applied sciences

MDPI

Article

Hybrid Model of Machine Learning Method and Empirical Method for Rate of Penetration Prediction Based on Data Similarity

Fei Zhou [1], Honghai Fan [1,*], Yuhan Liu [2], Hongbao Zhang [1,3] and Rongyi Ji [1]

[1] School of Petroleum Engineering, China University of Petroleum, Beijing 102249, China; 2018312012@student.cup.edu.cn (F.Z.); zhanghb.sripe@sinopec.com (H.Z.); jirongyi@cup.edu.cn (R.J.)
[2] CNPC Engineering Technology R&D Company Limited, Beijing 102206, China; liuyuhandr@cnpc.com.cn
[3] SINOPEC Research Institute of Petroleum Engineering, Beijing 102206, China
* Correspondence: fanhh@cup.edu.cn; Tel.: +86-010-89733221

Abstract: The rate of penetration (ROP) is an important indicator affecting the drilling cost and drilling performance. Accurate prediction of the ROP has important guiding significance for increasing the drilling speed and reducing costs. Recently, numerous studies have shown that machine learning techniques are an effective means to accurately predict the ROP. However, in petroleum engineering applications, its robustness and generalization cannot be guaranteed. The traditional empirical model has good robustness and generalization ability. Based on the quantification of data similarity, this paper establishes a hybrid model combining a machine learning method and an empirical method, which combines the high prediction accuracy of the machine learning method with the good robustness and generalization of the empirical method, overcoming the shortcomings of any single model. The AE-ED (the Euclidean Distance between the input data and reconstructed data from the autoencoder model) is defined to measure the data similarity, and according to the data similarity of each new piece of input data, the hybrid model chooses the corresponding single model to calculate. The results show that the hybrid model is better than any single model, and all the evaluation indicators perform better, making it more suitable for the ROP prediction in this field.

Keywords: machine learning; intelligence well; data similarity; rate of penetration (ROP)

Citation: Zhou, F.; Fan, H.; Liu, Y.; Zhang, H.; Ji, R. Hybrid Model of Machine Learning Method and Empirical Method for Rate of Penetration Prediction Based on Data Similarity. *Appl. Sci.* **2023**, *13*, 5870. https://doi.org/10.3390/app13105870

Academic Editors: Martin Ma, Hongbin Yang and Wei Zhou

Received: 17 April 2023
Revised: 6 May 2023
Accepted: 8 May 2023
Published: 10 May 2023

1. Introduction

Machine learning represented by deep learning has achieved great success in multiple fields [1], such as natural language processing [2], computer vision, quantitative transactions [3], and autonomous driving. These successful cases have attracted widespread attention in academia and industry, which has also promoted the advancement of machine learning technology. With the continuous development of basic theory, algorithm platforms, and computing power, the application of machine learning in the industry also has a good foundation. In the field of the petroleum industry, machine learning technology has also begun to be widely used [4], including formation parameter prediction, drilling equipment fault detection, the diagnosis of complex situations under the well, drilling parameter optimization, drilling fluid design, and hydraulic calculation [5,6]. These applications have provided important technical support and are a guarantee for the development and optimization of the oil industry.

The rate of penetration (ROP) is an important indicator affecting the drilling cost and drilling performance. Accurate prediction of the ROP has important guiding significance for increasing the drilling speed and reducing costs. However, drilling engineering operations are complex processes, and the formation, tools, and operating parameters are all important factors affecting the rock breaking by the drill bit [7], such as the formation drillability, drill-bit wear, revolutions per minute, mud performance, weight on bit, etc. [8]. Therefore, it is

difficult to accurately predict the ROP. Over the years, technologists around the world have proposed a variety of methods for ROP prediction using logging data, drilling data, and engineering records, mainly including empirical methods and machine learning methods. Compared with empirical methods, machine learning algorithms have obvious technical advantages, especially in high-dimensional nonlinear regression and prediction.

In 1974, Bourgoyne et al. [9] established a classic ROP prediction model, known as the B-Y model, based on previous experience, combined with field construction data, and considering factors such as the weight on bit, compaction effect, pressure difference, and rotary speed. In 1994, Harelend et al. [10] proposed an ROP prediction model considering various drill-bit feature parameters. In 2010, Motahhari et al. [11] proposed an ROP prediction method suitable for any PDC bit on the basis of Hareland's research, and they achieved good application results. In 2019, Xu et al. [12] and Su et al. [13] used integrated algorithms to establish ROP prediction models, and they used the goodness of fit as the evaluation indicator of the ROP prediction performance. Liu et al. [14] and Diaz et al. [15] used the artificial neural network (ANN) algorithm to establish an ROP prediction model for directional wells. Assuming that the amount of data is sufficient, the ROP prediction accuracy can meet the requirements of the field. In 2021, Hazbeh et al. [16] and Lawal et al. [17] used a variety of genetic algorithms to optimize the neural network algorithm, and compared and analyzed the effect of the algorithm from multiple aspects, such as the relative error, R square (R^2), and mean square error (MSE). Husam et al. [18] used the circular neural network algorithm to predict the rate of penetration based on well logging data and mud logging data, with an accuracy of up to 85%. In 2022, Zhou et al. [19] established a prediction model of the ROP based on support vector regression (SVR), and they proposed an improved algorithm to solve the nonconvex problem of determining the optimal values of the model hyperparameters. In 2023, Ren et al. [20] proposed an ROP prediction model based on stacking ensemble learning, and the prediction accuracy rate in specific oilfields reached 92.5%.

There are two main types of existing ROP prediction models: empirical models and machine learning models. Generally speaking, empirical models have better robustness and generalization, and machine learning models have higher prediction accuracy. Due to the strong nonlinear fitting ability of artificial intelligence algorithms, it has gradually become a major trend to use intelligent algorithms to predict the ROP. However, there is a problem that is easily overlooked: machine learning methods are usually based on the assumption of independent and identical distribution (IID) and test set data with high data distribution similarity, and the prediction accuracy of the machine learning model is very high; however, for data with low data distribution similarity, the prediction effect becomes unstable. In drilling engineering, due to the uncertainty of the downhole construction conditions, the complexity of the downhole geological conditions, and the uncertainty of drilling tool assemblies and drilling parameters, the similarity of the data distribution between the test set and training set cannot be guaranteed. Therefore, the prediction performance of the machine learning model is also good and bad.

In order to solve this problem, based on the quantification of data similarity, this paper establishes a hybrid model combining a machine learning method and an empirical method, which combines the high prediction accuracy of the machine learning model with the good robustness and generalization of the empirical model, overcoming the shortcomings of any single model. The AE-ED (the Euclidean Distance between the input data and reconstructed data from the autoencoder model) is defined to measure the data similarity, and according to the data similarity of each new piece of input data, the hybrid model chooses the corresponding single model to calculate. Data with high degrees of data similarity have smaller AE-ED values; on the contrary, data with low degrees of data similarity have larger AE-ED values. For the data to be predicted with higher data similarity, a machine learning model with a higher prediction accuracy is adopted, and for the data to be predicted with low data similarity, the empirical model with better robustness is adopted. The hybrid model overcomes the disadvantages of the poor generalization of the machine learning

model, combines the advantages of the empirical model and machine learning model, and has better prediction accuracy, robustness, and generalization.

2. Theory Background

2.1. Independent and Identical Distribution (IID) Assumption

The IID assumption is common in machine learning, especially in supervised learning, and it assumes an independent and identical distribution between each sampling [21]. The IID assumption is a basic guarantee that the model obtained through the training data can achieve good results in the test set. If we want to use a dataset to train a model, then we usually split the dataset into a training set and test set, assuming that both datasets are independent and identically distributed. This assumption allows us to consider a model's performance on the test set as a good estimate of its generalization ability on future data.

However, in drilling engineering, due to the complexity of the downhole geological conditions, the uncertainty of the downhole construction conditions, and the uncertainty of drilling tool assemblies and drilling parameters, the similarity of the data distribution between the test set and training set cannot be guaranteed. Therefore, the prediction performance of the machine learning model is also good and bad.

This research solves just this problem. In the newly input data waiting to be calculated, for data with high similarity, there is no doubt that the machine learning model should be used, and for data with low similarity, the IID assumption is not guaranteed, and the empirical model with better robustness and generalization should be selected. The method of measuring data similarity is given below.

2.2. Autoencoder (AE) and AE-ED Definition

An autoencoder (AE) is a common unsupervised learning algorithm [22]. It is able to learn a compressed representation from the input data while preserving as much information as possible from them.

As shown in Figure 1, an autoencoder (AE) is a special artificial neural network that has two main components: an encoder and a decoder. The function of the encoder is to extract and compress the original high-dimensional input data and map the input vector to a feature space. The function of the decoder is to reconstruct the high-dimensional data according to the feature information. The input data have the same dimensions as the output data reconstructed by the AE, and the label when training the network is the input vector itself.

The solution formula of the AE is as Equation (1):

$$f, g = \text{argmin}_{f,g} \| X - g[f(X)] \|^2 \tag{1}$$

where f is the mapping from the input space to the feature space, g is the mapping from the feature space to the input space, and X is the input data.

