Petroleum Science 22 (2025) 1596–1610

Contents lists available at ScienceDirect

Petroleum Science

journal homepage: www.keaipublishing.com/en/journals/petroleum-science

Original Paper

A novel drilling parameter optimization method based on big data of drilling

Chi Peng [a, b, *], Hong-Lin Zhang [a, b, **], Jian-Hong Fu [a], Yu Su [c], Qing-Feng Li [d], Tian-Qi Yue [e]

[a] State Key Laboratory of Oil & Gas Reservoir Geology and Exploitation, Southwest Petroleum University, Chengdu, 610500, Sichuan, China
[b] School of Petroleum Engineering, Southwest Petroleum University, Chengdu, 610500, Sichuan, China
[c] Engineering Technology Research Institute, PetroChina Southwest Oil & Gas Field Company, Chengdu, 610017, Sichuan, China
[d] CNPC Tarim Oilfield Branch Company Oil and Gas Engineering Research Institute, Korla, 841000, Xinjiang, China
[e] Safety, Environment and Technology Supervision Research Institute, PetroChina Southwest Oil and Gas Field Company, Chengdu, 610041, Sichuan, China

ARTICLE INFO

Article history:
Received 27 March 2024
Received in revised form
26 February 2025
Accepted 2 March 2025
Available online 4 March 2025

Edited by Jia-Jia Fei

Keywords:
Rate of penetration
Machine learning
Drilling parameter
Clustering analysis
Optimization

ABSTRACT

Rate of penetration (ROP) is the key factor affecting the drilling cycle and cost, and it directly reflects the drilling efficiency. With the increasingly complex field data, the original drilling parameter optimization method can't meet the needs of drilling parameter optimization in the era of big data and artificial intelligence. This paper presents a drilling parameter optimization method based on big data of drilling, which takes machine learning algorithms as a tool. First, field data is pre-processed according to the characteristics of big data of drilling. Then a formation clustering model based on unsupervised learning is established, which takes sonic logging, gamma logging, and density logging data as input. Formation clusters with similar stratum characteristics are decided. Aiming at improving ROP, the formation clusters are input into the ROP model, and the mechanical parameters (weight on bit, revolution per minute) and hydraulic parameters (standpipe pressure, flow rate) are optimized. Taking the Southern Margin block of Xinjiang as an example, the MAPE of prediction of ROP after clustering is decreased from 18.72% to 10.56%. The results of this paper provide a new method to improve drilling efficiency based on big data of drilling.

1. Introduction

Oil and gas drilling engineering is challenged by increased well depth, increased operation difficulty, and more complicated geological environment. It becomes more important to improve the drilling efficiency. Improving the rate of penetration (ROP) and reducing the drilling cycle are the most direct and effective ways to enhance drilling efficiency (Wang and Guang, 2022; Li et al., 2020). The drilling process is a complex process influenced by many factors, which can be divided into controllable factors and uncontrollable factors. Controllable factors (such as weight on bit (WOB), revolution per minute (RPM), standpipe pressure (SPP), flow rate, drilling fluid density, etc.) have considerable influence on ROP. Most of the previous drilling parameter optimization methods rely on empirical methods, where the drilling parameters are optimized by establishing empirical ROP prediction models through drilling data or experiments (Pessier, 1992; Rashidi, 2010; Cherif, 2012). The established models are closely related to drilling methods, rock breaking tools, drilling parameters, and rock mechanic parameters. However, many model parameters are difficult to accurately obtain, resulting in poor applicability of the models. With the complex drilling environment and massive drilling data, it is inefficient and difficult for conventional methods to meet the demand.

There are currently three main drilling parameter optimization methods (Cui et al., 2015; Huang and Gao, 2022; Sehsah et al., 2017; Hutchinson et al., 2018; Khadisov et al., 2020; Cayeux et al., 2019; Xiong et al., 2023; Hamzah et al., 2019; Armenta, 2008): 1) Optimization of drilling parameters based on mechanical specific energy (MSE); 2) Optimization of drilling parameters based on ROP; 3) Optimization of drilling process based on intelligent drilling system. The optimization of drilling parameters based on MSE is the most widely used. For example, Waughman et al. (2002) evaluated PDC bit wear in real-time by comparing the MSE values of new bits under the same conditions. Dupriest (2006) evaluated drilling

* Corresponding author.
** Corresponding author.
E-mail addresses: pengchiswpu@swpu.edu.cn (C. Peng), zhanghonglin_0503@163.com (H.-L. Zhang).

https://doi.org/10.1016/j.petsci.2025.03.002

efficiency based on MSE, and identified the causes of drill bit and system inefficiency. Rafatian et al. (2010) established a MSE model in high pressure and low permeability formation. Islam et al. (2018) discussed drilling problems encountered in granite and geothermal reservoirs by using the MSE model. Alsubaih et al. (2018) summarized the correlation between MSE and ROP, productivity, and other parameters by statistical analysis methods.

Since Teale (1965) first put forward the idea of MSE, various models have been established through continuous revision and improvement. Mohan and Adil (2009) proposed a new correlation for identifying inefficient drilling conditions based on MSE. Alali et al. (2012) studied the influence of axial vibration of drill string on ROP based on MSE. Meng et al. (2012) established a rock-breaking specific energy model under the condition of hydraulic parameters by analyzing the positive effect of hydraulic energy on rock-breaking and bottom hole cleaning based on the original specific energy theory of rock-breaking machines. Cui et al. (2014) brought about the theory of optimizing drilling parameters by MSE under the condition of compound drilling. Pinto and Lima (2016) presents a new real-time analysis of geomechanics capability based on MSE to optimize energy consumption and ROP in salt-layer drilling. Al-Sudani (2017) developed a model for predicting drilling performance by analyzing the real-time transmission drilling mechanical energy consumed by the drill bit. However there are many uncertainties in the existing drilling parameter optimization methods, and the uncontrollable factors such as formation rock characteristics and the big data of drilling are not fully considered.

Artificial intelligence and big data technology have the advantage of efficiently and quickly mining the deep hidden information in the drilling data, which is able to discover the trends ignored by the traditional theoretical models (Pang et al., 2023; Noshi, 2019; Pei et al., 2024; Chris, 2021; Zhang et al., 2021; Chen et al., 2023; Deng et al., 2023; Elmgerbi et al., 2021). In this work, the unsupervised learning is used in this paper to extract formation features in big data of drilling, and the drilling parameters are optimized through the established ROP prediction model.

2. Methodology

Fig. 1 is the frame of the ROP prediction model based on big data of drilling. The drilling data collected include well logging and mud logging data, which are processed through data cleaning, normalization, and dimension reduction. In formations with similar rock characteristics, the influence of drilling parameters on ROP is also similar. Therefore, taking sonic logging, density logging, and gamma ray logging data as input, formations with similar characteristics are classified into one cluster, and the cluster results provide a new feature parameter. Then, the ROP prediction model is established with formation cluster, WOB, RPM, flow rate, SPP, torque, drilling fluid density, and bit type as input parameters. Finally, the influence of different parameter combinations on ROP is analyzed through the established ROP prediction model to optimize drilling parameters.

2.1. Drilling data of southern margin block

The Southern Margin block is located in Junggar Basin, Xinjiang, China. The deep formation (vertical depth greater than 5000 m) has the characteristics of high rock compressive strength and poor drillability. Therefore, the ROP in deep formations is low and the drilling operation consumes more time than expected. The typical geological stratifications and well structure of the target reservoir are illustrated in Fig. 2. The combined strata in the southern margin are mainly mudstone, argillaceous siltstone and silty mudstone, with dense lithology and strong plasticity, which makes it difficult for bit cutting and drilling. According to the field data (Table 1), the average designed well depth in this area exceeds 6700 m, the average drilling cycle is 392 days, and the average ROP of deep formation is less than 2 m/h.

Data of 112,781 points from 15 wells (G101, G102, G103, etc.) in the Southern Margin block of Xinjiang are collected as the original field data. The data contains the complete geological formations in this area. Part of the data are shown in Table 2.

2.2. Pre-processing of big data of drilling

Big data of drilling is extracted and collated from different tool systems (such as logging while drilling, wire logging, geological logging, etc.), with complex types and huge quantities, so it is necessary to preprocess big data of drilling. Firstly, data cleaning is performed. The wrong and conflicting data are removed. The abnormal point refers to the unreasonable point in the drilling data set. Abnormal points include outliers, high leverage points, and strong influence points, all of which may cause serious deviation in data fitting and analysis during machine learning. Because the abnormal point can't be directly used, and the proportion of abnormal points is generally small, it has little impact on the scale of the data set. Therefore, abnormal points are directly deleted (Guo and Zhou, 2002).

In this paper, the box-plot is used to detect outliers. As shown in Fig. 3, the lower quartile (Q_1) and the upper quartile (Q_3) represent the data in 25% and 75% of the samples, respectively. The median quartile (Q_2) is the data in 50% of the samples. The upper and lower bounds are the threshold values for determining outliers.

$$X < Q_1 - 1.5 \times IQR \text{ or } X > Q_3 + 1.5 \times IQR \tag{1}$$

where X represents the outliers. IQR is the interquartile spacing, which is the difference between Q_3 and Q_1.

Furthermore, the Savitzky-Golay (SG) filter is employed to denoise the drilling data. The SG filter is a filtering method that achieves local polynomial fitting through least squares convolution, which can remove noise while preserving the shape and width of the signal. The principle of SG filter is shown in Eq. (2).

$$S_j^* = \frac{\sum_{i=-m}^{m} C_i S_{j+1}}{N} \tag{2}$$

where S is the original signal. S^* is the signal after noise reduction. C_i is the noise reduction coefficient of the i-th time. N is the sliding window width of $(2m+1)$ groups of data. j. is the j-th sample in the sample set.

Sometimes data sets contain non-operational data, which are called discrete features. In this way, it is necessary to transcode the non-operational data. The discrete features in drilling data mainly include formation and bit type. The values of these non-operational data need to be transferred as digital values. This paper uses vectors to encode non-operational feature values. For example, the formation can be defined by a binary numerical vector, as shown in Table 3.

The bit type is a kind of data with non-numerical characteristics. The different categories (PDC bit and Tri-Cone bit) contained in this feature parameter are in a parallel relationship without ordinal meaning. Therefore, one-hot encoding is used to transform bit type, making which suitable as input variables for the model. The encoded results are shown in Table 4.

