Petroleum Science 21 (2024) 3496–3516

Contents lists available at ScienceDirect

Petroleum Science

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

Review Paper

A systematic review of machine learning modeling processes and applications in ROP prediction in the past decade

Qian Li [a, b, *], Jun-Ping Li [c], Lan-Lan Xie [a]

[a] College of Environment and Civil Engineering, Chengdu University of Technology, Chengdu, 610059, Sichuan, China
[b] State Key Laboratory of Geohazard Prevention and Geoenvironment Protection (Chengdu University of Technology), Chengdu, 610059, Sichuan, China
[c] Institute of Exploration Technology, CAGS, Chengdu, 610059, Sichuan, China

ARTICLE INFO

Article history:
Received 15 May 2023
Received in revised form
20 May 2024
Accepted 21 May 2024
Available online 25 May 2024

Edited by Jia-Jia Fei

Keywords:
Drilling
Rate of penetration (ROP) prediction
Machine learning
Accuracy evaluation

ABSTRACT

Fossil fuels are undoubtedly important, and drilling technology plays an important role in realizing fossil fuel exploration; therefore, the prediction and evaluation of drilling efficiency is a key research goal in the industry. Limited by the unknown geological environment and complex operating procedures, the prediction and evaluation of drilling efficiency were very difficult before the introduction of machine learning algorithms. This review statistically analyses rate of penetration (ROP) prediction models established based on machine learning algorithms; establishes an overall framework including data collection, data preprocessing, model establishment, and accuracy evaluation; and compares the effectiveness of different algorithms in each link of the process. This review also compares the prediction accuracy of different machine learning models and traditional models commonly used in this field and demonstrates that machine learning models are the most effective technical means in current ROP prediction modeling.

1. Introduction

Deeply buried fossil fuels (oil and gas) are an important part of the current world energy supply. Drilling is the most critical link in the petroleum industry as it is the only technical method to realize underground exploration and development. The drilling process is complex, in general, as shown in Fig. 1, the rotary table or top drive from the drilling rig drives the drill bit to reach the target position through complex formations. During the whole process, the drill string (including drilling pipes, and bottom hole assembly) is responsible for power transmission, and the circulation system (composed of a mud pump, mud pit, and solid control systems) is mainly responsible for maintaining the stability of the wellbore, cooling the drill bit and transporting cuttings.

Drilling cost accounts for 50%–80% of the total cost of exploration and development (Sun, 2006). In the face of complex and unknown geological environmental conditions, to achieve overall control of the whole drilling process, relevant parameters must be collected during the drilling operation. However, because many fields are involved, such as geology, materials, machinery, mechanics, and fluids, it is extremely difficult to analyze more than 100 kinds of coupled real-time monitoring parameters, which leads to the frequent occurrence of drilling accidents (Gan, 2019).

In the evaluation system of drilling efficiency, the rate of penetration (ROP) is the most commonly used index for the high-precision prediction, control, and optimization of drilling efficiency (Barbosa et al., 2019); thus, the ROP must be kept within a reasonable range to ensure the smooth progress of the drilling process (Najjarpour et al., 2022). In the current common ROP models, theoretical models and statistical models are limited by complex boundary conditions, such as complex combinations among drilling parameters, equipment, and formations, all of them will impact ROP, and the same parameter may have significantly different effects in different drilling wells, and such flexible impact can hardly reflect in the fixed traditional models. Numerous influencing factors and a high degree of nonlinear coupling between different factors, and the accuracy and adaptability of ROP prediction are not guaranteed in most cases (Li et al., 2021a).

Therefore, machine learning models capable of high-accuracy nonlinear fitting are considered reasonable alternatives for

[*] Corresponding author. College of Environment and Civil Engineering, Chengdu University of Technology, Chengdu, 610059, Sichuan, China.
E-mail address: liqian2014@cdut.edu.cn (Q. Li).

https://doi.org/10.1016/j.petsci.2024.05.013

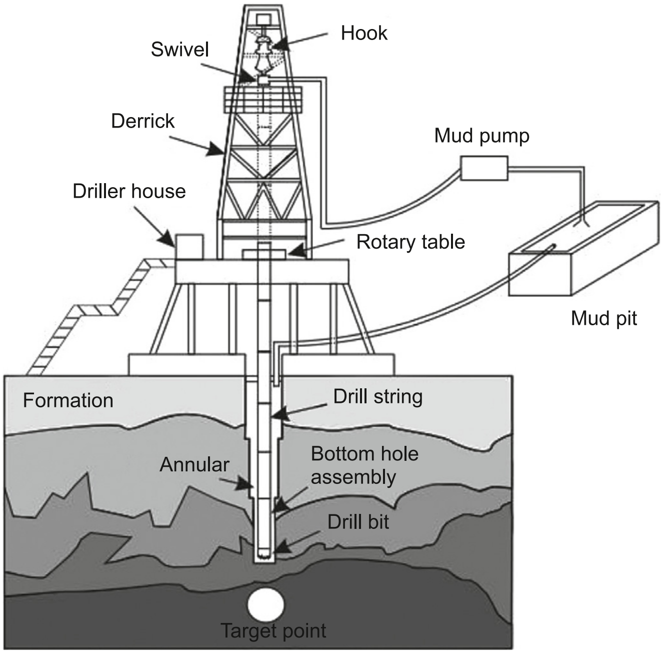

Fig. 1. Drilling process schematic (Gan et al., 2019a, b).

predicting the ROP (Soares and Gray, 2019; Brenjkar and Biniaz Delijani, 2022), and an increasing number of researchers pay more attention to this area. A typical machine learning model can be divided into three parts, including input parameters, core model, and output parameters, and the ROP is undoubtedly the regular output parameter. As for input parameters, present researchers preferred to select 6 to 8 technical parameters and 2 geological parameters (Section 2.2 for details) as the input parameters. For core modeling methods, despite the usage of ensemble algorithms increasing in recent years, single algorithms are still widely used by 80% of researchers (Section 4.1 for details), where artificial neural network and support vector machine occupy the greater proportion (over 75%). To reduce time consumption in training models, over 50% of researchers introduced optimization algorithms to find the suitable internal structure quickly. According to the proposed prediction accuracy, the machine learning algorithm significantly improves ROP prediction accuracy to over 90% (Sections 4.3 and 4.4 for details).

However, the research and development of machine learning models in the current drilling industry are too individualized, and there are obvious differences in the parameters, modeling processes, and modeling algorithms used by different researchers to build models, which sets up an immense barrier to the production application and commercial rollout of machine learning models. Specifically, the application barriers are mainly reflected in the following areas: 1) the number and range of input parameters used to build the model vary greatly and lack of explanation of the selection reason; 2) the focus on the model structure was significantly greater than how it was established and applied; 3) obvious differences in applied formation conditions; and 4) unpublished datasets due to strict data confidentiality requirements.

Accordingly, this paper presents a systematic and rule-based review of the parameters, modeling processes, and modeling algorithms used in recent years to build ROP prediction machine learning models and compares the computational differences and prediction accuracy of different algorithms. The results demonstrate that machine learning models can significantly improve the accuracy of ROP prediction and are a powerful tool to promote efficient and intelligent fossil fuel exploration operations.

2. Data collection

2.1. Target area

As a complex application system, there are a large number of factors that will affect the ROP performance. And geo-related parameters are important factors that cannot be ignored. The parameters of depth and geo-location are selected here to represent the impact of different formations on ROP prediction. It should be noted that the statistics here were not complete, because some of the papers did not indicate the location and depth of the drilling data source due to confidentiality requirements.

As shown in Fig. 2(a), a total of 63 papers provided explicit geolocation, in which there is more data collected from onshore drilling (48 samples), to a certain extent that the onshore drilling can achieve high-precision predictions more easily than offshore drilling. In terms of the depth (Fig. 2(b)), most of the researchers utilized a stable period of data to establish a prediction model, the data source of over 50% of papers were selected from depths from 1000 to 3500 m. As for the geo-location, as shown in Fig. 2(c) and (d), the data for offshore drilling were reactively evenly distributed in the China Sea, Persian Gulf, and Mexico Gulf, while the data for onshore drilling were mainly collected from China, Iran, and the United States. The data source here almost covers the main areas of oil and gas drilling globally, therefore the proposed ROP prediction performance was convincing and can be applied to most areas in the future.

2.2. Parameter types

Determining the appropriate input parameter type is the first step in establishing an accurate ROP prediction model, but there are obvious differences in the selection of modeling parameters for different models. For example, some researchers used 19 parameters to build a model (Abbas et al., 2019), and some researchers achieved accurate prediction using only two parameters (Yuswandari et al., 2019). Statistics reveal that at least 59 different parameters have been selected for ROP prediction modeling, as summarized in Table 1; 32 types of technical parameters and 27 types of geological parameters have been selected.

Early ROP prediction models only used technical parameters (Barbosa et al., 2019), but later studies found that the introduction of geological parameters could significantly improve the prediction accuracy of the model (Bezminabadi et al., 2017) and reduce the need for data samples (Hegde et al., 2017). However, due to the difficulty in obtaining accurate geological parameters, the number of geological parameters used in the current ROP models is significantly less than the number of technical parameters. As shown in Fig. 3(a), at least 50% of researchers (data comes from the statistics of 110 references) have used 5–8 technical parameters and no more than four geological parameters (averages of six and two, respectively) for modeling. From the details of the specific parameter types, as shown in Fig. 3(b) and (c), the top eight most frequently used technical parameters are WOB, RPM, Q, well depth, MW, SPP, T, and MV. The corresponding geological parameters with the highest usage frequency are only UCS and PPG because these two parameters characterize the strength of the rock.

Considering the collection time and modeling availability of the parameters listed in Table 1, the parameters can be divided into 6 groups. As shown in Fig. 4, the most direct difference is that the technical parameters can be used for establishing a prediction model without any calculation or experiments, however, none of the geological parameters can be used directly, despite some of

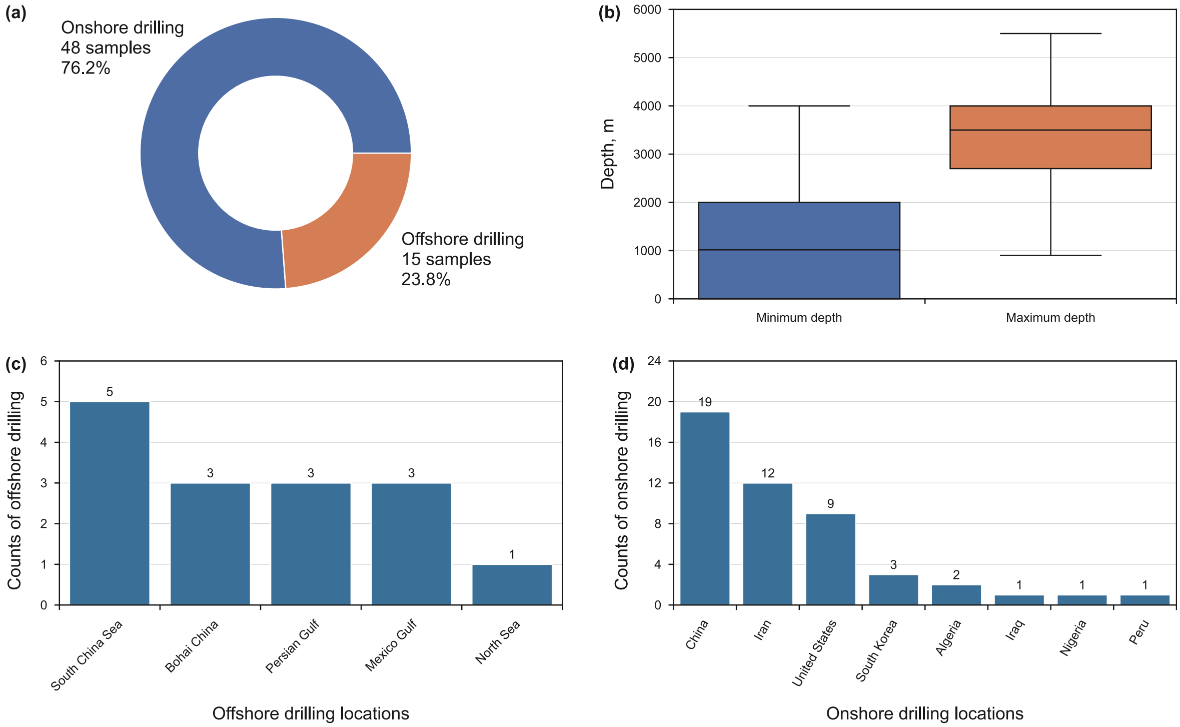

Fig. 2. Depth and geo-location distribution in this ROP prediction review ((**a**): distribution between onshore and offshore drilling; (**b**): distribution between depth statistics; (**c**)/(**d**): geo-location distribution for offshore and onshore drilling locations).

