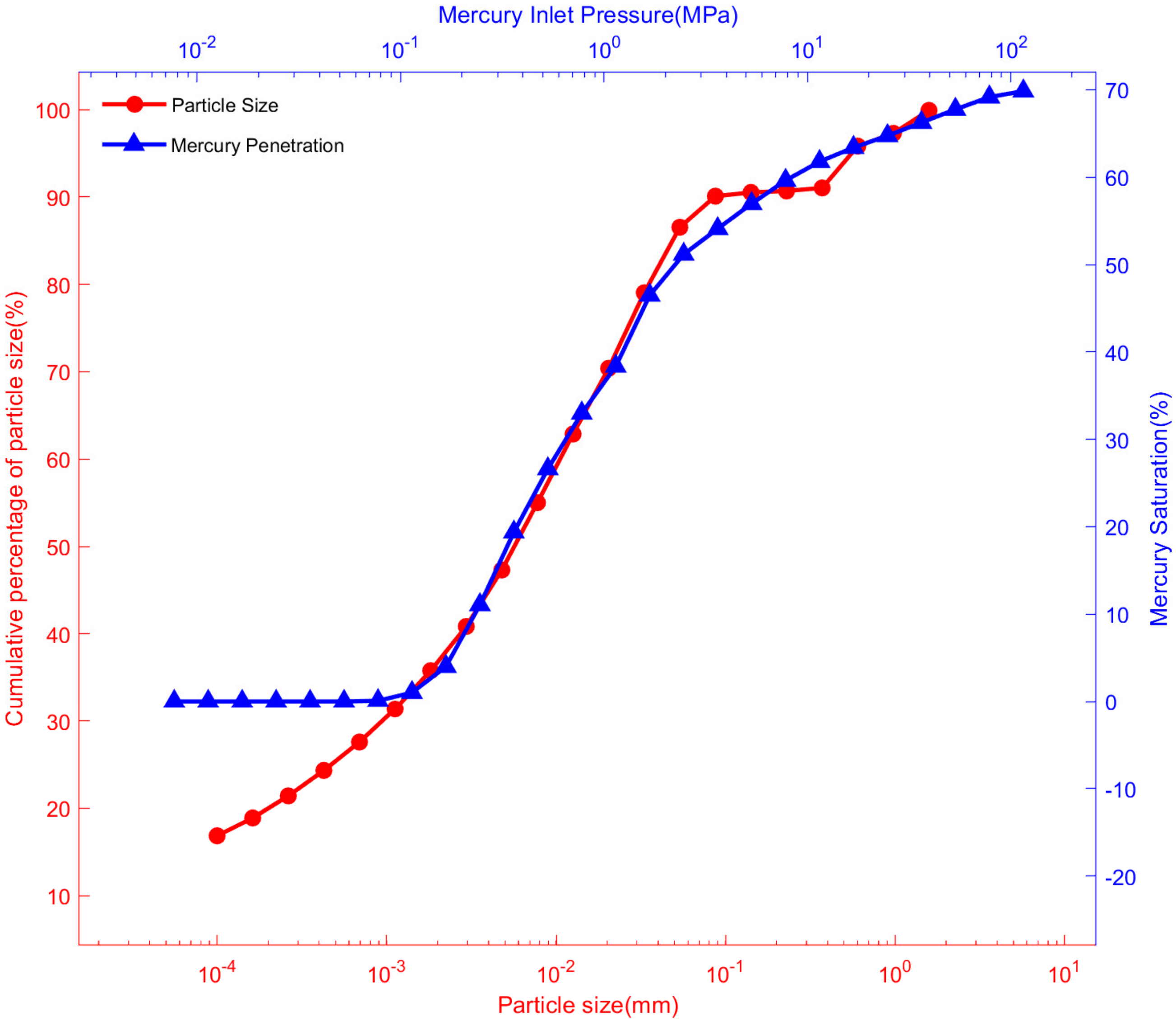


Figure 4. Fitting diagram illustrating the link between the particle size accumulation curve and the capillary pressure curve.

3.2. Research Ideas

In this study, 74 core samples from the study area were selected, of which 70 were used for training, and the remaining 4 were used for prediction. The four predicted samples exhibited different types of capillary pressure curves. The grain size accumulation curve of the core samples was measured using sieve analysis. This method uses standard screens arranged in descending order of pore size to classify the core samples [31]. By weighing the mass of each grain size fraction, the cumulative distribution of the sample was calculated. The capillary pressure curves of the core samples were measured using the mercury injection experiment, which involves injecting mercury into the core sample at varying injection pressures [32]. As the pressure gradually increases, mercury infiltrates pores of varying sizes to generate the capillary pressure curve. For this analysis, cumulative particle size data comprising 21 sets of features were used as input, whereas the capillary pressure data constituting 26 sets of features were used as output for regression learning. Traditional machine learning and artificial neural network methods implemented using MATLAB 2021b were employed to construct the mapping relationship between the particle size accumulation curve and the capillary pressure curve (Figure 5).

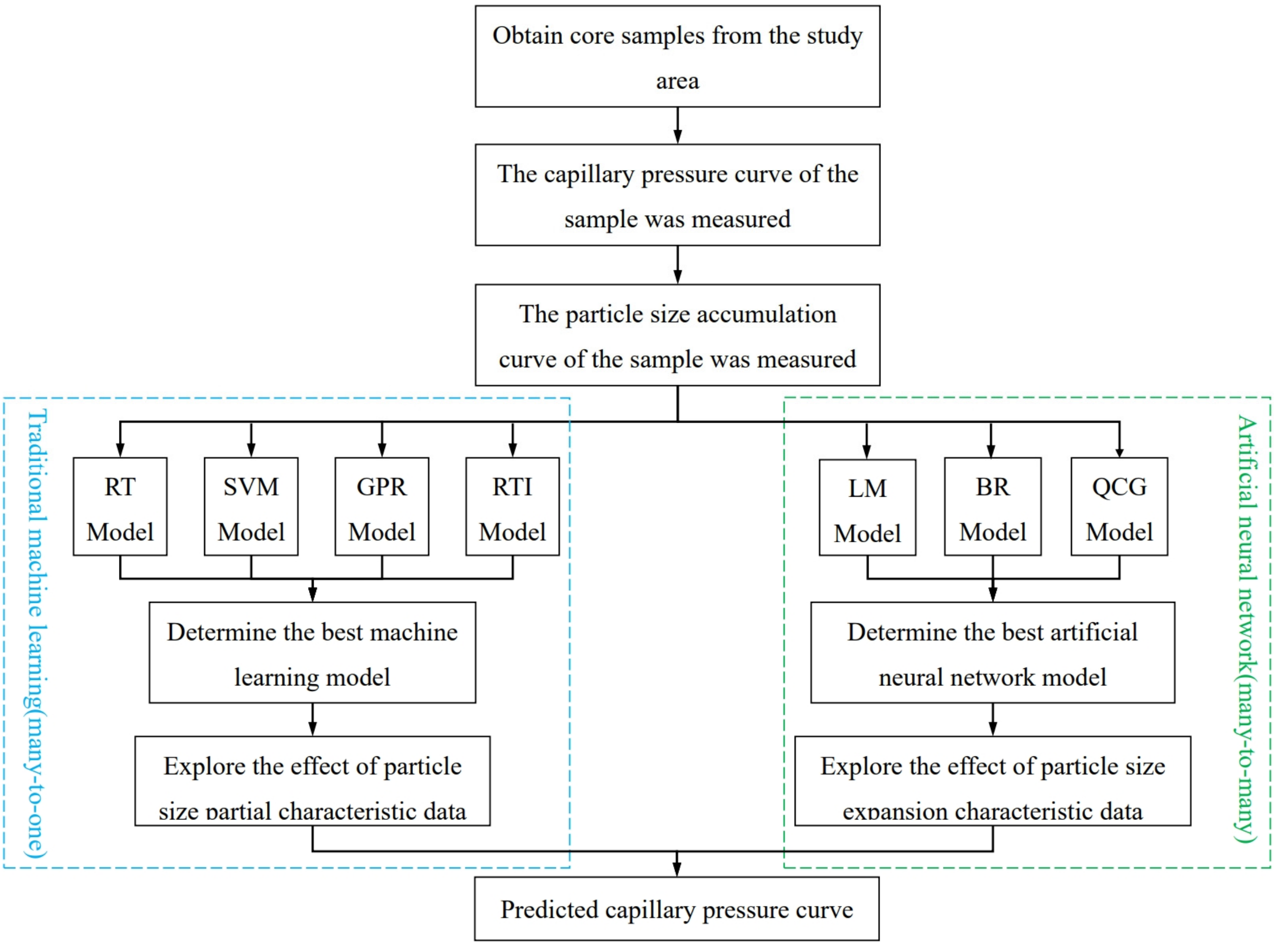

Obtain core samples from the study area
The capillary pressure curve of the sample was measured
The particle size accumulation curve of the sample was measured
Traditional machine learning(many-to-one)
RT Model
SVM Model
GPR Model
RTI Model
Determine the best machine learning model
Explore the effect of particle size partial characteristic data
LM Model
BR Model
QCG Model
Determine the best artificial neural network model
Explore the effect of particle size expansion characteristic data
Artificial neural network(many-to-many)
Predicted capillary pressure curve

Figure 5. Technology roadmap.