There are multiple feature parameters in the drilling data, and the range of each feature parameter is quite different. Therefore, it

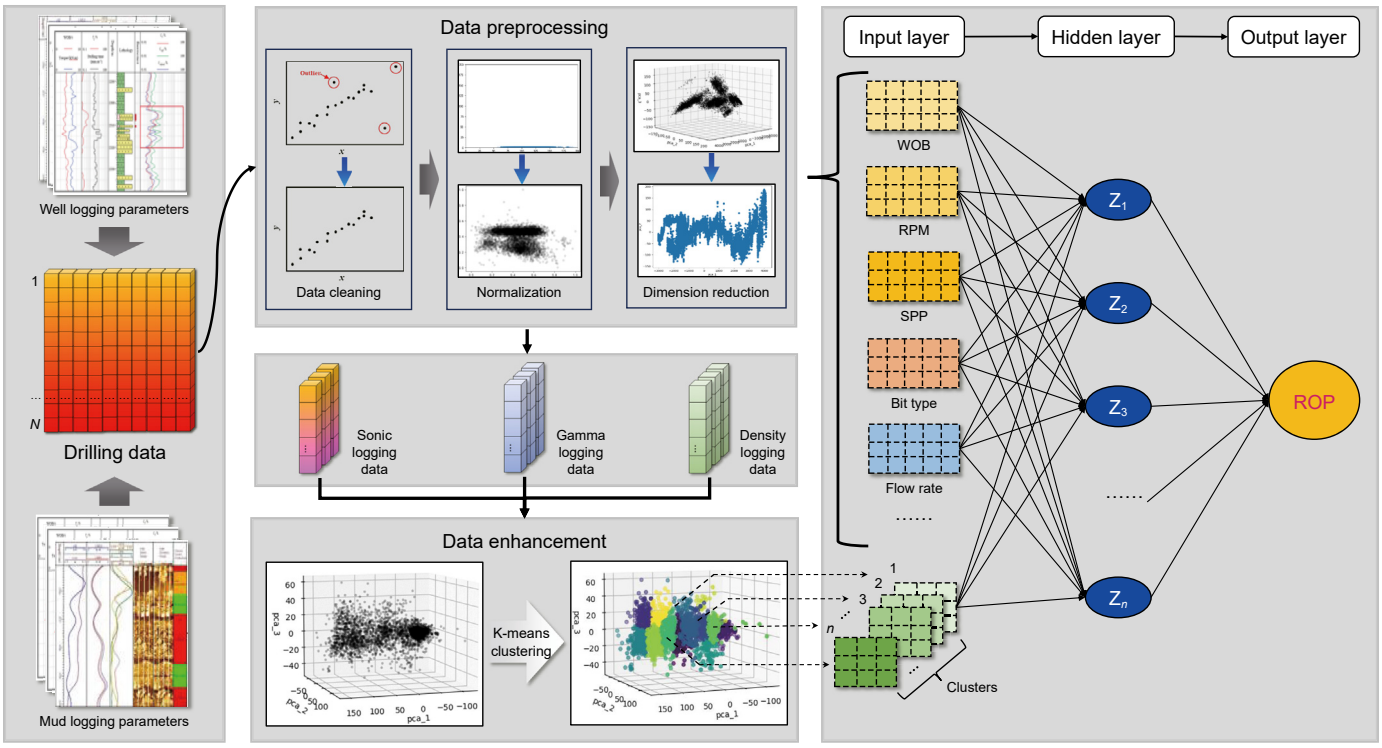
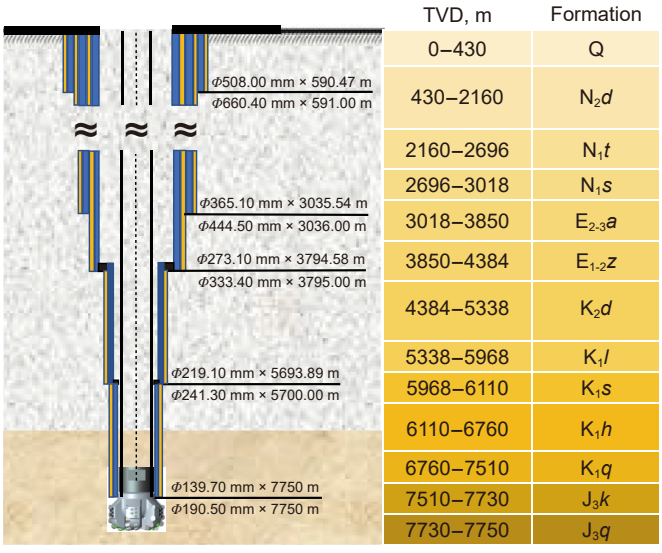


Fig. 1. Framework of the ROP prediction model based on big data of drilling.



TVD, m	Formation
0–430	Q
430–2160	N_2d
2160–2696	N_1t
2696–3018	N_1s
3018–3850	$E_{2\text{-}3}a$
3850–4384	$E_{1\text{-}2}z$
4384–5338	K_2d
5338–5968	K_1l
5968–6110	K_1s
6110–6760	K_1h
6760–7510	K_1q
7510–7730	J_3k
7730–7750	J_3q

Fig. 2. Typical geological stratifications and well structure in the target reservoir.

is necessary to normalize the data, eliminate the influence of dimensions, and make all data indicators comparable. This paper adopts the Min-Max normalization, and the conversion function is:

$$x_{ij}^{*} = \frac{x_{ij} - Min_j}{Max_j - Min_j} \quad (i = 1, 2, 3, \ldots; j = 1, 2, 3) \tag{3}$$

Fig. 4 shows the distribution of compressive strength and shale content before and after normalization. Without dimensional processing, the drilling parameters may be over-fitted, which will affect the prediction accuracy of ROP.

Besides, this paper adopts the popular data dimension reduction method of Principal Component Analysis (PCA). Through linear projection, the high-dimensional data is mapped into the low-dimensional space. With the larger variance of data on the projected dimension, fewer data dimensions are achieved (as shown in Fig. 5), which facilitate the training of ROP model.

2.3. Formation clustering based on unsupervised learning

Optimization of drilling parameters should be based on the matching relationship between drilling parameters and formation rock characteristics. For formations with similar rock characteristics, the influence of drilling parameters on ROP is similar. It is difficult to label complex and diverse formation characteristic parameters by establishing complicated standards. Unsupervised learning can be used to extract, screen and classify the features of

Table 1
Drilling cycle and average ROP of typical wells in Southern Margin block of Xinjiang.

Well	Well depth, m	Range of deep formation, m	Average ROP in deep formations, m/h	Drilling cycle, d
G101	7000	5900–7000	1.21	395
G102	6100	5600–6100	1.99	347
GQ5	6346	5951–6546	0.98	405
GQ6	6650	6181–6806	2.88	197
HT1	7601	5700–7601	1.41	453

Table 2
Part of the field data (drilling data of well G101 from 3476 to 3513 m).

#	TVD, m	Drilling time, min/m	WOB, kN	Torque, kN·m	RPM, r/min	SPP, MPa	Mud inlet density, g/cm^3	AC	GR	DEN
G101	3476	72.2	0.98	1	90	3.76	1.04	69.51	57.001	2.474
G101	3477	10.6	0.99	2	90	7.95	1.04	69.127	56.8	2.478
G101	3478	4.6	0.99	1	90	6.45	1.04	68.772	56.563	2.48
…	…	…	…	…	…	…	…	…	…	…
G101	3483	2.8	0.98	2	90	3.73	1.04	66.327	54.586	2.486
G101	3484	2.8	0.98	5	90	4.71	1.04	66.233	54.959	2.491
G101	3485	5.2	0.98	2	90	5.36	1.04	66.311	56.072	2.498
…	…	…	…	…	…	…	…	…	…	…
G101	3490	3.9	1	5	120	8.29	1.04	69.853	70.649	2.521
G101	3491	2.8	0.99	5	120	9.14	1.04	69.717	72.863	2.518
G101	3492	3.4	0.99	4	120	6.7	1.04	69.223	73.476	2.52
…	…	…	…	…	…	…	…	…	…	…
G101	3497	3.3	0.99	6	120	6.84	1.04	63.627	56.712	2.53
G101	3498	3.3	1	3	150	6.49	1.04	62.501	53.655	2.521
G101	3499	2.7	0.99	3	150	7.1	1.04	62.176	51.786	2.52
…	…	…	…	…	…	…	…	…	…	…
G101	3511	3.7	0.98	5	150	9.5	1.04	62.546	65.707	2.565
G101	3512	2.6	0.99	7	150	7.38	1.04	62.26	67.406	2.572
G101	3513	5	0.99	5	150	4.95	1.04	61.931	69.026	5.576

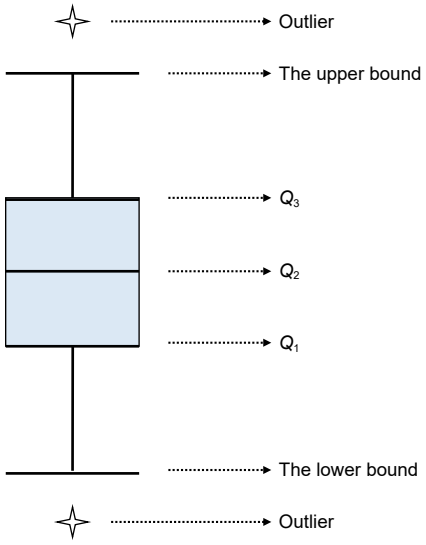


Fig. 3. Schematic diagram of box-plot.

unlabeled geological data of formation.

The complexity of formation requires multiple indicators to reflect the characteristics of formation. These features reflect the formation properties from different aspects, and each index has different emphases. According to the theory, the properties of rocks are closely related to the logging data. For example, acoustic time difference can reflect the tensile and compressive deformation characteristics and strength characteristics of rocks. The gamma ray can reflect the shale content and plasticity, the density logging can show the compaction degree of the rock, and the resistivity represent the compactness of the rock (Dong et al., 2022; Liang et al., 2006). Therefore, this paper takes three kinds of major logging data, namely density logging, gamma ray logging and sonic logging, as feature parameters characterizing the formation rock properties.

Cluster algorithms mainly include the K-means method and the Mean-shift method. Big data of drilling has the characteristics of multiple features and large quantity, so this paper uses the K-means method for clustering (Wang et al., 2018; Jing, 2019; Hou, 2018). The essence of the K-means method is to cluster a given unlabeled

Table 3
Example of independent transcoding of formation.

Well depth, m	Formation						
	Q	N_2d	N_1t	K_1tg	J_2t	J_2x	J_1b
543	1	0	0	0	0	0	0
1495	0	1	0	0	0	0	0
2019	0	1	0	0	0	0	0
3257	0	0	1	0	0	0	0
3894	0	0	1	0	0	0	0
5962	0	0	0	1	0	0	0
6050	0	0	0	0	1	0	0
6470	0	0	0	0	0	1	0
6905	0	0	0	0	0	0	1

Table 4
Example of one-hot encoding of bit types.

Well depth, m	Bit type	
	Tri-Cone	PDC
200	1	0
400	1	0
600	1	0
800	1	0
1000	1	0
1200	1	0
1400	0	1
1600	0	1
1800	0	1
2000	0	1
…	…	…

data set X into k (k < m) clusters (C_1, C_2, …, C_k):

$$X = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(m)} \end{bmatrix}, \quad x^{(i)} \in R^n \tag{4}$$

Fig. 6 is the flow chart of the K-means method. The K-means method needs to specify the value of k in advance, and the elbow method and contour coefficient method are used to determine the best k.

The key index of elbow method is the sum of squares due to error (*SSE*), which represents the sum of squares of the distance from the point of each cluster to the center of that cluster (Eq. (5)).