Table 1
Parameter categories selected for ROP modeling in recent years.

Parameter categories		Parameter types, abbreviation[a]
Operational	Technical	Well depth; well diameter; weight on bit, WOB; rotation speed, RPM; mud flowrate, Q; stand pipe pressure, SPP; torque, T; rotary time; incline angle, INC; azimuth angle, AZI; hook load, HL; differential pressure
Fluid	Technical	Mud weight, MW; mud viscosity, MV; equivalent circulating density, ECD; gel strengths; yield point, YP; mud temperature, TEMP; solid content, SC; filter loss, FL; fluid type; lag time; Reynolds number, R; mud tank volume; PH
Tool	Technical	Bit wear; bit hydraulics; bit types; bit structure; nozzle diameter; tool description; collar size
Mechanical	Geological	Uniaxial compressive strength, UCS; pore pressure gradient, PPG; formation drillability; formation abrasiveness; internal cohesion; internal friction angle; formation stress; vertical stress; formation brittleness; tensile strength; anisotropy index; Young's modulus; maximum horizontal stress; minimum horizontal stress; shear failure gradient; Poisson's ratio
Physical	Geological	Seismic wave speed; formation porosity; gamma ray, GR; formation density; rock quality designation, RQD; formation structure; formation type; formation content; resistivity; shale index; water content; permeability

[a] (only some of the conventional abbreviations are listed here).

them (such as physical parameters) can be obtained during drilling from LWD (logging while drilling). Before introducing to prediction model, all the geological parameters need to be processed (for example, the inversion or mechanical tests from the laboratory). The complexity was one of the main reasons why geological parameters are rarely used currently.

In terms of the collecting time, the parameters that can be collected during drilling are the most suitable parameters for establishing a prediction model, because they contain the most accurate information for the drilling at that time. And the top eight most frequently used technical parameters all belong to this category. For those parameters that accurate value only can be obtained after drilling (including UCS and PPG), they usually were introduced for evaluating ROP performance when designing in a similar area.

Notably, some studies have demonstrated that there is an upper limit on the number of modeling parameters (approximately 4–6) of an ROP prediction model. Before reaching the upper limit, increasing the types of input parameters can significantly improve the prediction accuracy of the machine learning model and reduce the training time (Ansari et al., 2017; Eskandarian et al., 2017; Ahmed A. et al., 2019; Ashrafi et al., 2019; Sabah et al., 2019; Mehrad et al., 2020; Li et al., 2021a); however, after exceeding the upper limit, new input parameters do not result in further improvement in the prediction accuracy (Ansari et al., 2017; Eskandarian et al., 2017; Ashrafi et al., 2019; Sabah et al., 2019; Mehrad et al., 2020; Li et al., 2021a), as shown in Fig. 5. Moreover, research indicates that the magnitude of the upper limit is associated with the correlation between the parameter and ROP (Liu et al., 2021).

In terms of the minimum number of input parameters required for accurate modeling, studies have indicated that the needed number decreases with the increasing correlation between parameters and ROP (Li et al., 2021a). As shown in Fig. 6 in the actual test dataset, when input parameters have a high correlation with ROP (correlation >0.6), no more than 6 parameters (regardless of

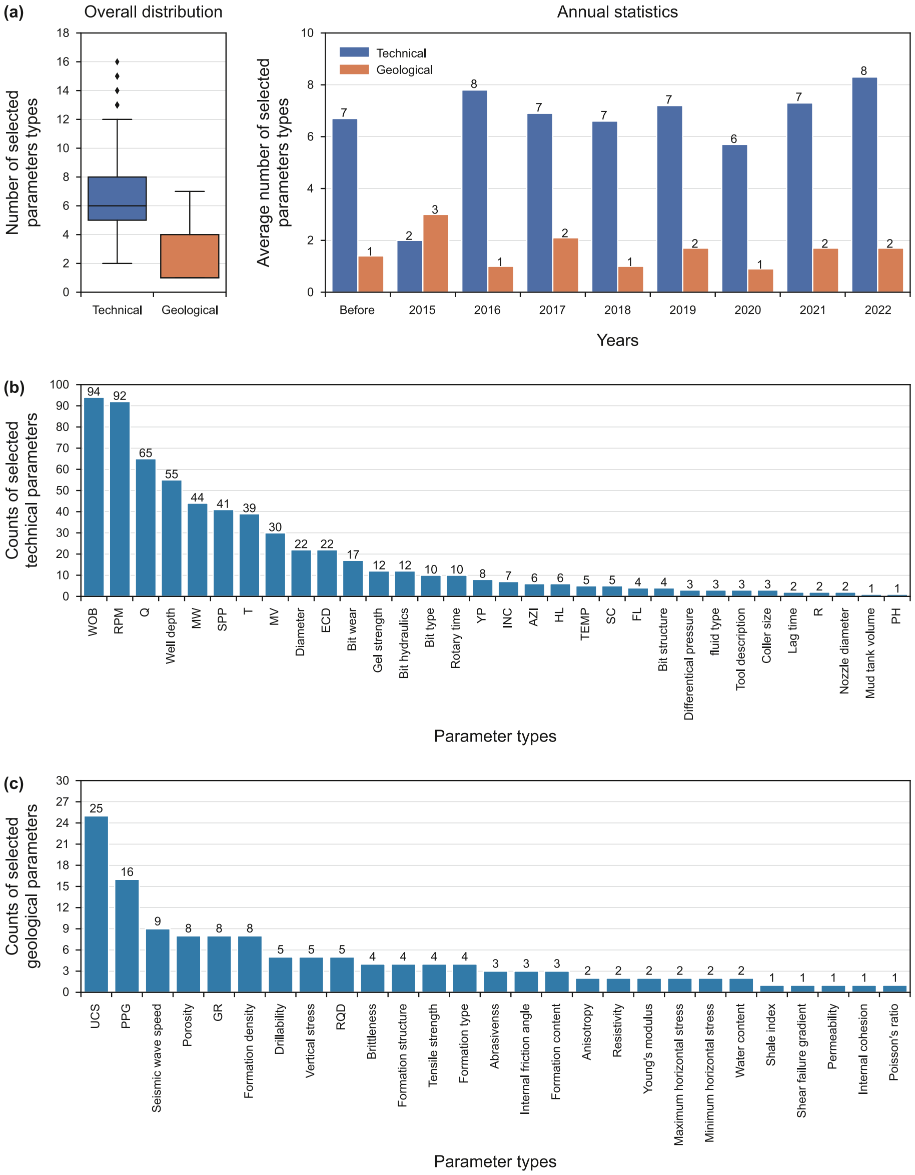

Fig. 3. Statistical comparison (**a**) and usage frequency of technical (**b**) and formation (**c**) parameters selected in ROP prediction (data comes from the statistics of 110 references).

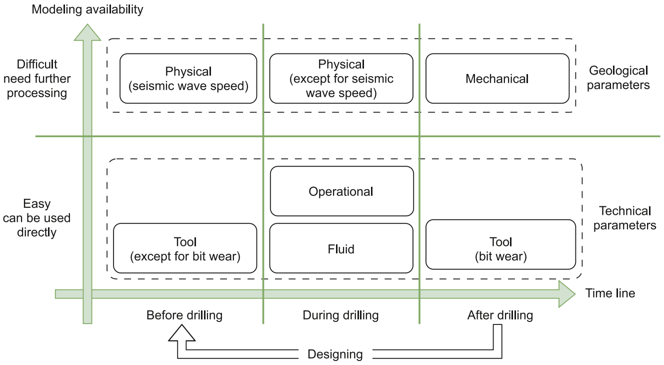

Fig. 4. Collecting timeline and modeling availability for different parameters.

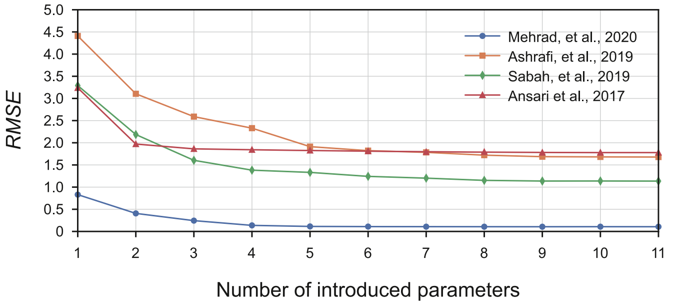

Fig. 5. Prediction accuracy improves with an increasing number of introduced parameters.

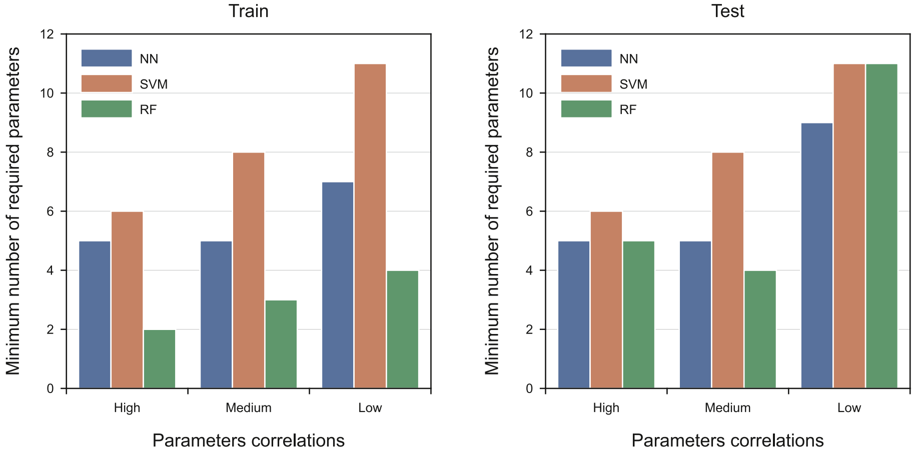

Fig. 6. Minimum number of input parameters for accurate modeling with different correlations between parameters and ROP (result data and modeling data was collected from South China Sea (Li et al., 2021a, d)).

the parameter types) can obtain accurate prediction results, while if the correlation is reduced to lower than 0.3, at least 9 to 11 input parameters will be required to achieve similar accuracy.

2.3. Dataset size

In addition to the type of modeling parameters, the accuracy of the ROP prediction model is affected by the size of the modeling dataset. Due to the efficient use of data, the amount of data required by a machine learning model to achieve the same ROP prediction accuracy as a statistical regression model is decreased by 70%–80% (Hegde et al., 2017), and with an increase in the amount of data, the prediction accuracy of an ROP machine learning model could be further improved (Hegde et al., 2017; Soares and Gray, 2019). Similar to the selection of input parameters, the dataset sizes used by different researchers are very different. The largest dataset comes from decades of collection in four regions, and the sample size is close to two million pieces of data (1,964,436 samples) (Zhu, 2021). In contrast, some researchers have pointed out that a minimum of 10 data points is already enough to yield satisfactory ROP prediction results (Soares and Gray, 2019).

For the dataset size distribution, based on the statistics shown in Fig. 7, more than 70% of the researchers used no more than 10,000 samples of data, 51.7% of the researchers used datasets between 1000 and 10,000 samples (median of 3250), 20.7% of the researchers used datasets with 100 to 1000 samples (median of 315), and datasets of other sizes were used less frequently. The possible reasons are as follows: 1) the prediction accuracy and generalization performance are not reliable when datasets that are too small (sample size <100), 2) for large datasets (sample size >10,000), it is not easily obtained for a single prediction.

3. Data preprocessing

3.1. Data cleaning

During the whole drilling process, due to the temporary failure of sensors and other reasons, the data collected on-site may be missing or incorrect to varying degrees (Zhu, 2021). The purpose of data cleaning is to eliminate the serious impact of missing data or errors on the accuracy of machine learning models before modeling (Qi, 2020).

Deletion and filling are common methods for addressing missing data. When the dataset is large enough, direct deletion of the rows with missing data is a common solution (Brenjkar et al., 2021; Fan et al., 2021; Encinas et al., 2022), but this could cause other data in rows with missing data to be lost and changes in the overall data structure, which may lead to the loss of key information, thus resulting in prediction results that do not match the actual results (Qu, 2021). When the size of a dataset is limited, researchers can fill in the missing values through numerical filling or regression filling. The conventional practice of numerical filling is to determine the filling window before and after the missing data and fill in the values according to a valid data value or mode (the value that appears most often in a set of data) before and after the missing data (Zhu, 2021). Regression filling uses existing data to construct a regression equation to fill in missing values (Zhu, 2021; Encinas et al., 2022). The choice of the type of regression equation depends on the basic characteristics of the data to be filled. Currently, the most commonly used regression equations are cubic spline interpolation (Diaz and Kim, 2020) and cubic Hermite interpolation (Gan et al., 2019b; Gan, 2019).

