

Article

A Random Forest-Based Method for Predicting Borehole Trajectories

Baoyong Yan ^{1,2}, Xiantao Zhang ^{1,2}, Chengxu Tang ³, Xiao Wang ³, Yifei Yang ³ and Weihua Xu ^{3,*}

¹ State Key Laboratory of the Gas Disaster Detecting, Preventing and Emergency Controlling, Chongqing 400039, China

² CCTEG Chongqing Research Institute, Chongqing 400039, China

³ College of Artificial Intelligence, Southwest University, Chongqing 400715, China

* Correspondence: chxuwh@swu.edu.cn; Tel.: +86-159-9892-1583

Abstract: Drilling trajectory control technology for near-horizontal directional drilling in coal mines is mainly determined empirically by manual skew data, and the empirical results are only qualitative and variable, meanwhile possessing great instability and uncertainty. In order to improve the accuracy and efficiency of drilling trajectory prediction, this paper investigates a random forest regression-based drilling trajectory prediction method after applying numerous machine learning methods to experimental data for comparison. In the selection of prediction features, this paper replaces all feature variables of the borehole trajectory with feature values with higher cumulative influence weights, and screens out three variables with high importance, azimuth, inclination and bend at the present moment of the drill, as the input variables of the model, and the increments in the borehole in a horizontal direction, left and right direction, and up and down direction at the present moment and the next moment as the output variables of the model. In the model training, the model in this paper uses the bootstrap self-service method resampling technique to collect training sample data, constructs a random forest regression model, and takes the mean value of the decision tree output as the result of the borehole trajectory prediction. To further improve the model accuracy, the paper improves the prediction performance of the model by adjusting the parameters of the random forest model such as tree, depth, minimum sample of leaf nodes, minimum sample number of internal node division, etc. The model is also evaluated by using common machine learning evaluation metrics R2 score, RAE, RMSE and MSE. The experimental results show that the average absolute error of the model drops to 1% on the prediction of the horizontal direction and up and down direction; the average absolute error of the model drops to 9% on the prediction of the left and right direction, and this error rate reaches the error requirement in the actual construction process, so the model can effectively improve the prediction accuracy of borehole trajectory while improving the safety of directional construction.



Citation: Yan, B.; Zhang, X.; Tang, C.; Wang, X.; Yang, Y.; Xu, W. A Random Forest-Based Method for Predicting Borehole Trajectories. *Mathematics* **2023**, *11*, 1297. <https://doi.org/10.3390/math11061297>

Academic Editors: Aleksandr Rakhmangulov and Jakub Nalepa

Received: 7 December 2022

Revised: 21 February 2023

Accepted: 3 March 2023

Published: 8 March 2023

Keywords: borehole trajectory prediction; random forest regression model; feature and predictor variable selection; parameter tuning

MSC: 68T30; 68U35



Copyright: © 2023 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

Coal mine target-application trajectory drilling has now become the main technical way of geological exploration and gas extraction at home and abroad. In the new era of development, China's energy structure continues to be optimized and the proportion of clean energy continues to rise, but coal is still a pillar energy source in China's energy structure at the current stage and for some time to come, and occupies an important position in the national economy and social development. With the increase in mining depth, mechanization and automation level, coal mining involves science and technologies that

are also constantly developing, underground gas management [1], water exploration and release, hole instead of alley and many other operational needs for coal mine underground near-horizontal directional drilling equipment, technology, process, etc. Thus, new requirements are put forward, the development of the underground automatic, autonomous, intelligent operation of the long drilling directional-precision construction function of the equipment and its supporting processes, collaborative operation line, related technologies and methods of directional drilling borehole-trajectory control, etc., has been the focus of the progress of coal mine underground-drilling construction in recent years and for the next few years, so directional drilling borehole-trajectory control technology can achieve targeted drilling [2].

At present, underground directional drilling in coal mines is carried out by the manual operation of drilling tools and equipment; moving, anchoring, opening, loading and unloading drilling tools, trajectory measurements, drilling control, data uploading and analysis are all manual operations, which completely rely on manual experience for the effective control of drilling construction to ensure that the drill hole operates according to the designed drilling trajectory. The drilling trajectory control technology applicable to the near-horizontal directional drilling in coal mines is mainly determined by manual experience on the measured inclination data, and the effectiveness of the adjustment of the bending head on the control of the drilling trajectory is determined by the prediction of the drilling trajectory trend at the next measurement point through manual experience based on the measured data and the approximate evaluation of the adjusted bending head value. Therefore, there are several problems:

(1) The empirical results are only qualitative and indeterminate, with great instability and uncertainty, and the trajectory control may appear unsmooth or even show drastic changes, and precise adjustment and control cannot be realized;

(2) The effect of manual experience on trajectory control under different geological conditions, construction equipment, construction parameters and operator conditions is unstable and even destructive leading to the original design not being implemented, high drilling resistance, overrunning in the direction of drilling, stringing holes, collapsing holes, falling out of drilling, difficulty in starting and pulling and blind areas in drilling;

(3) In addition, manual individual experience and group experience are not directly applicable and applied to qualitative drilling automation operations, and cannot be well adapted to the construction and development of intelligent mines.

In order to solve several problems mentioned above, we extracted useful and effective information from the data, reasoned computationally through the model to obtain the exact direction of the drilling trajectory control or construction elbow proposal with confidence level, completed a numerical over-prediction of the trajectory of the drilling, allowed the data to provide scientific theory and feasible methods to eliminate the uncertainty and instability of the artificial experience to a certain extent and summarized the artificial experience. The computer language, which represents the manual participation process to achieve no artificial individual differences in the computer processing of the target application drilling trajectory, can lay the foundation for the automatic control of the drilling trajectory, automatic directional drilling operations, and provides digital tools and methods for the achievement of intelligent drilling.

Since the introduction of drilling trajectory prediction technology into the petroleum field by the American scholar Lubinsk, drilling trajectory prediction technology has developed rapidly and entered a period of vigorous development, such as the introduction of the three-bending moment equation into the static analysis of drilling tools based on the longitudinal and transverse bending theory by Professor Bai Jiazhi and others in China [3]. Directional drilling technology in the coal-mine field was introduced from oil fields by researchers in the United States and Germany in the 1960s. Since then, the combination of single bending-screw drilling tools and the follow-on measurement system has become the mainstream of directional drilling construction in coal mines by virtue of its excellent slope-making rate and controllability [4], which provides equipment and a technical guarantee

for the operational needs of underground coal mine gas-management, water exploration and release, geological exploration, and hole replacement alley. The State Key Laboratory of Gas Disaster Monitoring and Emergency Technology and its supporting units have made certain achievements in related research, providing important support and guarantees for gas management and utilization in China's coal industry.