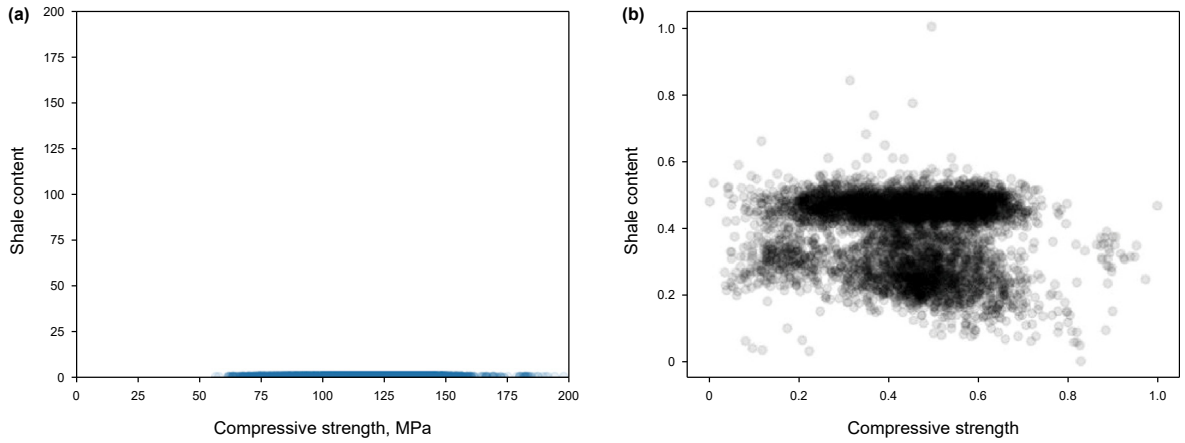


Fig. 4. Distribution map of compressive strength and shale content: (**a**) before normalization; (**b**) after normalization.

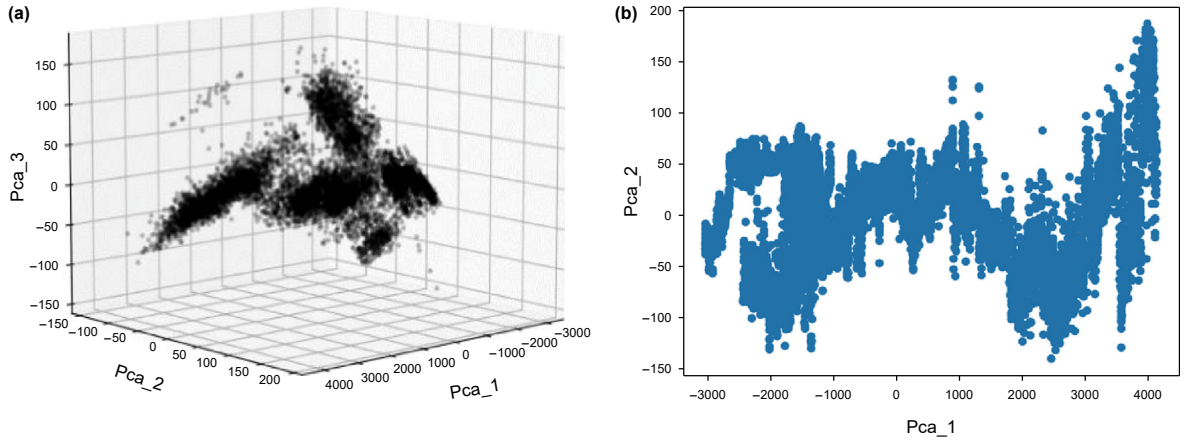


Fig. 5. Schematic diagram of dimension reduction of five-dimensional data: (**a**) reduced to three dimension; (**b**) reduced to two dimension.

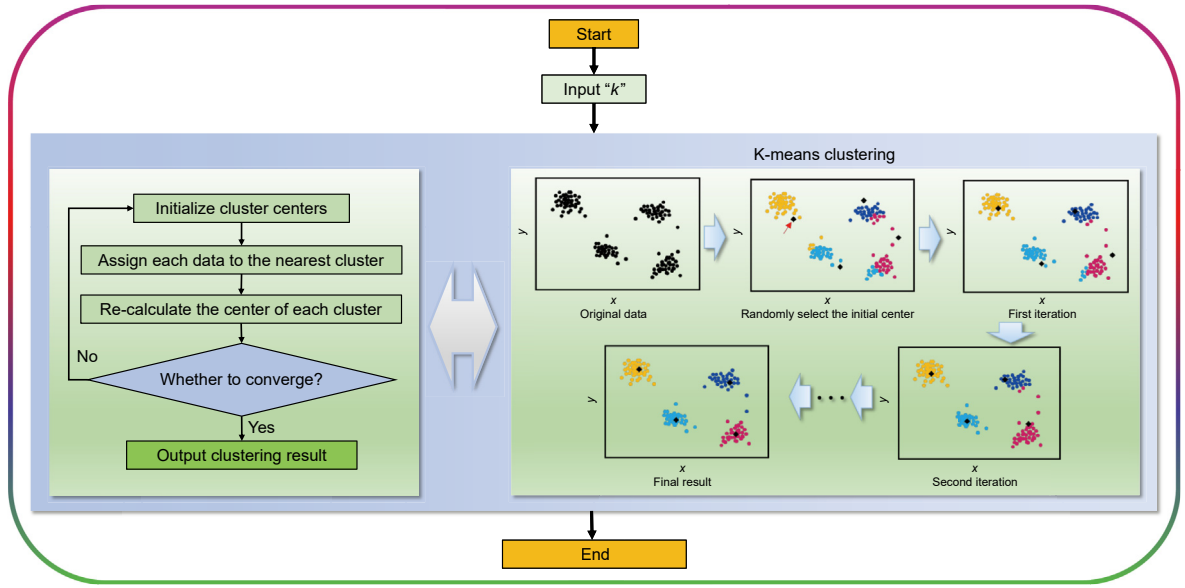


Fig. 6. Process of K-means clustering.

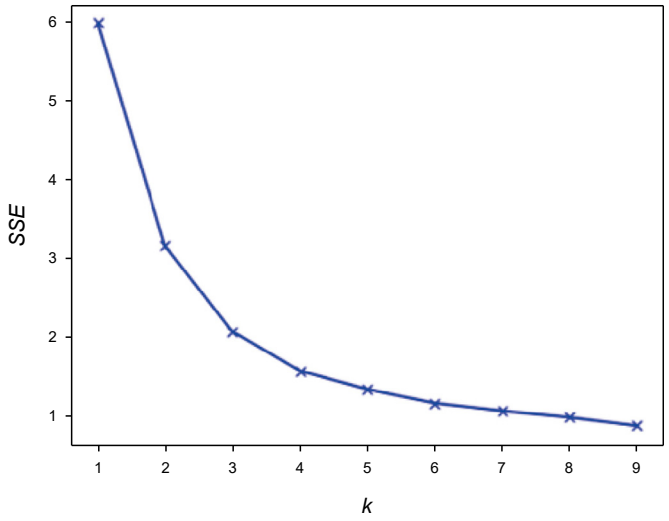


Fig. 7. Variation diagram of *SEE* of different cluster numbers (example).

$$SSE = \sum_{i=1}^{k} \sum_{p \in c^{(i)}} |p - m_i|^2 \tag{5}$$

When *SSE* drops suddenly and slowly, the k at the turning point is the best cluster number. As shown in Fig. 7, when the value of k is less than 4, the *SSE* obviously decreases with the increase of k, which indicates that the clustering effect is prominent. When k is greater than 4, the further reduction of *SSE* becomes unapparent, so $k = 4$ is the best cluster number in this case.

However, the elbow method may fail. As shown in Fig. 8, if *SEE* gradually decreases with the increase of k, it is difficult to find the turning point k for the best cluster number. Under such circumstance, the contour coefficient method is used as an auxiliary method to locate the optimal cluster number k.

The key index of contour method is the contour coefficient, which describes the clarity of contour between clusters after clustering. The calculation formulas are shown in Eqs. (6)–(9). The larger the contour coefficient, the better the clustering effect. In practical application, the elbow method is preferred to judge the optimal cluster number, while the contour coefficient method is used for auxiliary verification. If the conclusions obtained by elbow method and contour coefficient method are inconsistent, the result of contour method should be used as the final value.

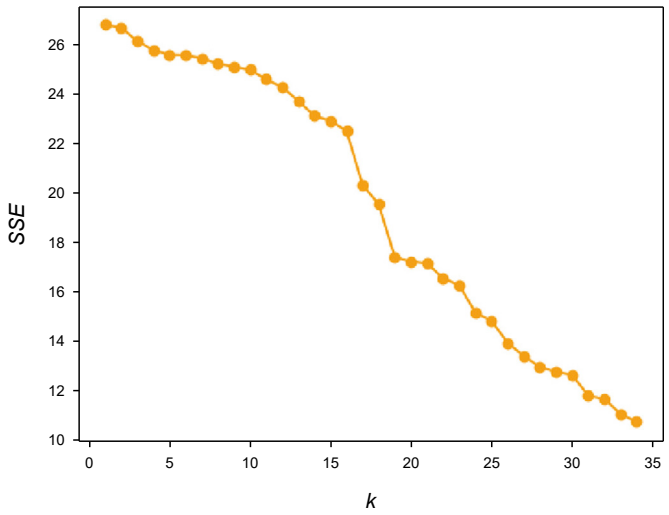


Fig. 8. Invalid judgment of elbow method (example).

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} = \begin{cases} 1 - \dfrac{a(i)}{b(i)}, & a(i) < b(i) \\ 0, & a(i) = b(i) \\ \dfrac{a(i)}{b(i)} - 1, & a(i) > b(i) \end{cases} \tag{6}$$

$$a_i = \frac{1}{n-1} \sum_{j \neq i}^{n} distance(i,j) \tag{7}$$

$$b_i = \frac{1}{n-1} \sum_{j \neq i}^{n} distance(i,j) \tag{8}$$

$$S = \frac{1}{n} \sum_{i=1}^{n} S(i) \tag{9}$$

The purpose of formation clustering is to classify formations with similar rock characteristics into the same cluster. The amplitude A, standard deviation σ, and variation coefficient V are used to describe the similarity of data sets:

$$A = \frac{Max - Min}{Min} \tag{10}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{m}} \tag{11}$$

$$V = \frac{\sigma}{\overline{X}} \tag{12}$$

2.4. Intelligent prediction model of ROP

At present there are two ways to build prediction methods of ROP: 1) Establishing a multivariate equation about ROP through a mathematical model, and predicting ROP by obtaining equation parameters; 2) Using the algorithm model where the ROP is predicted by data fitting. The second method has been greatly developed with the development and popularization of big data technology. To optimize drilling parameters, it is necessary to analyze the complicated drilling data and select the drilling parameters which are more favorable for the prediction of ROP.