The execution process of the autoencoder can refer to the pseudocode in Algorithm 1.

Algorithm 1: Autoencoder

Input: dataset $X^{(1)}, X^{(2)}, \ldots X^{(N)}$.
Output: encoder f_φ, decoder g_θ.

Random initialization parameters φ, θ
repeat
    calculate the reconstruction error $E = \sum_{i=1}^{N} \left\| X^{(i)} - g_\theta[f_\varphi(X^{(i)})] \right\|^2$
    minimizing reconstruction error using gradient descent
    update parameters φ, θ
until φ, θ parameters convergence.

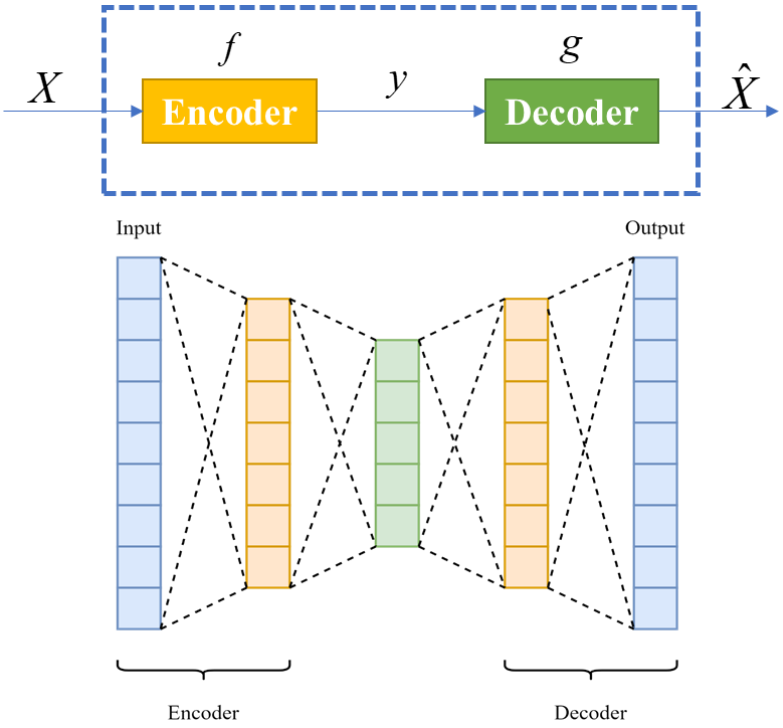

Figure 1. Autoencoder network structure.

The autoencoder (AE) can reconstruct data. In the process of the encoder mapping the input data to the feature space, the characteristic information of the data, including the data distribution, is stored in this feature space. The process of decoding is similar to reconstructing data according to the characteristic information.

Throughout the process, the AE plays the role of feature extractor. An AE is a data-dependent model; that is to say, it can only handle data similar to the training data, while dissimilar data cannot guarantee the model performance. For new input data with high data similarity, the AE can reconstruct the data very well, and the reconstructed data are very close to the original data. For new input data with low data similarity, the AE performs poorly because the features it learns are about the dataset used to train the AE.

In order to quantify the difference between the new input data and reconstructed data, and also to quantify the data similarity, we define a metric named the AE-ED, which stands for "the Euclidean Distance between input data and the reconstructed data from the autoencoder model".

The calculation method of the AE-ED is as Equation (2):

$$\text{AE-ED} = \sqrt{\left(X^R - \hat{X}^R\right)^2} \tag{2}$$

where X^R is a piece of new input data with the dimension R, and $\hat{X}^R$ is the R-dimensional data reconstructed by the autoencoder (AE).

Data with high degrees of data similarity have smaller AE-ED values; on the contrary, data with low degrees of data similarity have larger AE-ED values.

2.3. Empirical Models

Hareland and Motahhari, two esteemed researchers, have proposed accurate and effective ROP prediction models that are widely used. Whether it is the Hareland model or Motahhari model, they are both applicable to any PDC drill bit. The Hareland model is based on theoretical considerations of single-cutter rock interactions, lithology coefficients, and bit wear. Motahhari improved the model based on Hareland's research. He simplified the calculation of the model coefficients related to the drill-bit structure and optimized the

calculation formula of the bit-wear coefficient, which makes the model more universal and accurate. These two esteemed scholars have had a major impact on the understanding of how PDC bits act on rock and the associated wear and ROP. Therefore, we chose these two models as typical representatives of empirical models.

2.3.1. Hareland Model

In 1994, Harelend et al. [10] proposed an ROP prediction model considering various drill-bit feature parameters. The model is suitable for all types of drag bits, including natural diamond bits (NOB) and polycrystalline diamond compact (PDC) bits.

$$ROP = \frac{a_c}{RPM^b WOB^c} * \frac{14.14 N_c RPM}{D_B} A_v \tag{3}$$

$$A_v = \cos\theta * \sin\theta \left[\begin{array}{c} \left(\frac{d_c}{2}\right)^2 \cos^{-1}\left(1 - \frac{2P_c}{\cos\theta * d_c}\right) - \\ \left(\frac{d_c P_c}{\cos\theta} - \frac{P_c^2}{(\cos\theta)^2}\right)^{0.5} \left(\frac{d_c P_c}{2\cos\theta}\right) \end{array} \right] \tag{4}$$

$$P_c = \frac{2W_{mech}}{\pi d_c \sigma_c} \tag{5}$$

where N_c is the number of cutting teeth; A_v is the rock-breaking area of a single cutting tooth, in ^{2}; α is the side rake angle, $^\circ$; θ is the back rake angle, $^\circ$; d_c is the diameter of the cutting tooth, in; σ_c is the uniaxial compressive strength, psi; W_f is the wear degree of the bit; a_c, b, c are the correction coefficients of the cutting-tooth structure.

This model considers many factors in the characteristics of the bit, and it can theoretically optimize the bit structure. However, the relatively detailed structural parameters of the drill bit increase the quality requirements of the field data, which limits the application and promotion of this method.

2.3.2. Motahhari Model (Mota)

In 2010, Motahhari et al. [11] proposed an ROP prediction method suitable for any PDC bit on the basis of Hareland's research, and they achieved good application results. Compared with the actual drilling data, the error was small, and the field application effect was very good.

$$ROP = W_f\left(\frac{\beta * RPM^\gamma * WOB^\alpha}{D_b * CCS}\right) \tag{6}$$

where β is a model coefficient related to the bit structure, and γ and α are general model coefficients.

2.4. Machine Learning Model

The following are the three machine learning models established in this paper: RF, ANN, and SVM.

2.4.1. Random Forest (RF)

As shown in Figure 2, random forest is an ensemble method based on decision trees. The basic idea is to combine multiple weak classifiers into a strong classifier [23]. In the regression and prediction performance of industrial data, random forest has high prediction accuracy, strong generalization ability, low sensitivity to outliers and noise, few hyperparameters, and easy-to-adjust parameters, and it is widely used in various industrial scenarios.

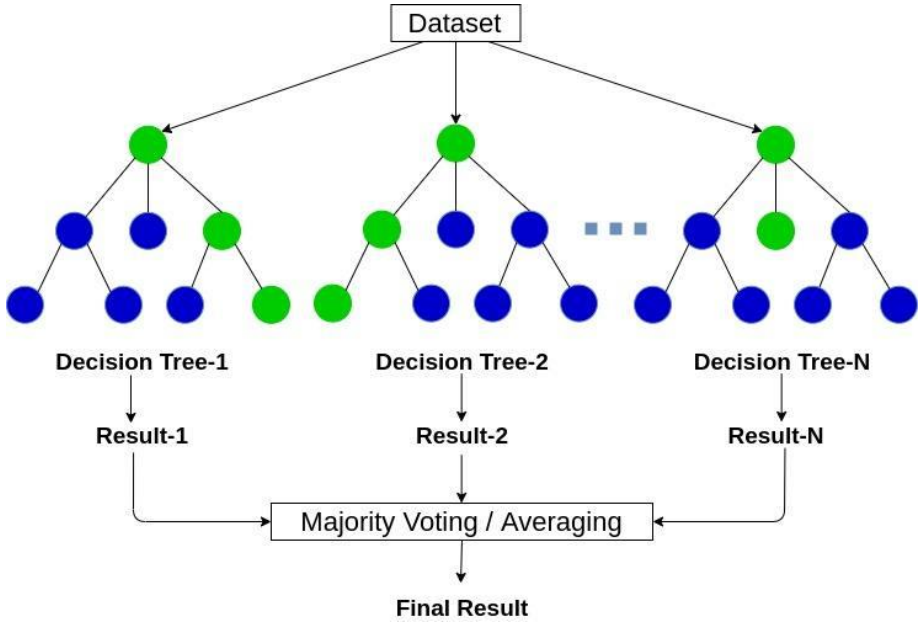

Figure 2. Random forest structure.