In recent years, with the continuous development in the fields of data mining, machine learning and artificial intelligence, some relatively novel trajectory prediction methods have appeared. Some scholars have studied the prediction model in combination with these emerging fields and achieved many good results. Machine learning is becoming more and more widespread in engineering forecasting, and its role is constantly being proven with the development of the information age. There are many examples of machine learning applied to forecasting. For example, Zhang Jinchuan's machine learning algorithm for predicting oil and gas production status [5]; the performance prediction of nuclear-power structural materials based on a machine learning algorithm studied by Wang Zhuo et al. [6]; Wang Zixuan studied the optimal scheduling of an electric heating comprehensive energy system based on a machine learning prediction algorithm [7]. At the same time, there is a lot of research on trajectory prediction by machine learning. For example, Song Lujie et al. studied the moving object position-prediction algorithm combined with a Markov model and trajectory similarity [8]. Chen Weihua et al. studied a model based on a deep learning algorithm to predict the cutting track of the shearer [9]. Meng Qinghua et al. established a corresponding borehole trajectory prediction model by combining a wavelet and neural network [10]. Liu Leijun et al. proposed a sparse trajectory-prediction method combining iterative network partitioning and entropy estimation [11]. However, there is little research on whether these machine learning trajectory-prediction methods are applicable to the theory and application practice of directional drilling trajectory-prediction in coal mine. The existing prediction methods of the directional drilling trajectory in underground coal mines mainly form a set of theoretical methods by analyzing the complex geological conditions of the near-horizontal directional drilling in coal seam construction, summarizing the characteristics of the drilling trajectory and drilling accident experience. For example, Sun Rongjun et al. proposed a near-horizontal directional drilling technology for underground coal mine [12]. Research on prediction and updating the method of logging while drilling was proposed by Guo Yongheng et al. [13]. Feng Dahui put forward research on the directional pore-forming technology of the drainage holes in tunneling roadways with downhole precision [14]. Sun Tao et al. improved the traditional three-point circle-fixing method [15]. However, there is little research [16–21] on model prediction related to machine learning, so applying machine learning predictions to directional drilling trajectory-prediction in coal mining is a promising research direction. The main contributions of this paper are as follows:

(1) The establishment of a scientific, effective and applicable trajectory prediction model to determine the distance that the drilling tool may deviate from the target point, so that timely manual or automatic adjustment can be made. Drill hole trajectory prediction and control aims to reduce the numerical gap between the actual trajectory of the drilling tool and the designed trajectory. Improving the accuracy of trajectory prediction can provide a numerical basis for accurate control, and control and adjust according to the predicted reference results so that the actual trajectory of the drilling tool will be constructed according to the designed trajectory;

(2) The drilling trajectory is influenced by a variety of operational factors, and a variety of factors need to be considered in the final segmentation study. As the drilling data of the completed construction, the comprehensive influence of each factor has been formed, i.e., the result of the comprehensive effect of each factor has been determined and will not change, and these influences are fused into a final result without the influence of factor variables, and the drilling trajectory prediction can be carried out with this influence result as the research object;

(3) The drilling trajectory control technology for near-horizontal directional drilling in coal mines is mainly determined empirically by manual skew data, and the empirical results are only qualitative and indeterminate, with great instability and uncertainty. We used the machine learning method (the random forest algorithm has the advantages of few parameters [22], a fast learning speed, applicable to high-dimensional samples, that can effectively avoid overfitting, has high prediction accuracy and has been widely used in classification and regression problems) to predict the next measurement point borehole-trajectory trend based on the measured data, and the key lies in how to select effective features and a suitable prediction model to fit the existing data to meet the project prediction accuracy requirements (the absolute value of deviation between construction results and prediction results for a single reference tool face of each 3 m is lower than 0.2 m).

Briefly, this paper proposes a random forest-based method for predicting borehole trajectories, and Section 2 introduces the basic theory of random forest and directional drilling model. In Section 3, the drilling trajectory data and preprocessing methods are analyzed. In Section 4, a random forest-based borehole-trajectory prediction method is proposed. Section 5 of this paper presents the prediction results of our random forest regression model. In Section 6, the comparison with other prediction models shows that our proposed model is more accurate in predicting the results. Section 8 concludes and outlines future work.

2. Basic Theory

2.1. Calculation of the Predicted Values of the Attitude of the Measurement Points

The three-point circularization method is used to define the slope of the drilling tool $k = \frac{2v}{l_1+l_2}$ (where k is the size of the structural bend; l_1 is the distance from the drill bit to the lower stabilizer, and l_2 is the distance from the lower stabilizer to the upper stabilizer). The predicted values of the inclination, azimuth and three-axis coordinates (corresponding to the 3D coordinate system of the wellhead are three directions: horizontal (x axis), left and right (y axis) and up and down (z axis)) for the next measurement point are calculated based on the measured data. Each piece of data for the next measurement point can be calculated by the following equation.

When the slope of creation is determined, the predicted inclination and azimuth of the next measurement point are calculated.

Inclination of the next measurement point:

$$\theta_{n+1} = \theta_n + \arcsin(\sin(k\Delta L)\cos\Omega_n) \tag{1}$$

Azimuth of next measurement point:

$$\alpha_{n+1} = \alpha_n + \arcsin(\sin(k\Delta L)\sin\Omega_n) \tag{2}$$

where θ_{n+1} , α_{n+1} are the inclination and azimuth of the next measurement point; θ_n , α_n are the inclination and azimuth of the current measurement point; k is the slope, ΔL is the length of the drill pipe and Ω_n is the tool face angle used for the next drill pipe.

In order to more intuitively represent the closeness of the predicted trajectory to the actual trajectory, the predicted results need to be expressed in the form of trajectory coordinates. Based on the equilibrium tangent method, the predicted coordinate value of the next measurement point in the orifice coordinate system can be calculated from the predicted attitude data and the drilled borehole trajectory data.

Predicted value of the triaxial coordinates of the next measurement point:

$$x_{n+1} = \frac{1}{2}\Delta L(\cos\theta_{n+1}\cos\alpha_{n+1} + \cos\theta_n\cos\alpha_n) + x_n \tag{3}$$

$$y_{n+1} = \frac{1}{2}\Delta L(\cos\theta_{n+1}\sin\alpha_{n+1} + \cos\theta_n\sin\alpha_n) \tag{4}$$

$$z_{n+1} = \frac{1}{2} \Delta L (\sin \theta_{n+1} + \sin \theta_n) \quad (5)$$

where x_{n+1} , y_{n+1} , z_{n+1} are the triaxial coordinates of the next measurement point and x_n , y_n , z_n are the triaxial coordinates of the current measurement point. The trajectory deviation of each borehole can be determined by comparing the predicted values with the trajectory values of the actual construction.