As shown in Fig. 9, the data affecting ROP mainly include mechanical parameters (WOB, RPM, torque), hydraulic parameters (flow rate, SPP), drilling fluid properties, drill assembly (drill bit, speed-up tools), rock properties (Wang and Guang, 2022; Li et al., 2020; Pessier, 1992). The rock characteristics can be represented by logging data. According to the correlation coefficient between ROP and drilling data (Fig. 10), hook height barely has any correlation with ROP and other parameters. Well depth, mud temperature, and mud conductivity only have weak correlation (<0.6) with ROP. Hook weight has a strong correlation with ROP but it is similar as WOB. On the other hand, WOB, RPM, SPP, torque, flow rate, drilling fluid density, and borehole diameter are highly correlated to ROP (correlation coefficient>0.6). Therefore in this paper, formation clusters, WOB, RPM, SPP, torque, flow rate, drilling fluid

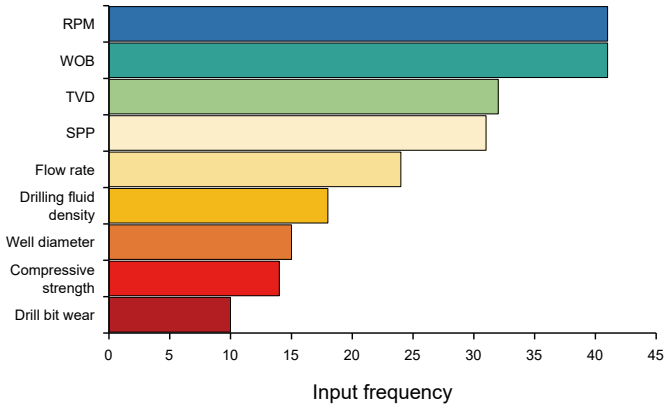


Fig. 9. Input frequency of some parameters in ROP prediction models.

density, bit type and borehole diameter are selected as input parameters of the ROP prediction model.

Artificial Neural Network (ANN) is a kind of analog logic algorithm that simulates human brain neurons to transmit and convert information cooperatively. ANN has many advantages, such as adaptability, generalization and easy realization, and is suitable for fitting drilling data and predicting ROP. The early ANN model is extremely difficult to train because of its unique structure. After the network circulates many times, in most cases, the problem of gradient disappearance or gradient explosion will occur, that is, the network training has not reached the pre-set conditions and has to be terminated in advance. The appearance of the Long Short Term Memory Network (LSTM) and Grated Recurrent Unit Network (GRU) solved the problem of gradient explosion (Gao et al., 2021). Recurrent neural network (RNN) has become a mature machine learning algorithm. The LSTM, as an improved model of RNN, makes up for some problems, but its neuron structure is too complicated and its operation process is complicated. GRU simplifies the network structure on the basis of LSTM, improves the operation speed, and solves the over-fitting problem of LSTM. Therefore, in this paper, GRU is selected to predict ROP.

Fig. 11 shows the structure of neurons in the hidden layer of GRU network (Aemail et al., 2022). The final output value and signal output h_t calculated by this neural network are as follows:

$$h_t = (1 - z_t) \times h_{t-1} + z_t \times \tilde{h}_t \tag{13}$$

$$z_t = \sigma'\left(W_z \cdot [h_{t-1}, x_t] + b_f\right) \tag{14}$$

$$r_t = \sigma'(W_r \cdot [h_{t-_1}, x_t]) \tag{15}$$

$$\tilde{h}_t = \tanh(W_{\tilde{h}} \times [r_t \cdot h_{t-1}, x_t] + b_{\tilde{h}}) \tag{16}$$

where z_t refers to the update gate. h_{t-1} refers to the output signal of a neuron on the same layer. h_t is the output signal of this neuron. x_t refers to the input of this neuron. W_z refers to the weight of the update gate. σ' refers to sigmoid function. r_t is the reset gate. W_r refers to the weight of the reset gate. $\tilde{h}_t$ is a pending output value. $W_{\tilde{h}}$ refers to the weight of the pending output value. $b_{\tilde{h}}$ refers to the compensation value of the pending output value.

Although the increase in the number of hidden layers may be more conducive to calculation, it also increases the complexity of the network and reduces the efficiency of the algorithm, making it more prone to over-fitting. Therefore, to prevent over-fitting caused by too many hidden layers, the number of hidden layers is set as one layer.

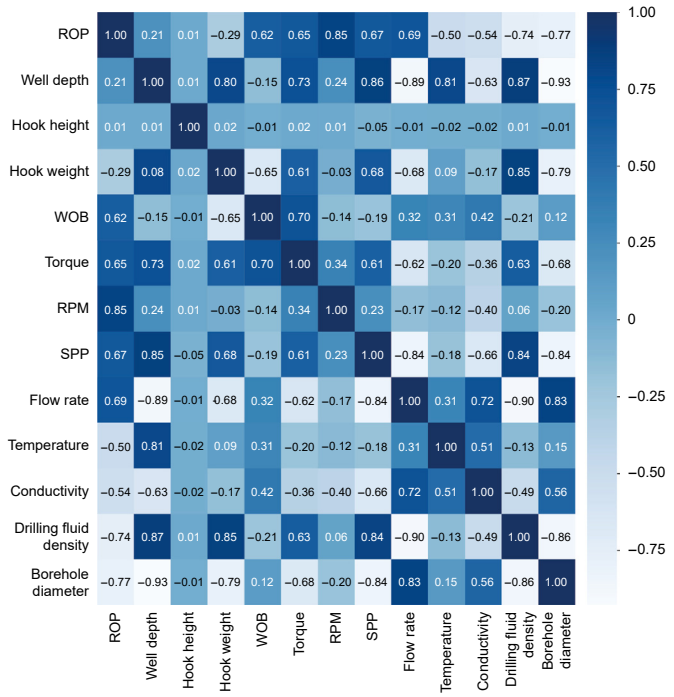


Fig. 10. The correlation coefficient between the ROP and drilling data.

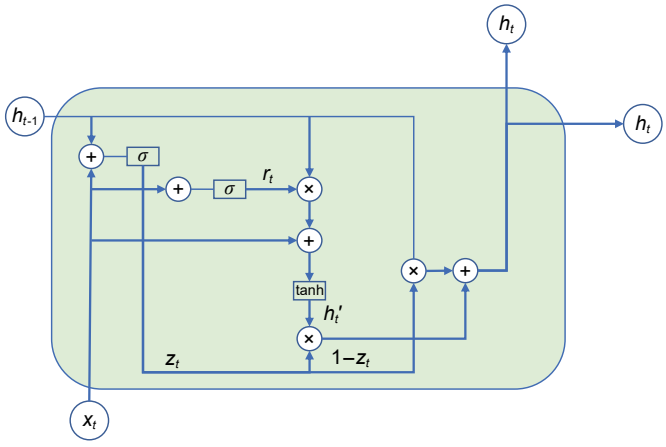


Fig. 11. Structure of neurons in the hidden layer of GRU network.

The determination of the number of neurons in the hidden layer has always attracted much attention. At present, there is no rigorous theory to verify the number of neurons in the hidden layer. The golden section method is the most widely used method (Xia et al., 2005).

$$h' = \sqrt{m' + n'} + a' \tag{17}$$

where h' is the number of neurons in the hidden layer. m' is the number of neurons in the input layer. n' is the number of neurons in the output layer. a' is the adjustment constant, and its value range is [1,10].

The root mean square error (*RMSE*) and the coefficient of determination (R^2) are used to judge the optimal neural network structure with the average value of multiple calculation results as the final judgment condition. *RMSE* is used as the main judgment index, and R^2 is used as the auxiliary judgment index.

$$RMSE = \frac{RSS}{j} \tag{18}$$

$$R^2 = \frac{RSS}{SST} \tag{19}$$

$$SST = \sum_{i=1}^{j} (Y_i - \overline{Y})^2 \tag{20}$$

$$RSS = \sum_{i=1}^{j} (\widehat{Y}_i - \overline{Y})^2 \tag{21}$$

where Y_i is the actual value. $\overline{Y}$ is the average value. $\widehat{Y}_i$ is the fitting value. *SST* is the sum of squares of total deviation. *RSS* is the sum of regression squares. j is the total number of samples.

The closer the *RMSE* is to 0, the better the neural network structure. The closer R^2 is to 1, the better the neural network structure.

According to the golden section method (Eq. (17)), when the number of neurons in the input layer is 9 and the number of neurons in the output layer is 1, the optimal number of neurons in the hidden layer ranges from 5 to 14. Table 5 shows *RMSE* and R^2 of GRU network with different numbers of neurons. It can be found that when the number of neurons in the hidden layer is 12, $RMSE = 0.179$ and $R^2 = 0.94$, and the training effect of the model is the best, so the number of neurons in the neural network model is set to 12.

In addition, the batch-size value also has a great influence on the training model. At present, there is no unified view on the optimal selection of batch-size value. According to previous studies, batch-size is set to be less than or equal to 16 when the data set is less than 100,000, and set to be 32 or 64 when the data set is more than 100,000. The data sets in this paper are all more than 100,000, so the batch-size value is initially set to 32. Finally, the ROP prediction model based on the GRU neural network is constructed (as shown in Fig. 12). In this study, a training/testing data ratio of 8:2 is used for the ANN model. The focus is primarily on optimizing the model's performance with the available data. An unseen dataset beyond the training and testing data is not included.

2.5. Evaluation index

The performance of ROP prediction model is evaluated by mean absolute error (*MAE*) and mean absolute percentage error (*MAPE*). *MAE* is the absolute error between the predicted ROP y^{pre} and the actual ROP y, and the formula is as follows:

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|y_i^{pre} - y_i\right| \tag{22}$$

where m is the number of data points. y_i^{pre} is the i-th predicted ROP, m/h. y_i is the i-th real ROP, m/h.

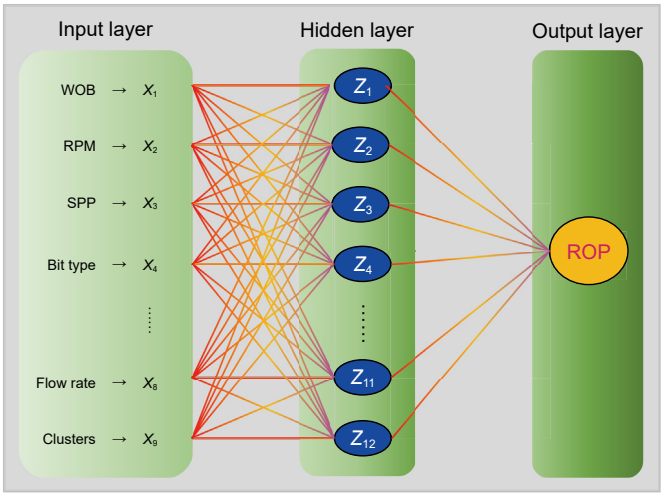


Fig. 12. GRU neural network model of ROP prediction.

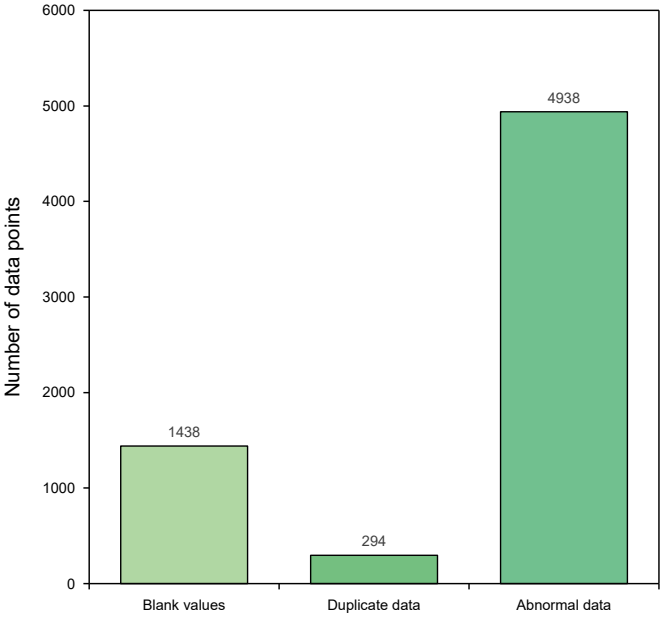


Fig. 13. Amount of invalid data in the original drilling data.