To compare the filling effect, experiments were carried out

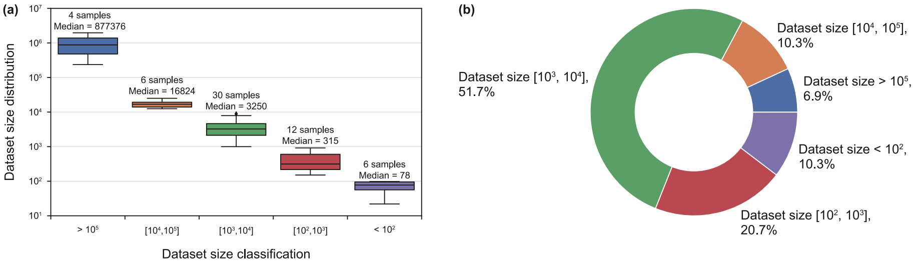

Fig. 7. Dataset size distribution (**a**) and proportion (**b**) in ROP prediction.

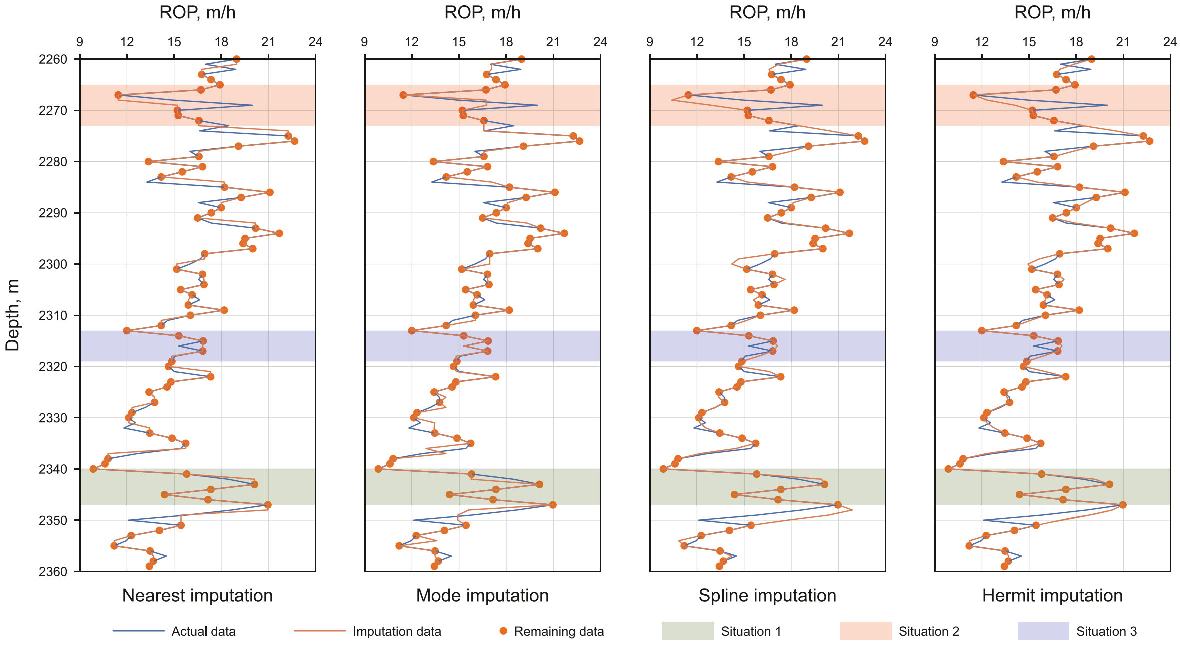

Fig. 8. Data imputation example for different methods (data were collected from the South China Sea (Li et al., 2021a)).

based on a group of 100 consecutive actual ROP samples collected from the South China Sea (Li et al., 2021a). Then, 30 actual samples were randomly deleted and replaced by the value calculated from four different imputation methods. As shown in Fig. 8, to select a suitable method for data imputation, the following three situations should be considered:

- Situation 1: When the missing point is not a valley or a peak and the data trend has no obvious turning, such as the green area in Fig. 8, each method proposed can obtain acceptable filling results;
- Situation 2: When large peaks (or valleys) are missing, especially when the missing data are outside the data fluctuation range, as shown in the red area in Fig. 8, no method can recover the correct value. Among the four methods, nearest and mode imputation were slightly better than the other two methods, as they can show a small trend of missing values; however, the trend is not particularly precise.
- Situation 3: When the missing peaks (or valleys) are not sharp or when they do not fluctuate within the mean range, as shown in the blue area in Fig. 8, mode imputation is the only method to obtain the accurate value of missing points.

For the correction of erroneous data, data assimilation for measurement correction is currently mainly used. When there is a significant deviation between the measured value and the estimated value due to sensor failure and other reasons, on the premise of ensuring independence between the measured value and the estimated value, the following method (Eq. (1)) can be used (McLaughlin, 2014; Law et al., 2015; Geekiyanage et al., 2018; Encinas et al., 2022):

$$x_c = \frac{\frac{x_1}{\sigma_1{}^2}}{\frac{1}{\sigma_1{}^2 + \frac{1}{\sigma_2{}^2}}} + \frac{\frac{x_2}{\sigma_2{}^2}}{\frac{1}{\sigma_1{}^2 + \frac{1}{\sigma_2{}^2}}} \tag{1}$$

where x_c is the measurement correction value, x_1/x_2 is the measured/estimated value, and σ_1/σ_2 is the standard error of the measured value/estimated value.

3.2. Outlier removal

After data cleaning, a considerable proportion of outliers still exist in the original data, and the main reasons for these outliers including human errors (inevitable mistakes during observations and data collection), equipment malfunctions (failures in the drilling rod or pump), sudden changes in downhole conditions (wellbore collapse), machine or BHA vibrations, and data transmission interference. The determination and deletion of outliers are effective means to improve the accuracy of machine learning models (Tan, 2019; Zhu, 2021). Studies have shown that the prediction error after outlier removal is only approximately 10% of the prediction error before deletion (Tan, 2019; Tan et al., 2019).

The core of the outlier removal approach is to determine the threshold of the data, and the data exceeding the threshold are regarded as outlier data and deleted. The current conventional threshold determination methods include the Z score method (the common threshold is 3 (Al-AbdulJabbar et al., 2020; Elkatatny, 2020; Soares et al., 2020; Fan et al., 2021; Alsaihati et al., 2022), and some researchers set the threshold to 2 (Soares and Gray, 2019) or 5 (Hassan et al., 2020)), the Tukey method (threshold value is calculated by the interquartile range (IQR)) (Mehrad et al., 2020; Al-AbdulJabbar et al., 2021; Fan et al., 2021; Alsaihati et al., 2022; Encinas et al., 2022), and methods based on the original distribution (OD) (Bani Mustafa et al., 2021) and average deviation (AD) (Diaz et al., 2019), as shown in Table 2. In addition, because of the significant correlation between the actual mechanical specific energy of drilling and UCS, some researchers regard data with a poor correlation between the mechanical specific energy and UCS as outliers (Al-AbdulJabbar et al., 2021). However, this method is affected by the field of mechanical specific energy and UCS data collection. For more complex drilling-related data, a variety of outlier determination methods can be used. If a sample is determined to be an outlier by more than half of the methods, the sample should be deleted (Tan, 2019; Tan et al., 2019).

To clearly show the difference among the methods listed in

Table 2
Common methods of outlier detection in ROP prediction.

Methods	Threshold calculation	Criteria for determining outliers
Z score	$Z = (x_i - \overline{x})/\sigma$	$\lvert Z\rvert \geq 2/3/5$
Tukey (IQR)	$IQR = P_{75} - P_{25}$	$x_i \geq P_{75} + 1.5IQR$ or $x_i \leq P_{25} - 1.5IQR$
Based on the original distribution	—	$x_i \geq P_{95}$ or $x_i \leq P_5$
Based on the average deviation	$AD = \sum_{i=1}^{n} \lvert x_i - \overline{x}\rvert / n$	$\lvert x_i - \overline{x}\rvert \geq 6AD$

In the table, x_i is the initial value, $\overline{x}$ is the mean value of the dataset, σ is the standard error of the initial dataset, n is the sample number, and $P_5/P_{25}/P_{75}/P_{95}$ are the 5th/25th/75th/95th percentiles.

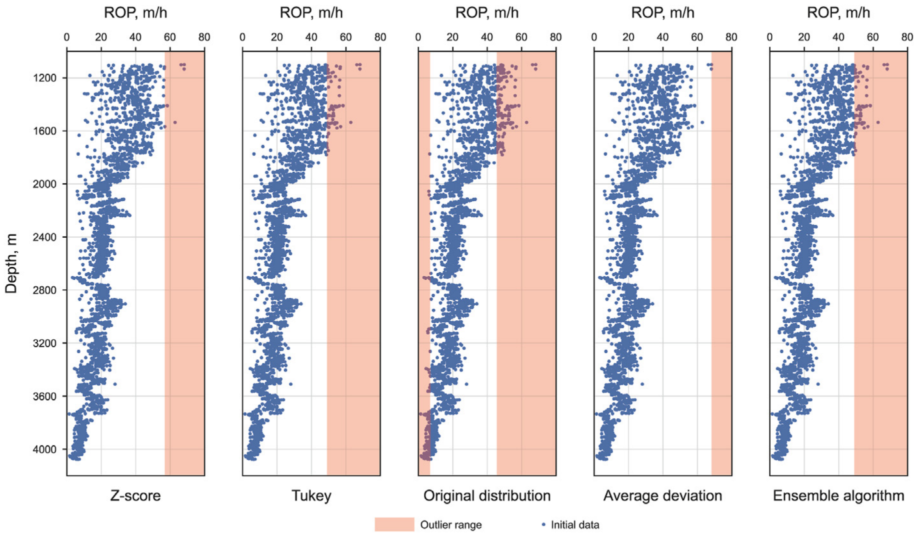

Fig. 9. Outlier detection in an actual ROP prediction task with different methods (data were collected from the South China Sea (Li et al., 2021a)).

Table 2, data from a gas well with 2978 actual ROP samples between 1100 and 4077 m that were also collected from the South China Sea (Li et al., 2021a) were introduced. As shown in Fig. 9, except for the average deviation-based algorithm that does not detect outliers (the initial judgment indicates that the requirement of 6 is too low), all other algorithms focus on outliers in the areas of high ROP values. Among them, only a few outliers are detected by the Z score method. However, the algorithm based on the original distribution directly determines that some of the data at the beginning and the end of the entire data interval are outliers, and the number of outliers is large. The number of outliers determined by the IQR-based Tukey method is moderate. The ensemble algorithm integrates the characteristics of different algorithms, and its determination range is close to that of the Tukey algorithm.

3.3. Data filtering

Deleting outliers does not eliminate the data noise caused by the instability of the monitoring system and other factors, and the research results indicate that the data fluctuation caused by noise could significantly reduce the ROP prediction accuracy (Sabah et al., 2019; Brenjkar and Biniaz Delijani, 2022). The current mainstream methods for denoising drilling data are divided into moving window and frequency domain transform methods. For the moving window method, as shown in Fig. 10(a), a window of specified length moves over the data, and the data within the window are calculated and filtered. According to different calculation methods, the main current moving window methods include the moving average filter, i.e., the average value of all data in the window is used as the filtered data result (Encinas et al., 2022), and the main parameter affecting the filtering effect is the selected window length; the envelope filter, i.e., the window includes the entire dataset, the upper and lower envelopes are drawn according to the distribution range of the ROP curve, and the filtering result is the average of the envelopes at the calculated depth (Diaz et al., 2019); and the Savitzky–Golay (SG) filter, i.e., after performing polynomial fitting on the data in the moving window, the fitted data are regarded as the filtered data (Savitzky and Golay, 1964; Ashrafi et al., 2019; Sabah et al., 2019; Brenjkar et al., 2021; Liu et al., 2021; Zhu, 2021; Brenjkar and Biniaz Delijani, 2022), and the main parameters that affect the SG filtering effect are the selected polynomial degree and window size (increasing the polynomial degree and reducing the window size can reduce the smoothness).