2.4.2. Artificial Neural Network (ANN)

The ANN can improve the data representation and function fitting ability of the model by increasing the number of hidden layers [24]. The multilayer network structure is helpful for the extraction and representation of input features, but it also leads to a sharp increase in the number of parameters of the model, which greatly increases the training time of the model, which is not conducive to industrial applications.

As shown in Figure 3, in order to avoid too many model parameters and overly long model training times, this paper chooses the back propagation (BP) neural network with four hidden layers.

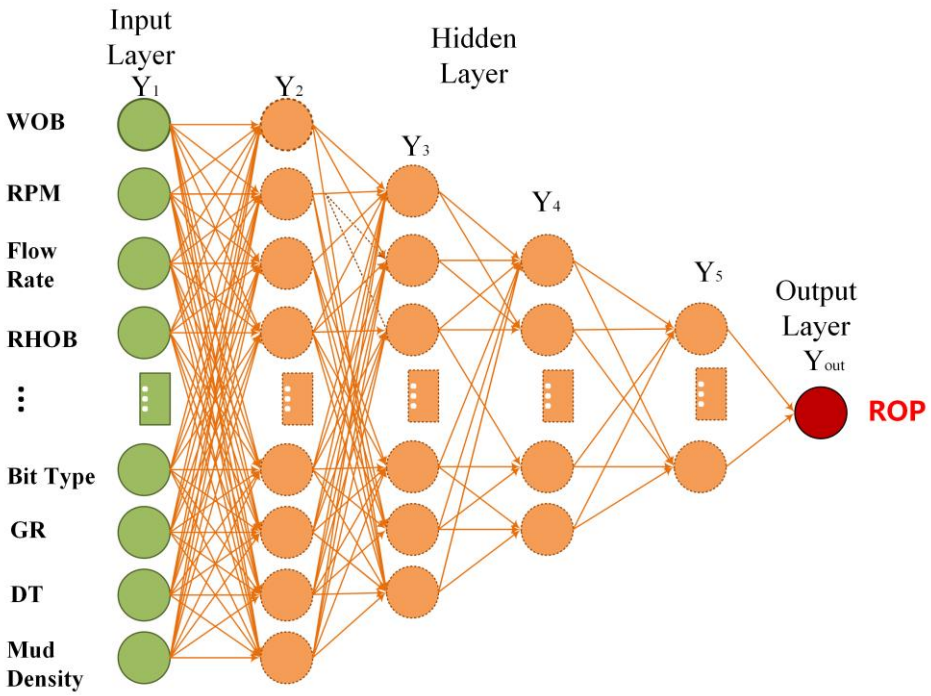

Figure 3. ANN network structure.

The calculation is shown in Equation (7):

$$y_i = \varphi_{active}(w^T x_i + b) \tag{7}$$

where φ_{active} is the activation function, w^T is the weight matrix, and b is the bias matrix.

Combined with Figure 3, the formula has a more understandable form:

$$\begin{aligned} Y_1 &= X_{in} \\ Y_2 &= \varphi_{active}(W_2 Y_1 + B_1) \\ Y_i &= \varphi_{active}(W_i Y_{i-1} + B_{i-1}) \\ Y_{out} &= Y_6 = W_6 Y_5 + B_5 \end{aligned} \tag{8}$$

where Y_i is the output matrix of the i-th layer.

The execution process of the ANN can refer to the pseudocode in Algorithm 2.

Algorithm 2: ANN

Input: Training set $D = \left\{(x^{(n)}, y^{(n)})\right\}$, validation set V, learning rate η.
Output: weight matrix W, bias matrix b.

```
Random initialization parameters W, b
Repeat for D
    random reordering of training set samples.
    feedforward computation
    error backpropagation
    update parameters W, b
until the error of the neural network on the validation set does not drop anymore
```

2.4.3. Support Vector Machine (SVM)

SVM is a supervised learning algorithm that is widely used in image classification, text classification, bioinformatics, and other fields [25].

As shown in Figure 4, SVM can convert the input feature space to a high-dimensional space through the nonlinear transformation of the kernel function, and it can find the optimal classification hyperplane in the high-dimensional space so as to maximize the distance between different types of data points and the hyperplane.

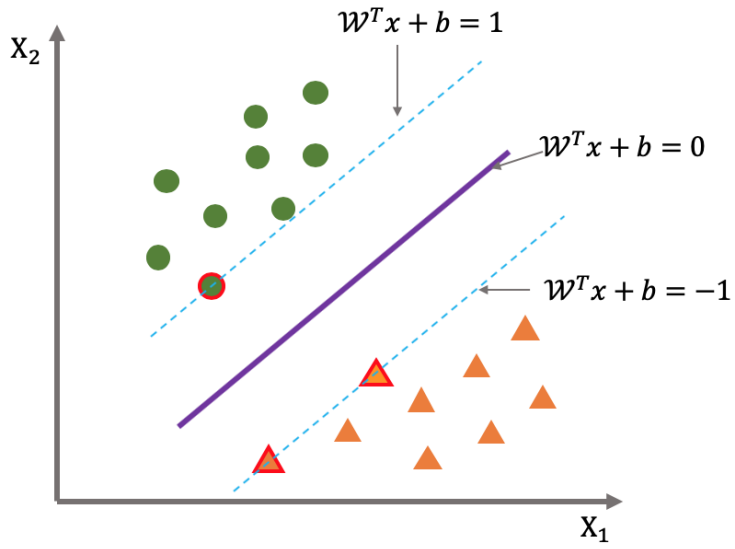

Figure 4. Support vector machine overview.

The SVM of the kernel function has a large computational complexity under the large amount of data. In order to reduce the calculation time, this paper chooses the linear support vector machine.

The solution of the SVM is to solve the optimization problem expressed by Equation (9):

$$\min \frac{1}{2}\|w\|^2 s.\ t.\ y_i(w^T x_i + b) \geq 1 \tag{9}$$

3. Data Description and Data Preprocessing

3.1. Data Description

In this paper, we used data from 70 wells from a carbonate field in the Middle East to train and test the model, including the drill-bit parameters, mud logging data, and geomechanical parameters, such as the standpipe pressure, bit type, density, uniaxial compressive strength, pore pressure, gamma rays, porosity, well depth, weight on bit, torque, revolutions per minute (RPM), mud density, equivalent circulating density, friction resistance, and flow rate.

Table 1 shows part of the data display used in this paper. There are 348,702 pieces of data in total. The means, standard deviations (std), minimum values (min), maximum values (max), and 25%, 50%, and 75% quantiles are shown in the table. The minimum depth is 1504 m, and the maximum is 4102.9 m. The minimum ROP is 0.1758 m/h, the maximum is 50 m/h, and the average is 8.12864 m/h. Among them, the data of the ROP were obtained through calculation. The original data include the time spent drilling per meter, and we calculated its reciprocal as the average ROP per meter.

Table 1. Data description.

	Depth	**DT (Interval Transit Time)**	**GR (Gamma Ray)**	**RHOB (Rho Bulk)**	**WOB (Weight on Bit)**	**ROP**	... ...
Unit	**(m)**	**(us/ft)**	**(API Unit)**	**(g/cm^3)**	**(Equivalent Weight, tons)**	**(m/h)**	
count	348,702	348,702	348,702	348,702	348,702	348,702	... ...
mean	2742.849	67.98826	22.39776	2.498545	4.939172	8.12864	... ...
std	350.797	9.625563	14.5044	0.11308	2.063821	5.39589	... ...
min	1504	41.8162	0.59623	1.72089	0.01	0.1758	... ...
25%	2505.7	62.2781	10.2852	2.42873	3.51	4.61538	... ...
50%	2732.3	66.6066	20.4368	2.52209	4.93	7.01877	... ...
75%	2946.7	71.0860	31.5935	2.57843	6.2	10.02	... ...
max	4102.9	119.963	398.004	3.00439	41.17	50	... ...

3.2. Data Preprocessing

Due to the large differences in the sizes and value ranges of the different parameters and the inability to represent categorical variables, the original data cannot be directly input into the model, and data preprocessing is required. As shown in Figure 5, the data preprocessing process includes data normalization, one-hot encoding, and feature selection.

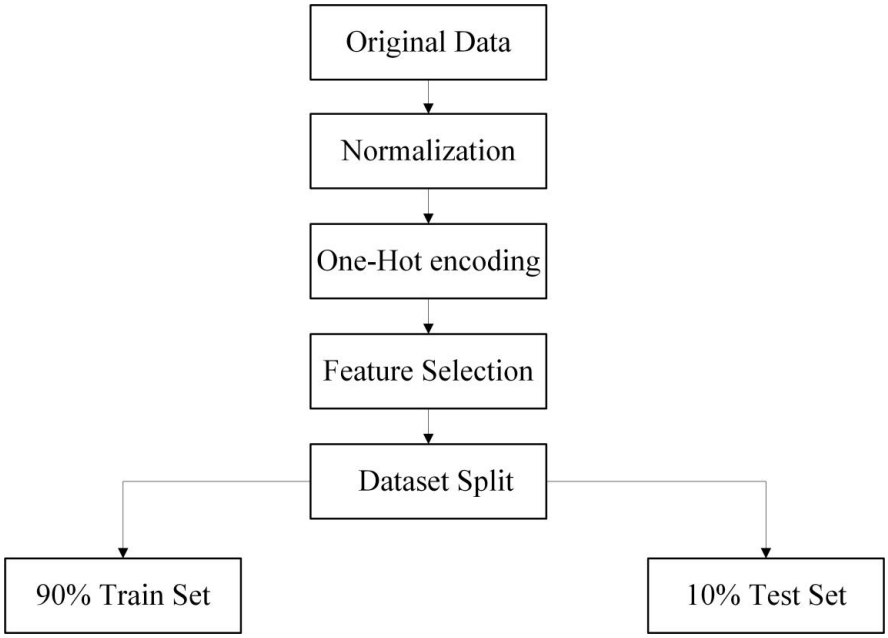

Figure 5. Flowchart of data preprocessing.