MAPE measures the relative error between y^{pre} and y (Eq. (23)). Generally, the smaller the *MAE* and *MAPE*, the higher the performance of the model.

$$MAPE = \frac{1}{m}\sum_{i=1}^{m}\frac{|y_i^{pre} - y_i|}{y_i} \times 100\% \tag{23}$$

Table 5
Error table of GRU neural network in different models.

Number of neurons	5	6	7	8	9	10	11	12	13	14
RMSE	0.34	0.427	0.274	0.255	0.453	0.342	0.424	0.179	0.187	0.269
R^2	0.657	0.266	0.82	0.846	0.083	0.676	0.733	0.94	0.938	0.84

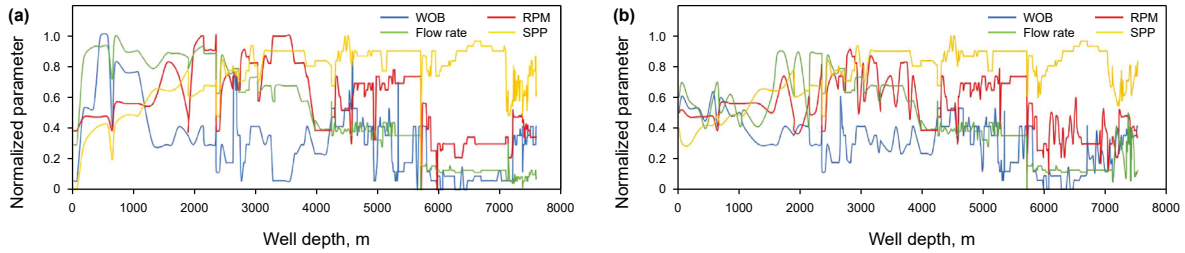


Fig. 14. Distribution of some parameters with well depth: (**a**) before data preprocessing; (**b**) after data preprocessing.

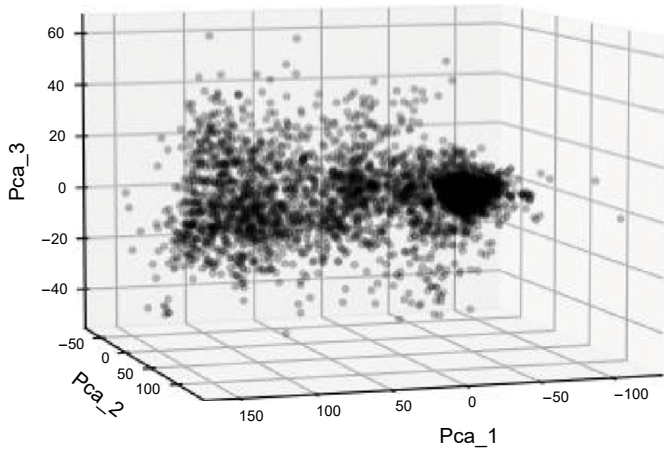


Fig. 15. Spatial distribution diagram of data after dimension reduction of data set.

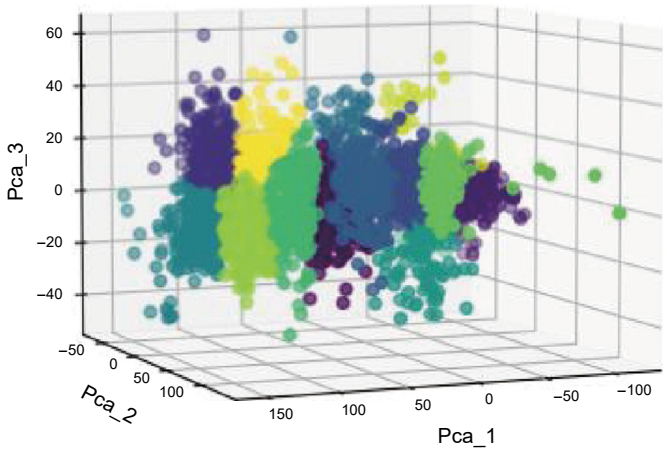


Fig. 17. Distribution of formation clusters.

3. Results and discussion

3.1. Preprocessing results of data sets in Southern Margin block

Based on the proposed method, the drilling parameters in the Southern Margin block are processed. Data sets without data preprocessing are filled with a lot of vacancy values, duplicate values, and various abnormal points. The data to be processed is summarized in Fig. 13. There are 1438 blank values, 294 duplicate data, and 4938 abnormal data. Total of 6670 pieces of data will negatively affect the algorithm operation, accounting for 6.1% (less than 10%) of the original data set. Thus, direct deletion is adopted to reduce the impact of invalid data.

Fig. 14 shows the distribution of parameters of data sets with well depth before and after deletion of invalid data. The normalized parameters are more concentrated after deletion, which indicates that the quality of data sets has been improved.

There are 95,452 data sets after data preprocessing. Fig. 15 is a three-dimensional scatter diagram of the data set. It can be seen that most of the data are aggregated to some extent, and a few of them are scattered. The data set after data preprocessing and dimension reduction can improve the quality of data processing.

3.2. Cluster analysis results based on formation characteristics in Southern Margin block

The density logging data, sonic logging data, and gamma logging data of the Southern Margin block are used as input parameters for

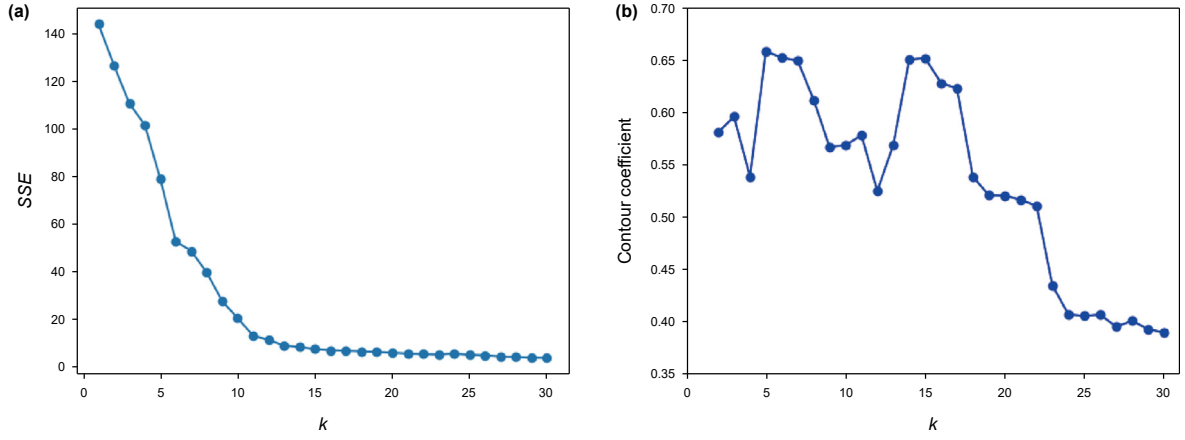


Fig. 16. Determination of optimal *k*: (**a**) elbow method; (**b**) contour coefficient method.

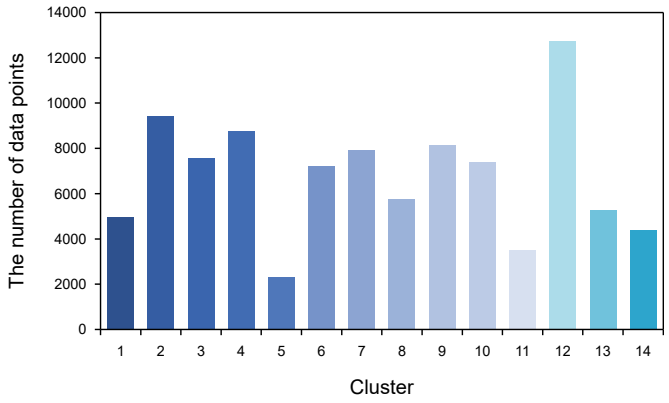


Fig. 18. The number of data points in each formation cluster.

cluster analysis. The best range of k judged by elbow method is 12~14 (as shown in Fig. 16(a)). The contour coefficient method is used as an auxiliary method to verify the elbow method, and the optimal cluster number is decided to be 14 (Fig. 16(b)).

Fig. 17 shows the distribution of formation clusters in a 3D space after clustering. The characteristics of formations in the Southern Margin block are clustered into 14 clusters. Fig. 18 shows the comparison of the number of data points in each cluster. It can be seen that the number of data points in each cluster is quite different. Cluster 12 has the most data points, with 12,753 pieces of data. Cluster 5 has the fewest data points, with only 2321 pieces of data.

3.3. Verification of formation cluster

Take the 5700–7601 m deep formation of Well HT1 in the

Table 6
Cluster results of deep formation in Southern Margin block.

Clusters	Main geological formation	Characteristics of rocks	Data volumes
Cluster 1	K_1h (lower part)	Compressive strength is 150–160 MPa, internal friction angle is 31.0–34.6.	3001
	K_1q (upper part)	Compressive strength is 151–200 MPa, internal friction angle is 32.0–35.4	1981
Cluster 3	K_1l (upper part)	Compressive strength is 135–148 MPa, internal friction angle is 39.0–43.4	3185
	K_1q (lower part)	Compressive strength is 139–147 MPa, internal friction angle is 39.0–43.4	4387
Cluster 4	K_1l (middle)	Compressive strength is 132–155 MPa, internal friction angle is 38.1–39.7	8763
Cluster 6	K_1h (upper part)	Compressive strength is 172–191 MPa, internal friction angle is 37.3–38.9	4221
	J_3k	Compressive strength is 175–200 MPa, internal friction angle is 36.7–38.7	2981
Cluster 8	K_1h (upper part)	Compressive strength is 124–142 MPa, internal friction angle is 38.1–40.4	5761
Cluster 9	K_1l (lower part)	Compressive strength is 140–150 MPa, internal friction angle is 37.0–39.9	8132
Clusters 2, 5, 7, 10, 12, 13	Mainly distributed in other upper geological structures		45114
Clusters 11, 14	Scattered distribution and accounting for less than 8% in total		7925