The frequency domain transformation method, as shown in Fig. 10(b), regards drilling data as a time domain signal with depth as a scale, converts it into a frequency domain signal, filters (cuts) the high-order spectrum of the original signal, and then converts it back to the time domain to achieve the filtering effect. The key to this method lies in the time-frequency transformation method and the filter cutoff frequency. Among them, the most common methods of time-frequency transformation in drilling data filtering are the Fourier transform (Diaz et al., 2018; Zielinski, 2021) and wavelet transform (Gan et al., 2019a; Gan et al., 2020; Li et al., 2021b, c). The difference between them is the basis function, i.e., a trigonometric function of infinite length is used in the Fourier transform and a wavelet basis of finite length that can decay is used

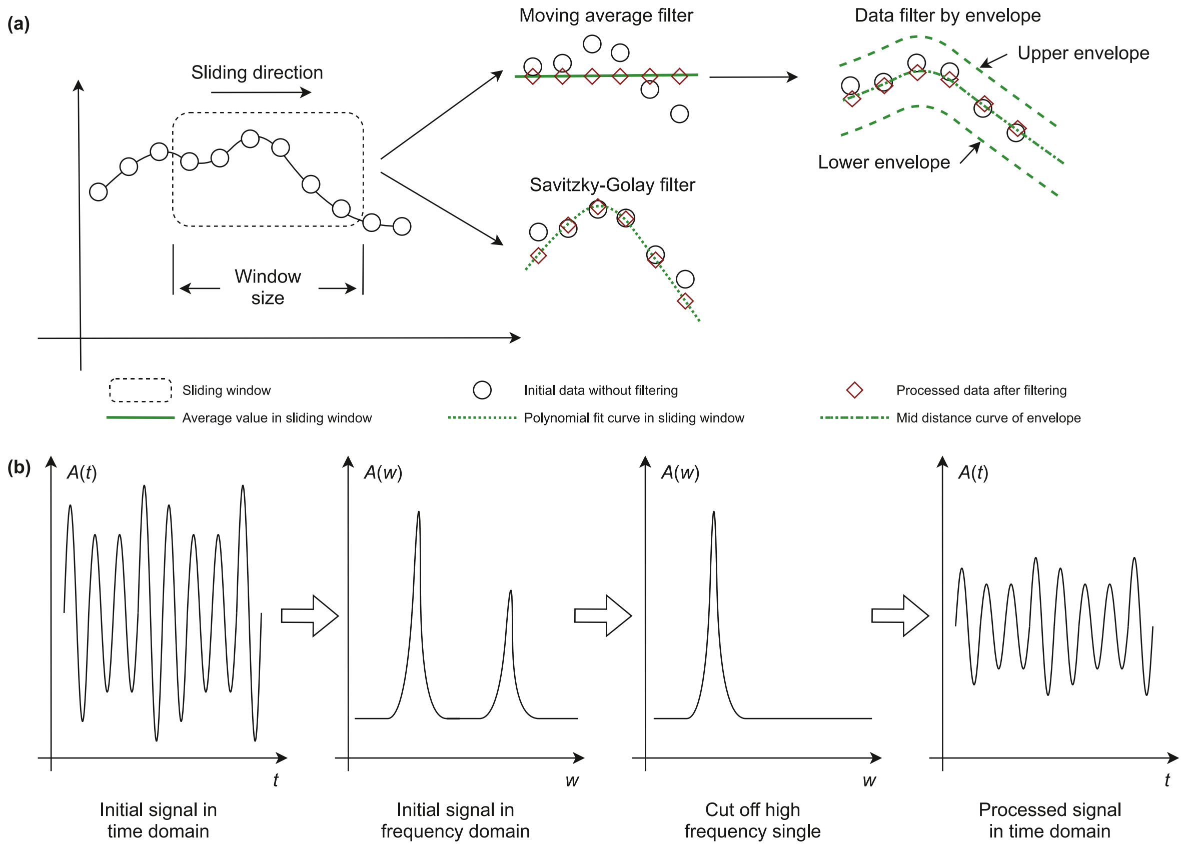
(a)
Sliding direction
Window size
Moving average filter
Savitzky-Golay filter
Data filter by envelope
Upper envelope
Lower envelope
Sliding window
Average value in sliding window
Initial data without filtering
Polynomial fit curve in sliding window
Processed data after filtering
Mid distance curve of envelope
(b)
A(t)
t
A(w)
w
A(w)
w
A(t)
t
Initial signal in time domain
Initial signal in frequency domain
Cut off high frequency single
Processed signal in time domain

Fig. 10. Data filter principle of moving sliding windows (**a**) and frequency transforms (**b**).

in the wavelet transform, as shown in Eq. (2).

$$\begin{cases} F(w) = \int_{-\infty}^{\infty} f(t)b(t)\mathrm{d}t \\ b(t) = \begin{cases} e^{-iwt}, \text{Fourier transfrom} \\ \frac{1}{\sqrt{a}}\psi\left(\frac{(t-\tau)}{a}\right), \text{Wavelet transform} \end{cases} \end{cases} \tag{2}$$

where w is the frequency, $b(t)$ is the basis function, $\Psi(t)$ is the basic wavelet function, α is the wavelet scaling factor, and τ is the wavelet translation factor.

After frequency domain transformation, the smoothness of the filter is completely determined by the cutoff frequency of the filter f_c. For drilling data, the cutoff frequency f_c is determined by the number of points (n) considered and the minimum depth of the sampling interval of the dataset ΔD, as shown in Eq. (3) (Diaz et al., 2018).

$$f_c = \frac{1}{n\Delta D} \tag{3}$$

The data filtering effects of four common filtering algorithms based on real data are shown in Fig. 11, which indicates that in the comparison of the two moving average window methods, the SG filter can better reflect the volatility of the data when the data changes are large. For the frequency domain transformation methods, the Fourier transform reduces the high-frequency fluctuation of the ROP. In the shallow well section (<1800 m) with a high ROP, the filtered result is lower than the real value, while in the deep well section (>3700 m) with a sudden drop in the ROP, an unrealistically high ROP is obtained after filtering. The wavelet transform filter is more sensitive to the fluctuation of the ROP. In sections with significant ROP fluctuations (<1800 m), the fluctuations are also pronounced in the filtered results, and in sections with less significant ROP fluctuations (>2400 m), smoother filtered results can be obtained.

3.4. Data normalization

Data normalization is a critical step before machine learning. In the modeling process, the immense differences in dimension and order of magnitude among the many monitoring parameters during drilling (up to four orders of magnitude, such as 10^{-1}–10^{3}) (Li et al., 2021a) could mislead the machine learning algorithm when setting the weight of each parameter, which could lead to enormous errors in the modeling efficiency and accuracy (Liu, 2021; Qu, 2021). However, it should be noted that for some specific machine learning algorithms (such as decision trees (DTs)), unnormalized data are better for the visualization of tree growth and the splitting criteria (Sabah et al., 2019).

The core of data normalization is to dedimensionalize the original data and scale them to a specified data interval. At present, in the modeling process of ROP prediction, there are five main normalization methods, as shown in Table 3. Among them, the min–max algorithm compresses the sample set to the [0, 1] interval, but recalculation is required when the sample set changes (especially when the maximum or minimum value changes) (Wang

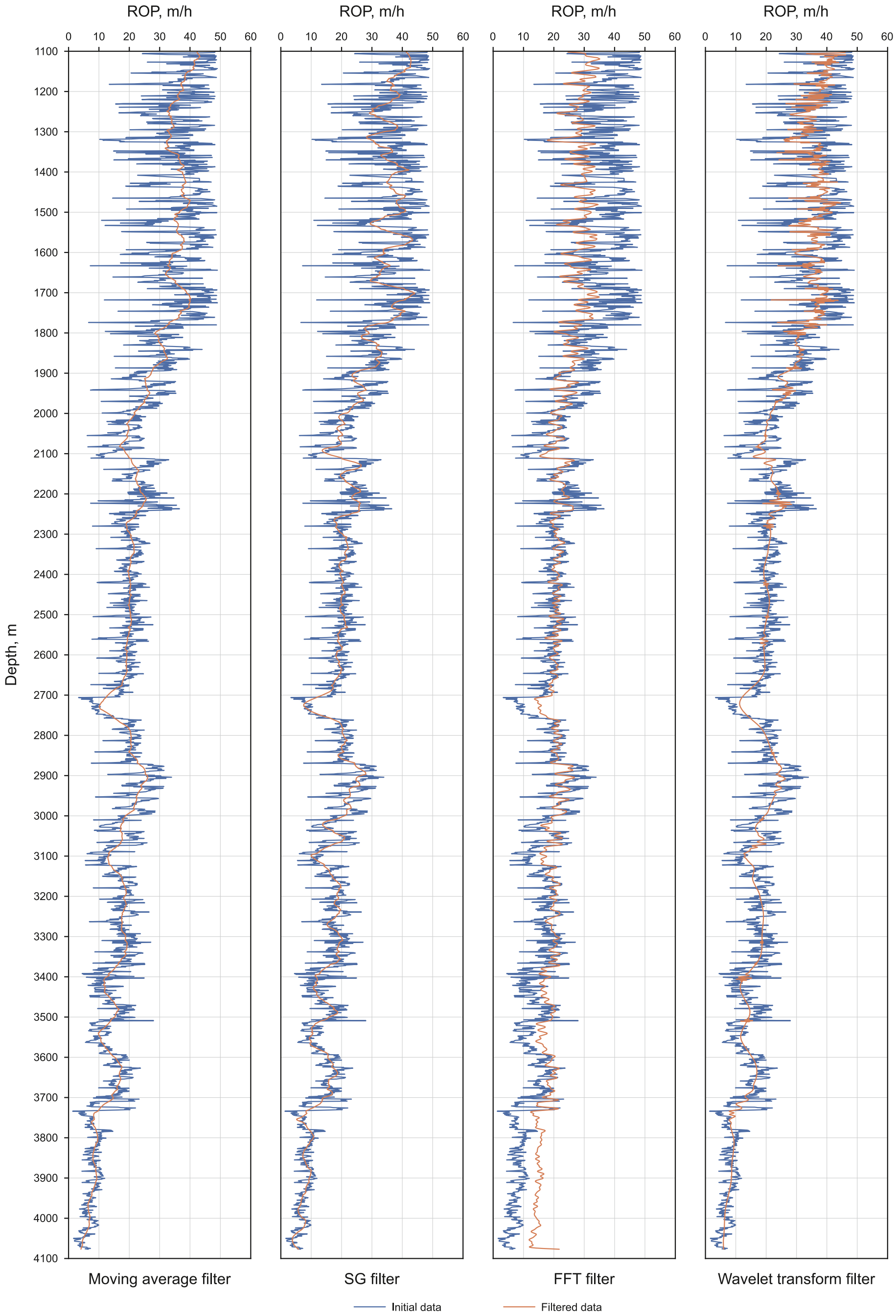

Fig. 11. Data filtering in an actual ROP prediction task with different methods.

Table 3
Common methods of data normalization in ROP prediction.

Method	Calculation formula	Scaled range
Min–Max	$x_n = (x_i - X_{\min})/(X_{\max} - X_{\min})$	$[0,1]$
Z score	$x_n = (x_i - \overline{x})/\sigma$	$[-3,3]$ (after outlier removal)
By a logarithmic function	$x_n = \log_{10}(x_i)/\log_{10}(X_{\max})$	$[0,1]$
By an arctangent function	$x_n = 2\arctan(x_i)/\pi$	$[-1,1]$
Decimal scaling	$x_n = x_i/10^m$	Depends on m

In the table, $X_{\min}/X_{\max}$ is the minimum/maximum value in the initial dataset, and m is the smallest integer that satisfies the needed scaling condition.

et al., 2018; Abbas et al., 2019; Liu et al., 2019; Youcefi et al., 2020; Delavar et al., 2021; Deng et al., 2021; Liu, 2021); the Z score method maps the data to a normal distribution with a mean of 0 and a variance of 1, is suitable for datasets with changes, and can

Table 4
The normalized range from different methods.