2.2. A Study of the Generalizability of Trajectory Data Prediction Algorithms

In order to study the prediction accuracy and applicability of the theoretical model during actual construction, an attempt was made to study a parametric model with universal applicability to multiple boreholes, and to improve the prediction accuracy and applicability of the drilling trajectory of coal seam drilling tools by adjusting the parameters under the condition of existing data.

The pre-existing data structure can be formulated as follows:

$$x_n = \{a_1, a_2, \dots, a_n\} \quad (6)$$

where x_n represents the trajectory data at the time of n and $\{a_1, a_2, \dots, a_n\}$ represents the data of n variables such as horizontal depth, azimuth, inclination, etc., at the time of n .

2.3. Borehole Trajectory Prediction Based on Random Forest Regression

The random forest algorithm is an integrated learning algorithm built on the idea of the Bagging algorithm for the classification and regression studies of data [23]. When it is used to study regression problems with continuous variables, it is called the random forest regression problem.

The main steps of random forest regression-based prediction of borehole trajectories are as follows:

- (1) A new training set of n samples with the same number of samples as the initial training set is repeatedly selected at random with put-backs in the initial training set, and the unselected data are called unselected data;
- (2) Randomly select m features from the M input features affecting the borehole trajectory as the set of alternative feature variables for the branching nodes of the decision tree, and select the optimal features from the set for node splitting according to the branching superiority criterion to construct the decision tree;
- (3) n borehole-trajectory decision trees were constructed to form a random forest regression model. The average of the output result values of the n decision trees is taken as the prediction result of the borehole trajectory, and the unselected data are used to evaluate the prediction performance of the regression model.

3. Analysis of Drilling Trajectory Data

3.1. Description of the Data Set

The data source used in this paper was mainly the on-site construction data of the China Coal Science and Technology Group Chongqing Research Institute Company Limited, which provided 14,694 pieces of borehole data collected from drilling wells across the country from August 2014 to May 2022. The observed elements of the trajectory data file of the drilling process included the coordinates of the drill bit position at each moment, the geology of the location, the inclination angle of the tool facing angle and the branch where the drilling point is located.

3.2. Data Cleaning

Since directional drilling has the characteristic of making the borehole achieve directional bending changes by changing the orientation of the elbow during drilling, in each piece of drilling data, corresponding to one borehole branch, the data will show a certain continuity and the hole depth will show the same magnitude increment (3 m). The track of underground drilling is shown in Figure 1. During the construction process, the

data of drilling depth, top Angle and azimuth Angle are measured by special measuring instruments. After calculation and drawing, the spatial coordinates and track graphs of each measuring point of drilling are obtained. The experiment requires a machine learning algorithm where the input data from the previous moment can estimate the data from the next moment. However, some of the hole depth data do not show a continuous increment due to various uncertainties or characteristics of the data sampling time. In this paper, the data cleaning is for this problem [24].

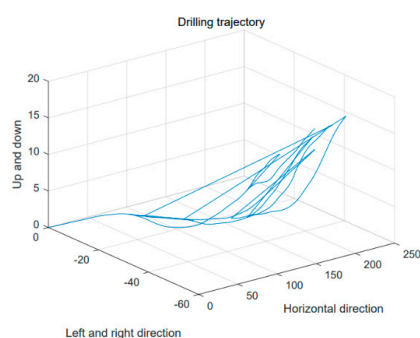


Figure 1. Visualization of drilling trajectory.

If the data labeled as “Xmain” have a value of 0 in the horizontal, left and right directions, or up and down directions, and are irregularly distributed in the overall data, then these data should be deleted.

Each branch of the drilling trajectory corresponds to a path and the hole depth increases by the same magnitude (3 m), then it is possible to determine whether it is the same branch based on whether the hole depth difference is constant or not; if the difference between the two adjacent data hole depths is constant, then it is kept, if it is not, then it should be deleted.

3.3. Feature Selection

The purpose of feature selection work is to combine the knowledge related to directional drilling to select the feature set, and then analyze and evaluate the importance of each feature among the known features to select a subset of features that can comprehensively and effectively characterize the original data [25], and reduce the feature dimension of the sample to fit a suitable machine learning model to achieve high prediction accuracy. The subsequent prediction of the trajectory of the drilling application process can be carried out accordingly.

According to the theoretical knowledge related to directional drilling, the azimuth and inclination angle in the tool-facing angle at the same moment combined with the hole depth can calculate the horizontal direction, left and right direction and up and down direction coordinates at that moment; the elbow angle at the previous moment combined with the horizontal direction, left and right direction, and up and down direction coordinates can calculate the horizontal direction, left and right direction, and up and down direction coordinates at the next moment; additionally, consider that when drilling, geological factors can also affect the trajectory of the drill bit; therefore, the bend angle, azimuth angle, inclination angle and geological material of the previous moment can be included in the selection of the character variables, and the azimuth angle and inclination angle of the next moment and the up and down direction, left and right direction and horizontal direction of the next moment and their incremental values can be included in the selection of the character root variables. However, since the machine learning method does not accept character-type input, the geological features should be converted into numerical features using unique thermal coding [26].

According to machine learning algorithms, the degree of correlation between variables can be calculated qualitatively. For the borehole trajectory prediction task, this paper

sets a threshold value for calculating the correlation degree between each feature in the training set and the prediction variables, and excludes the features with a small correlation degree. In the random forest regression prediction model, it automatically builds different random forests according to the variables to be predicted, and assigns corresponding correlation weights to the different feature variables. In this paper, using the random forest regression algorithm, the correlation weight of geological factors on the predictor variables is calculated to be less than 0.5%, so the geological materials are removed from the category of feature variables.

4. A Random Forest-Based Method for Predicting Borehole Trajectories

4.1. Problem Definition

The coordinates of the drill bit in the three-dimensional space at each moment of the drilling process constitute a trajectory sequence (as shown in Figure 1), and the drill bit trajectory prediction task requires predicting the coordinates of the space where the drill bit is located after a single drilling step (3 m) based on the azimuth and inclination of the current drilling point as well as the curved head.