In traditional machine learning, models such as regression trees (RTs), support vector machines (SVMs), Gaussian process regression (GPR), and regression tree integration (RTI) are used to train the many-to-one regression learning strategy. The best algorithm is determined based on the training performance, which is then used for the corresponding prediction analysis. For artificial neural networks, the dataset is divided into training sets (70% of the total data), validation sets (15%), and test sets (15%). Training is conducted using the Levenberg–Marquardt (LM) method, Bayesian regularization (BR) method, and quantized conjugate gradient (QCG) method through a many-to-many regression learning strategy. In the many-to-one strategy, 21 particle size characteristics are used as inputs, and 26 capillary pressure characteristics are used as outputs. In each regression study, multiple particle size characteristics correspond to capillary pressure characteristics on a one-to-one basis, resulting in the establishment of 26 independent models to complete the prediction task. The many-to-many strategy involves the simultaneous selection of multiple particle size characteristics and multiple capillary pressure characteristics for regression learning, allowing for the development of a comprehensive model to complete the prediction task.

In addition, within the scope of machine learning, the prediction performance of regression learning using partial particle size data features was analyzed based on the model considered best. Meanwhile, particle size data in artificial neural networks were interpolated and expanded to determine the influence of data expansion on the prediction performance. Finally, through comparative analysis, the most effective artificial intelligence method was identified and the optimal particle size data characteristics were determined to accurately predict the capillary pressure curve.

4. Theoretical Basis

4.1. PCHIP Interpolation Method

Piecewise cubic Hermite interpolation polynomial (PCHIP) [33,34] is expressed as follows:

$$H_3(x) = \left[\left(1+2\frac{x-x_0}{x_1-x_0}\right)y_0 + (x-x_0){y'}_0\right]\left(\frac{x-x_1}{x_0-x_1}\right)^2 + \left[\left(1+2\frac{x-x_1}{x_0-x_1}\right)y_1 + (x-x_1){y'}_1\right]\left(\frac{x-x_0}{x_1-x_0}\right)^2 \tag{1}$$

where x_0 and x_1 denote the measurement points before and after the points to be interpolated; y_0 and y_1 correspond to the measurements before and after the points to be interpolated; and ${y'}_0$ and ${y'}_1$ denote the first derivative values calculated from known data.

4.2. Principles of Machine Learning

4.2.1. Regression Tree

The Classification And Regression Tree (CART) is a machine learning model that uses a tree data structure to illustrate decision rules [35,36]. A regression tree (RT), a specific type of CART decision tree, mainly functions to predict unknown data by using the structural characteristics of the data. The calculation process involves the recursive construction of a binary tree.

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \tag{2}$$

First, the optimal segmentation variable j and the segmentation point s are determined to obtain the minimum error (j, s).

$$min_{j,s}\left[min_{c_1}\sum_{x_i \in R_1(j,s)}(y_i - c_1)^2 + min_{c_2}\sum_{x_i \in R_2(j,s)}(y_i - c_2)^2\right] \tag{3}$$

The sample features are then divided into two nodes based on the optimal (j, s), and the aforementioned steps are repeated until the termination condition is satisfied. The input space is ultimately divided into m regions to generate a decision tree.

$$c_m = average(y_i | x_i \in R_m(j,s)) \tag{4}$$

$$f(x) = \sum_{m=1}^{M} c_m I, X \in R_m \tag{5}$$

4.2.2. Support Vector Machine

Support vector machines (SVMs) are supervised learning algorithms designed to find the optimal hyperplane that divides different classes of data points. For nonlinear problems, kernel functions can be introduced to perform inner product operations in high-dimensional spaces [37]. In the current study, both the cubic kernel function and the Gaussian kernel function in the polynomial kernel function are used, which are expressed as follows:

$$K(x_i, x_j) = \left(x_i^T x_j + c\right)^d \tag{6}$$

where d represents the order of the polynomial kernel.

$$K(x_i, x_j) = exp(-\frac{|x_i - x_j|^2}{2\sigma^2}) \tag{7}$$

where $K(x_i, x_j)$ is the inner product kernel, where x_i and x_j are two different samples; σ is a free parameter.

4.2.3. Gaussian Process Regression

Gaussian process regression (GPR) is a probability-based machine learning algorithm that implements the prediction function by modeling the data as a series of Gaussian processes. For an input sample set $D = \{x_i, y_i, i = 1, 2 \ldots, n\}$, where i is the sample sequence, GPR aims to determine the relationship f between the independent variables x and y to complete the prediction. Let f obey the GPR $F \sim GPR(m, k)$, where m is the mean value function, and k is the covariance function. The following formula is thus given:

$$y = f(x) + \varepsilon(1) \tag{8}$$

$$\varepsilon = N\left(0, \delta_n^2\right) \tag{9}$$

where ε is random noise and variance is δ_n^2 Gaussian distribution.

The prior distribution of y is expressed as

$$y \sim N\left(0, K + \delta_n^2 I\right) \tag{10}$$

where $K = K(X, X)$ is a symmetric positive definite covariance matrix of order $n \times n$.

4.2.4. Regression Tree Integration

Regression tree integration (RTI) involves combining multiple independent regression trees into a single model. The LSBoost algorithm, an ensemble learning method, integrates multiple weak models to build more powerful prediction models [38,39]. The current study uses the boosting tree model based on the LSBoost algorithm. Its estimation and regression are expressed as follows:

$$\hat{f}_0(t) = argmin \sum_{j=1}^{n} (y_j - \gamma)^2 \tag{11}$$

where $\hat{f}_0(t)$ represents the base model—that is, the initial model in the set. For predicting the target value y, a constant γ is chosen to minimize the square error sum.

4.3. Principles of Artificial Neural Network

Neural network fitting refers to the application of an artificial neural network (ANN) to fit a function. A neural network is a machine learning model that simulates a neural network of the human brain, consisting of multiple neurons that fit a given function by adjusting parameters such as weights and biases. In neural network fitting, the back-propagation algorithm is typically used to optimize the parameters of the neural network, thus minimizing the error between the predicted value and the actual value. Gradually increasing the number of hidden layers and neurons improves the fitting ability of the neural network.

The artificial neural network model employed in this study is essentially a deep learning-based machine learning model with few layers. The network structure of the model comprises an input layer, a hidden layer, and an output layer (Figure 6). A sigmoid transfer function is used in the hidden layer, whereas a linear transfer function is used in the output layer. The number of neurons in the hidden layer is determined by the size of the layer.