Parameters	Initial range	Min–Max	Z score	Logarithmic	Arctangent	Decimal scaling
MW	[1.05, 2.02]	[0, 1]	[−1.26, 2.79]	[0.06, 1]	[0.52, 0.73]	[1.05, 2.02]
ROP	[0, 45]	[0, 1]	[−1.84, 2.33]	[0.35, 1]	[0.83, 0.99]	[0, 4.5]
HL	[103.4, 167.2]	[0, 1]	[−2.09, 1.56]	[0.91, 1]	[0.994, 0.996]	[1.03, 1.67]
SPP	[681, 3230]	[0, 1]	[−3, 1.09]	[0.81, 1]	[0.999, 1]	[0.68, 3.23]

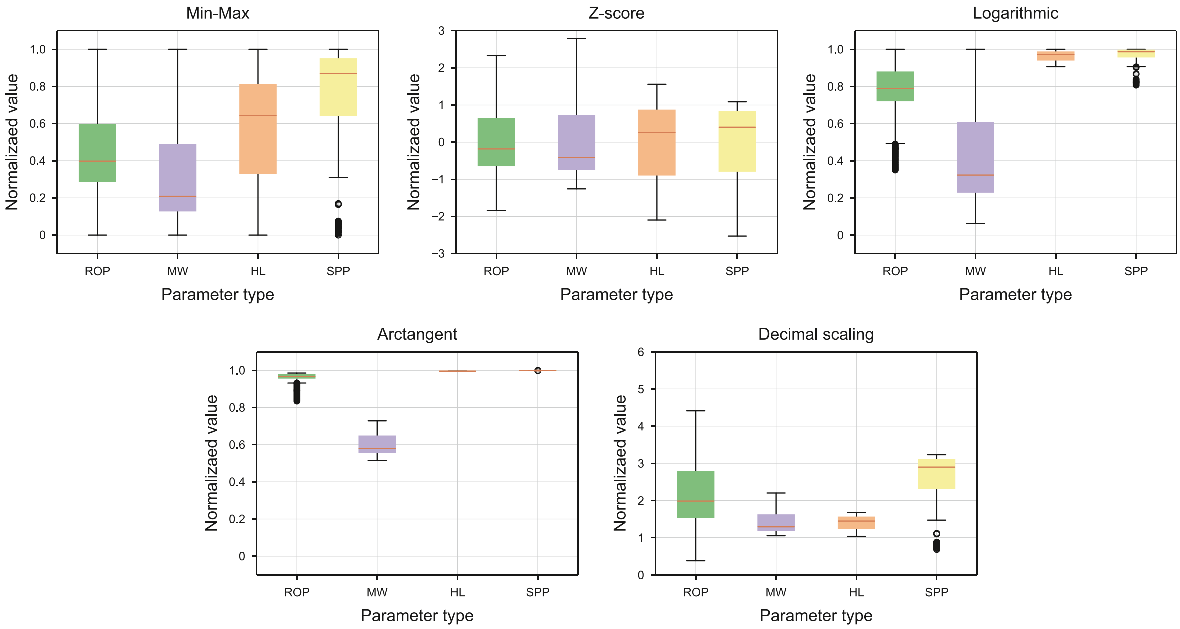

Fig. 12. Data normalization with actual drilling data using different methods.

compress the sample set to the [−3, 3] interval with the deletion of outlier points (Xiong and Li, 2018; Tan, 2019; Zhu, 2021); the logarithmic transformation algorithm limits the range of sample set compression to [0, 1] but is suitable only if the original data are all greater than 0 (Zhu, 2021); the arctangent algorithm also maps the raw data to [0, 1], but the actual interval after normalization could be smaller (Zhu, 2021); and the decimal scaling algorithm normalizes the data directly by moving the decimal point of the data, but it is not easy to control the mapping interval of the final result (Qu, 2021).

Four real parameters with different original ranges are selected for normalization, including drilling fluid density MW, drilling ROP, hook load (HL), and SPP. The data range before and after normalization are listed in Table 4, and the normalized curves are shown in Fig. 12. It can be indicated that the min−max and Z score algorithms can truly preserve the fluctuations in the original data; the algorithms based on logarithmic and arctangent functions are affected by the order of the original data, i.e., the larger the order of the original dataset is, the smaller the mapped normalized interval. Taking HL and SPP with a large range of raw data as an example, the range is reduced to [0.8, 1] and [0.999, 1] after normalization by the logarithmic and arctangent functions, respectively, and the latter can no longer meet the actual application requirements. The scaling interval of the decimal scaling algorithm is affected by the original dataset. For example, the scaling effect of the ROP is better than that of the other parameters. In summary, the min−max and Z score algorithms can scale the original data to the same interval while retaining the fluctuation characteristics of the original data, which is more appropriate for drilling data processing.

3.5. *Feature selection*

To better control the drilling process, many parameters are usually monitored at the drilling site. However, too many parameters could dilute the information density contained in the drilling data (Geng, 2021), which increases the probability of overfitting in machine learning modeling, making the model less easy to interpret and apply (Eskandarian et al., 2017; Sabah et al., 2019). This is the main reason why it is necessary to use feature selection to reduce the dimensionality of the modeling parameters before ROP modeling. Some studies have demonstrated that after feature selection, the accuracy of the ROP prediction model can be increased by 25.3%, while the training time can be reduced by 48.9% (Zhu, 2021).

As shown in Fig. 13, the current methods of feature selection can be divided into two main categories: feature filtering and feature merging. The core of the feature filter algorithm is to directly delete the features that are less correlated with the ROP. The conventional method is to calculate the correlation indexes between each feature and the ROP, sort the features involved in the selection, and delete the features with low rankings as irrelevant features. The main commonly used ranking indicator is currently the correlation coefficient (Bezminabadi et al., 2017; Xiong and Li, 2018; Zuo, 2018; Darbor et al., 2019; Zhao et al., 2019; Kor and Altun, 2020; Elkatatny, 2021; Kor et al., 2021). Generally, there are three types of correlation coefficients, Pearson, Spearman, and Kendall methods. Among the three methods, the Kendall method is more suitable for parameters with finite data values (which can hardly be found in drilling data), and the Spearman method is more suitable to

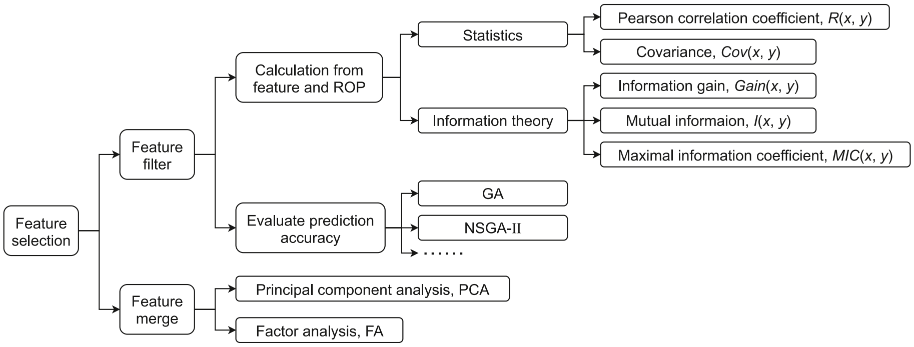
Feature selection
Feature filter
Feature merge
Calculation from feature and ROP
Evaluate prediction accuracy
Statistics
Information theory
Pearson correlation coefficient, $R(x, y)$
Covariance, $Cov(x, y)$
Information gain, $Gain(x, y)$
Mutual informaion, $I(x, y)$
Maximal information coefficient, $MIC(x, y)$
GA
NSGA-II
……
Principal component analysis, PCA
Factor analysis, FA

Fig. 13. Classification of methods used in feature selection for ROP prediction.

represent the data whether the correlation relationship exists, instead of indicating correlation degree, which can be calculated from Pearson method. Hence the Pearson method was widely used in drilling feature selection. To reflect the nonlinear correlation between parameters, some researchers have used three information theory-based indicators, i.e., information gain (Conradie et al., 2019; Li et al., 2013), mutual information (Gan et al., 2019a, b; Gan, 2019; Leng et al., 2020; Zhang, 2020; Zhou et al., 2021a, b; Li et al., 2021b, c), and the maximal information coefficient (Zhang et al., 2021; Reshef et al., 2011), for filtering.

Another type of feature filter algorithm directly evaluates the impact of different feature combinations on the ROP prediction accuracy. Through some replaceable optimization algorithms (including but not limited to the Fscaret package (Eskandarian et al., 2017; Abbas et al., 2019), genetic algorithm (GA) (Ashrafi et al., 2019; Sabah et al., 2019) and GA for multi-objective optimization (such as the nondominated sorting genetic algorithm II (NSGA-II) (Mehrad et al., 2020)), different quantities and types of features are randomly extracted from the original dataset to establish an ROP prediction model. After modeling, the original features are sorted based on the prediction accuracy, and the top-ranked features are selected according to the needs.

The feature merging algorithm does not remove features but converts all features into a relatively small number of composite variables (or factors) by calculation, where the calculation process of each composite variable involves all the original features. The main types of algorithms currently include principal component analysis (PCA) (Samaei et al., 2020; Deng et al., 2021; Alsaihati et al., 2022), and factor analysis (FA) (Xiong and Li, 2018). Although the number of new variables or new factors is less than that of the original features, at least 85% of the information of the original features is still included (Mariani et al., 2021), and the new variables that are uncorrelated with each other eliminate multicollinearity and the prediction error caused by the mutual coupling between the original features, which is especially effective in eliminating the correlation between multiple parameters in the ROP prediction process.

Comparing with different methods for drilling feature selection was listed in Table 5. For the feature filter, the core is to evaluate whether the parameters are important or useful (by different calculation methods), keep the suitable parameters, and remove those inappropriate. Then introducing the selected parameters directly into the model without dealing with their internal correlation relationship, will lead to potential prediction error. As for feature merge, both PCA and FA can eliminate the correlation between parameters, but the information loss (maybe less than 15%) is the main reason for reducing prediction accuracy. Therefore, the recommendation for feature selection depends on the number of features. When there are plenty of features, the correlation between features should pay more attention and feature merge is the better choice. On the contrary, the information loss cannot be ignored when there are fewer features, and feature filter methods are preferred.

3.6. Data splitting

Through data splitting, the ROP modeling dataset is divided into a training set and a test set. The training set allows the machine learning algorithm to grasp the changing patterns of the existing dataset, and the test set reduces the generalization error when the model encounters new samples (Zhou, 2016). To ensure the prediction accuracy of the algorithm when faced with new data, the

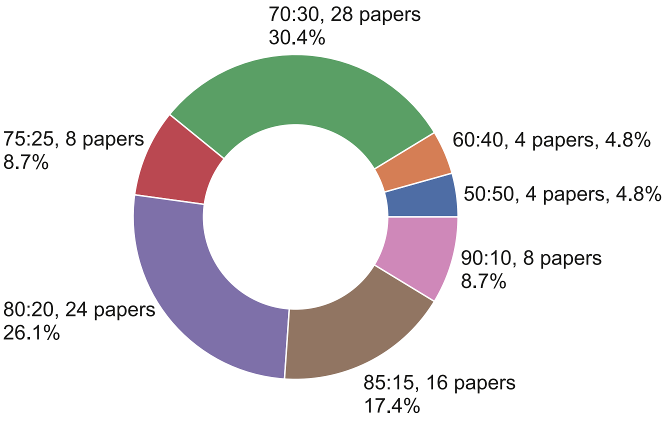

Fig. 14. Static data splitting distribution for ROP prediction.

Table 5
Limitations and recommendations for the feature selection method.

Selection method		Limitation	Recommendation
Calculation from feature and ROP	Feature filter	Cannot deal with correlation between parameters	Lack of features
Evaluate prediction accuracy	Feature filter	(1) Cannot deal with correlation between parameters; (2) Time consumption	Lack of features
PCA	Feature merge	(1) Information loss; (2) Physical meaning is not clearly	Plenty of features
FA	Feature merge	Information loss	Plenty of features

test set is strictly required to be mutually exclusive with the training set. Therefore, the ratio of the training set and test set is a control indicator affecting the accuracy of ROP prediction by machine learning models (Hegde et al., 2017; Li, 2020).

Based on whether the dataset used by the ROP prediction model is changed, current data splitting methods are divided into two modes: static and dynamic. When the dataset remains unchanged, more researchers tend to use static splitting, according to the statistics in 92 papers, and more than 73% of the researchers have used three ratios of 70:30 (30.4%), 80:20 (26.1%) and 85:15 (17.4%) for data splitting, as shown in Fig. 14. When the original dataset is large enough, some researchers have even used the extreme ratio of 96:4 (the number of samples in the original dataset of that study is 1,964,436, and the number of samples in the test set is still as high as 78,577) (Zhu, 2021). However, when the dataset is small, the *k*-fold cross validation method is the most popular (Bodaghi et al., 2015; Zhou, 2016; Ansari et al., 2017; Eskandarian et al., 2017; Gan et al., 2019a; Liu et al., 2019; Fan et al., 2021; Yu et al., 2021; Alsaihati et al., 2022), and the value of *k* is usually between 4 and 10. Considering that the ROP is related to the formation, some researchers have classified the original dataset according to the formation characteristics (formation type, rock strength, etc.), and high prediction accuracy has been achieved (Al-AbdulJabbar et al., 2019; Liao et al., 2020; Najjarpour et al., 2020; Oyedere and Gray, 2020; Soares et al., 2020).