Assuming that the azimuth, inclination, bend and the three-dimensional coordinates $s = (x, y, z)$ of the current drilling point of a certain drilling, i.e., horizontal position, left and right position and up and down position, the predicted increment of the drill bit in each direction after three meters of drilling is $\hat{s} = (\hat{x}, \hat{y}, \hat{z})$, the coordinates of the drill bit after three meters of drilling are $s + \hat{s} = (x + \hat{x}, y + \hat{y}, z + \hat{z})$.

4.2. Feature and Model Selection

A borehole-trajectory prediction model system was constructed using multiple regression analysis combined with a random forest model to predict borehole trajectories. The prediction in this paper is based on the borehole data at the present moment to predict the borehole data at the next measurement point. Before bringing in the model for solution, the data are cleaned by removing invalid and anomalous data, then feature variables and predictor variables are constructed and machine learning regression models are fitted using the processed data.

The problem studied in this paper is a regression problem, and the key to solving the problem lies in the selection of the feature and predictor variables and the regression model. As shown in Figure 2, by replacing the feature and predictor variables and fitting different machine learning models, the highest experimental accuracy was obtained by using the chosen random forest regression model to predict the increments of up and down, horizontal, left and right directions of the next measurement point using the bend, orientation and inclination of the drill, from which the predictor and feature variables and the machine learning method were determined.

4.3. Hyperparameter Determination

The random forest parameter tuning contains: decision trees of tree Estimators, the minimum number of samples required for internal node repartitioning `Min_samples_split`, decision tree depth `Max_depth`, leaf node minimum number of samples `Min_samples_leaf`, the hyperparameters of the model can be selected by specific data to further improve the model prediction accuracy [27]. (Since there are different hyperparameters for predicting horizontal, up-down, and left-right directions increments, the horizontal direction is used here as an example).

From Figure 3, the decision tree Estimators = 197.0; the minimum number of samples required for internal node repartitioning `Min_samples_split` = 2.0; the minimum number of samples of leaf nodes `Min_samples_leaf` = 2.0; the maximum depth of the tree `Max_depth` = 16.0, the random forest regression model in response to the horizontal position of the borehole. The prediction parameters of the increment reached the optimum.

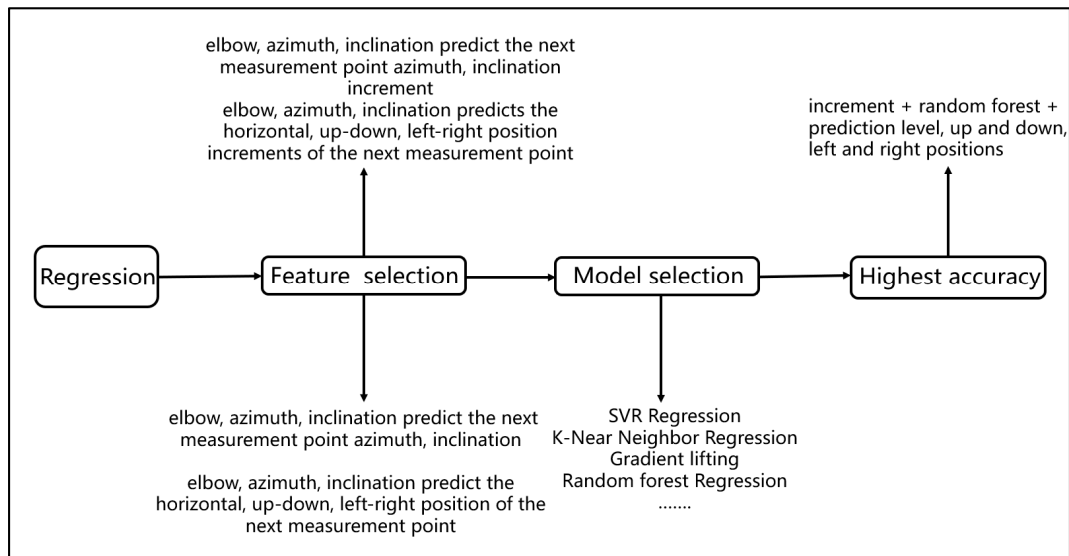


Figure 2. Feature and model selection flowchart.

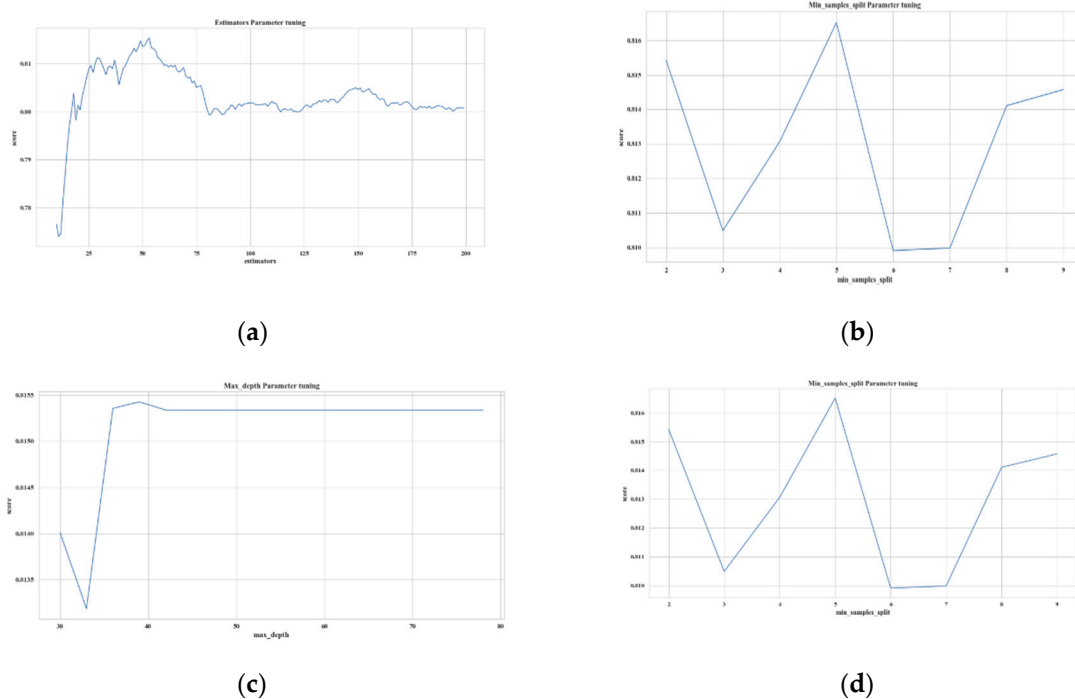


Figure 3. Model hyperparameter determination. (a) Estimators = 197.0. (b) Min_samples_split = 2.0. (c) Max_depth = 16.0. (d) Min_samples_leaf = 2.0.