3.2.1. Data Normalization

The dimensions and value ranges of the different parameters vary greatly, and so they cannot be compared horizontally, nor can they be directly input into the model.

In order to ensure that the dimensions and value ranges between different features will not have an adverse effect on the model training, the input data need to be normalized before they are input into the model for training. This paper chooses the min–max normalization method to process the data, thus mapping the data to the range of [0, 1].

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{10}$$

where x is the original data, x^* is the data after normalization, and $x_{\min}$ and $x_{\max}$ are the minimum and maximum values of the data, respectively.

Processing data with minimum–maximum standardization solves the problem of large differences in dimensions and value ranges between different features, making different features comparable and helping to improve the training speed and model performance.

3.2.2. One-Hot Encoding

The original data contain categorical variables that cannot be directly input into the model, such as the bit type and formation lithology. Categorical variables need to be processed. One-hot encoding can convert categorical variables into a numerical form that can be used in machine learning algorithms, and it can enable algorithms to better understand categorical variables.

However, it has the potential to greatly increase the data dimensionality, especially for categorical variables with large numbers of categories. To avoid such problems, this paper only uses one-hot encoding for the bit types.

The bits mentioned in this article are all PDC bits, but they have different specifications, sizes, and nomenclatures. They need to be encoded, and the encoded forms are as follows:

$$\begin{aligned} \text{T1665R} &\rightarrow [1,0,0,0,0] \\ \text{KM1662} &\rightarrow [0,1,0,0,0] \\ \text{M1665S} &\rightarrow [0,0,1,0,0] \\ \text{ST915} &\rightarrow [0,0,0,1,0] \\ \text{ST615RS} &\rightarrow [0,0,0,0,1] \end{aligned} \tag{11}$$

where these bits are named differently by the IADC. They come from different manufacturers and are named according to the manufacturers' own rules. The first (or first two letters) in the bit code refer to the manufacturer, and the rest refer to the bit properties, such as the cutter size and number of blades.

3.2.3. Feature Selection

During the drilling process, there are many factors affecting the ROP, including the drilling parameters, drilling fluid properties, formation properties, wellbore geometry, drill-bit properties, drilling tool properties, etc. We know that if there are too many features, then the complexity of the model will be greatly increased. At the same time, adding features with low correlation will affect the accuracy of the model instead. Feature selection before training the model can improve the accuracy and efficiency of the model while reducing the number of features, reducing the complexity of the model, and avoiding overfitting. Therefore, we need to select features with higher correlation.

Through correlation analysis, we only select parameters that have a greater correlation with the ROP to input into the model, filter or reduce variables with low correlation, and avoid introducing multiple parameters with high similarity at the same time.

For proper feature selection, this paper chooses to use the Pearson correlation coefficient for correlation analysis. It can be used to measure the correlation between two variables, with values ranging from -1 to 1.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \tag{12}$$

where $\rho_{X,Y}$ is the Pearson correlation coefficient; X and Y are two random variables; σ_X and σ_Y are the standard deviations of X and Y, respectively; $\operatorname{cov}(X,Y)$ is the covariance between X and Y; $\overline{X}$ and $\overline{Y}$ are the means of X and Y, respectively; X_i and Y_i are the samples of X and Y, respectively.

Generally, features with a Pearson correlation coefficient greater than 0.3 are considered correlated. We filtered out the features with a Pearson correlation coefficient less than 0.3, which finally left 13 input parameters, as shown in Table 2.

Table 2. The input parameters of the model.

Categories	Input Parameters
Mechanical parameters	Well depth, weight on bit (WOB), revolutions per minute (RPM)
Hydraulic parameters	Inlet flow rate, mud density
Formation parameters	Density, gamma rays
Bit parameters	Bit type
Other logging parameters	interval transit time (DT), caliper log (CAL), Standpipe pressure (SPP), Rho bulk (RHOB), hookload

3.2.4. Dataset Split

In this paper, the data after data preprocessing are divided into training sets and test sets, of which 90% are training sets and 10% are test sets. The training sets were used to establish and train the models, and the test sets were used to test and evaluate the models.

4. Model Establishment and Evaluation

The research in this paper is based on the data similarity.

As mentioned in Section 2.1, machine learning methods are usually based on the IID assumption. If the test set data have high data distribution similarity, then the prediction accuracy of the machine learning model is very high; however, for the data with low data

distribution similarity, the prediction effect becomes unstable. In drilling engineering, due to the complexity and uncertainty of the downhole geological conditions, the similarity of the data distribution between the test set and training set cannot be guaranteed.

Therefore, for newly input data, it is very important to select an appropriate calculation model according to the data similarity. For the data to be predicted with higher data similarity, the machine learning model with higher prediction accuracy should be adopted, and for the data to be predicted with low data similarity, the empirical model with better robustness should be adopted. Therefore, how to measure the data similarity is the key issue.

4.1. Model Establishment

4.1.1. AE Model and AE-ED

As mentioned in Section 2.2, an autoencoder (AE) can reconstruct data. In the process of the encoder mapping the input data to the feature space, the characteristic information of the data, including the data distribution, is stored in this feature space. The process of decoding is similar to reconstructing data according to the characteristic information.

The AE is the feature extractor, which can extract representative features from the original data, used to describe the properties, structures, and characteristics of the data. After training the AE model using the training set, the characteristic information of the training set is saved in the AE model, which can reconstruct the training set data very well. For each new piece of input data of the test set, we can feed it into the AE model to reconstruct the data. If the reconstructed data is very different from the new input data, then the data similarity between the new input data and the training set is low; otherwise, the data similarity is high. In order to quantify the difference between the new input data and reconstructed data, we define a metric named the AE-ED, which stands for "the Euclidean Distance between input data and the reconstructed data from the autoencoder model". The calculation formula of the AE-ED is given in Section 2.2.

We chose continuous well interval data to test and evaluate the model, and the data used are all from Well X in the Middle East. First, we established the AE model and calculated the data similarity of Well X. The results are shown in Figure 6.

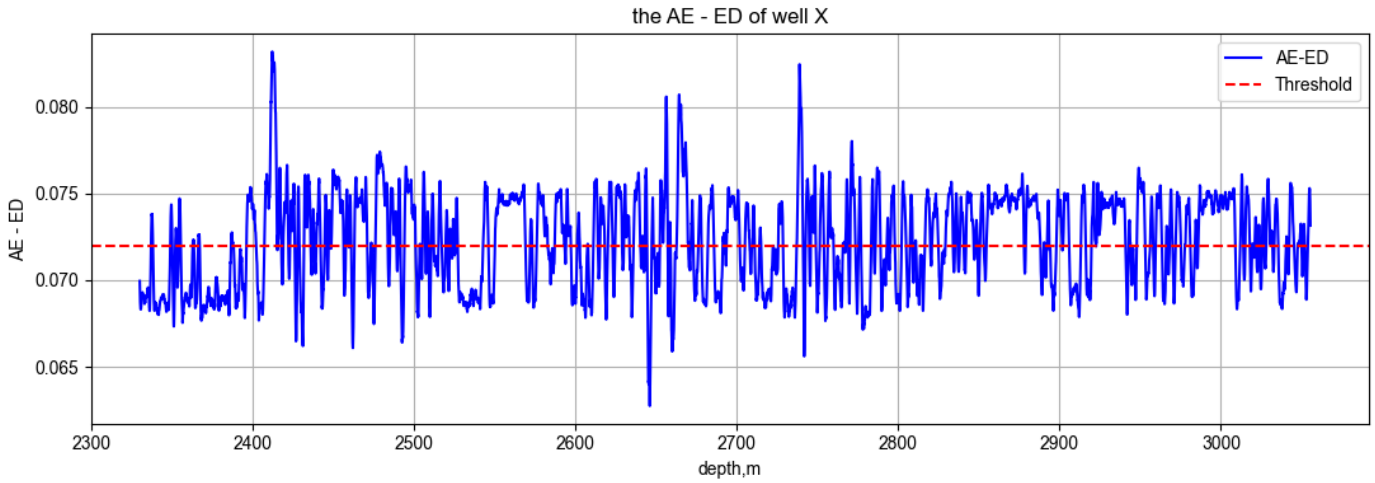

Figure 6. The AE-ED value of Well X.