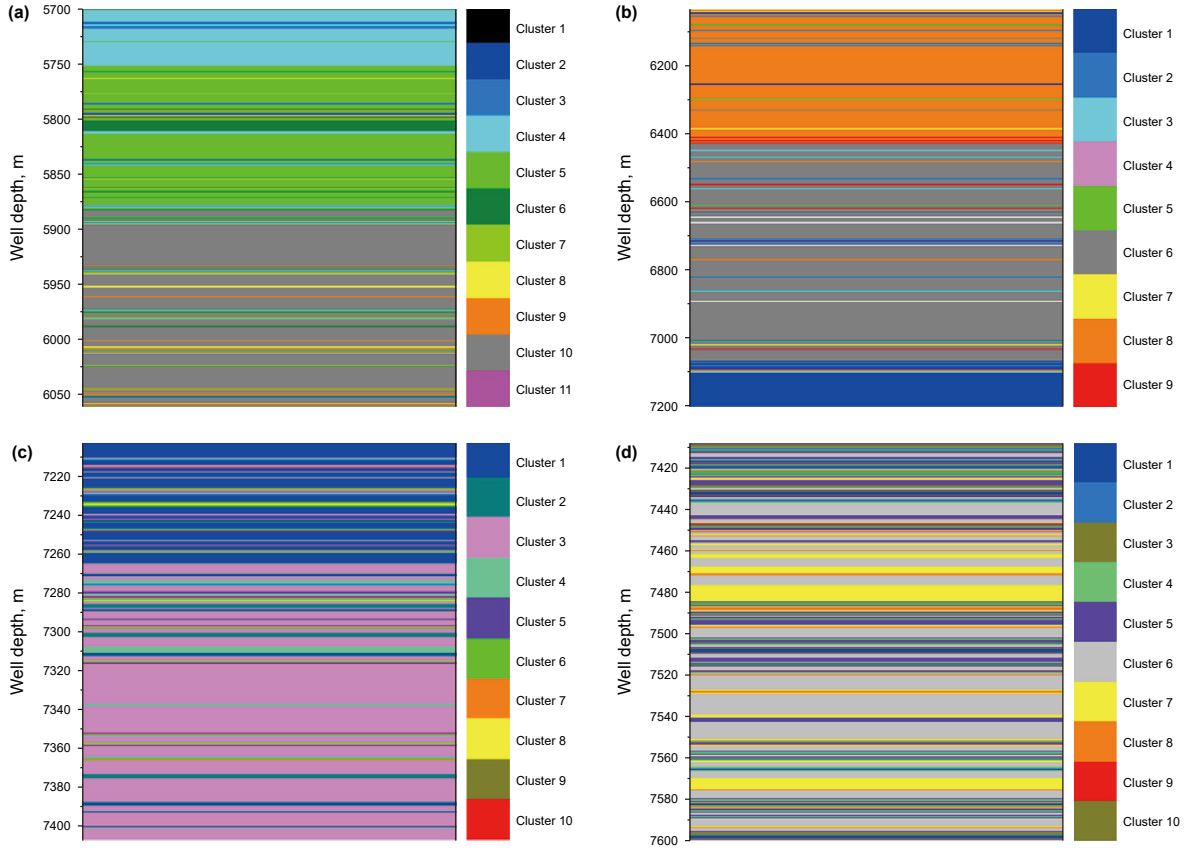


Fig. 19. Cluster results of deep formations: (**a**) K_1l; (**b**) K_1h; (**c**) K_1q; (**d**) J_3k.

Table 7
Variation coefficient of compressive strength of deep formation in Southern Margin block before and after clustering.

	Category	Data size	Average compressive strength, MPa	Standard deviation of compressive strength	Variation coefficient	Weight of variation coefficient
Before clustering	K_1l	8268	140.2	9.35	0.06669	0.00944
	K_1h	18650	134.1	12.84	0.09574	0.03287
	K_1q	9750	140.9	21.28	0.15102	0.05983
	J_3k	5744	176.6	14.51	0.08216	0.0036
	Weighted average			**0.11933**		
After clustering	Cluster 1	4982	138.34	12.01	0.08681	0.02577
	Cluster 3	7572	186.13	13.34	0.07167	0.00499
	Cluster 4	8763	100.98	11.59	0.11477	0.00291
	Cluster 6	7202	149.48	13.23	0.08850	0.01393
	Cluster 8	5761	129.79	15.25	0.11749	0.03022
	Cluster 9	8132	168.22	16.23	0.09648	0.00754
	Weighted average			**0.09129**		

Southern Margin block as an example, there are four formations: Lianmuqin Formation (K_1l), Hutubi Formation (K_1h), Qingshuihe Formation (K_1q), Kalaza Formation (J_3k). Table 6 shows the results of formation cluster.

Fig. 19 shows the clustering results of each data point in the deep formation. The depth of Lianmuqin Formation ranges from 5700 to 6062 m, with a total of 362 data points. Cluster 3, cluster 4, and cluster 9 include the most data, about 91%. Thus, Lianmuqin Formation is separately classified into cluster 3, cluster 4, and cluster 9 to optimize drilling parameters. The depth of Hutubi Formation ranges from 6062 to 7203 m. Cluster 1, cluster 6, and cluster 8 account for the most data, about 95%. Therefore, Hutubi Formation is classified into cluster 1, cluster 6, and cluster 8. The depth of Qingshuihe Formation ranges from 7203 to 7408 m, and cluster 1 and cluster 3 account for the most data, about 79%. So Qingshuihe Formation is classified into cluster 1 and cluster 3. The well depth of Kaladza Formation ranges from 7408 to 7601 m. Cluster 6 accounted for 59% of data points, and none of the other cluster data points exceeded 10%. Therefore, Kaladza Formation belongs to cluster 6.

Table 7 shows the standard deviation and variation coefficient of deep formation in the Southern Margin block before and after clustering. The weighted average variation coefficient of each formation represents the dispersion degree of formation characteristics. It can be found that the weighted average variation coefficient of formation before clustering is 0.11733, which is greater than 0.1, and it is classified as medium variation. After clustering, the weighted average variation coefficient of formation is 0.09129, which is a statistically weak variation degree. The weighted average variation coefficient decreases by 30.8%. The results show that the data of the same cluster is more concentrated after clustering, and the rock characteristics of the same cluster are more similar. The influence of formation rock characteristics on ROP can be regarded as the same in the same cluster. Thus, it can potentially improve the accuracy of the prediction of ROP.

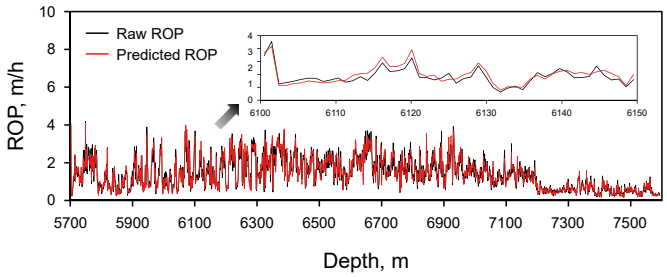


Fig. 20. Raw and predicted ROPs vs. measure depth.

3.4. Influence of formation cluster on ROP prediction model

Previous researches on ROP prediction by neural networks usually utilize the feature parameters in a direct way. That is, the models take all the drilling and logging parameters as input. This paper uses the clustering algorithm to cluster the formations, and distinguishes the data points with different geological features. Furthermore, the cluster results are used as one of the input parameters for the ROP prediction model. In this way, the geological factors on ROP is considered, which ensures the accuracy of ROP model with less feature parameters than the direct way. The major difference between previous and present models lies in whether the formation cluster is used as an input parameter of the ROP prediction model.

The ROP of 5700–7601 m in Well HT1 is predicted by the two methods. The data set contains 1901 pieces of data and is divided into 6 clusters. Both methods adopt the same GRU neural network structure and change the number of neurons in the hidden layer to get the best result. Table 8 shows the *MAE* and *MAPE* of the prediction results of two ROP prediction methods. As can be found, for the prediction results of the six clusters, the *MAE* ranges from 0.43 to 0.58, and the *MAPE* ranges from 9.12% to 12.06%. The *MAE* and *MAPE* of the ROP prediction results based on the clustering results (as shown in Fig. 20) are 0.49 and 10.56% respectively. For the prediction of ROP with all parameter input, the *MAE* is 0.82, and the

Table 8
The *MAE* and *MAPE* of the prediction results of two ROP prediction methods.

Evaluation index	Cluster						Prediction of ROP based on clustering results	Prediction of ROP with all parameters input
	1	3	4	6	8	9		
MAE	0.43	0.49	0.51	0.48	0.58	0.47	**0.49**	**0.82**
MAPE	9.12%	10.12%	11.45%	10.34%	12.06%	10.05%	**10.56%**	**18.72%**

MAPE is 18.72%.

The *MAPE* of prediction of ROP after clustering is decreased from 18.72% to 10.56%, which indicates that the similarity of formation and rocks in a cluster is high. Therefore, the method adopted in this paper improves the accuracy of ROP prediction.

3.5. Field application scenario of the model

The model can be utilized for both predrilling design and real-time optimization of drilling parameters. By collecting mud logging and formation logging data from drilled wells in the target block, the ROP model can be effectively trained and tested. This allows to identify optimal drilling parameters (WOB, RPM, flow rate, and SPP) tailored for each formation cluster in the predrilling design. For real-time optimization, real-time mud logging and downhole logging data can be used as model inputs to determine the most favorable drilling parameters dynamically.

It is noted that the present ROP optimization method includes mud logging and downhole data as inputs. Although these data are fully available in the Southern Margin block of Xinjiang, these inputs are not always available in every well since many companies drop them to save cost except for few logging parameters such as Gamma Ray. The use of mud logging data requires a full circulation which is not guaranteed in every well or section. The use of downhole logging is mainly attributed with the reservoir section to land the well properly. In this case the ROP is controlled to preserve the logging quality and mitigate the chance of having a poor logging data. Also, the placement of logging tools is far behind the bit which creates a lag in the model epically if a new formation is encountered and the model needs to await that formation to pass by the logging tool and identify it.

Table 9
Optimized drilling parameters.

Cluster	Optimal value of drilling parameters			
	WOB, kN	RPM, r/min	Flow rate, L/s	SPP, MPa
1	150–160	100–120	18–20	35–38
3	140–160	220–240	18–20	35–38
4	140–160	220–240	20–25	35–38
6	60–80	220–240	20–25	35–38
8	30–50	220–240	22–25	35–40
9	140–160	100–120	22–25	30–32

For cases where downhole logging data are not available in a certain well, the logging data from neighboring wells in the same geological structure can be used as substitution. In the Southern Margin block of Xinjiang, the mud logging data are recorded in ever 30 s. Ideally, the drilling fluid density can be provided in a real-time manner, along with WOB, RPM, SPP, torque, and flow rate. If drilling fluid density is not provided in real-time, the drilling density can be decided according to the drilling fluid density of drilling design at the specific measure depth. In this study, we focus on an all-in-one model to demonstrate the feasibility of integrating multiple drilling parameters into a cohesive framework. However, it is acknowledged developing separate models for different well-bore section/size has the potential benefits of reducing model inputs (especially bit size, formation type, and bit type). Moving forward, we will consider this approach in future research to further refine our