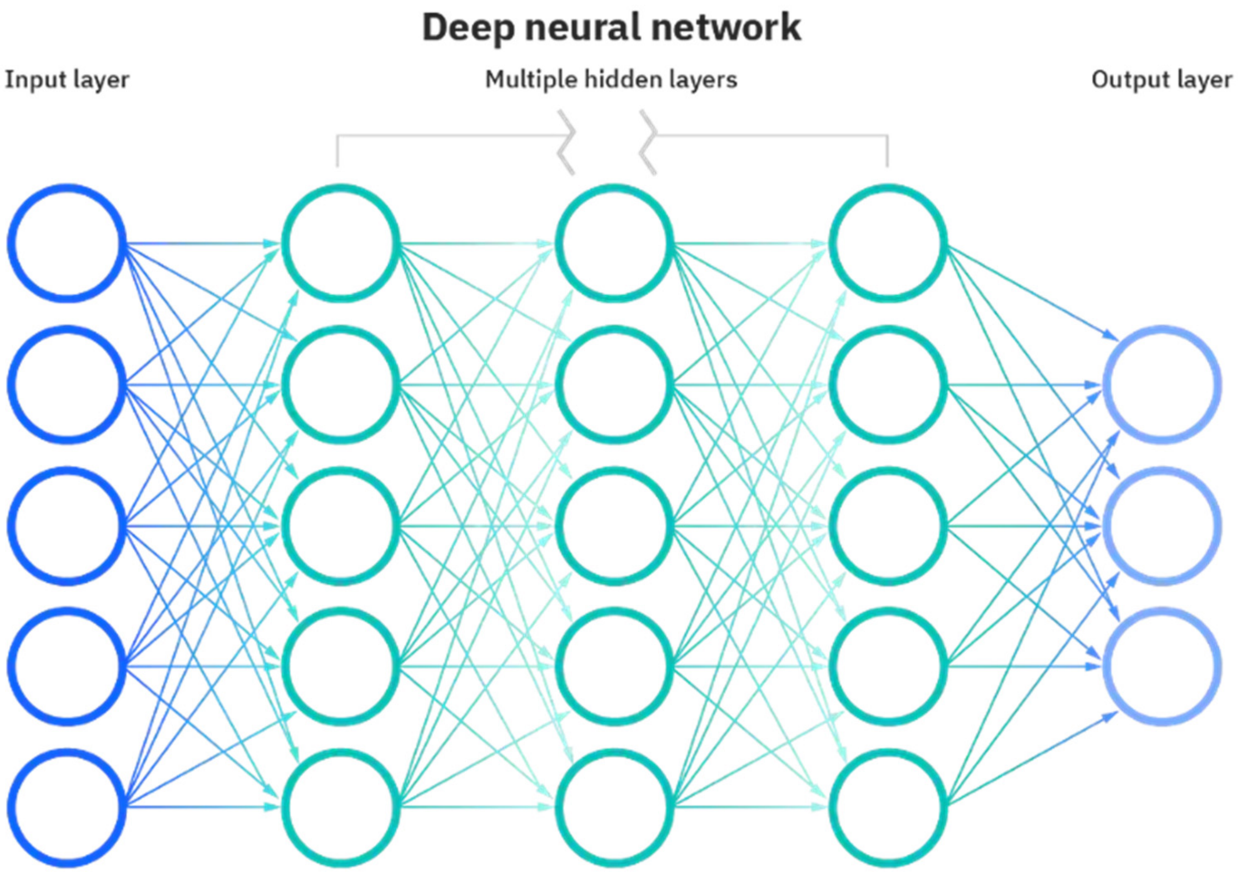


Figure 6. Neural network structure diagram.

4.4. Evaluation Indicators

4.4.1. Root Mean Square Error

The root mean square error (*RMSE*) is a statistical index used to measure the deviation between observed and actual values. In this study, *RMSE* was mainly used to evaluate the training performance and the degree of deviation between the predicted and actual values [40–42], calculated as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2} \tag{12}$$

where x_i represents the input value, $\hat{x}_i$ denotes the predicted value, and n is the number of training data.

4.4.2. Correlation Coefficient

Correlation coefficients are statistical indicators that quantify the degree of linear correlation between variables, typically applied to numerical data. In this study, correlation coefficients are primarily used to evaluate the degree of fit of the model, thereby assessing the training performance and predictive accuracy of the model. Numerous methods for calculating the correlation coefficient have been developed, with the Pearson correlation coefficient being the most widely used [42,43] and given as follows:

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - (E(X))^2}\sqrt{E(Y^2) - (E(Y))^2}} \tag{13}$$

Equation (8) indicates that the Pearson correlation coefficient ranges between [−1, 1]. A correlation coefficient closer to 1 or −1 suggests a stronger correlation, whereas for values nearer and vice versa, the correlation is weaker.

In addition to the Pearson correlation coefficient, the determination coefficient [41], Kendall correlation coefficient [44], and Spearman correlation coefficient [45,46] are used in the current study. The formulas for calculating the three are presented below:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}{\sum_{i=1}^{n}(x_i - \overline{x}_i)^2} \tag{14}$$

where R^2 represents the coefficient of determination, x_i denotes the input value, $\overline{x}_i$ indicates the average value, n is the number of training data points, $\sum_{i=1}^{n}(x_i - \hat{x}_i)^2$ denotes the sum of squared differences between the actual and predicted values, and $\sum_{i=1}^{n}(x_i - \overline{x}_i)^2$ represents the sum of squares of the actual values and the mean.

$$\tau = \frac{2}{n(n-1)}\sum_{i<j} sgn(X_i - X_j)sgn(Y_i - Y_j) \tag{15}$$

where τ denotes the Kendall correlation coefficient, n is the sample size, and $sgn(X_i - X_j)$ and $sgn(Y_i - Y_j)$ represent $(X_i - X_j)$ and $(Y_i - Y_j)$, respectively. If $(X_i - X_j) > 0$, then $sgn(X_i - X_j) = 1$; otherwise, $sgn(X_i - X_j) = -1$. The same applies to Y.

$$\gamma_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{16}$$

where γ_s represents the Spearman correlation coefficient, $d_i = X_i - Y_i$ denotes the difference between two variables, and n is the number of samples.

Similar to Pearson's correlation coefficient, both the Kendall and Spearman correlation coefficients have values in the [−1, 1] range, with the correlation criteria remaining consistent. However, the determination coefficient falls within the [0, 1] range. A result closer to 1 indicates a better fit, whereas that closer to 0 suggests a poorer fit. The Pearson correlation coefficient applies to variables with a non-zero standard deviation and is typically used for variables with linear relationships or normal distributions. By contrast, both the Kendall and Spearman correlation coefficients can reflect the linear and nonlinear relationships of the variables that do not follow a normal distribution, rendering them suitable for a wider range of applications [47].

5. Research Results

5.1. Machine Learning Method

5.1.1. Machine Learning Model Fitting

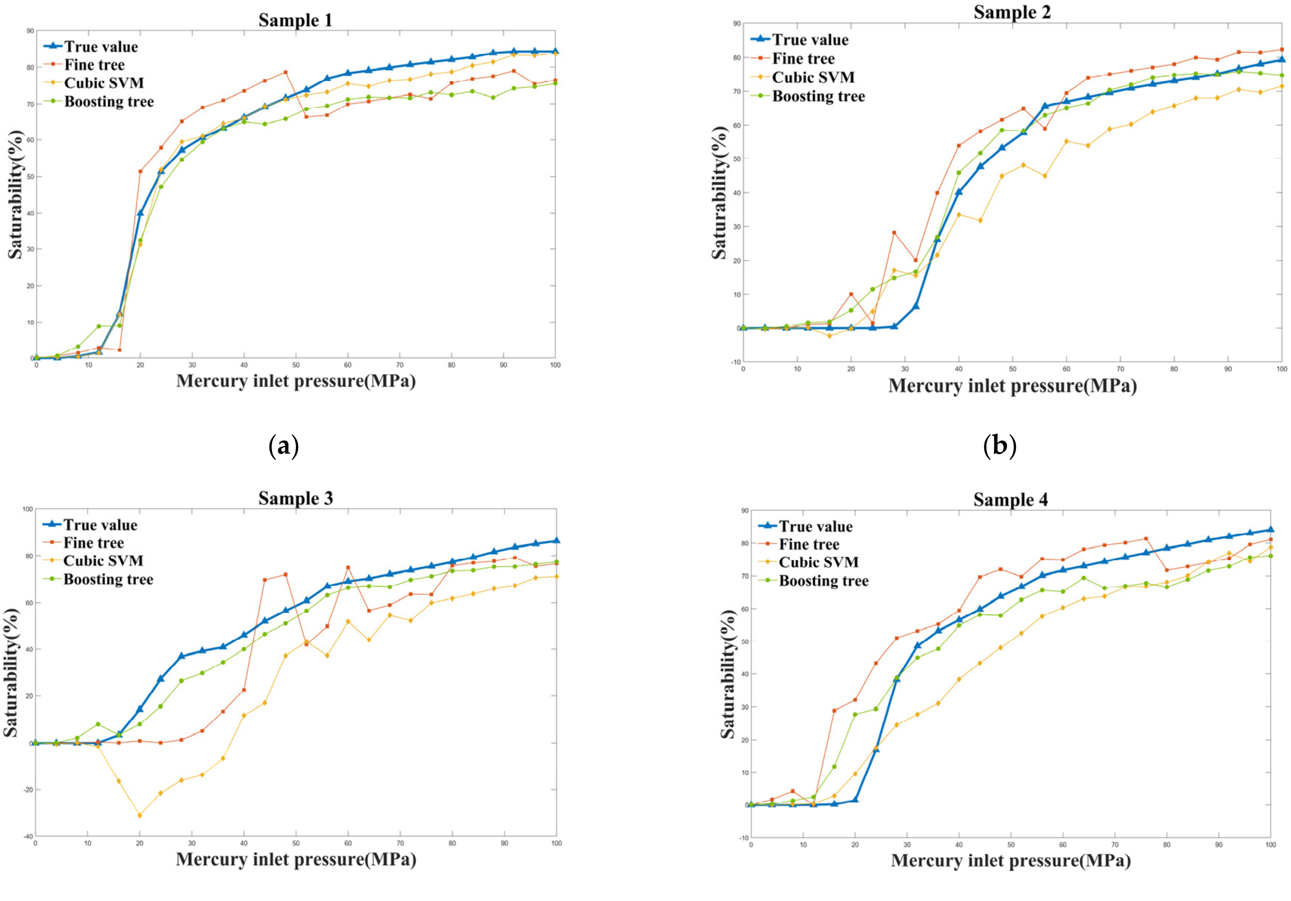
To determine the most suitable machine learning model, this study adopted the resubstitution verification method based on all the particle size characteristic data. The RT, SVM, GPR, and RTI machine learning models were employed to perform regression learning on 26 groups of capillary pressure characteristic data individually.