The dynamic mode is more suitable for building accurate models when real-time data streams during drilling cause changes to the modeling dataset. In this mode, the size of the test set is always fixed, the real-time data are introduced directly into the test set, and the same amount of old data from the original test set is moved to the training set. As shown in Fig. 15, the training set is adjusted in two ways:

1) the size of the training set is maintained (excluding the same amount of old data after the entry of new data) (Zhou et al., 2021a, b; Encinas et al., 2022; Zhang et al., 2022, 2023), and the size of the maintained training set generally can be calculated from Figs. 7 and 14, where the recommended size was 315 (the median value of dataset size below 1000) × 0.7 (the most popular distribution for data splitting)≈220. Instead of selecting a dataset size below 10000, using the dataset below 1000 here is mainly considering the flexibility for model updating.
2) the training set is expanded (retaining the old data). According to the expansion method, the training set can be proportionally expanded; that is, the training set is expanded as a whole after the size of the new data reaches that of the original test set (Soares and Gray, 2019; Brenjkar and Biniaz Delijani, 2022). The training set can be expanded in real-time; that is, the training set is expanded once with each new piece of data added (Zhou et al., 2021a, b).

4. Model establishment

4.1. Modeling algorithm selection

The main intelligent algorithms used in the establishment of ROP prediction models include modeling algorithms and optimization algorithms. The main task of the modeling algorithm is to achieve accurate ROP prediction. According to the number of algorithms used, current ROP prediction methods can be divided into two modes: single algorithms and ensemble algorithms. As shown in Fig. 16, the current mainstream single algorithm methods are dominated by k-nearest neighbor (KNN) algorithms, artificial neural networks (ANNs), support vector regression (SVR) algorithms, and DT algorithms, and the total usage frequency reaches 80%. The remaining 20% of researchers have started to explore the application of ensemble algorithms based on bagging, boosting, and stacking in ROP prediction. Notably, the proportion of ensemble algorithms has begun to rise in the past two years (2021 and 2022), and its application proportion has reached 44% (11 of the 25 papers have used ensemble algorithms).

Among the single algorithm methods, ANNs are the most popular, and they currently occupy the largest share. More than half of the researchers (51.4%) in all ROP prediction studies have used this method. The statistics on ANN algorithms used in all ROP prediction studies (Fig. 17(a)) indicate that the multilayer perceptron (MLP) is the most frequently used, accounting for 61.1% of all ANN algorithms. In addition, radial basis function neural networks (RBFNNs), extreme learning machines (ELMs), and adaptive network-based fuzzy inference systems (ANFISs) are also popular ANN algorithms in ROP prediction, which their usage proportions can reach 15.8%, 8.4%, and 8.4%, respectively.

The second-ranked single algorithm is SVR, which is chosen by nearly a quarter of researchers (24.3%). SVR for predicting the ROP is based on a kernel function, as shown in Fig. 17(b), and nearly half of SVR users (42.1%) believe that the radial basis function (RBF) can achieve better ROP prediction. For the remaining single algorithms, the KNN algorithm is used for classification before building a prediction model (Zhou et al., 2021a, b) and is rarely used in ROP prediction, while the DT algorithm is more often used as the basis function of ensemble algorithms, with an extremely low usage frequency as a single algorithm.

According to the different ensemble methods, ensemble algorithms are divided into three types: bagging, boosting, and

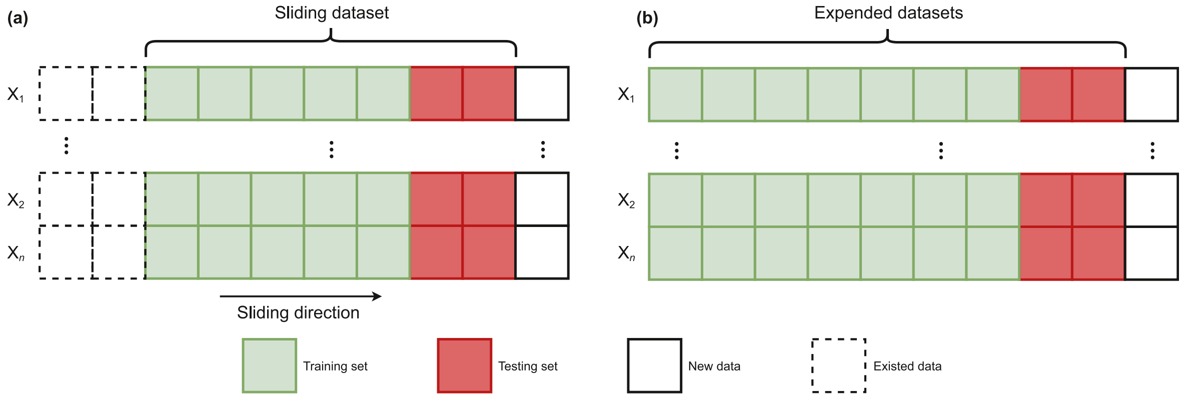

Fig. 15. Two types of dynamic data splitting ((**a**): fixed size for both the training set and testing set; (**b**): expanding the training set and fixing the testing set).

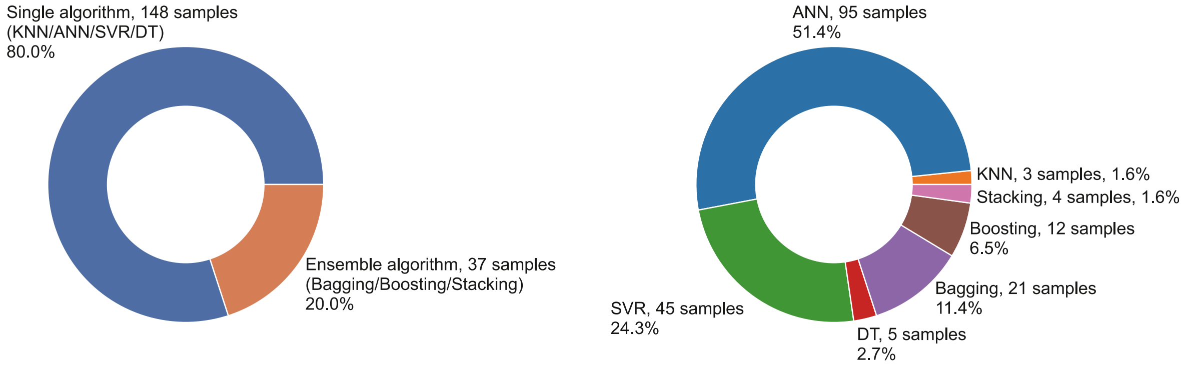

Fig. 16. Modeling algorithm selection distributions in ROP prediction.

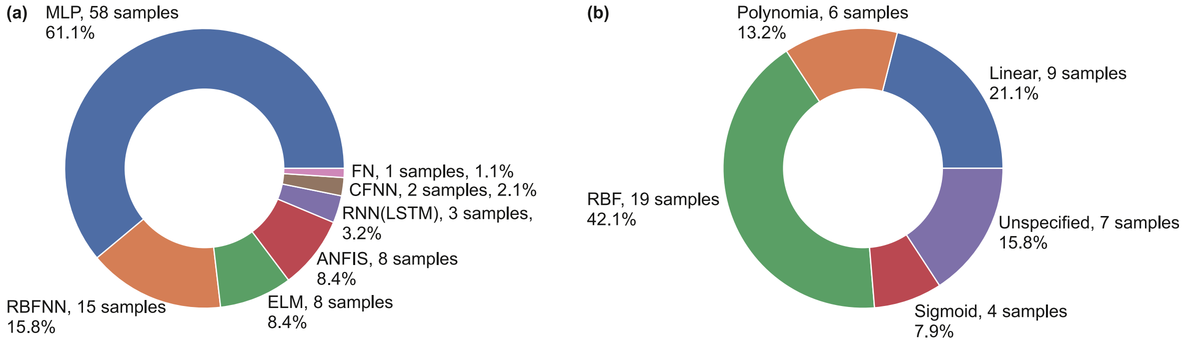

Fig. 17. Modeling algorithm selection distributions of ANNs (**a**) and SVM algorithms (**b**) in ROP prediction.

stacking. More than half of the ensemble algorithms use bagging (11.4%), and the most representative algorithm is the random forest (RF) algorithm obtained by ensembles of multiple DTs. The boosting algorithm and stacking algorithm have become more popular in other fields of machine learning in recent years, and the computational complexity is slightly higher than that of the bagging algorithm. Although most researchers believe that the accuracy of a single algorithm and the bagging algorithm can meet the current requirements, the usage frequency of boosting and stacking algorithms is not high, only 6.5% and 1.6%, respectively.

4.2. Optimization algorithm selection

In contrast to modeling algorithms, the goal of optimization algorithms is to better establish or apply prediction algorithms rather than building predictive models directly. As the complexity of current machine learning algorithms increases, optimization algorithms are increasingly used. According to statistics in 79 papers, as shown in Fig. 18 left, more than half (43 papers, 54.4%) of researchers have used optimization algorithms when building machine learning ROP prediction models. The current optimization algorithms are based on three main aspects.

- First, the optimization of the internal structure of the specified machine learning model can allow fast determination of the hyperparameter values within the algorithm;
- Second, the optimization of the application effect of the established machine learning model can allow a fast search for the model input parameter combination that can achieve the maximum ROP;
- Third, the optimization of the traditional statistical regression model can allow an accurate search for the regression coefficient.

As shown in Fig. 18 right, among the above three aspects, the optimization of the machine learning structure was widely utilized

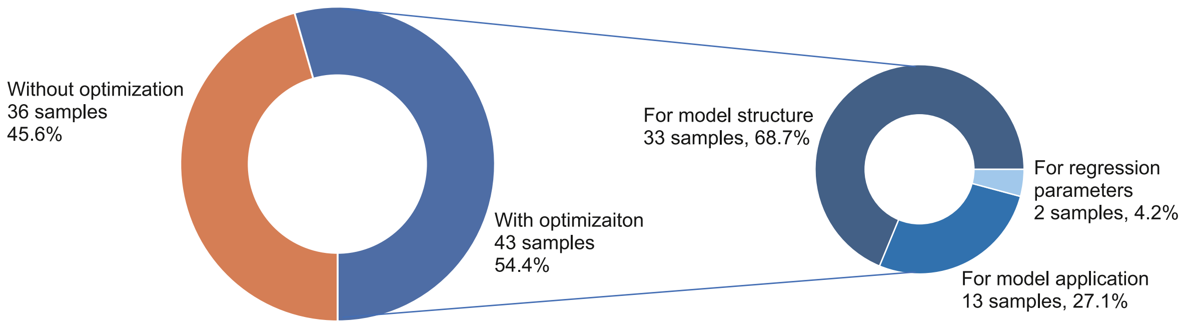

Fig. 18. Optimization algorithm selection distributions in ROP prediction.

Table 6
Differences between two major optimization schemes in ROP prediction.

Difference	ROP prediction model establishment	ROP prediction model application
Optimization target	To reduce the training consumption and increase the prediction accuracy	To increase the ROP
Objective function	Loss function (defined to evaluate the difference between predicted and actual ROP)	ROP prediction model
Decision variable	Hyperparameters (such as the number of hidden layers, the weights, and the threshold for each neuron in ANN)	Input parameters of the ROP prediction model (such as WOB, RPM, T, and UCS)

in 33 papers, which occupied the largest proportion, reaching 68.7%, while the remaining two directions only account for 27.1% and 4.2%, respectively. The main difference between the first and second optimization schemes is the optimization target and objective function, as listed in Table 6. The distribution of the usage of optimization methods indicates that the current machine learning methods for ROP prediction modeling still focus on the efficient construction of high-precision models, not the practical application of the models.

The classification of optimization algorithms can be divided into three categories based on different search methods: exact algorithm, heuristic algorithm, and meta-heuristic algorithm as follows.

- Exact algorithm converts the problem to be solved into a mathematical planning problem and solves it accurately. It can obtain the optimal solution in the entire domain. However, when the variable domain is large, the solution time increases exponentially. For ROP prediction, an overly complex variable domain is not suitable;
- Heuristic algorithms are problem-specific, and aim to find approximate solutions within an acceptable time and space range (not necessarily the global optimal solution). However, the heuristic algorithm relies too much on the optimization problem itself, making it less versatile.
- Metaheuristic algorithms focus on combining random algorithms and local search to create independent search strategies derived from related phenomena in nature or social production. Because the search strategy is problem-independent, current optimization algorithms focus more on this direction.