The blue line represents the AE-ED value. For each new piece of input data of Well X, we can calculate its AE-ED value according to the above method. The red dotted line is the AE-ED threshold. The AE-ED value quantitatively expresses the data similarity, and the AE-ED threshold is the key judgment condition of the hybrid model. They will be mentioned in the establishment of the hybrid model based on data similarity below.

4.1.2. Single Models

In this paper, we established five single models for ROP prediction. Among them were three machine learning models (ANN, SVM, and RF) and two empirical models (Hareland and Motahhari).

In order to keep the models comparable, all the machine learning models used the same input parameters, and both physical models used the parameters involved in the formula as much as possible. The principles of the models are mentioned in Sections 2.3 and 2.4.

4.1.3. Hybrid Models Based on Data Similarity

As shown in Figure 7, we trained the autoencoder (AE) model, machine learning model, and empirical model with the test set data. The AE-ED was used to measure the data similarity, and the machine learning model and empirical model were used for ROP prediction. We calculated the AE-ED value through the autoencoder (AE) model to measure the similarity between each newly input piece of test set data and training set data. According to the value of the AE-ED, we judged and selected the corresponding model to form the hybrid model.

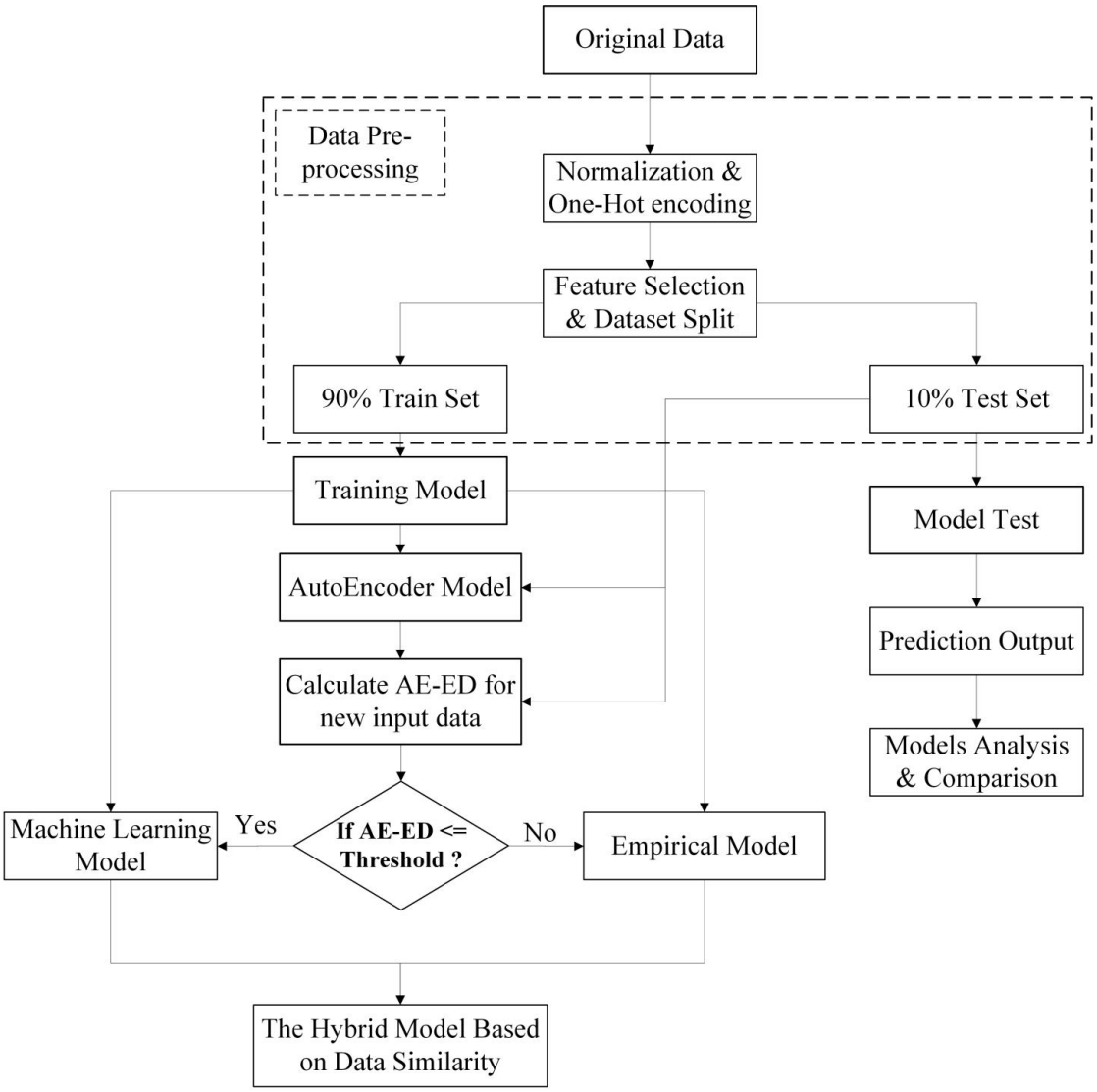

Figure 7. Flowchart of hybrid model establishment.

Combined with the AE-ED value in Figure 6, for newly input data with AE-ED values less than the threshold value (that is, below the threshold line), the machine learning model with higher prediction accuracy is adopted; and for newly input data with AE-ED values greater than the threshold value (that is, above the threshold line), the empirical model with better robustness is adopted. According to the automatic switching of different prediction models under the condition of the AE-ED threshold judgment, a new hybrid model based on data similarity is formed.

In this paper, the main function of the empirical model is to make up for the loss of accuracy of the machine learning model in the case of low data similarity, and to improve the overall robustness and generalization ability. We just need to select a better empirical model. Therefore, we chose the selected model as a representative of empirical models, and we combined it with three machine learning models (SVM, ANN, RF) to establish three hybrid models.

4.2. Model Evaluation Indicators

In order to comprehensively and objectively evaluate the performances of the models, we selected the root mean square error (RMSE), mean relative error (MRE), and predictive error variance (PEV) as the evaluation indicators of the models.

The RMSE is calculated as Equation (13):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{13}$$

The MRE is calculated as Equation (14):

$$\mathrm{MRE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i|} \tag{14}$$

The PEV is calculated as Equations (15)–(17):

$$\mathrm{PEV} = \frac{\sum_{i=1}^{m}(e_i - \bar{e})^2}{m} \tag{15}$$

$$e_i = y_i - \hat{y}_i \tag{16}$$

$$\bar{e} = \frac{1}{m}\sum_{i=1}^{m} e_i \tag{17}$$

where y_i is the real value, $\hat{y}_i$ is the predicted value, e_i is the prediction error, and $\bar{e}$ is the mean of the prediction errors.

The RMSE can visually express the error of the model. However, it is greatly affected by outliers and needs to be comprehensively evaluated in combination with other indicators.

The MRE can avoid the error amplification problem caused by the large gap between the predicted value and real value to a certain extent.

The PEV refers to the variance of the prediction error, which can not only measure the size of the prediction error as a whole to judge whether the model algorithm is scientifically feasible, but also measure the stability and dispersion of the prediction error. Generally speaking, more robust models have smaller PEV values.

5. Results and Discussion

5.1. Results

In this section, we compare and analyze the performances of a single model and hybrid model on the Well X data from the test set. We verified that the hybrid model has better

accuracy, robustness, and generalization ability, and we finally selected and evaluated the best hybrid model for ROP prediction.

The model performances of two empirical models are shown in Figure 8, and the model evaluation indicators are shown in Table 3.