The results presented in Table 1 indicate that the optimal algorithm for each model is selected using the criteria that the *RMSE* is less than 9, and the coefficient of determination (R^2) is greater than 0.5. Among the algorithms, the fine tree algorithm demonstrates the best performance within the regression tree model. Meanwhile, the cubic SVM is identified as the most optimal algorithm in the SVM model. The training performance of the GPR model indicates an overall deviation, with the optimal model identified as the quadratic rational GPR. In the regression tree integration model, the boosting tree algorithm is considered optimal. Among the evaluated algorithms, the boosting tree algorithm exhibits the best training fitting performance, achieving a maximum R^2 of 0.579 and a minimum *RMSE* of 8.291. The cubic SVM follows, with an R^2 of 0.541. The fine tree ranks third, with an R^2

of 0.521. The other machine learning models generally yield poor fitting results, with R^2 below 0.5. Through statistical analysis, three algorithms with the best training performance are identified: the fine tree, cubic SVM, and boosting tree. Subsequently, a prediction study was conducted using these three algorithms (Figure 7).

Table 1. Training performance of each machine learning model.

Model	**Regression Tree Model**			**Support Vector Machine Model**		**Gaussian Process Regression Model**		**Regression Tree Integration Model**	
Algorithm	Rough tree	Medium tree	Fine tree	Fine Gaussian SVM	Cubic SVM	Quadratic rational GPR	Square index GPR	Bagging tree	Boosting tree
RMSE	13.452	11.681	8.786	9.089	8.4399	11.326	12.896	10.543	8.291
R^2	0	0.217	0.521	0.462	0.541	0.238	0.052	0.352	0.579



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

(**c**) (**d**)

Figure 7. Truth–prediction comparison graph: prediction performance of three machine learning algorithms on (**a**) Sample 1, (**b**) Sample 2, (**c**) Sample 3, and (**d**) Sample 4.

The boosting tree algorithm performs well across the four prediction samples, and the fine tree and cubic support vector machine fail to maintain stable prediction performance with different types of capillary pressure curves, resulting in large fluctuations (Figure 7).

5.1.2. Fitting of Partial Particle Size Characteristics

In this section, regression learning training was conducted based on the boosting tree model. A part of particle size data from 21 groups of features was selected to perform fitting training with capillary pressure data from 26 groups of features. This approach aimed to evaluate the influence of some feature data on the fitting performance of the model. Figure 8 presents the regression learning process with a three-to-one example: the first, second, and third particle size features correspond to the first capillary pressure feature, arranged from left to right, continuing until the ninth capillary pressure feature. The 19th, 20th, and 21st particle size features correspond to the 26th capillary pressure feature, arranged from right to left, continuing until the 18th capillary pressure feature. The median of the particle size characteristic data was determined to be 11, and the 10th, 11th, and 12th particle size characteristics were subsequently selected to correspond to the remaining capillary pressure characteristics (denoted by the red dotted frame in Figure 8). This selection was made because the particle size accumulation curve and capillary pressure curve exhibited greater consistency in the middle region.

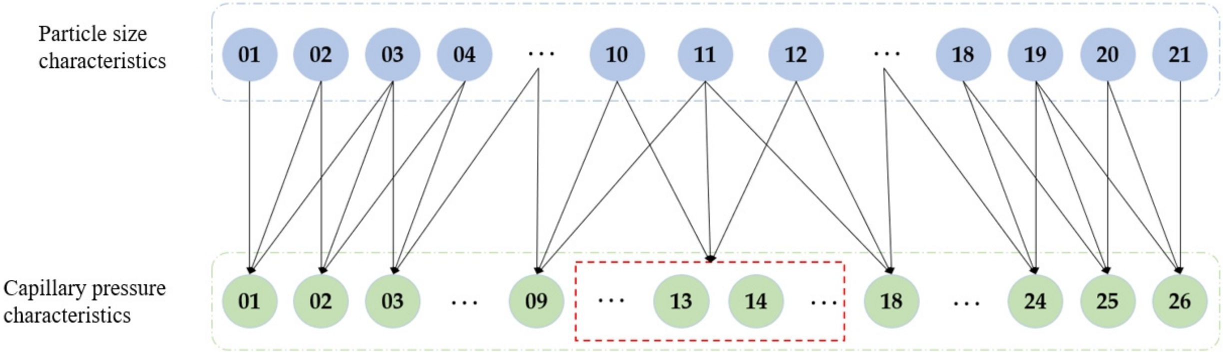

Particle size characteristics
01
02
03
04
···
10
11
12
···
18
19
20
21
Capillary pressure characteristics
01
02
03
···
09
···
13
14
···
18
···
24
25
26

Figure 8. Three-to-one regression learning diagram.

As the number of features increases, the *RMSE* value gradually decreases, whereas the R^2 value gradually rises, indicating an improvement in the training performance, as shown in Table 2. However, when 19 groups of feature data are used for regression learning, the R^2 value is 0.556, which is lower than that of all the feature data for regression learning. At the same time, the *RMSE* value of 19 groups of feature data is lower than that of all the feature data.

Table 2. Training performance of partial particle size characteristics.