Metaheuristic algorithms have developed rapidly in recent years, and almost all ROP prediction-related optimization algorithms reviewed in this review belong to this category, such as genetic algorithm (Brenjkar and Biniaz Delijani, 2022), differential evolution (Brenjkar and Biniaz Delijani, 2022), particle swarm optimization (Ashrafi et al., 2019), biogeography-based optimization (Brenjkar et al., 2021), imperialist competitive algorithm (Samaei et al., 2020), cuckoo optimization algorithm (Bodaghi et al., 2015), artificial bee colony (Zhao et al., 2020), whale optimization algorithm (Youcefi et al., 2020), and beetle antennae search (Li et al., 2021b), etc.

4.3. Accuracy evaluation indicators

The accuracy evaluation indicator is important for selecting the most suitable algorithm to predict ROPs among various algorithms. As listed in Table 7, the current accuracy evaluation indicators for ROP prediction modeling algorithms can be divided into four categories.

- First, a certain measure, such as the mean absolute percentage error (*MAPE*) and root mean square error (*RMSE*), is used to represent the error between the predicted ROP and the real ROP, and the smaller the measure is, the better the accuracy. These indicators can help researchers know the deviation between predicted ROP and true ROP, which can be used to fix the predicted value.
- Second, when calculating the degree of fit, such as R and R^2, between the prediction results and the real dataset, the evaluation results are usually in [0, 1], and the closer the evaluation result is to 1, the closer the prediction result is to the true result. These indicators allow researchers to estimate the accuracy of the model and provide convenience for comparing prediction accuracy between different models.
- Third, the error distribution of the entire prediction result set, such as the normalized error rate (*NER*) and error distribution (*ED*), can be calculated. These indicators show the error across the entire dataset, which can help researchers observe the prediction error in different depths for ROP, providing direction for subsequent revisions or improvements of the model.
- Fourth, secondary judgment can be used; such as the performance index (*PI*) and the pseudo coefficient of determination (P_R^2), which uses multiple deterministic indicators to achieve accuracy comparison and judgment between different algorithms.

According to statistics (Fig. 19), the most commonly used indicators in all current ROP prediction studies are *RMSE*, R^2, *MAPE*, R, mean absolute error (*MAE*), and mean square error (*MSE*). Among them, *RMSE*, *MAPE*, *MAE*, and *MSE* belong to the first category of evaluation indicators that give the magnitude of the error between the predicted and actual ROPs, while R^2 and R belong to the second category of indicators that give the degree of fit between the predicted and actual ROP curves. To better demonstrate the prediction effect, most researchers often choose two or three evaluation indicators and combine the error value and the degree of fit to evaluate the established model in multiple aspects.

4.4. Comparison among various ROP algorithms

Since there are many types of algorithms, the comparison in this paper focuses on the comparison of algorithm types. Therefore, in the following comparison, the single algorithm machine learning model (MLs), ensemble algorithm machine learning model (MLe), optimized machine learning algorithm (MLO), traditional statistical regression algorithm (C), and optimized statistical regression algorithm (CO) are considered.

The key point here is to compare the prediction accuracy, but various indicators have been used in different papers, as shown in Fig. 19. The indicators used to evaluate error, such as *RMSE* or *MAPE,* may cause misleading comparisons among papers when there is a large difference in the actual ROP. Instead, the indicators used to evaluate the degree of fit, such as R^2 or R, can prevent this problem. For instance, Al-AbdulJabbar et al. (2020) proposed a novel application of ANNs, in which the performance in the test datasets achieved an R^2 of 0.93 and an *RMSE* of 0.062. In another ANN prediction study, there was a lower *RMSE* (0.0104) but also a lower R^2 (0.86) (Bezminabadi et al., 2017). The indicator of R^2 is more

Table 7
Common algorithm evaluation indicators.

Classification	Indicators	Formula
By evaluating error	Mean absolute percentage relative error	$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\lvert Y_i - Y_\mathrm{p}\rvert}{Y_i}\right) \times 100\%$
	Average percentage relative error	$APRE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(Y_i - Y_\mathrm{p})}{Y_i}\right) \times 100\%$
	Symmetric mean absolute percentage error	$SMAPE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\lvert Y_i - Y_\mathrm{p}\rvert}{(Y_i + Y_\mathrm{p})}\right) \times 100\%$
	Mean absolute error	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert Y_i - Y_\mathrm{p}\rvert$
	Maximum absolute error	$MAE_\mathrm{max} = \max(\lvert Y_i - Y_\mathrm{p}\rvert)$
	Root mean square error	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - Y_\mathrm{p})^2}$
	Mean square error	$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - Y_\mathrm{p})^2$
	Normalized root mean square error	$N\_RMSE = RMSE/\frac{1}{n}\sum_{i=1}^{n}Y_i$
	The sum of square error	$SSE = \sum_{i=1}^{n}(Y_i - Y_\mathrm{p})^2$
By evaluating the degree of fit	Correlation coefficient	$R = \mathrm{cov}(Y_i, Y_p)/(\sigma_{Y_i}\sigma_{Y_\mathrm{p}})$
	Coefficient of determination	$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - Y_\mathrm{p})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$
	Variance account	$VAF = \left(1 - \frac{\mathrm{var}(Y_i - Y_\mathrm{p})}{\mathrm{var}(Y_i)}\right) \times 100\%$
By evaluating the error distribution	Normalized error rate	$NER(i) = \frac{\lvert Y_\mathrm{p} - Y_i\rvert}{Y_i}$
	Error distribution	$ED(i) = Y_\mathrm{p} - Y_i$
Comprehensive evaluation	Performance index	$PI = (R + VAF/100) - RMSE$
	Pseudo coefficient of determination*	$P\_R^2 = 1 - \frac{SSE_\mathrm{current}}{SSE_\mathrm{compare}}$

In the table, Y_i is the true value, Y_p is the predicted value, $\overline{Y}$ is the average value of the true values, n is the number of samples, $\mathrm{cov}(x, y)$ is the covariance between x and y, and σ_x is the standard deviation of x.

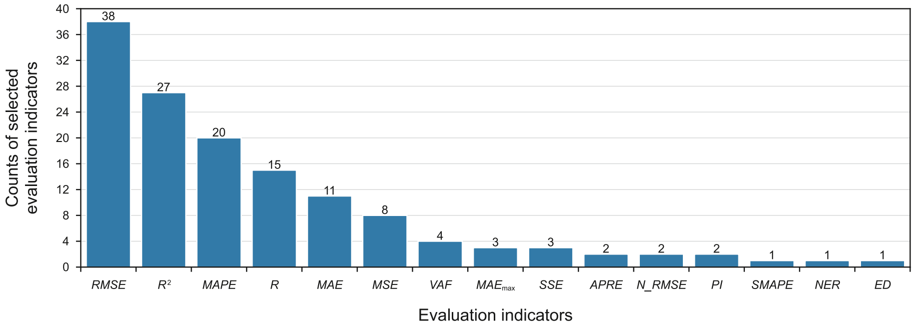

Fig. 19. Usage frequency of different evaluation indicators in ROP prediction.

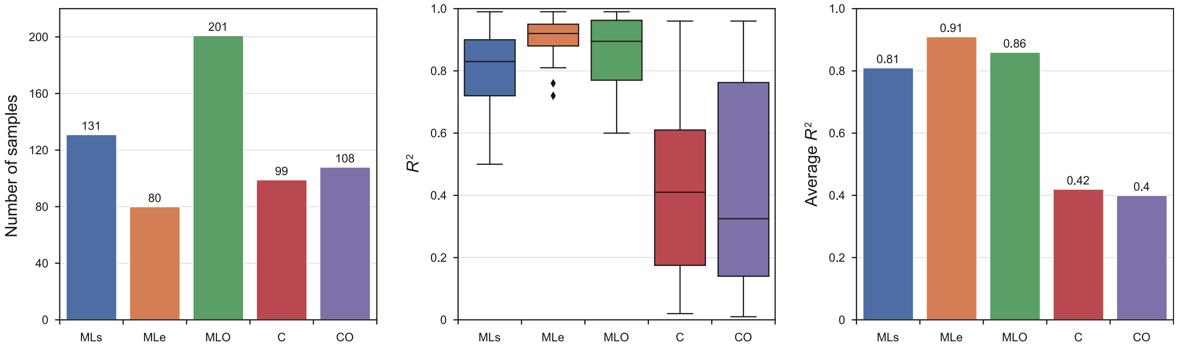

Fig. 20. Accuracy statistics for different kinds of ROP prediction models.

reliable, and it is selected here to evaluate the accuracy of the algorithms.

The collected number of each algorithm is shown in Fig. 20 left, and the comparison results are shown in Fig. 20 middle and right, which indicates that the prediction accuracy of the machine learning model is significantly higher than that of the traditional

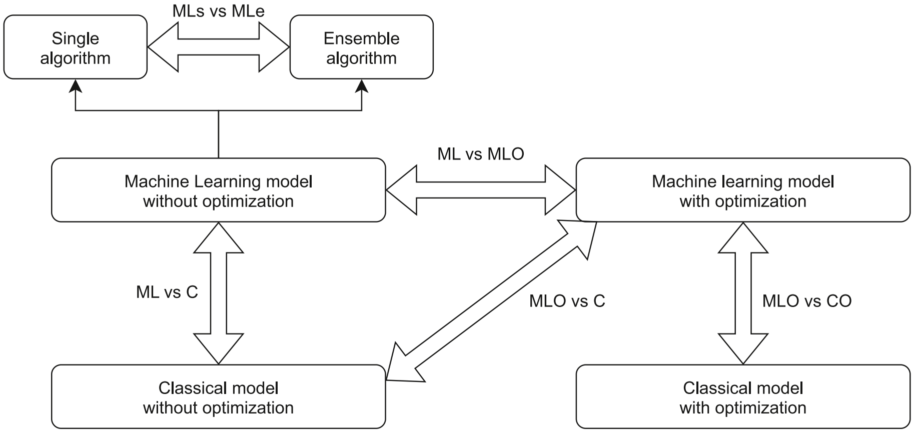
MLs vs MLe
Single algorithm
Ensemble algorithm
ML vs MLO
Machine Learning model without optimization
Machine learning model with optimization
ML vs C
MLO vs C
MLO vs CO
Classical model without optimization
Classical model with optimization

Fig. 21. Five comparisons to evaluate the accuracy of different models.

model. The average degree of fit of the prediction accuracy of the three machine learning algorithms is greater than 0.8, while that of the two traditional models is only approximately 0.4. For the machine learning algorithms, the average prediction accuracy of the ensemble algorithm (0.91) is higher than that of a single algorithm (0.81). Notably, although the addition of an optimization algorithm can improve the training efficiency of machine learning, it does not necessarily bring about a significant improvement in the prediction accuracy. This conclusion applies to both machine learning models and statistical models.

In addition to the prediction accuracy, this paper conducts a comparison between different types of algorithms. As shown in Fig. 21, five types of comparisons are carried out as follows:

- ML vs C: The comparison between the machine learning model (ML) and traditional statistical model (C) reflects the prediction effect of the machine learning model.
- MLs vs MLe: The comparison between single algorithm machine learning model (MLs) and ensemble algorithm machine learning model (MLe) reflects whether an ensemble algorithm is more advanced than a single algorithm.
- ML vs MLO: The comparison between the machine learning model without optimization (ML) and the optimized machine learning algorithm (MLO) reflects whether the introduction of the optimization algorithm will increase the prediction accuracy.
- MLO vs C: The comparison between the optimized machine learning algorithm (MLO) and traditional statistical model (C) reflects the improvement effect of the high-precision model on the traditional model.
- MLO vs CO: The comparison between the optimized machine learning algorithm (MLO) and optimized statistical regression algorithm (CO) reflects whether the optimization of the traditional model can achieve the same accuracy as the machine learning model.

To eliminate the difference between different indicators used in different studies, two unified dimensionless percentage, including the reduced error rate (*RER*) and the increased accuracy rate (*IAR*), was calculated to characterize the modeling effect for different algorithms, as shown in Eq. (4).

$$\begin{cases} RER = \dfrac{I_{\text{E}-\text{base}} - I_{\text{E}-\text{compare}}}{I_{\text{E}-\text{base}}} \times 100\% \\ IAR = \dfrac{I_{\text{A}-\text{compare}} - I_{\text{A}-\text{base}}}{I_{\text{A}-\text{base}}} \times 100\% \end{cases} \tag{4}$$

where, $I_{\text{E-base}}/I_{\text{E-compare}}$ is the base/comparison indicator to indicate the magnitude of error, such as the *RMSE* and *MAPE*, and $I_{\text{A-base}}/I_{\text{A-compare}}$ is the base/comparison indicator to indicate the degree of fit, such as R^2 and *R*.