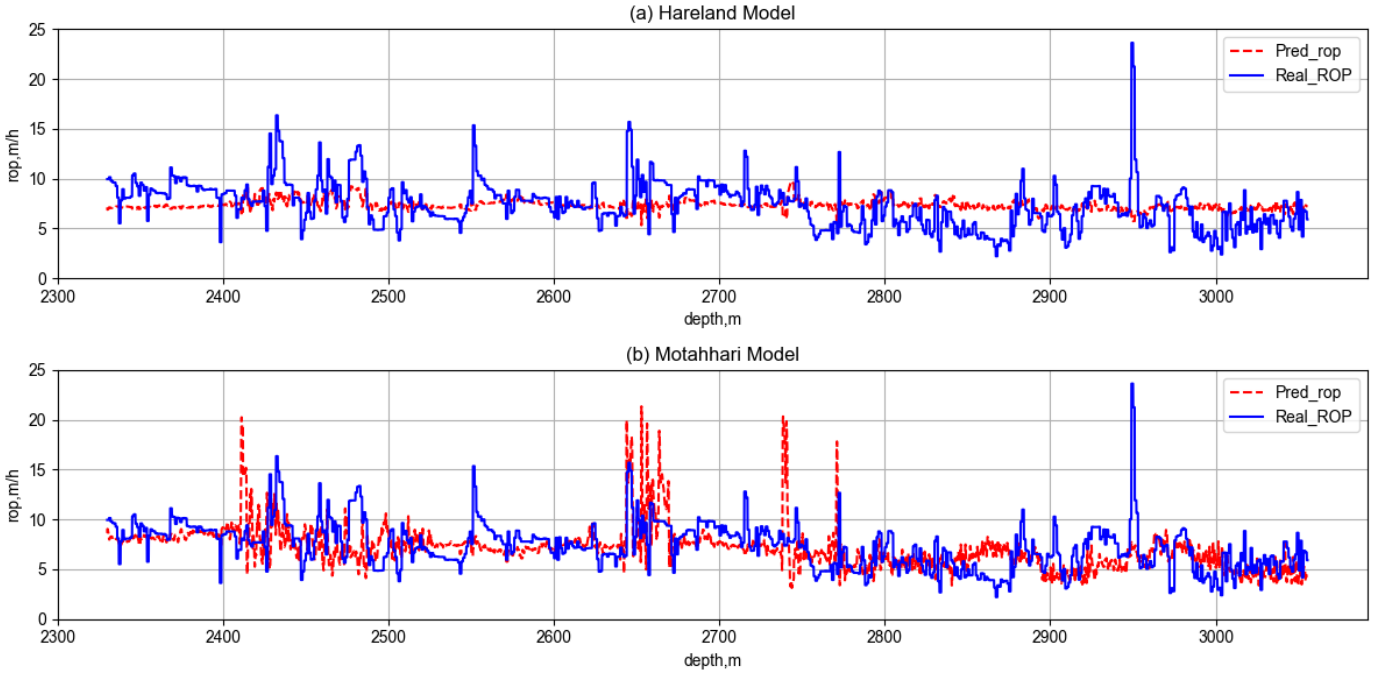

Figure 8. Comparison results of real and predicted ROP: (**a**) prediction results of Hareland model; (**b**) prediction results of Motahhari model.

Table 3. Comparison of Hareland's and Motahhari's model evaluation indicators.

Evaluation Indicators	Hareland	Motahhari
RMSE	7.25	5.59
MRE	0.40	0.25
PEV	4.73	3.02

Observing the prediction results, as shown in Figure 8, reveals that the predicted curves of the Motahhari model fit the original data better than the Hareland model. As shown in Table 3, the three evaluation indicators of the Motahhari model are much better than those of the Hareland model. Generally speaking, the performance of the Motahhari model is better than that of the Hareland model.

In this paper, the main function of the empirical model is to make up for the loss of accuracy of the machine learning model in the case of low data similarity, and to improve the overall robustness and generalization ability. In this process, the empirical model is used to deal with data with low data similarity, which cannot be handled stably and accurately by the machine learning model. Therefore, before building the hybrid model, we need to choose the better empirical model to ensure the performance of the hybrid model. After the comparison, we chose the Motahhari (Mota) model with a better model performance as the representative empirical model, and we combined it with three machine learning models (SVM, ANN, RF) to establish three hybrid models: Mota-SVM, Mota-ANN, and Mota-RF.

The depth of the test well X ranges from 2330 m to 3055 m. In this continuous well section, the drill bit is a PDC drill bit named KM1662.

The model performances are shown in Figures 9–11.

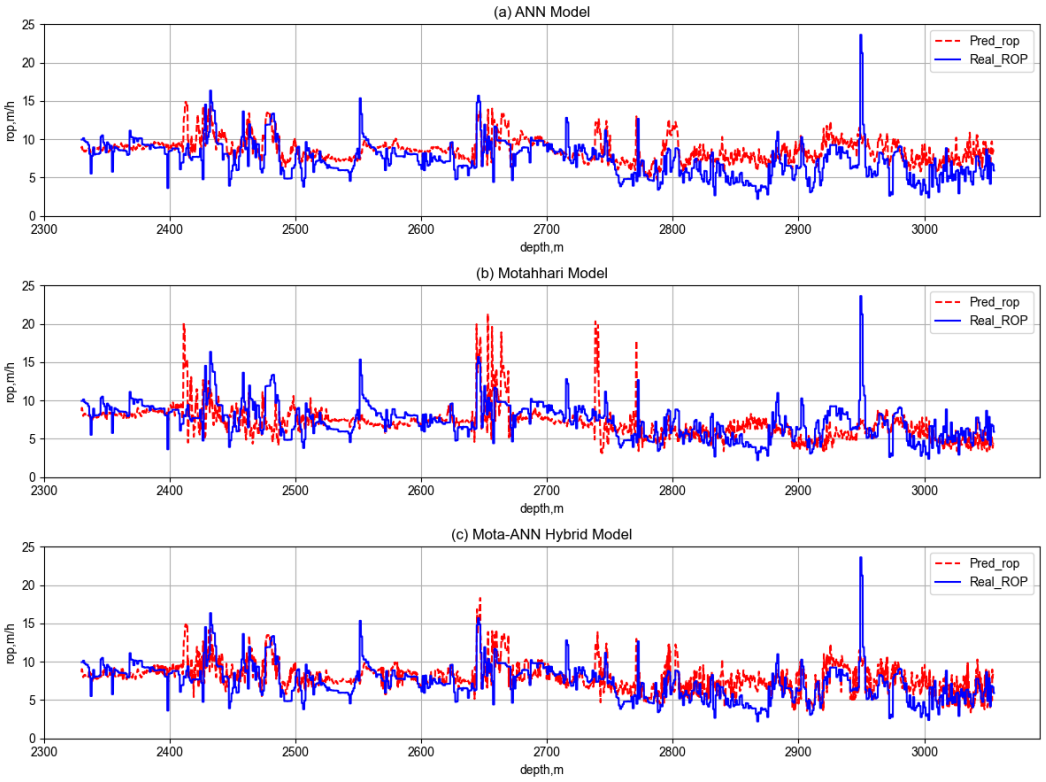

Figure 9. ROP prediction results of Mota-ANN hybrid model and single models: (**a**) ROP prediction results of ANN model; (**b**) ROP prediction results of Motahhari model; (**c**) ROP prediction results of Mota-ANN hybrid model.

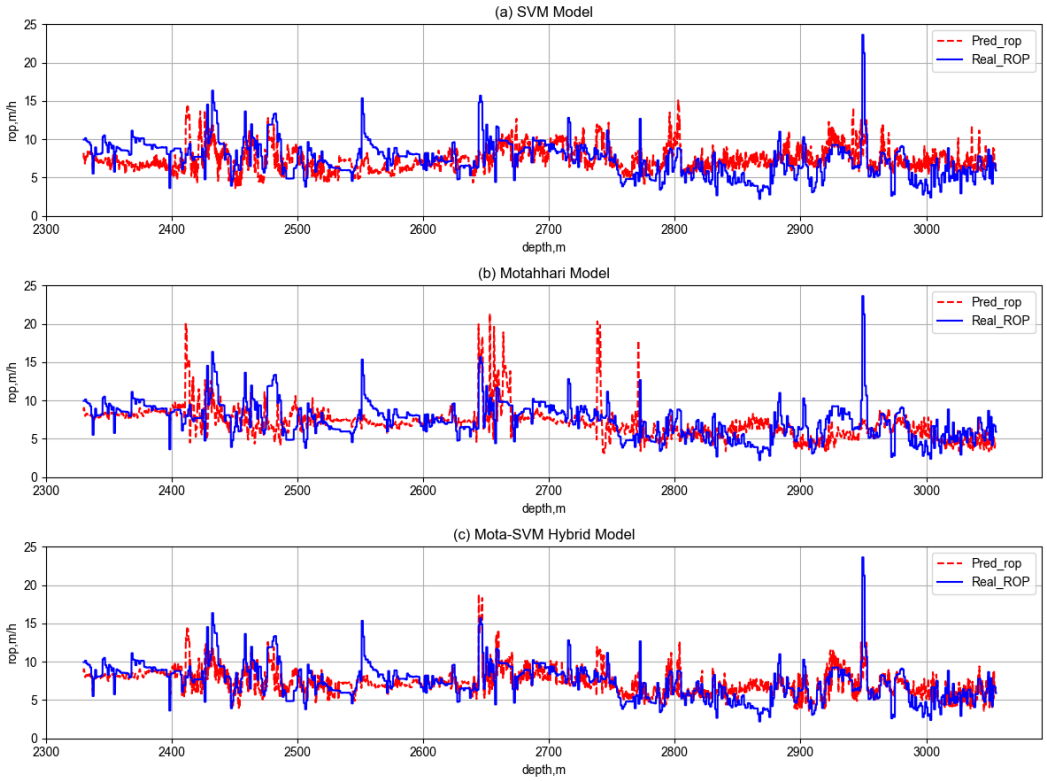

Figure 10. ROP prediction results of Mota-SVM hybrid model and single models: (**a**) ROP prediction results of SVM model; (**b**) ROP prediction results of Motahhari model; (**c**) ROP prediction results of Mota-SVM hybrid model.