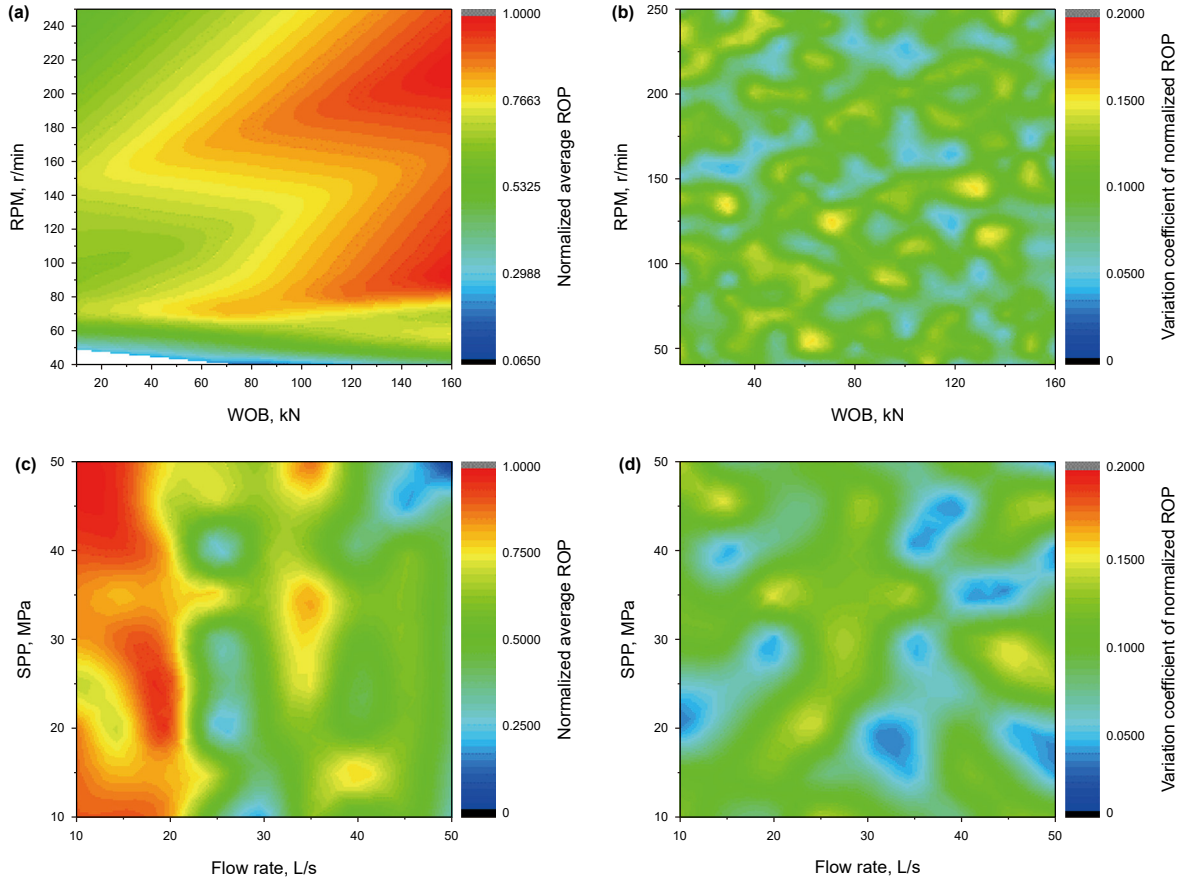


Fig. 21. Effects of different parameter combinations on ROP in cluster 1: (**a**) Normalized average ROP at different WOB and RPM; (**b**) Variation coefficient of normalized ROP under different WOB and RPM; (**c**) Normalized average ROP at different flow rate and SPP; (**d**) Variation coefficient of normalized ROP under different flow rate and SPP.

Table 10
Comparison of drilling parameter optimization results between MSE and big data analysis.

Formation	Parameters of big data training optimization					Parameters recommended by MSE				
	WOB, kN	RPM, r/min	Flow rate, L/s	SPP, MPa	Drilling mode	WOB, kN	RPM, r/min	Flow rate, L/s	SPP, MPa	Drilling mode
K_1l (upper part)	140–160	220–240	18–20	35–38	Composite drilling	70–80	70–90	18–20	30–32	Torsional impact
K_1l (middle)	60–80	220–240	20–25	35–38	Composite drilling					
K_1l (lower part)	140–160	100–120	22–25	30–32	Conventional drilling					
K_1h (upper part)	30–50	220–240	18–20	35–38	Composite drilling	60–80	140–160	25–27	34–36	Composite drilling
K_1h (middle)	60–80	220–240	20–25	35–40	Composite drilling					
K_1h (lower part)	150–160	100–120	22–25	35–40	Conventional drilling					
K_1h (upper part)	150–160	100–120	18–20	35–38	Conventional drilling	40–60	140–160	16–18	30–32	Composite drilling
K_1h (lower part)	140–160	100–120	18–20	35–38	Conventional drilling					
J_3k	60–80	220–240	20–25	35–40	Composite drilling	50–55	60–90	16–18	36–38	New tools

models and enhance predictive performance.

4. Optimization of drilling parameters in Southern Margin block

4.1. Optimization of drilling parameters (5700–7601 m deep formation)

After the ROP model is trained and tested, the influence of different parameter combinations on ROP is analyzed to optimize the drilling parameters. As in Section 3.3, take the 5700–7601 m deep stratum of Well HT1 in the southern margin as an example. The drilling parameters (WOB, RPM, SPP, flow rate) are optimized based on the parameter combination to obtain the maximum ROP.

Fig. 21 shows the effects of different parameter combinations on ROP in cluster 1. As can be seen, the normalized average ROP of all data points in cluster 1 under different WOB and RPM (as shown in Fig. 21(a)), which reflects the relationship between ROP and mechanical parameters in cluster 1. It can be seen that when WOB is constant, increasing RPM has no obvious change in ROP, but when RPM is constant, increasing WOB will increase ROP. Therefore, the formation in cluster 1 is suitable for the parameter combination of high WOB and proper RPM. It is recommended that WOB should be 150–160 kN and RPM should be 100–120 r/min. The conventional drilling method with no PDM (positive displacement motor) should be adopted. Fig. 21(b) shows the variation coefficient of the normalized ROP of each point in cluster 1 at different WOB and RPM. It can be seen that the coefficient of variation under most parameter combinations is less than 0.1, and only a few are greater than 0.2. The average coefficient of variation under each parameter combination is 0.0871, which is a weak variation, indicating that different WOB and RPM have similar influence on the ROP of each data point in cluster 1. Fig. 21(c) is the normalized average ROP of each data point of cluster 1 under different flow rate and SPP, reflecting the relationship between ROP of each data point of cluster 1 and hydraulic parameters. It can be seen that the combination of high SPP and appropriate flow rate can achieve higher ROP. Therefore, the recommended flow rate should be 18–20 L/s and the SPP should be 35–38 MPa. Fig. 21(d) shows the variation coefficient of the normalized ROP of each point in cluster 1 under different flow rate and SPP. It can be seen that the coefficient of variation under most parameter combinations is less than 0.1, and only a few are greater than 0.2. The average coefficient of variation under each parameter combination is 0.0643, which is a weak variation, indicating that different flow rate and SPP have similar influences on ROP.

Similarly, the optimized drilling parameters of cluster 3, cluster 4, cluster 6, cluster 8 and cluster 9 can be obtained through analysis, as shown in Table 9.

4.2. Comparison with MSE

At present, the optimization of drilling parameters based on MSE is the most widely used. In recent years, scholars have continuously optimized and improved the Teale model, and many new models have been formed on this basis, such as the Armenta model, Rashidi model, Mohan model and MSE model under compound drilling conditions (Eq. (24)). In this paper, the research focus is not on the MSE model, so it is not detailed.

$$\mathrm{MSE} = E_f W\left[\frac{4}{\pi D_b^2} + \frac{0.16\mu(qn+60Q)}{D_b\nu}\right] = E_f W\left[\frac{4}{\pi D_b^2} + \frac{0.16\mu(qn+K_N Q)}{D_b\nu}\right] \tag{24}$$

This section aims to compare the proposed intelligent optimization method of drilling parameters with MSE method. Take the 5700–7601 m deep formation in the Southern Margin of Well HT1 as an example. The optimization results of drilling parameters of deep formation in the Southern Margin block obtained by MSE and big data analysis are shown in Table 10.

In the table, composite drilling refers to adding a PDM in bottom hole assembly, and torsional impact refers to adding a torsion impactor in bottom hole assembly. It should be noted that drilling parameters optimization model based on big data re-divides the formation, so that drilling parameters can be optimized in a more delicate way for each formation. On the other hand, the MSE model used to optimize hydraulic parameters has many parameters that are difficult to obtain directly, so it is difficult to optimize hydraulic parameters. The drilling parameter optimization model based on big data machine learning can fully combine drilling data to optimize hydraulic parameters.

5. Conclusions

This paper establishes a drilling parameter optimization method based on big data of drilling and machine learning, and optimizes the drilling parameters in drilling of deep formations at the Southern Margin block of Xinjiang. The main conclusions are as follows:

(1) A clustering model of formation features based on the K-means algorithm is established. Elbow method and contour coefficient method are used to decide the best cluster number. The cluster data of the 5700–7601 m deep formation of Well HT1 in the Southern Margin block shows that each cluster has more uniform formation and rock characteristics.
(2) The *MAE* and *MAPE* of the ROP prediction results based on the clustering results are 0.49 and 10.56% respectively. For the prediction of ROP with all parameter input, the *MAE* is 0.82, and the *MAPE* is 16.52%. The *MAPE* of prediction of ROP after clusteringdecreased from 18.72% to 10.56%.
(3) Mechanical parameters and hydraulic parameters of deep formation in the Southern Margin block are optimized. The model proposed in this paper provides more detailed instructions than the conventional MSE method. This method can be used to optimize drilling parameters in a delicate way for deep formations.

CRediT authorship contribution statement

Chi Peng: Writing – original draft, Supervision, Project administration, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Hong-Lin Zhang:** Writing – original draft, Validation, Software, Resources, Investigation, Formal analysis. **Jian-Hong Fu:** Supervision, Resources, Project administration, Methodology, Funding acquisition. **Yu Su:** Resources, Methodology, Conceptualization. **Qing-Feng Li:** Writing – original draft, Resources, Project administration. **Tian-Qi Yue:** Investigation, Data curation.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is financially supported by Sichuan Science and Technology Program (No. 2025ZNSFSC0373), National Natural Science foundation of China (Grant No. 52104006), and Science and Technology Cooperation Project of the CNPC-SWPU Innovation Alliance (Grant No. 2020CX040202). The authors would also like to thank Professor Michael C. Sukop of Florida International University for his suggestions and help.

Nomenclature

X	the outliers
IQR	the interquartile spacing
S	the original signal
S^*	the signal after noise reduction
C_i	the noise reduction coefficient of the i th time
N	the sliding window width of $(2m + 1)$ groups of data
$j.$	the j-th sample in the sample set
x_{ij}^*	data after normalization
x_{ij}	data before normalization
Max_j	maximum value of data of the j-th feature
Min_j	minimum value of the data of the j-th feature
$x^{(i)}$	elements in the data set
SSE	error sum of squares, dimensionless
$c^{(i)}$	i-th cluster
p	point in $c^{(i)}$
m_i	average of all samples in $c^{(i)}$
$S(i)$	contour coefficient, dimensionless
$a(i)$	average value of the dissimilarity of sample i to other points in the same cluster, dimensionless
A	amplitude, dimensionless
σ	standard deviation, dimensionless
X_i	sample i
X	average value of sample, dimensionless
m	number of samples, dimensionless
V	coefficient of variation, dimensionless
z_t	the update gate
$h_{t\text{-}1}$	the output signal of a neuron on the same layer
h_t	the output signal of this neuron
x_t	the input of this neuron
W_z	the weight of the update gate
σ'	the sigmoid function
r_t	the reset gate
W_r	the weight of the reset gate
$\tilde{h}_t$	the pending output value
$W_{\tilde{h}}$	the weight of the pending output value
$b_{\tilde{h}}$	the compensation value of the pending output value
h'	the number of neurons in the hidden layer
m'	the number of neurons in the input layer
n'	the number of neurons in the output layer
a'	the adjustment constant, its value range is [1,10]
Y_i	the actual value
$\overline{Y}$	the average value
$\widehat{Y}_i$	the fitting value
SST	the sum of squares of total deviation
RSS	the sum of regression squares
j	the total number of samples
MAE	the mean absolute error
$MAPE$	the mean absolute percentage error
y_i^{pre}	the i-th predicted ROP, m/h
m	the number of data points
y_i	the i-th real ROP, m/h
MSE	modified mechanical specific energy, MPa
E_f	coefficient, dimensionless
n	rotation speed, r/min
v	ROP, m/h
W	WOB, N
Q	flow rate, L/s
D_b	bit diameter, mm
μ	sliding friction coefficient of drill bit, dimensionless
q	displacement of drilling tool per revolution, L/r
K_N	speed-flow ratio of power drilling tools, r/L

References

Aemail, M.S., Bartoš, V., Lee, B., 2022. GRU-Based deep learning approach for network intrusion alert prediction. Future Gener. Comput. Syst. 128, 235–247. https://doi.org/10.1016/j.future.2021.09.040.