4.4. Algorithmic Principles of the Random Forest Regression Model (Algorithm 1)

Three feature variables were chosen as input variables for the prediction of the drill trajectory: inclination $x_i^{(1)}$, azimuth $x_i^{(2)}$, elbow $x_i^{(3)}$ and incremental values as output variables: increments in the horizontal direction of the drill $y_i^{(1)}$, increments in the up-down direction of the drill $y_i^{(2)}$, and increments in the left-right direction of the drill $y_i^{(3)}$. We set the number of samples brought into training to m (random forest auto-division); the number of feature variables selected for training to g (random forest auto-division); the cut feature variable to j , $j \in \{x_i^{(1)}, x_i^{(2)}, x_i^{(3)}\}$; the cut point to s , $s \in \{s_1, s_2 \dots s_n\}$; the left subtree mean to c_1 ; the right subtree mean to c_2 ; and the function satisfies $R_1(j, s) = \{x | x^{(j)} \leq s\}$, $R_2(j, s) = \{x | x^{(j)} > s\}$.

Algorithm 1. Pseudocode for random forest prediction principles

```

1 : for  $t \leftarrow 1$  to  $n\_estimators$  do
2 : input : traing set  $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$ , among them  $x_i$ 
            $= (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(g)})$ ,  $y_i = (y_i^{(1)}, y_i^{(2)}, y_i^{(3)})$ 
3 : process : function  $TreeGenerate(D)$ 
4 : generate node
5 :  $PredictTotal \leftarrow 0$ 
6 :  $MinMse \leftarrow 0$ 
7 : fork  $\leftarrow 1$  to  $n$  do
8 : for  $f \leftarrow 1$  to  $g$  do
9 : if  $(\min_{x_i \in R_1(s_{k+1}, f+1)} \sum (y_i - c_1)^2 + \min_{x_i \in R_1(s_{k+1}, f+1)} \sum (y_i - c_i)^2)$ 
            $> (\min_{x_i \in R_1(s_k, f)} \sum (y_i - c_i)^2 + \min_{x_i \in R_1(s_k, f)} \sum (y_i - c_i)^2)$ 
10 : then  $MinMse \leftarrow \min_{x_i \in R_1(s_k, f)} \sum (y_i - c_i)^2 + \min_{x_i \in R_1(s_k, f)} \sum (y_i - c_i)^2$ 
11 :  $s \leftarrow k, j \leftarrow f$ ; generate a branch node  $node1$ ; through  $s, j$  turn  $D$  to  $D^1, D^2$ 
12 : endif
13 : endfor
14 : endfor
15 : if  $D^1, D^2 \neq \emptyset$  then
16 :  $TreeGenerate(D^1)$ 
17 :  $TreeGenerate(D^2)$ 
18 : else
19 : marks as a leaf node and assign a value of  $MinMse$ 
20 : endif
21 : // get single tree prediction  $PredictSingle$ 
22 :  $PredictTotal \leftarrow PredictSingle + PredictTotal$ 
23 :  $MeanPredictTotal \leftarrow PredictTotal / n$ 
24 : output :  $MeanPredictTotal$ 

```

5. Prediction Results of the Random Forest Regression Model

5.1. Evaluation Indicators

Let the difference between the actual coordinates of the horizontal, left and right, and up and down directions after drilling three meters forward and before drilling be $y = (y_1, y_2, y_3)$, and the random forest model obtains the predicted coordinate difference based on the azimuth and inclination of the current drilling point and the prediction of the bend as $\hat{y} = (\hat{y}_1, \hat{y}_2, \hat{y}_3)$, and this paper calculates the difference between y and \hat{y} R2 score, mean absolute error (MAE), mean square error (MSE) and root mean square error (RMSE) between the models to evaluate the prediction performance of the model [28].

The R2 score can be popularly expressed using the mean in each direction as the error reference to test whether the prediction error is greater than or less than the mean reference error:

$$R2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y}_i - y_i)^2} \tag{7}$$

The numerator represents the sum of the squared differences between the actual and predicted coordinate differences, and the denominator represents the sum of the squared differences between the actual and the mean values of the actual coordinate differences. The goodness of the model is judged according to the value of R-Squared, which is in the range of [0, 1]: if the calculated result is 0, the model cannot be correctly fitted to the data; if the calculated result is 1, the model is fitted to the data without error.

In general, a larger R-Squared indicates a better model fit. R-Squared reflects how approximately accurate it is. The R-Squared reflects how accurate the model is, and because the R-Squared is bound to increase as the sample size increases, it is not possible to truly quantify the degree of accuracy, but only approximately. Therefore, the MAE, MSE and

RMSE need to be considered together, and each indicator is calculated as follows, where m indicates the number of regression samples

$$\text{MAE} = \frac{1}{m} \sum_{i=1}^m |(y_i - \hat{y}_i)| \quad (8)$$

$$\text{MSE} = \frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2 \quad (9)$$

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2} \quad (10)$$

5.2. Experimental Prediction Results

In the following experimental results, the random forest method is adopted to model in three directions, and the mesh parameters of the model in three directions are optimized. Because the prediction task in each direction is equivalent to an independent problem, the construction of the three random forests is independent and has different parameters. All three models used the same experimental environment and the same experimental data set in the simulation process.

The prediction results in this paper are based on the 80 records generated during the actual construction process from 9 August to 20 August 2018. Under the condition of the elbow, inclination angle and azimuth of the current drilling position, the random forest regression model trained according to the historical data was applied, and on the basis of the optimal adjustment of the parameters, the horizontal position, left and right directions and the upper and lower positions of the next moment after turning into 3 m were obtained, and the specific results are shown in Figure 4 and Table 1.

Table 1. Data on machine learning evaluation indicators.

	Horizontal Direction	Up and Down Direction	Left and Right
R2 score	0.90	0.99	0.42
MAE	0.02	0.01	0.12
MSE	0.01	0.01	0.09
RMSE	0.06	0.04	0.30

The comparison between predicted and actual values revealed that the random forest regression model showed good predictions in all three directions, with a distribution of residuals similar to the standard normal distribution, concentrated between ± 0.2 m. The predicted residuals in both the upper and lower directions were concentrated between ± 0.05 m, meeting the requirement of < 0.2 m absolute value of deviation between construction results and predicted results for a single reference tool surface of 3 m.