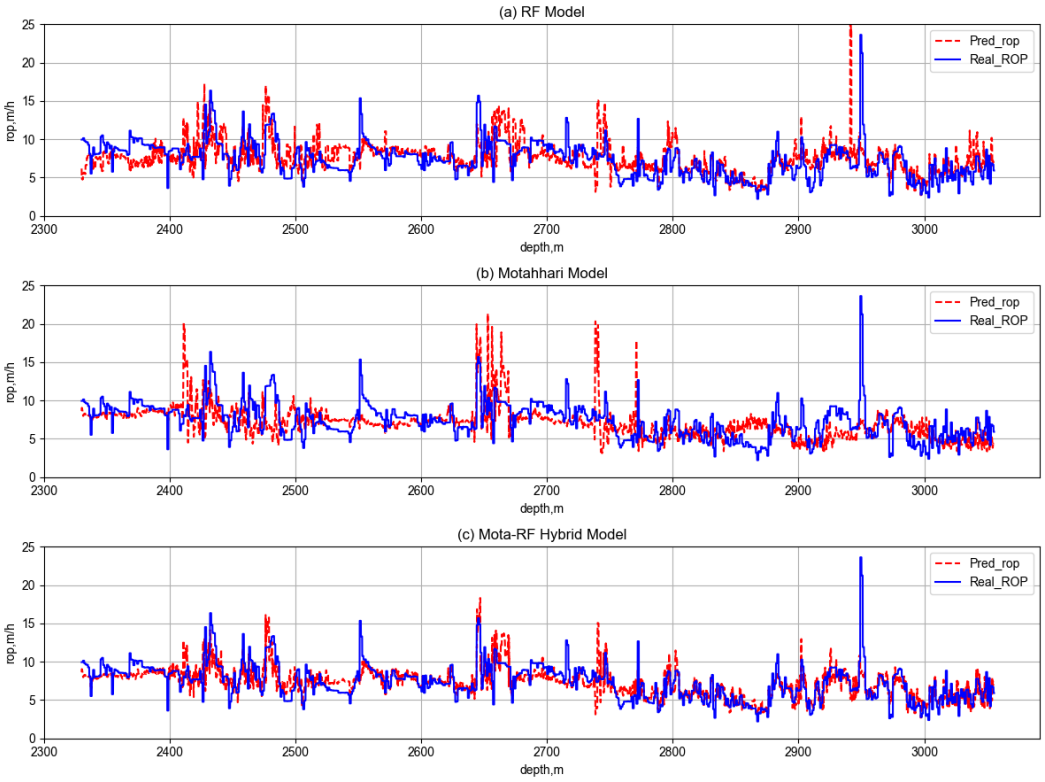

Figure 11. ROP prediction results of Mota-RF hybrid model and single models: (**a**) ROP prediction results of RF model; (**b**) ROP prediction results of Motahhari model; (**c**) ROP prediction results of Mota-RF hybrid model.

As shown in Figures 9–11, the prediction curves of the three hybrid models are closer to the real data curve than the single models. Specifically, the hybrid model can better predict with the help of an empirical model when the data similarity of the input data is low compared to the machine learning model. The hybrid model compensates for the shortcomings of the machine learning model with the help of the empirical model.

5.2. Discussion

In order to more objectively and comprehensively evaluate the eight ROP prediction models established in this paper, we calculated their evaluation indicators and obtained Table 4 and Figure 12.

Table 4. Evaluation indicators of different prediction models.

Evaluating Indicators	Hareland	Motahhari	ANN	SVM	RF	Mota-ANN	Mota-SVM	Mota-RF
Runtime(s)	44.14	20.97	27.24	65.67	14.63	48.21	86.64	35.60
RMSE	7.25	5.59	6.12	5.71	5.14	4.39	4.22	2.65
MRE	0.40	0.25	0.29	0.24	0.21	0.17	0.15	0.09
PEV	4.73	3.02	3.61	3.32	3.16	2.12	2.01	1.51

Among the three evaluation indicators, the smaller the RMSE and MRE, the higher the prediction accuracy of the model; the smaller the PEV, the better the stability, robustness, and generalization of the model.

The hybrid model outperformed the two single models that compose it, with average reductions of 33% in the RMSE, 45% in the MRE, and 41% in the PEV. Compared with the Motahhari model, the Mota-ANN model reduced the three indicators by 21.47%, 32.00%, and 29.80%, respectively, and decreased them by 28.27%, 41.38%, and 41.27%, respectively,

compared with the ANN model. Secondly, compared with the Motahhari model, the three indicators of the Mota-SVM model were reduced by 24.51%, 40.00%, and 33.44%, respectively, and compared with the SVM model, they were reduced by 26.09%, 37.50%, and 39.46%, respectively. The best-performing model was the Mota-RF hybrid model, which reduced the indicators by 52.59%, 64.00%, and 50.00%, respectively, compared with the Motahhari model, and reduced them by 48.44%, 57.14%, and 52.22%, respectively, compared with the RF model. A fair comparison should not ignore the model runtime, at which point the performance of the hybrid model becomes mediocre. Because the hybrid model is a combination of the physical model and empirical model, it takes longer to run than the two single models that compose it. This is an unavoidable problem.

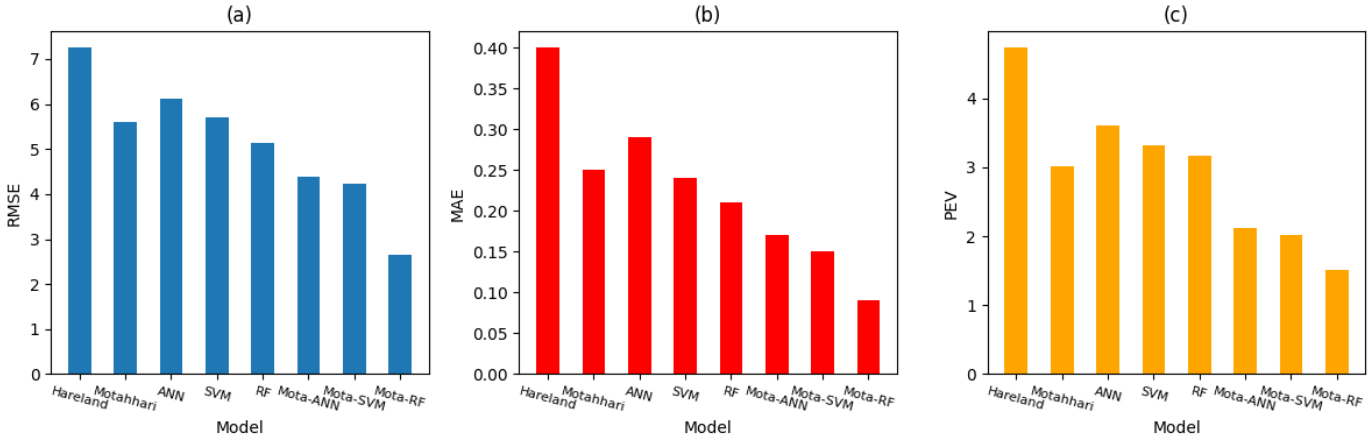

Figure 12. Evaluation indicators of different prediction models: (**a**) RMSEs of different prediction models; (**b**) MAEs of different prediction models; (**c**) PEVs of different prediction models.

By comparing the evaluation indicators and through further analysis, it can be seen that the three indicators of the hybrid model have obvious advantages, especially the smaller PEV, which was reduced by up to 41% compared with the single model. This demonstrates that the hybrid model is more stable and has the robustness and generalization that the machine learning model lacks. In addition, compared with the empirical model, the RMSE and MRE of the hybrid model were reduced by 33% and 45%, respectively, which proves that the hybrid model has higher accuracy than the empirical model. Further analysis shows that both the empirical model and machine learning model play very important roles in the hybrid model. Based on the data similarity analysis, the data with high data similarity are input into the machine learning model for processing, giving full play to its high prediction accuracy. At the same time, the data with low data similarity are input into the optimized empirical model to ensure the robustness and prediction stability while improving the accuracy as much as possible, as the main function of the empirical model is to make up for the loss of stability of the machine learning model. Therefore, we can conclude that the hybrid model combines the high prediction accuracy of the machine learning method with the good robustness and generalization of the empirical method, overcoming the shortcomings of any single model.

Among the three hybrid models, the Mota-RF hybrid model performs better and has the shortest runtime, which has obvious advantages compared to the other two hybrid models. Compared with the single model, its advantages become even greater, except for the runtime. These comparisons show that the Mota-RF hybrid model is most suitable for this field when the drill-bit type is KM1662.

6. Conclusions

For machine learning methods, the IID assumption is a basic guarantee that the model obtained through the training data can achieve good results in the test set, and it requires a high degree of data similarity between the test set and training set. However, the complex

petroleum engineering environment cannot satisfy its assumptions, which leads to the instability of the machine learning method on the test set.