Feature Quantity	3-1	5-1	7-1	15-1	17-1	19-1
RMSE	10.814	10.297	9.895	8.756	8.615	8.503
R^2	0.340	0.390	0.425	0.528	0.548	0.556

5.2. Artificial Neural Network Method

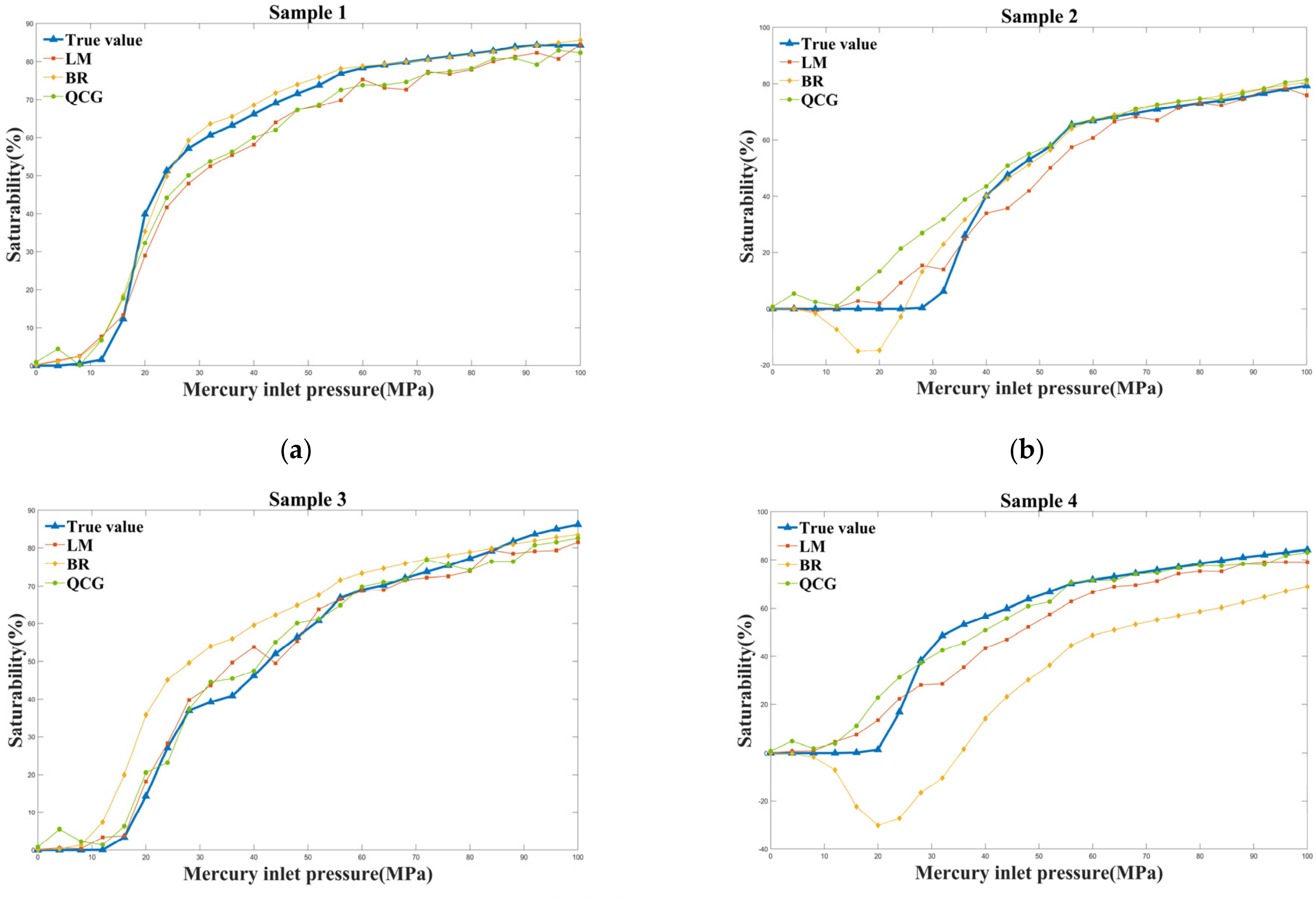
5.2.1. Artificial Neural Network Fitting

This section shows the use of all the particle size characteristic data and three artificial neural network algorithms for conducting fitting learning on 26 groups of capillary pressure characteristic data. The aim is to assess the effects of different algorithms and the number of hidden layer neurons on training fitting and prediction performance to determine the optimal artificial neural network model. Through systematic experiments, the number of hidden layer neurons is adjusted step by step, and the optimal number of hidden layer neurons for achieving the best training performance for each algorithm is determined. The LM, BR, and QCG methods performed best with 15, 10, and 15 hidden layer neurons, respectively (Table 3).

Table 3. Training performance of three kinds of artificial neural networks.

Algorithm	LM Algorithm	BR Algorithm	QCG Algorithm
Training	0.93754	0.99247	0.89952
Verify	0.88275	——	0.85773
Test	0.82719	0.75711	0.89228
All	0.91111	0.94887	0.89110

Overall, the highest fitting R-value of 0.948 was obtained using the BR method, followed by 0.911 obtained using the LM method, and 0.891 obtained using the QCG method (Figure 9). Specific analysis indicates that when the number of hidden layer neurons was 10, the BR method demonstrated the best-fitting performance; however, the test set for this method exhibited greater volatility during training. The LM method performed best when the number of hidden layer neurons reached 15. With the same number of neurons, the QCG method also performed well; however, increasing the number of neurons did not significantly influence its fitting performance. In addition, the training time for both the LM and QCG methods was maintained at about 5 s, which was exceeded by the BR method. The training time was gradually extended as the number of hidden layer neurons increased. When the number of hidden layer neurons reached 30, the training time could extend to approximately 300 s.



(**a**)

(**b**)

(**c**)

(**d**)

Figure 9. Truth–prediction comparison graph: prediction performance of three artificial neural network algorithms on (**a**) Sample 1, (**b**) Sample 2, (**c**) Sample 3, and (**d**) Sample 4.

The prediction performance of the LM method is slightly superior to that of the QCG method. Meanwhile, the BR method is significantly inferior to the other two methods (Figure 9). Specifically, the LM method yielded better prediction results across different types of capillary pressure curves, highlighting its robust generalization performance.

5.2.2. Interpolation Particle Size Characteristic Fitting

Given the significant superiority of the LM method to the machine learning algorithm and other artificial neural network algorithms in prediction, this section selects the LM method with 15 hidden layer neurons as the basis for the model. PCHIP interpolation was employed to interpolate and expand the particle size features from 21 groups (the original number) to 30, 40, 50, 60, and 70 groups (Figure 10).

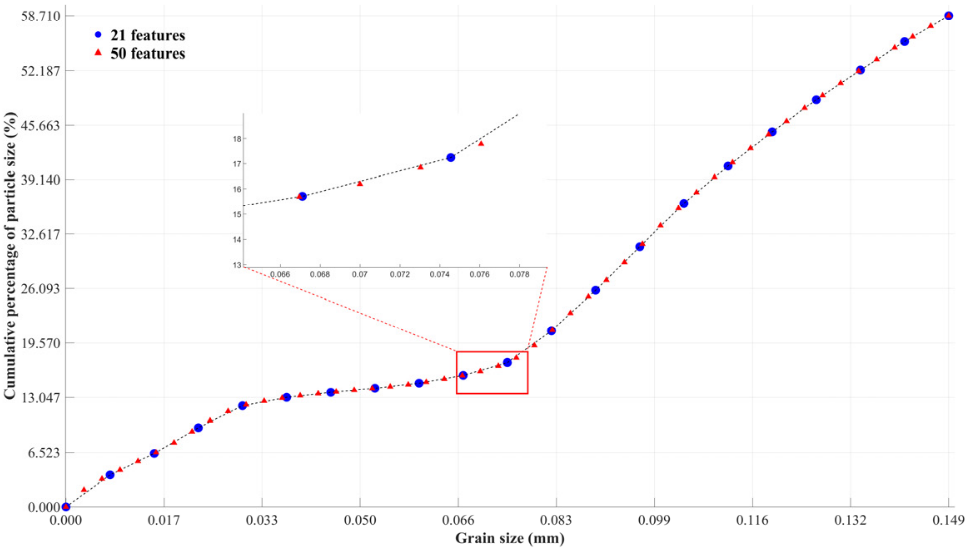


Figure 10. Interpolation diagram of 21 particle size characteristic groups extended to 50 groups.

This section evaluates the R^2 value for each feature count during training and the average Pearson correlation coefficient during prediction to indicate the optimal number of features. The aim is to further investigate the effect of the number of particle size features on training and prediction results. As shown in Figure 11, as the number of feature data rises, the training performance decreases in volatility, Meanwhile, the prediction performance is relatively optimal with 50 groups of features. Accordingly, this section presents the prediction results using 50 feature groups across four prediction samples, as shown in Figure 12.