The evaluation metric calculation by machine learning was able to quantify the prediction effectiveness of the random forest regression model in all three directions. According to the calculated values of the R2 score, MAE, MSE and RMSE, it can be obtained that overall the model has a good prediction effect, and the mean square error of the model in the horizontal direction, up and down direction, and left and right direction is less than 10%. The accuracy of the prediction in the horizontal direction and the up and down direction is above 90%, and the prediction in both directions is better than the prediction in the left and right directions.

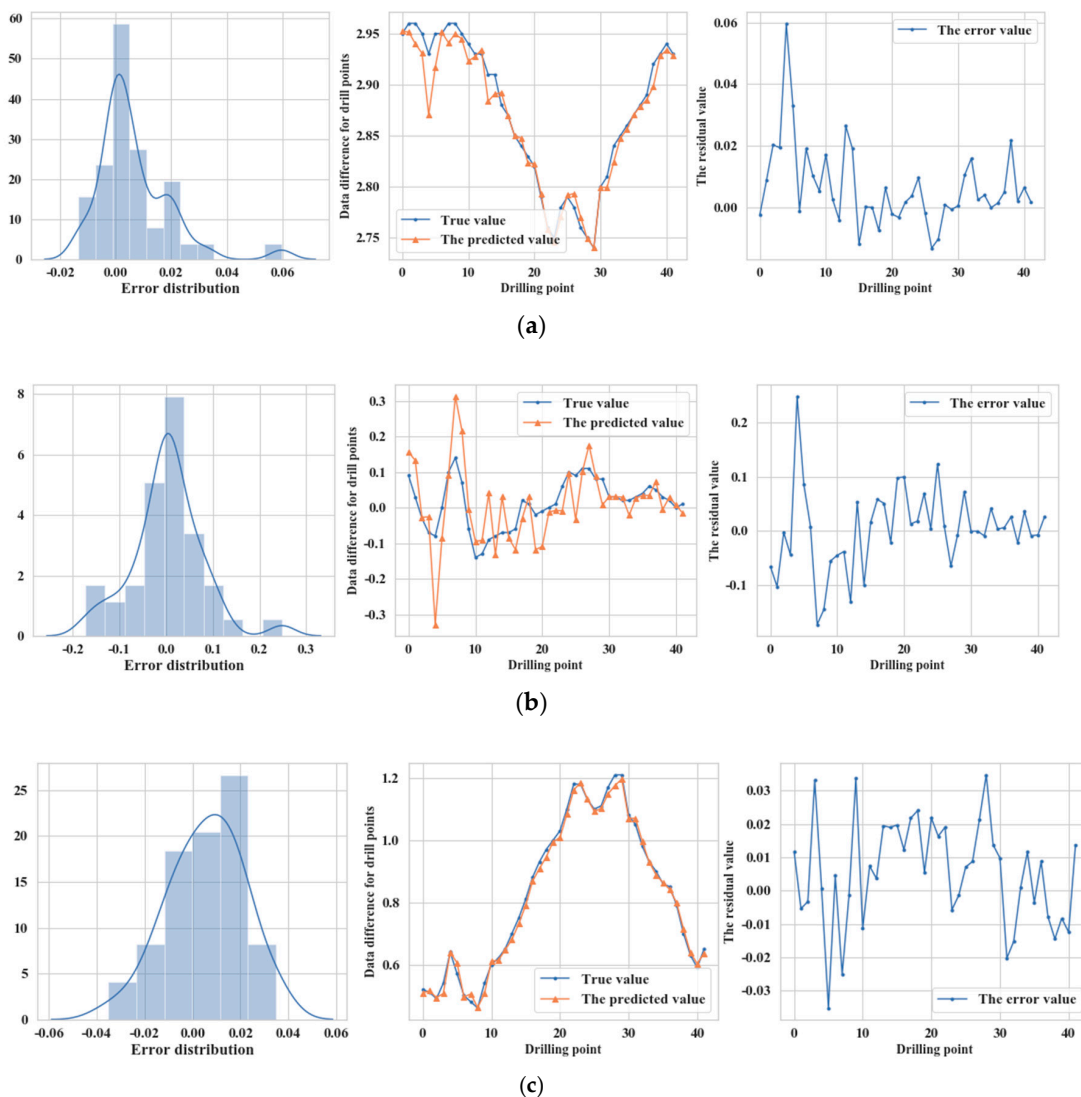


Figure 4. Residuals between predicted and actual values. (a) Horizontal. (b) Left and right direction. (c) Up and down direction.

5.3. Predicting Feature Importance

Due to the characteristics of the random forest regression model, through the model visualization, it can be intuitively concluded that the random forest regression model has different independent variable importance in three directions, as shown in Figure 5 below:

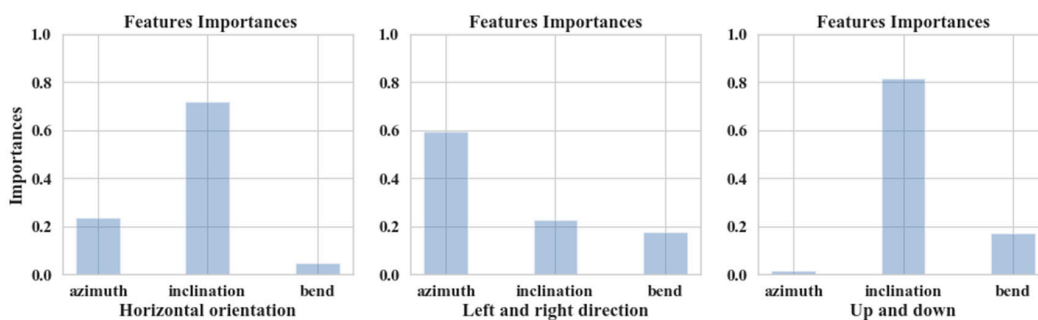


Figure 5. Predicted feature importance.

In the prediction process of the three directions, the random forest regression model automatically builds different random forests according to the different dependent variables that need to be predicted, and assigns corresponding weights to different independent variables [29], so as to achieve better prediction results.

6. Comparison of Random Forest Model Prediction Results with Other Models

6.1. Comparative Model for Borehole Trajectory Prediction

6.1.1. Gradient Boosting Regression

Gradient boosting regression (GBR) is a technique that learns from its mistakes. It is essentially integrating a bunch of poorer learning algorithms for learning, this algorithm was first proposed by Friedman in 2001 [30], and has been evolving in recent years and has been widely used in many fields.

6.1.2. K-Nearest Neighbor Regression

K-nearest neighbor regression is a common nonparametric regression method that finds the k nearest neighbors of a sample by searching a historical database [31], and by assigning the mean of some attribute(s) of these neighbors to the sample, the value of the corresponding attribute(s) of the sample can be obtained.