Alsubaih, A., Albadran, F., Alkanaani, N., 2018. Mechanical specific energy and statistical techniques to maximizing the drilling rates for production section of mishrif wells in southern Iraq fields. SPE/IADC Middle East Drilling Technology Conference and Exhibition. UAE, Abu Dhabi. https://doi.org/10.2118/189354-MS.

Al-Sudani, J.A., 2017. Real-time monitoring of mechanical specific energy and bit wear using control engineering systems. J. Petrol. Sci. Eng. 149, 171–182. https://

doi.org/10.1016/j.petrol.2016.10.038.
Alali, A., Akubue, V.A., Barton, S.P., et al., 2012. Agitation tools enables significant reduction in mechanical specific energy. SPE Asia Pacific Oil and Gas Conference and Exhibition. Perth, Australia. https://doi.org/10.2118/158240-MS.
Armenta, M., 2008. Identifying inefficient drilling conditions using drilling-specific energy. SPE Annual Technical Conference and Exhibition. Denver, Colorado, USA. https://doi.org/10.2118/116667-MS.
Cayeux, E., Daireaux, B., Saadallah, N., et al., 2019. Toward seamless interoperability between real-time drilling management and control applications. SPE/IADC International Drilling Conference and Exhibition. The Hague, The Netherlands. https://doi.org/10.2118/194110-MS.
Chen, Y.T., Zhang, D.X., Zhao, Q., et al., 2023. Interpretable machine learning optimization (InterOpt) for operational parameters: a case study of highly-efficient shale gas development. Pet. Sci. 20 (3), 1788–1805. https://doi.org/10.1016/j.petsci.2022.12.017.
Cherif, H., 2012. FEA Modelled MSE/UCS values optimise PDC design for entire hole section. North Africa Technical Conference and Exhibition. Cairo, Egypt. https://doi.org/10.2118/149372-MS.
Chris, C., 2021. Big data and machine learning optimize operational performance and drill-bit design. J. Petrol. Technol. 73, 49–50. https://doi.org/10.2118/1221-0049-JPT.
Cui, M., Li, J.J., Ji, G.D., et al., 2014. Optimize method of drilling parameter of compound drilling based on mechanical specific energy theory. Petroleum Drilling Techn. 42, 66–70. https://doi.org/10.3969/j.issn.1001-0890.2014.01.013 (in Chinese).
Cui, M., Zhao, J.Y., Wang, H.G., 2015. Optimizating drilling operating parameters with real-time surveillance and mitigation system of downhole vibration in deep wells. Adv. Petrol. Explor. Dev. 10, 22–26. https://doi.org/10.3968/7386.
Deng, S., Pan, H.Y., Wang, H.G., et al., 2023. A hybrid machine learning optimization algorithm for multivariable pore pressure prediction. Pet. Sci. 21 (1), 535–550. https://doi.org/10.1016/j.petsci.2023.09.001.
Dong, S.Q., Zeng, L.B., Che, X.H., et al., 2022. Application of artificial intelligence in fracture identification using well logs in tight reservoirs. J. China Univ. Geosci. 48, 1–23. https://doi.org/10.3799/dqkx.2022.088 (in Chinese).
Dupriest, F.E., 2006. Comprehensive drill rate management process to maximize ROP. SPE Annual Technical Conference and Exhibition. San Antonio, Texas, USA. https://doi.org/10.2118/102210-MS.
Elmgerbi, A.M., Ettinger, C.P., Tekum, P.M., et al., 2021. Application of machine learning techniques for real time rate of penetration optimization. SPE/IADC Middle East Drilling Technology Conference and Exhibition. UAE, Abu Dhabi. https://doi.org/10.2118/202184-MS.
Gao, Y., Wang, R., Zhou, E., 2021. Stock prediction based on optimized LSTM and GRU models. Sci. Program. 2021, 1–8. https://doi.org/10.1155/2021/4055281.
Guo, Z.M., Zhou, A.Y., 2002. Research on data quality and data cleaning: a survey. J. Software 13, 2073–2082 (in Chinese).
Hamzah, M., Erge, O., Chambon, S., 2019. Automated drilling narratives: a scalable workflow to measure the effectiveness of drilling procedures. SPE/IADC International Drilling Conference and Exhibition. The Hague, The Netherlands. https://doi.org/10.2118/194129-MS.
Hou, K., 2019. Research on Bit Selection Method Based on Big Data. Master Thesis. Southwest Petroleum University (in Chinese).
Huang, W.J., Gao, D.L., 2022. Analysis of drilling difficulty of extended-reach wells based on drilling limit theory. Pet. Sci. 19 (3), 1099–1109. https://doi.org/10.1016/j.petsci.2021.12.030.
Hutchinson, M., Thornton, B., Theys, P., et al., 2018. Optimizing drilling by simulation and automation with big data. Conference Paper. Proceedings-SPE Annual Technical Conference and Exhibition. Dallas, Texas, USA. https://doi.org/10.2118/191427-MS.
Islam, N., Vijapurapu, R., Jones, M., et al., 2018. Application of mechanical specific energy and at-the-bit measurements for geothermal drilling applications in hot, high strength, high modulus reservoirs. 52nd U.S. Rock Mechanics/Geomechanics Symposium. Seattle, Washington.
Jing, B.J., 2019. Research on Bit Selection Based on Deep Learning. Master Thesis. Southwest Petroleum University (in Chinese).
Khadisov, M., Hagen, H., Jakobsen, A., et al., 2020. Developments and experimental tests on a laboratory-scale drilling automation system. J. Pet. Explor. Prod. Technol. 10, 605–621. https://doi.org/10.1007/s13202-019-00767-6.
Li, G.S., Song, X.Z., Tian, S.C., 2020. Intelligent drilling technology research status and development trends. Petrol. Drilling Techn. 48, 1–8. https://doi.org/10.11911/syztjs.2020001 (in Chinese).
Liang, Q.M., Zou, D.Y., Zhang, H.W., et al., 2006. Predicting rock drillability by well logging: an experimental research. Petrol. Drilling Techn. 34, 17–19 (in Chinese).
Meng, Y.F., Yang, M., Li, G., et al., 2012. New method of evaluation and optimization of drilling efficiency while drilling based on mechanical specific energy theory. J. China Univer. Petrol. 36, 110–114. https://doi.org/10.3969/j.issn.1673-5005.2012.02.018 (in Chinese).
Mohan, K., Adil, F., 2009. Tracking drilling efficiency using hydro-mechanical specific energy. SPE/IADC Drilling Conference and Exhibition. Amsterdam, The Netherlands. https://doi.org/10.2118/119421-MS.
Noshi, C.I., 2019. Application of data science and machine learning algorithms for ROP optimization in west Texas: turning data into knowledge. Offshore Technol. Confer. Houston, Texas, USA. https://doi.org/10.4043/29288-MS.
Pang, H.W., Wang, H.Q., Xiao, Y.T., et al., 2023. Machine learning for carbonate formation drilling: mud loss prediction using seismic attributes and mud loss records. Pet. Sci. 21 (2), 1241–1256. https://doi.org/10.1016/j.petsci.2023.10.024.
Pei, Z.J., Song, X.Z., Wang, H.T., et al., 2024. Interpretation and characterization of rate of penetration intelligent prediction model. Pet. Sci. 21 (1), 582–596. https://doi.org/10.1016/j.petsci.2023.10.011.
Pessier, R.C., 1992. Quantifying common drilling problems with mechanical specific energy and a bit-specific coefficient of sliding friction. SPE Annual Technical Conference and Exhibition. Washington, D.C., USA. https://doi.org/10.2118/24584-MS.
Pinto, C.N., Lima, A.L.P., 2016. Mechanical specific energy for drilling optimization in deepwater brazilian salt environments. IADC/SPE Asia Pacific Drilling Technology Conference. Singapore. https://doi.org/10.2118/180646-MS.
Rafatian, N., Miska, S., Ledgerwood, L.W., et al., 2010. Experimental study of MSE of a single PDC cutter interacting with rock under simulated pressurized conditions. SPE Drill. Complet. 25, 10–18. https://doi.org/10.2118/119302-PA.
Rashidi, B., 2010. Comparative study using rock energy and drilling strength models. 44th U.S. Rock Mechanics Symposium and 5th U.S. Canada Rock Mechanics Symposium. Salt Lake City, Utah.
Sehsah, O., Ghazzawi, A., Vie, G.J., et al., 2017. Intelligent drilling system: expanding the envelope of wired drill pipe. Abu Dhabi International Petroleum Exhibition & Conference. Abu Dhabi, UAE. https://doi.org/10.2118/188321-MS.
Teale, R., 1965. The concept of specific energy in rock drilling. Int. J. Rock Mech. Min. Sci. Geomech. Abstracts 2, 57–73. https://doi.org/10.1016/0148-9062(65)90022-7.
Wang, M.S., Guang, X.J., 2022. Status and development trends of intelligent drilling technology. Acta petrolei 41, 505–512. https://doi.org/10.7623/syxb202004013 (in Chinese).
Wang, J.R., Ma, X., Duan, G.L., 2018. Improved K -means clustering k -Value selection algorithm. Computer Eng. Applica. 55, 27–33. https://doi.org/10.3778/j.issn.1002-8331.1810-0075 (in Chinese).
Waughman, R.J., Kenner, J.V., Moore, R.A., 2002. Real-time specific energy monitoring reveals drilling inefficiency and enhances the understanding of when to pull worn PDC bits. SPE Drill. Complet. 18, 59–68. https://doi.org/10.2118/81822-PA.
Xia, K.W., Li, C.B., Shen, J.Y., 2005. An optimization algorithm on the number of hidden layer nodes in feed-forward neural network. Computer Science 32, 143–145 (in Chinese).
Xiong, C., Huang, Z.W., Shi, H.Z., 2023. Performances of a stinger PDC cutter breaking granite: cutting force and mechanical specific energy in single cutter tests. Pet. Sci. 20, 1087–1103. https://doi.org/10.1016/j.petsci.2022.10.006.
Zhang, H., Lu, B., Liao, L., et al., 2021. Combining machine learning and classic drilling theories to improve rate of penetration prediction. SPE/IADC Middle East Drilling Technology Conference and Exhibition. UAE, Abu Dhabi. https://doi.org/10.2118/202202-MS.