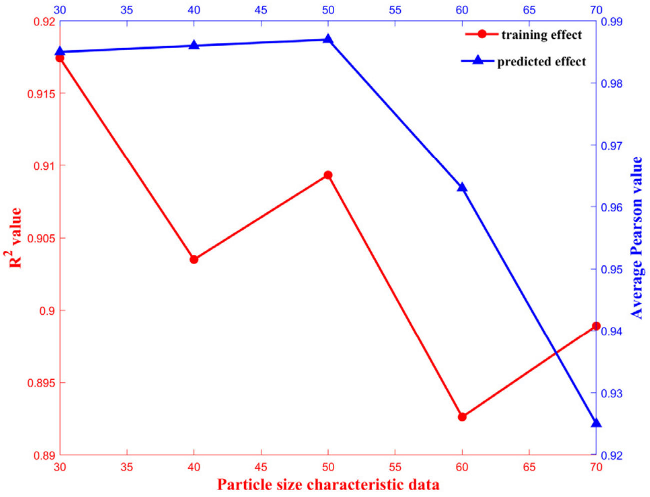


Figure 11. Training–prediction visualizations based on the number of features.

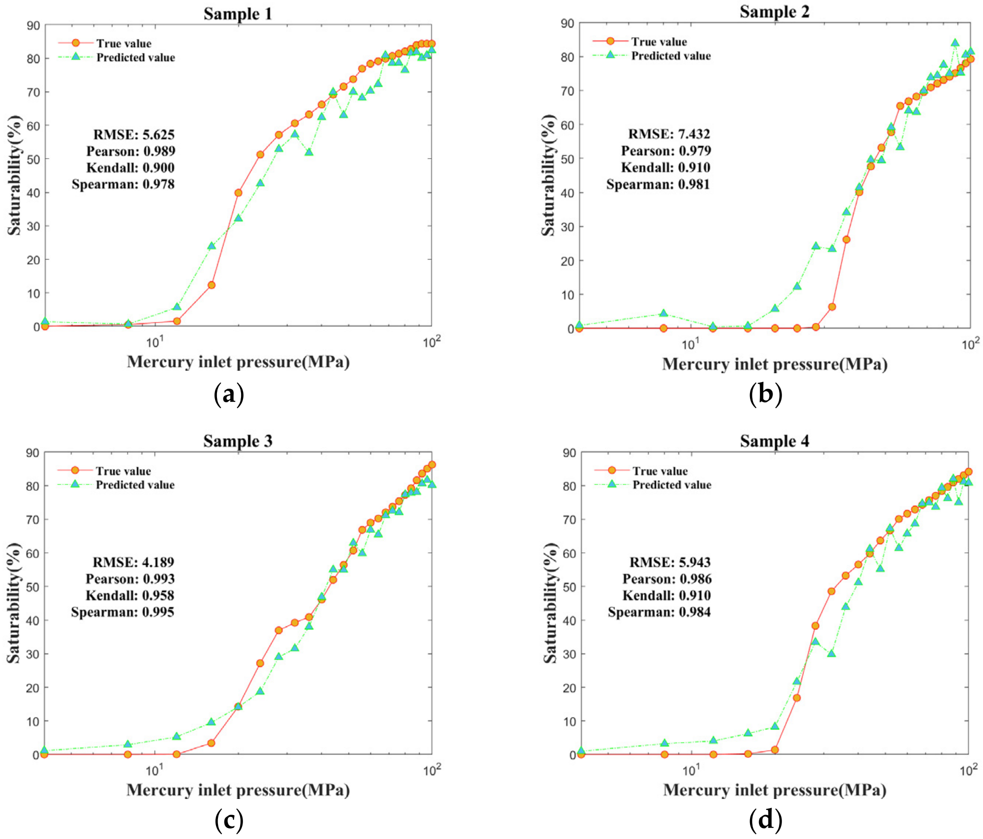


Figure 12. Truth–prediction comparison: prediction performance of 50 groups of particle size characteristics on (**a**) Sample 1, (**b**) Sample 2, (**c**) Sample 3, and (**d**) Sample 4.

6. Discussion

In this study, two artificial intelligence models—traditional machine learning and ANN—were used to construct a prediction model for the capillary pressure curve. The best algorithm for each method was determined based on the research findings. In traditional machine learning, the boosting tree was identified as the best algorithm. Regression learning was then performed using partial particle size characteristic data based on the aforementioned algorithm. Among the ANN methods, the LM method demonstrated the best performance. This algorithm was thus used as the basis for performing the regression learning of the interpolated expanded particle size characteristic data. The four evaluation indicators of these four algorithms across the four prediction samples were statistically analyzed. The average and cumulative values of each indicator were then calculated (Table 4 and Figure 13).

Table 4. Average values of each evaluation index for four prediction samples evaluated using four algorithms.

Model	Traditional Machine Learning		Artificial Neural Network	
Algorithm	BT (21-1)	BT (19-1)	LM (21-26)	LM (50-26)
Average *RMSE*	6.645	8.128	5.866	5.797
Average Pearson	0.990	0.983	0.987	0.987
Average Kendall	0.944	0.931	0.960	0.920
Average Spearman	0.986	0.984	0.993	0.985

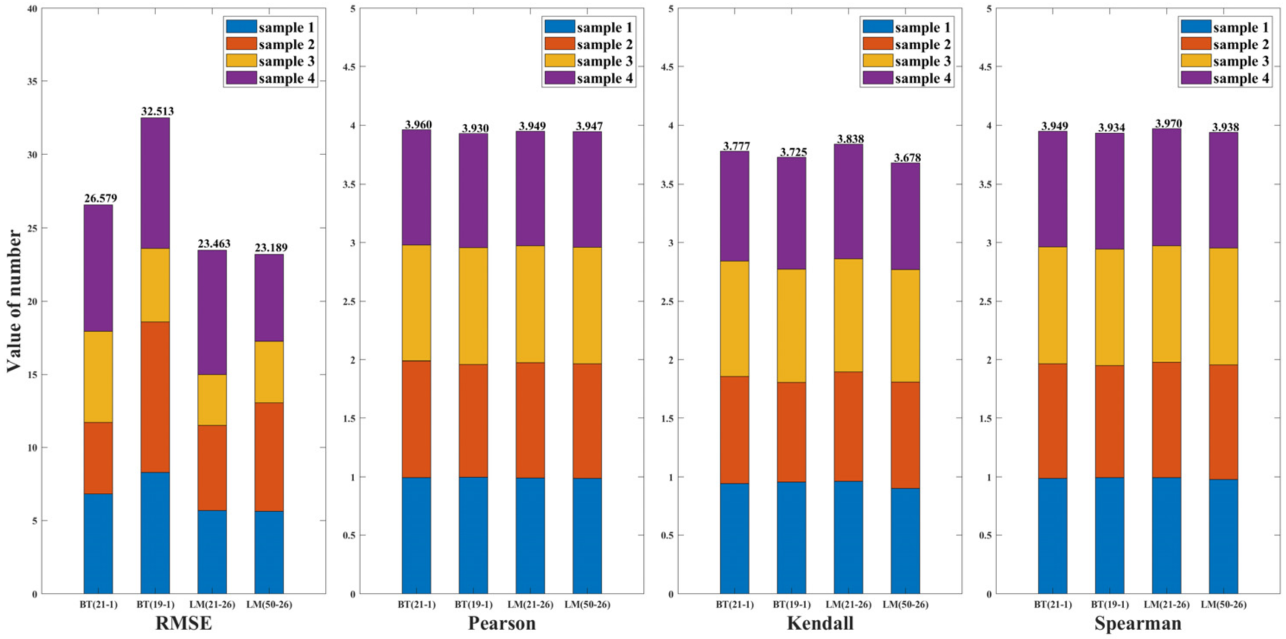


Figure 13. Stacked histogram of each evaluation index for four prediction samples evaluated using four algorithms.

Although the correlation coefficients of the traditional machine learning algorithm are on par with those of the artificial neural network algorithm, its *RMSE* is notably inferior to that of the artificial neural network. The interpolation expansion algorithm based on the LM method performs best when the average *RMSE* value is 5.797; however, its correlation coefficient is not ideal (Table 4). Without interpolation expansion, the mean Pearson, mean Kendall, and mean Spearman correlation coefficients for this algorithm exceed 0.95. Moreover, while the cumulative *RMSE* values of the two methods are similar, the Pearson cumulative and Spearman cumulative values are also comparable. However, the Kendall cumulative value of the former is slightly better than that of the interpolation method (Figure 13). This finding indicates that while interpolation can slightly reduce *RMSE*, the process may lead to a reduction in the Kendall correlation coefficient.