6.1.3. SVR Regression

For the SVR regression model [32], which has been widely used in recent years for regression prediction problems, this paper uses the kernel function of the SVR model as the Rbf kernel, and the parameters are adjusted by the grid search method

6.2. Comparison of Predicted Results

Since this article studies a typical regression problem, the training of regression models to achieve the purpose of prediction and the selection of models occupies its important position among the factors affecting the accuracy of prediction results, after considering the nature and principle of this research problem, four machine learning models with good generalization learning ability were selected: random forest regression model, gradient improvement model, k-nearest neighbor regression model and SVR regression model [33]. The models were trained using the existing data. In order to make a horizontal comparison of the four models, this paper uses the same empirical data to train the four models, and adjusts and optimizes the parameters of the above four models to ensure that the parameter combination of each model is optimized and reasonable when performing the regression task. The R2 scores of the prediction results of the training models and the residual mean absolute error (MAE) were mainly used as evaluation indicators to evaluate the models, and the results were as follows (Figure 6).

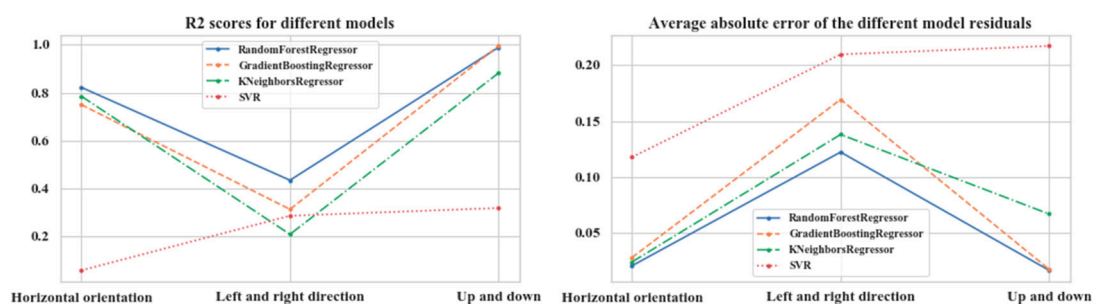


Figure 6. Comparison of regression model prediction results.

According to the results of the evaluation index, the random forest model has a higher R2 score and a lower mean absolute error of the residuals in the three directions of prediction compared to the other three regression models, indicating that for this problem under study, the random forest regression model has a better degree of explanation of

the dependent variable by the independent variables and a smaller residual between the predicted data and the true value, making it one of the better regression models for the research problem.

7. Limitations of the Approach in This Article

7.1. Algorithm Limitations

The random forest model involves many parameters, and the parameter value range and its broad, parameter tuning process can only try different parameter combinations and select the optimal parameter combination. This paper involves the grid parameter optimization method which is based on experience of different parameters after delineating the value range through the enumeration combination to find the optimal parameter combination method; this strategy of adjusting parameters will consume a lot of time which is one of the limitations of the random forest algorithm. In addition, although the random forest algorithm is fast enough, when there are many decision trees in the random forest, model training will consume a lot of time and space costs.

7.2. Practical Application Restrictions

From the perspective of the principle of the random forest algorithm, the random forest algorithm is an algorithm that summarizes experience from a large amount of data rather than a more essential principle analysis. In the actual drilling process, there may be different experiences reflected in the corresponding drilling trajectory data due to the different nature of rock formations at the construction site and the difference in geological structure, and training the random forest model with these different data together will lead to the weakening of important experience and reduce the accuracy of prediction. Therefore, it is also necessary to train models for different characteristics of rock formations according to different situations to improve accuracy. However, there are three main reasons for the actual realization of this work and its difficulties: 1. Complex and diverse rock formations and geological structures are difficult to classify and model; 2. Modeling each rock formation and the geological structure data will consume a lot of time and space; 3. The amount of data of some specific rock formations is too small to train an effective model.

8. Conclusions

Compared with the theoretical prediction calculation method proposed by traditional technicians, this paper proposes a trajectory prediction method based on random forest regression combined with machine learning, and compares it with the prediction results of other models such as gradient boosting regression, k-nearest neighbor regression and SVR regression. The results show that the proposed model is effective and accurate. The research of this paper is as follows:

(1) Study the method of predicting borehole trajectories based on random forest regression, and quantitatively determine the accuracy of random forest regression prediction by calculating evaluation indicators such as MAE, MSE, RMSE and R2 score;

(2) Based on the feature variable importance analysis method and the underlying theory of directional drilling, three feature variables, azimuth, inclination and elbow, and three key predictor variables, horizontal increment, left–right increment and up–down increment, were preferentially selected from all feature variables;

(3) The random forest regression model has a good fitting and prediction effect on the experimental data. After tuning the random forest parameters, the optimal parameters for fitting the data were found, so that the absolute value of the deviation between the construction results of the model's single reference tool surface of 3 m and the predicted results was <0.2 m. The random forest regression model based on the preferential selection of the characteristic variables has a high prediction accuracy and is suitable for borehole trajectory prediction.

For future research and work, we first look forward to continuing to refine the accuracy of the proposed model so that it can predict more accurate results. Secondly, we hope to be able to apply our model to more complex environments, because in actual orbit operation, geology is not only coal, but also possesses some other impurities, so we hope to propose a model that can adapt to more complex environments as soon as possible, greatly improving its practicality.

Author Contributions: Funding acquisition, W.X.; Investigation, B.Y., X.Z., C.T., X.W., Y.Y. and W.X.; Methodology, C.T., X.W., Y.Y. and W.X.; Project administration, B.Y., X.Z. and W.X.; Software, C.T., X.W. and Y.Y.; Supervision, W.X.; Writing—review and editing, C.T., X.W., Y.Y. and W.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Open Fund for State Key Laboratory of Gas Disaster Detecting, Preventing and Emergency Controlling, China (No. 2021SKLKF09).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