Based on the quantification of data similarity, this paper establishes a hybrid model combining a machine learning method and an empirical method, and it defines the AE-ED to quantify data similarity. According to the AE-ED value of each new piece of input data, the hybrid model chooses the corresponding single model to calculate. We apply the empirical model to deal with data with low data similarity, which cannot be processed stably and accurately by the machine learning model. The physical model guarantees accuracy, while the empirical model guarantees robustness and generalization. Therefore, the hybrid model combines the high prediction accuracy of the machine learning method with the good robustness and generalization of the empirical method, overcoming the shortcomings of any single model.

(1) Calculate the AE-ED value of each new piece of data in the test set, and we find that the AE-ED values of many of the data exceed the threshold, which does not satisfy the IID assumption. This proves that quantifying the data similarity is very necessary;
(2) Among the empirical models established in this paper, the Motahhari model performs better. Meanwhile, RF is the best-performing machine learning model;
(3) The hybrid model outperformed the two single models that compose it, with average reductions of 33% in the RMSE, 45% in the MRE, and 41% in the PEV. It has high prediction accuracy, as well as good robustness and generalization;
(4) Because the hybrid model is a combination of the physical model and empirical model, it takes longer to run than the two single models that compose it;
(5) Among the three hybrid models, the Mota-RF hybrid model has the best evaluation indicator and model performance, and it is the most suitable for the ROP prediction of the field when the drill-bit type is KM1662;
(6) In the hybrid model, the main role of the empirical model is to make up for the low robustness and generalization of the machine learning model when the data similarity is low.

According to the research results of this paper, the best-performing empirical model is the Motahhari model, the best-performing machine learning model is the RF model, and the best-performing hybrid model is the Mota-RF hybrid model that combines them. We do not believe that simply combining the best-performing physics model with the machine learning model will result in the best hybrid model. For future work, we plan to focus on this issue. In addition, the model in this paper mainly considers the interaction between the drill bit and the rock, while ignoring the motion state of the drilling tool. Therefore, we plan to use the wellbore trajectory and the stress state of the drilling tools in the wellbore as input features to establish a more stable and accurate ROP prediction model to cope with complex working conditions in the field, which is applicable to the ROP prediction of both vertical wells and directional wells.

Author Contributions: Conceptualization, F.Z. and H.F.; methodology, F.Z., Y.L., and H.Z.; software, Y.L.; validation, F.Z. and R.J.; formal analysis, H.Z. and F.Z.; investigation, H.Z. and R.J.; resources, Y.L.; data curation, F.Z.; writing—original draft preparation, F.Z.; writing—review and editing, H.F.; visualization, R.J.; supervision, Y.L.; project administration, H.F.; funding acquisition, H.F. and H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: The authors gratefully acknowledge the financial support from the Natural Science Foundation of China (Basic Research on Marine Deep Oil and Gas Enrichment Mechanism and Key Engineering Technology, grant number: U19B6003), and the joint fund for enterprise innovation and development of the National Natural Science Foundation of China (Basic Theory and Control Method of Marine Deep High Temperature and High-Pressure Drilling and Completion Engineering, grant number: U19B6003-05).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are not publicly available due to the involvement of the information of oil fields and need to be kept confidential.

Acknowledgments: The author would like to thank the SINOPEC Research Institute of Petroleum Engineering for its support, and Hongbao Zhang for providing constructive suggestions and ideas, which helped improve the quality of the paper.

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

References

1. Larrañaga, P.; Atienza, D.; Diaz-Rozo, J.; Ogbechie, A.; Puerto-Santana, C.E.; Bielza, C. *Industrial Applications of Machine Learning*; CRC Press: London, UK, 2018; pp. 133–166.
2. Edgcomb, J.B.; Zima, B. Machine learning, natural language processing, and the electronic health record: Innovations in mental health services research. *Psychiatr. Serv.* **2019**, *70*, 346–349. [CrossRef] [PubMed]
3. Emerson, S.; Kennedy, R.; O'Shea, L.; O'Brien, J. Trends and applications of machine learning in quantitative finance. In Proceedings of the 8th International Conference on Economics and Finance Research (ICEFR 2019), Lyon, France, 18–21 June 2019.
4. Noshi, C.I.; Schubert, J.J. The role of machine learning in drilling operations; a review. In Proceedings of the SPE/AAPG Eastern Regional Meeting, Pittsburgh, PA, USA, 5 October 2018.
5. Zhong, R.; Salehi, C.; Johnson, R., Jr. Machine learning for drilling applications: A review. *J. Nat. Gas Sci. Eng.* **2022**, *108*, 104807. [CrossRef]
6. Barbosa, L.F.F.; Nascimento, A.; Mathias, M.H.; de Carvalho, J.A., Jr. Machine learning methods applied to drilling rate of penetration prediction and optimization-A review. *J. Pet. Sci. Eng.* **2019**, *183*, 106332. [CrossRef]
7. Mazen, A.Z.; Rahmanian, N.; Mujtaba, I.; Hassanpour, A. Prediction of Penetration Rate for PDC Bits Using Indices of Rock Drillability, Cuttings Removal, and Bit Wear. *SPE Drill. Complet.* **2021**, *36*, 320–337. [CrossRef]
8. Lin, Y.; Zong, Y.; Zheng, L. The Developments of ROP Prediction for oil Drilling. *Pet. Drill. Tech.* **2004**, *32*, 10–13.
9. Bourgoyne, A.T., Jr.; Young, F.S., Jr. A multiple regression approach to optimal drilling and abnormal pressure detection. *Soc. Pet. Eng. J.* **1974**, *14*, 371–384. [CrossRef]
10. Hareland, G.; Rampersad, P.R. Drag-bit model including wear. In Proceedings of the SPE Latin America/Caribbean Petroleum Engineering Conference, Buenos Aires, Argentina, 27–19 April 1994.
11. Motahhari, H.R.; Hareland, G.; James, J.A. Improved drilling efficiency technique using integrated PDM and PDC bit parameters. *J. Can. Pet. Technol.* **2010**, *49*, 45–52. [CrossRef]
12. Xu, M.; Wei, M.; Deng, S.; Cai, W. New application of multi-model ensemble learning in ROP prediction. *Comput. Sci.* **2019**, *48*, 619–622.
13. Su, X.; Sun, J.; Gao, X.; Wang, M. Prediction Method of Drilling Rate of Penetration Based on GBDT Algorithm. *Comput. Appl. Softw.* **2019**, *36*, 87–92.
14. Liu, S.; Sun, J.; Gao, X.; Wang, M. Analysis and Establishment of ROP Prediction Model of Drilling Machinery Based on Artificial Neural Network. *Comput. Sci.* **2019**, *46*, 605–608.
15. Diaz, M.B.; Kim, K.Y.; Shin, H.; Zhuang, L. Predicting rate of penetration during drilling of deep geothermal well in Korea using artificial neural networks and real-time data collection. *J. Nat. Gas Sci. Eng.* **2019**, *67*, 225–232. [CrossRef]
16. Hazbeh, O.; Aghdam, S.K.; Ghorbani, H.; Mohamadian, N.; Alvar, M.A.; Moghadasi, J. Comparison of accuracy and computational performance between the machine learning algorithms for rate of penetration in directional drilling well. *Pet. Res.* **2021**, *6*, 271–282. [CrossRef]
17. Lawal, A.I.; Kwon, S.; Onifade, M. Prediction of rock penetration rate using a novel antlion optimized ANN and statistical modelling. *J. Afr. Earth Sci.* **2021**, *182*, 104287. [CrossRef]
18. Alkinani, H.H.; Al-Hameedi, A.T.T.; Dunn-Norman, S. Data-driven recurrent neural network model to predict the rate of penetration. *Upstream Oil Gas Technol.* **2021**, *7*, 100047. [CrossRef]
19. Zhou, Y.; Lu, C.; Zhang, M.; Chen, X. A novel rate of penetration model based on support vector regression and modified bat algorithm. *IEEE Trans. Ind. Inform.* **2022**, *19*, 3205374. [CrossRef]
20. Ren, Y.; Lu, B.; Zheng, S.; Bai, K.; Cheng, L.; Yan, H.; Wang, G. Research on the Rate of Penetration Prediction Method Based on Stacking Ensemble Learning. *Geofluids* **2023**, *2023*, 6645604. [CrossRef]
21. Major, P. On the invariance principle for sums of independent identically distributed random variables. *J. Multivar. Anal.* **1978**, *8*, 487–517. [CrossRef]
22. Zhai, J.; Zhang, S.; Chen, J.; He, Q. Autoencoder and its various variants. In Proceedings of the 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Miyazaki, Japan, 7–10 October 2018.
23. Parmar, A.; Katariya, R.; Patel, V. A review on random forest: An ensemble classifier. In Proceedings of the International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018, Coimbatore, India, 7–8 August 2018.

24. Abiodun, O.I.; Jantan, A.; Omolara, A.E.; Dada, K.V.; Mohamed, N.A.; Arshad, H. State-of-the-art in artificial neural network applications. A survey. *Heliyon* **2018**, *4*, 938. [CrossRef] [PubMed]
25. Brereton, R.G.; Lloyd, G.R. Support vector machines for classification and regression. *Analyst* **2010**, *135*, 230–267. [CrossRef] [PubMed]