In this study, a capillary pressure curve prediction model was developed based on particle size data by using traditional machine learning and artificial neural network algorithms. The major difference from predecessors lies in the innovative use of rock particle size as input data for prediction [11–14]. However, the particle size data in the current study were obtained by sieve analysis, and the research object was sand conglomerate, limiting the applicability of the findings to other lithologies, such as igneous and metamorphic rocks. The sieve analysis method can also damage samples while extracting particle size information, potentially limiting subsequent core experiments. However, particle size information can also be obtained using other nondestructive or low-loss techniques, such as laser diffraction or image analysis [48]. These techniques may be more suitable for measuring particle size across different rock types while preserving sample integrity for further study. The diameter of the rock may be strongly correlated with the pore diameter distribution as the capillary pressure curve reflects the pore diameter distribution of the core. However, this hypothesis is not supported by conclusive evidence in this study. In future research, we will further investigate the feasibility of a predictive relationship between rock particle size and pore size distribution.

7. Conclusions

This study compares traditional machine learning and artificial neural network models to evaluate the significant ability of artificial intelligence methods in predicting capillary pressure curves based on particle size data.

(1) Machine learning uses the boosting tree model as the optimal algorithm, achieving an average *RMSE* of 6.645. Fitting training on selected particle size features shows improved training and prediction performances as the number of features increases; however, performance remains inferior to that of the full feature data.
(2) The optimal configuration for the ANN model is the LM method with 15 hidden layer neurons, achieving an average *RMSE* of 5.797. This finding indicates better performance than that of the best machine learning algorithm. After the interpolation of the original particle size data, the prediction performance improved when expanded to 50 groups of features, resulting in a cumulative *RMSE* of 23.189. However, this interpolation method led to a slight reduction in the Kendall correlation coefficient, indicating that interpolation introduces volatility that affects the prediction results.
(3) A comparison of the two artificial intelligence methods—machine learning and artificial neural networks—demonstrates superior performance in training and fitting. In addition, its prediction process is more efficient and quick, resulting in the highest predictive performance.

Author Contributions: Conceptualization, X.Q. and Y.W.; methodology, Z.W; software, S.W.; validation, M.Z., Y.W., and S.W.; writing—original draft preparation, Y.W.; writing—review and editing, X.Q. All the authors have read and agreed to the published version of the manuscript.

Funding: This study was financially supported by the Xinjiang backbone talent training program: 2024; National Natural Science Foundation of China: 41790453 and 41874135; Student Innovation and Entrepreneurship Project of Xinjiang Institute of Engineering: 202310994008.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ma, W.G.; Wang, P.; Wang, Y. The Study Of Daqing Oilfield Class II Reservoirs Pore Throat Ratio Using Conventional Pressure Mercury Method. *Appl. Mater. Technol. Mod. Manuf.* **2013**, *423–426*, 622–625. [CrossRef]
2. Lai, J.; Wang, G.W. Fractal analysis of tight gas sandstones using high-pressure mercury intrusion techniques. *J. Nat. Gas Sci. Eng.* **2015**, *24*, 185–196. [CrossRef]
3. Kenvin, J.; Jagiello, J.; Mitchell, S.; Perez-Ramírez, J. Unified Method for the Total Pore Volume and Pore Size Distribution of Hierarchical Zeolites from Argon Adsorption and Mercury Intrusion. *Langmuir* **2015**, *31*, 1242–1247. [CrossRef] [PubMed]
4. Chen, J.; Li, E.; Luo, J. Characterization of Microscopic Pore Structures of Rock Salt through Mercury Injection and Nitrogen Absorption Tests. *Geofluids* **2018**, *2018*, 9427361. [CrossRef]
5. Zhang, F.; Jiang, Z.X.; Sun, W.; Li, Y.H.; Zhang, X.; Zhu, L.; Wen, M. A multiscale comprehensive study on pore structure of tight sandstone reservoir realized by nuclear magnetic resonance, high pressure mercury injection and constant-rate mercury injection penetration test. *Mar. Pet. Geol.* **2019**, *109*, 208–222. [CrossRef]
6. Fu, S.S.; Fang, Q.; Li, A.; Li, Z.P.; Han, J.L.; Dang, X.; Han, W.C. Accurate characterization of full pore size distribution of tight sandstones by low-temperature nitrogen gas adsorption and high-pressure mercury intrusion combination method. *Energy Sci. Eng.* **2021**, *9*, 80–100. [CrossRef]
7. Wu, B.; Xie, R.H.; Jin, G.W.; Liu, J.L.; Wang, S.; Fan, W.S. Investigation on the Pore Structure and Multifractal Characteristics of Tight Sandstone Using Nitrogen Gas Adsorption and Mercury Injection Capillary Pressure Experiments. *Energy Fuels* **2022**, *36*, 262–274. [CrossRef]
8. Zhou, Y.Q.; Xu, J.; Lan, Y.Y.; Zi, H.; Cui, Y.L.; Chen, Q.X.; You, L.Z.; Fan, X.Q.; Wang, G.W. New insights into pore fractal dimension from mercury injection capillary pressure in tight sandstone. *Geoenergy Sci. Eng.* **2023**, *228*, 212059. [CrossRef]
9. Xiao, L.; Zhang, W. A new method to construct reservoir capillary pressure curves using NMR log data and its application. *Appl. Geophys.* **2008**, *5*, 92–98.
10. Eslami, M.; Kadkhodaie-Ilkhchi, A.; Sharghi, Y.; Golsanami, N. Construction of synthetic capillary pressure curves from the joint use of NMR log data and conventional well logs. *J. Pet. Sci. Eng.* **2013**, *111*, 50–58. [CrossRef]
11. Liang, X.; Zou, C.C.; Wang, H. Comments on "Construction of synthetic capillary pressure curves from the joint use of NMR log data and conventional well logs". *J. Pet. Sci. Eng.* **2015**, *135*, 429–432.
12. Xiao, L.; Mao, Z.; Zou, C.; Jin, Y.; Zhu, J.C. A new methodology of constructing pseudo capillary pressure (Pc) curves from nuclear magnetic resonance (NMR) logs. *J. Pet. Sci. Eng.* **2016**, *147*, 154–167. [CrossRef]
13. Zhang, H.T.; Li, G.R.; Guo, H.P.; Zhang, W.J.; Wang, Y.M.; Li, W.B.; Zhou, J.Y.; Wang, C.S. Applications of nuclear magnetic resonance (NMR) logging in tight sandstone reservoir pore structure characterization. *Arab. J. Geosci.* **2020**, *13*, 572. [CrossRef]