- Guo, H.; Wang, H.; Cheng, X. Data-driven fine-grained control of gas extraction boreholes based on data. *Min. Saf. Environ. Prot.* **2022**, *49*, 125–130. [[CrossRef](#)]
- Liang, Q.; Liu, X.; Shi, L.; Tao, Y.; Li, L.; Hu, G. Application of combined drilling mode of guide eye and expansion eye in directional well trajectory control. *Pet. Drill. Prod. Technol.* **2015**, *37*, 9–11. [[CrossRef](#)]
- Su, Y.; Bai, J. Three-dimensional analysis of bent-joint-downhole power drilling tool combinations by longitudinal and transverse bending method. *Acta Pet. Sin.* **1991**, *12*, 110–120.
- Wang, J. A mining with drilling measurement system. *Shandong Coal Sci. Technol.* **2016**, *34*, 132–133. [[CrossRef](#)]
- Huang, J.; Zhang, J. Overview of oil and gas production forecasting by machine learning. *Pet. Reserv. Eval. Dev.* **2021**, *11*, 613–620. [[CrossRef](#)]
- Wang, Z.; Zhu, H.; Xu, B.; Yan, D.; Du, H.; Luo, L.; Cui, Y. Mechanical Properties prediction of nuclear structural Materials based on Machine Learning Algorithm. *Shanghai Met.* **2022**, *44*, 102–110. [[CrossRef](#)]
- Wang, Z. Optimal Scheduling of Electric Heating Integrated Energy System Based on Machine Learning Prediction Algorithm. Master's Thesis, Shanxi University, Taiyuan, China, 2022.
- Song, L.; Meng, F.; Yuan, G. Moving object location prediction algorithm based on Markov model and trajectory similarity. *Comput. Appl.* **2016**, *36*, 39–43+65. [[CrossRef](#)]
- Chen, W.; Nan, P.; Yan, X.; Peng, J. Prediction and model optimization of coal mining machine cut-off trajectory based on deep learning. *J. Coal* **2020**, *45*, 4209–4215. [[CrossRef](#)]
- Meng, Q.; Liu, Q. A mathematical model for borehole trajectory prediction based on wavelet-neural network. *Mech. Des.* **2008**, *25*, 25–27. [[CrossRef](#)]
- Liu, L. Research on Sparse Trajectory Prediction Method Based on Entropy Estimation. Master's Thesis, China University of Mining and Technology, Xuzhou, China, 2017.
- Sun, R. Description and calculation method of near-horizontal directional borehole trajectory in underground coal mine. *China Coalbed Methane* **2010**, *7*, 30+36–39. (In Chinese) [[CrossRef](#)]
- Guo, Y. Prediction and Update of LWD Curve While Drilling. *Pet. Drill. Tech.* **2010**, *38*, 25–28. [[CrossRef](#)]
- Feng, D. Research on directional pore-forming technology of accurate drainage hole of tunneling roadway in underground coal mine. *Ore Explor. Eng. (Rock Soil Drill. Eng.)* **2018**, *45*, 33–36+41. [[CrossRef](#)]
- Sun, T.; Lin, L.; Liu, Z.; Song, J.; Wang, X. Research on trajectory 28prediction method of directional borehole in underground coal mine. *Coal Mine Min.* **2019**, *24*, 15+22–25. [[CrossRef](#)]
- Yang, X.; Chuan, Y.; He, S.; Jiang, D.; Cao, B.; Wang, S. Machine learning prediction of specific capacitance in biomass derived carbon materials: Effects of activation and biochar characteristics. *Fuel* **2023**, *331*, 125718. [[CrossRef](#)]
- Sheikhi, A.; Mesiar, R.; Holeňa, M. A dimension reduction in neural network using copula matrix. *Int. J. Gen. Syst.* **2022**. [[CrossRef](#)]
- Chen, Z.; Zhang, L.; Li, K.; Xue, X.; Zhang, X.; Kim, B.; Li, C.Y. Machine-learning prediction of aerodynamic damping for buildings and structures undergoing flow-induced vibrations. *J. Build. Eng.* **2023**, *63*, 105374. [[CrossRef](#)]
- Xu, W.; Guo, D.; Mi, J.; Qian, Y.; Zheng, K.; Ding, W. Two-way concept-cognitive learning via concept movement viewpoint. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, 1–15. [[CrossRef](#)]

20. Xu, W.; Guo, D.; Qian, Y.; Ding, W. Two-way concept-cognitive learning method: A fuzzy-based progressive learning. *IEEE Trans. Fuzzy Syst.* **2022**, *1*–15. [[CrossRef](#)]
21. Zhang, X.; Guo, D.; Xu, W. Two-way Concept-Cognitive Learning with Multi-source Fuzzy Context. *Cogn. Comput.* **2023**. [[CrossRef](#)]
22. Pareek, R. Technology and healthcare (machine learning). *PC Quest* **2018**, *31*, 50–51.
23. Chen, S.; Sun, W.; He, Y. Application of Random Forest Regressions on Stellar Parameters of A-type Stars and Feature Extraction. *Astron. Astrophys. Res.* **2022**, *22*, 189–194. [[CrossRef](#)]
24. Li, H.; Xia, D.; Wang, Q. A regression model-based cleaning technique for acquisition data. *Electro-Opt. Control.* **2022**, *29*, 117–120. [[CrossRef](#)]
25. Liu, Z.; Li, Z.; Wang, L.; Wang, T.; Yu, H. Enhancement and extension of forest optimization feature selection algorithm. *J. Softw.* **2020**, *31*, 1511–1524. [[CrossRef](#)]
26. Liu, H.; Tao, J.; Qiu, L. Implementation of Python-based One-hot coding. *J. Wuhan Shipbuild. Vocat. Technol. Coll.* **2021**, *20*, 136–139. [[CrossRef](#)]
27. Liu, D.; Sun, K. Random forest solar power forecast based on classification optimization. *Energy* **2019**, *187*, 115940.1–115940.11. [[CrossRef](#)]
28. Handelman, G.S.; Kok, H.K.; Chandra, R.V.; Razavi, A.H.; Huang, S.; Brooks, M.; Lee, M.J.; Asadi, H. Peering into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods. *Am. J. Roentgenol.* **2019**, *212*, 38–43. [[CrossRef](#)]
29. Li, N.; Wang, Y.; Zhou, L.; Zou, C.; Tian, Y.; Guo, N. A random forest detection method for DDoS attacks based on secondary filtering of feature importance. *Comput. Sci.* **2021**, *48*, 464–467+476. [[CrossRef](#)]
30. Friedman, J. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [[CrossRef](#)]
31. Tanveer, M.; Shubham, K.; Aldhaifallah, M.; Ho, S. An efficient regularized K-nearest neighbor based weighted twin support vector regression. *Knowl.-Based Syst.* **2016**, *94*, 70–87. [[CrossRef](#)]
32. Lin, Y. An airport track prediction method based on SVR regression. *Inf. Commun.* **2019**, *32*, 58–59. [[CrossRef](#)]
33. Chen, R.; Chen, X.; Dong, M.; Liang, J.; Li, Y.; Wang, G.; Liu, X. Machine learning-based lifetime and trajectory prediction of marine buoys. *Ocean. Bull.* **2021**, *40*, 262–273. [[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.