14. Wu, B.; Xie, R.; Liu, M.; Jin, G.; Xu, C.Y.; Liu, J.L. Novel Method for Predicting Mercury Injection Capillary Pressure Curves of Tight Sandstone Reservoirs Using NMR T2 Distributions. *Energy Fuels* **2021**, *35*, 15607–15617. [CrossRef]
15. Zhou, Y.Q.; You, L.Z.; Zi, H.; Lan, Y.Y.; Cui, Y.L.; Xu, J.; Fan, X.Q.; Wang, G.W. Determination of pore size distribution in tight gas sandstones based on Bayesian regularization neural network with MICP, NMR and petrophysical logs. *J. Nat. Gas Sci. Eng.* **2022**, *100*, 104468. [CrossRef]
16. Wang, W.H.; Wang, Z.W.; Han, R.Y.; Xu, F.H.; Qi, X.H.; Cui, Y.T. Lithology classification of volcanic rocks based on conventional logging data of machine learning: A case study of the eastern depression of Liaohe oil field. *Open Geosci.* **2021**, *13*, 1245–1258.
17. Yu, Z.C.; Wang, Z.Z.; Zeng, F.C.; Song, P.; Baffour, B.A.; Wang, P.; Wang, W.F.; Li, L. Volcanic lithology identification based on parameter-optimized GBDT algorithm: A case study in the Jilin Oilfield, Songliao Basin, NE China. *J. Appl. Geophys.* **2021**, *194*, 104443. [CrossRef]
18. Song, Z.J.; Xiao, D.S.; Wei, Y.B.; Zhao, R.X.; Wang, X.C.; Tang, J.F. The Research on Complex Lithology Identification Based on Well Logs: A Case Study of Lower 1st Member of the Shahejie Formation in Raoyang Sag. *Energies* **2023**, *16*, 1748. [CrossRef]
19. Wang, J.; Cao, J.X. A Lithology Identification Approach Using Well Logs Data and Convolutional Long Short-Term Memory Networks. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 3322677. [CrossRef]
20. Wang, J.; Cao, J.X.; Zhao, S.; Qi, Q.M. S-wave velocity inversion and prediction using a deep hybrid neural network. *Sci. China-Earth Sci.* **2022**, *65*, 724–741. [CrossRef]
21. Alenizi, F.A.; Mohammed, A.H.; Alizadeh, S.M.; Gohari, O.M.; Motahari, M.R. Appraisal of rock dynamic, physical, and mechanical properties and forecasting shear wave velocity using machine learning and statistical methods. *J. Appl. Geophys.* **2023**, *223*, 105216. [CrossRef]
22. Hong, Y.; Li, S.M.; Wang, H.L.; Liu, P.C.; Cao, Y. Quantitative Prediction of Rock Pore-Throat Radius Based on Deep Neural Network. *Energies* **2023**, *16*, 7277. [CrossRef]
23. Li, J.; Xu, T.; Zhang, W.T.; Liu, H.N.; Kang, Y.; Lv, W.J. A borehole porosity prediction method with focusing on local shape. *Geoenergy Sci. Eng.* **2023**, *228*, 211933. [CrossRef]
24. Li, M.; Zhang, J.X.; Zhou, N.; Huang, Y.L. Effect of Particle Size on the Energy Evolution of Crushed Waste Rock in Coal Mines. *Rock Mech. Rock Eng.* **2017**, *50*, 1347–1354. [CrossRef]
25. Li, M.; Zhang, J.X.; Song, W.J.; Germain, D.M. Recycling of crushed waste rock as backfilling material in coal mine: Effects of particle size on compaction behaviours. *Environ. Sci. Pollut. Res.* **2019**, *26*, 8789–8797. [CrossRef]
26. You, S.; Zhang, C.H.; Ji, H.G. Mesostructure failure mode of compacted rock medium in deep strata. *Emerg. Mater. Res.* **2020**, *9*, 460–471.
27. Feng, Z.T.; Fan, X.M.; Ni, T.; Deng, Y.; Zou, C.B.; Zhang, J.; Xu, Q. How Ice Particles Increase Mobility of Rock-Ice Avalanches: Insights From Chute Flows Simulation of Granular Rock-Ice Mixtures by Discrete Element Method. *J. Geophys. Res.-Earth Surf.* **2023**, *128*, e2023JF007115. [CrossRef]
28. Furuichi, M.; Chen, J.; Nishiura, D.; Arai, R.; Yamamoto, Y. Thrust formation using a numerical granular rock box experiment. *Tectonophysics* **2023**, *862*, 229963. [CrossRef]
29. Shuai, W.; Ying, X.; Yanbo, Z.; Xulong, Y.; Peng, L.; Xiangxin, L. Effects of sandstone mineral composition heterogeneity on crack initiation and propagation through a microscopic analysis technique. *Int. J. Rock Mech. Min. Sci.* **2023**, *162*, 105307. [CrossRef]
30. Wu, J.B.; Hu, Y.T.; Zhang, H.R. A Method of Inverting Rock Grain Size Based on Nuclear Magnetic Resonance Logging Data and Application. *Geofluids* **2023**, *2023*, 7941695. [CrossRef]
31. *GB/T 29172-2012*; Practices for Core Analysis. Standardization Administration of China (SAC): Beijing, China, 2012.
32. *GB/T 29171-2023*; Rock Capillary Pressure Measurement. Standardization Administration of China (SAC): Beijing, China, 2023.
33. Rabbath, C.A.; Corriveau, D. A comparison of piecewise cubic Hermite interpolating polynomials, cubic splines and piecewise linear functions for the approximation of projectile aerodynamics. *Def. Technol.* **2019**, *15*, 741–757. [CrossRef]
34. Barker, P.M.; McDougall, T.J. Two Interpolation Methods Using Multiply-Rotated Piecewise Cubic Hermite Interpolating Polynomials. *J. Atmos. Ocean. Technol.* **2020**, *37*, 605–619. [CrossRef]
35. Saidi, A.Y.N.; Ramli, N.A.; Muhammad, N.; Awalin, L.J. Power outage prediction by using logistic regression and decision tree. *J. Phys. Conf. Ser.* **2021**, *1988*, 012039. [CrossRef]
36. Yousefmarzi, F.; Haratian, A.; Kalatehno, J.M.; Kamal, M.K. Machine learning approaches for estimating interfacial tension between oil/gas and oil/water systems: A performance analysis. *Sci. Rep.* **2024**, *14*, 858. [CrossRef]
37. Ikeagwuani, C.C. Determination of Unbound Granular Material Resilient Modulus with MARS, PLSR, KNN and SVM. *Int. J. Pavement Res. Technol.* **2022**, *15*, 803–820. [CrossRef]
38. Hoseini, B.; Jaafari, M.R.; Golabpour, A.; Momtazi-Borojeni, A.A.; Karimi, M.; Eslami, S. Application of ensemble machine learning approach to assess the factors affecting size and polydispersity index of liposomal nanoparticles. *Sci. Rep.* **2023**, *13*, 18012. [CrossRef]
39. Dehghani, M.; Jahani, S.; Ranjbar, A. Comparing the performance of machine learning methods in estimating the shear wave transit time in one of the reservoirs in southwest of Iran. *Sci. Rep.* **2024**, *14*, 4744. [CrossRef]
40. Fang, Z.C.; Cheng, J.; Xu, C.; Xu, X.Y.; Qajar, J.; Rastegarnia, A. Comparison of machine learning and statistical approaches to estimate rock tensile strength. *Case Stud. Constr. Mater.* **2024**, *20*, e02890. [CrossRef]

41. Lin, Y.W.; Zhou, Z.Y.; Song, Z.X.; Shi, Q.; Hao, Y.C.; Fu, Y.Q.; Li, T.; Zhang, Z.S.; Wu, J.Y. Insights into the mechanical stability of tetrahydrofuran hydrates from experimental, machine learning, and molecular dynamics perspectives. *Nanoscale* **2024**, *16*, 6296–6308. [CrossRef]
42. Shuai, G.; Zhou, Y.; Shao, J.L.; Cui, Y.L.; Zhang, Q.L.; Jin, C.W.; Xu, S.Y. Comparison of Multiple Machine Learning Methods for Correcting Groundwater Levels Predicted by Physics-Based Models. *Sustainability* **2024**, *16*, 653. [CrossRef]
43. Nia, A.M.; Misra, D.; Kashani, M.H.; Ghafari, M.; Sahoo, M.; Ghodsi, M.; Tahmoures, M.; Taheri, S.; Jaafarzadeh, M.S. Runoff and Sediment Yield Processes in a Tropical Eastern Indian River Basin: A Multiple Machine Learning Approach. *Land* **2023**, *12*, 1565. [CrossRef]
44. Yürüşen, N.Y.; Uzunoğlu, B.; Talayero, A.P.; Estopiñán, A.L. Apriori and K-Means algorithms of machine learning for spatio-temporal solar generation balancing. *Renew. Energy* **2021**, *175*, 702–717. [CrossRef]
45. Zhang, G.Y.; Shao, F.; Yuan, W.; Wu, J.Y.; Qi, X.; Gao, J.; Shao, R.; Tang, Z.; Wang, T. Predicting sepsis in-hospital mortality with machine learning: A multi-center study using clinical and inflammatory biomarkers. *Eur. J. Med. Res.* **2024**, *29*, 156. [CrossRef] [PubMed]
46. Abbasimaedeh, P. Soil liquefaction in seismic events: Pioneering predictive models using machine learning and advanced regression techniques. *Environ. Earth Sci.* **2024**, *83*, 189. [CrossRef]
47. Khan, S.; Alzaabi, A.; Ratnarajah, T.; Arslan, T. Novel statistical time series data augmentation and machine learning based classification of unobtrusive respiration data for respiration Digital Twin model. *Comput. Biol. Med.* **2024**, *168*, 107825. [CrossRef]
48. Krawczykowski, D. Unification of particle size analysis results, part 1—Comparison of particle size distribution functions obtained by various measurement methods. *Measurement* **2024**, *238*, 115403. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.