As none of the reviewed papers (over 110 papers) provided the original data, therefore the accuracy comparison results in Fig. 22 were calculated from the reporting value in the literature. Without considering the introduction of the optimization algorithm, compared with the traditional model, the machine learning model can significantly improve the prediction accuracy, i.e., the prediction error rate decreases by 65.7% on average compared to that of the traditional model, and the prediction accuracy is improved by more than 40% on average (the machine learning model even improves the prediction accuracy by nearly 1.5-fold, while the traditional model is less accurate). In terms of the comparison between machine learning models, the ensemble algorithm is more accurate than a single algorithm in most cases (the error rate decreases by an average of 35%, and the degree of fit increases by an average of 21%), but not absolutely. Some studies have demonstrated that a single algorithm with an improved internal structure can achieve higher accuracy than conventional ensemble algorithms (such as RF).

After considering the introduction of the optimization algorithm, the optimization algorithm does not significantly improve the prediction accuracy of the machine learning model, and the improvement is only 6.6%. This small improvement is due to the already high prediction accuracy of the unoptimized machine learning models. Comparing the optimized machine learning algorithm with the traditional model, regardless of whether the traditional model is optimized, the average prediction efficiency of the machine learning algorithm is significantly improved, and the average accuracy improvement is as high as 63%, 76%, and even 200%.

Notably, the data in Fig. 22 are not very consistent. For example, the decrease in the error rate does not match the increase in accuracy because the data sources for comparison come from

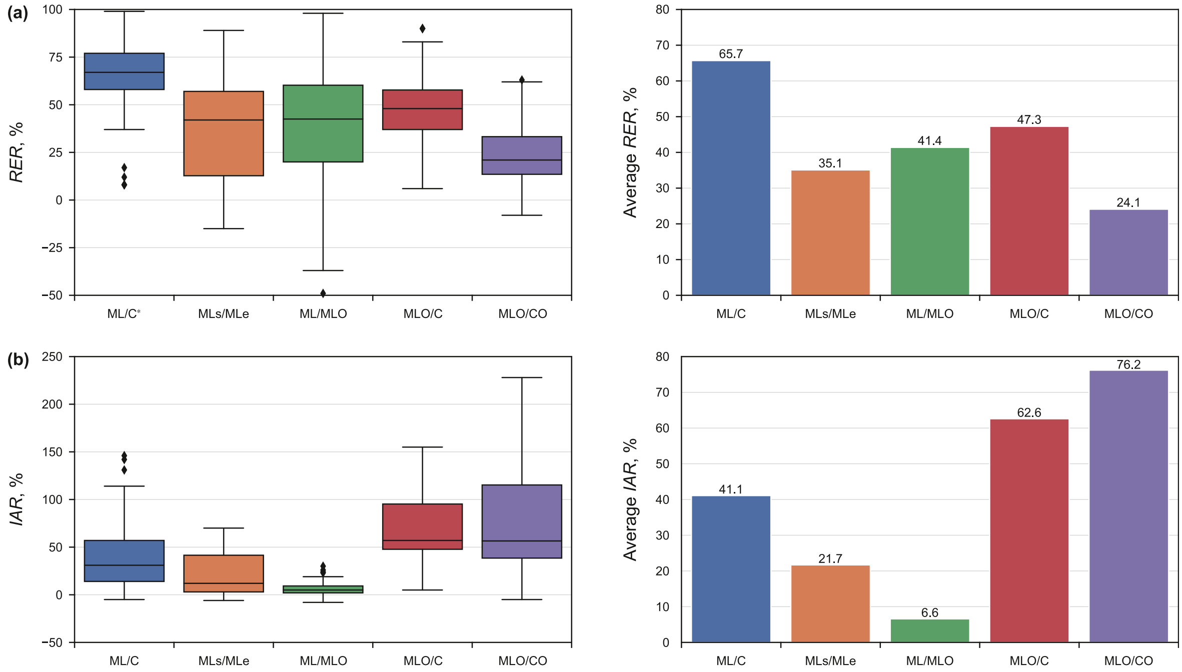

Fig. 22. Accuracy statistics for different kinds of ROP prediction models ((**a**): the distribution and average value of *RER*; (**b**): the distribution and average value of *IAR*) (*in the horizontal variable of *X*/*Y*, where *X* is the comparison indicator and *Y* is the base indicator.).

different papers, and different papers use different parameters, but the obtained patterns are reliable.

5. Discussion about ML modeling in ROP prediction

5.1. Challenge

(1) Data isolation

Data is the most important core basis for machine learning applications. With years of technological development, massive amounts of data have been accumulated in all aspects of the drilling industry. Typical characteristics of these data include:

- Complex data sources, including pre-drilling data, design data, construction records, well logging and project management reports, etc.;
- Timely changing data, and the types and contents of various data change as the operation progresses. The volume will expand in real time;
- Low data value density, most of the core information required for operations needs to be integrated and analyzed to obtain multiple types of data;
- Diverse data forms, including various static structured tables, unstructured videos, pictures, reports, and various industrial data format standards;
- Poor data manageability, various data generation cycles, and collection methods are different, and there is a lack of a unified data model for summary and organization.

The above characteristics create obvious data isolation, which makes researchers difficult to collect similar data in different regions, and the data cannot be shared and utilized efficiently. Without solid data support, the ML model not only will reduce prediction accuracy, but also limit the adaptability in actual application. It is also the reason that most of the ML models at present can only reported useful within a certain region, and few papers have reported examples of successful prediction across regions. How to break data isolation under the conditions of data confidentiality requirements is the primary challenge in the current application of machine learning models.

(2) Model generalization

The essence of machine learning algorithms is data-driven, and the model accuracy is greatly affected by modeling data. Most machine learning models perform poorly when faced with untrained data sets. However, in a drilling area, facing unknown stratigraphic structures (especially the complexity of deep stratigraphic structures combined with abnormal ground temperature, complex geo-stress, and geological structures), the probability of encountering data that have never appeared in modeling training sets is extremely high. Therefore, the requirements for model generalization and updating are extremely high.

As mentioned in Section 3.6, the generalization performance for the ROP prediction model at present depends on the test sets from data splitting, and the evaluation of model generalization performance should be divided into four levels based on how to select the test set. For instance, suppose there is a dataset containing several wells of actual in the same area, and the four levels can be divided as follows.

- Level one, mixing all the data from all wells together, then randomly selecting 30% (or other proportions as shown in Fig. 14) of them to put into a test set. Considering the density and continuity of drilling data collection, it is likely that there will be very little difference between a set of consecutive data. If some of the test data are selected from this consecutive data, the rest is put into a training set. In this situation, for the model, part of the test data is already equivalent to the training data, which will reduce the credibility of the test result and generalization performance;

- Level two, based on improved k-fold cross validation, divides the data of each well in the area into uniform k groups along the depth, and then the test set consists of a group of data from different depth intervals of each well. This method can avoid the problem of extracting test sets from continuous data, and the resulting generalization performance is more accurate than the first level. If some of the wells are very close together, the data from different well sections may be similar, and the problem of data continuity cannot be avoided;
- Level three, selecting one of 10 wells as the test set, and putting all the remaining 9 wells into the training set. It is the true method that can represent the generalization performance of the prediction model in the assuming area;
- Level four, uses all data of the assumed area as the training set, and selects test data from a new well in another area, which the test result will be the true performance of model generalization.

The generalization performance of most current ROP prediction models is concentrated at the first and second levels, which is also the main reason for the limited application of the model. Although a small number of papers had tested their model in a new well through dynamic data splitting, it is a weakened version of the third level, as the data from the new well was also involved in the modeling. Hence, how to improve the generation performance for the ML model is a key challenge for future model applications.

(3) Model interpretability

In general, the use of ML models for predicting ROP can receive good accuracy, however, these models are complex and always referred to as black-box, which brings difficulties in understanding its internal structure. As alternatives, some researchers started to extract the ML model to figure out the influence of each input parameter. Compared to other ML modeling methods, ANN is not only the most widely used at present (Fig. 16), but it is also relatively easier to extract. The basic structure of ANN includes the weights and biases that control the connections between input/hidden layers and hidden/output layers, hence the established ANN can be extracted as Eq. (5). According to Eq. (5), researchers have realized the extracting the ANN focused on the prediction of ROP (Elkatatny, 2018, 2019; Al-AbdulJabbar et al., 2020, 2021), UCS (Gowida et al., 2021), ECD (Abdelgawad et al., 2019), and formation pressure (Ahmed O.S. et al., 2019).

$$y=\left[\sum_{i=1}^{N} w_{2i}\left(f\left(\sum_{j=1}^{M} w_{1j}x_j+b_j\right)\right)\right]+b_2 \tag{5}$$

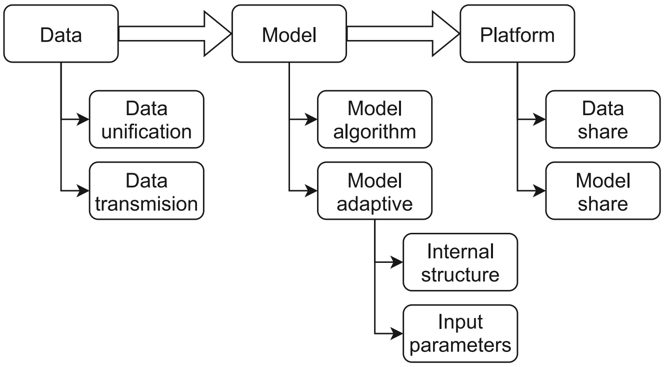

Fig. 23. Future roadmap for ML model in ROP prediction.

where y is the output, x is the input parameter, $f(x)$ is the activation function of ANN, N is the number of neurons in the hidden layer, M is the number of input parameters, w_1/w_2 is the weights between input/hidden and hidden/output layers, b_j/b_2 is the biases associated with the hidden and output layer.

In addition to extracting the ML model, some researchers use another way to interpret the model. With the established model, varying some of the input parameters, while others remain unchanged, then observing the changing trend of model output to perform sensitive analysis, which indicates the relationship and impact among input parameters (Barbosa et al., 2019). By analyzing sensitive plots, a suitable value for input variables can be obtained to reach the maximum value of ROP. Interpreting the ML model is an effective way to deepen the internal understanding of the model, and is also the key to improving its generalization performance.

5.2. Future roadmap

The future roadmap of the ML model for ROP prediction, as shown in Fig. 23, can be divided into three steps, including data, model and platform:

For data, data unification needs to eliminate the differences in data media, form, and structure, reduce data errors through automatic data preprocessing algorithms, and improve the value density and manageability of data. Data unification is the basic way to eliminate data isolation. Then, fast and low-error data transmission will be a key link affecting the further development of the ML model. Although there is already a WITS (wellsite information transfer specification) standard, combined with the development of data unification, there may be more undefined new types of data that need to be transmitted, and the improvement to the WITS standard is foreseeable.

For the model, improving model interpretability and generalization are two directions that must be developed. Rather than being considered as a black box, understanding the internal structure of the model and knowing the impact on the output of each input parameter will better unleash the potential of the model, and researchers can select the most suitable model for different situations, instead of blindly building complex model structures to improve tiny accuracy. In addition, being able to adapt the model itself, for instance when external conditions change (such as technical or formation), the model can autonomously adjust the internal structure or input parameters, which is also an effective measure to improve the generalization performance of the ML model. Some of the researchers have realized this and launched related explorations, such as using sliding window method to update both the training datasets and model in real-time (Zhang et al., 2022) and the higher the update frequency will bring the more accurate predicting performance (Zhang et al., 2023).

For the platform, when data unification is completed and model generalization performance is excellent enough, rapid growth in the amount of both data and model is foreseeable. Then, establishing a cloud-based intelligent platform can improve the efficiency of data and model applications. Data and model sharing will undoubtedly help reduce the cost of data collection and model establishment, while also helping to improve the accuracy and application scope of modeling.

6. Conclusion

The introduction of ML algorithms has greatly improved the ROP prediction accuracy and has great potential for development and application. This review systematically sorts out and analyses the whole process of ROP prediction using ML models. The conclusions of this review are as follows:

- A very wide range of input parameter types have been used, an average of six engineering parameters (including WOB, RPM, Q, well depth, MW, and SPP, sorted by usage frequency) and two geological parameters (UCS and PPG) were used. The most common sample size was less than 10,000 pieces of data.
- Data cleaning, outlier removal, data filtering, data normalization, feature selection, and data splitting are necessary pre-processing procedures for ML modeling, and using these procedures in an orderly manner can significantly improve the modeling accuracy. This review also compares and validates different algorithms used in each step combined with real datasets.
- Single ML algorithm is still the mainstream method for ROP prediction, and the ANN algorithm represented by MLP is the most popular single algorithm. ML modeling algorithms can produce a major improvement in accuracy compared with traditional algorithms. And the introduction of optimization algorithms currently only plays a role in improving training efficiency.
- Data isolation, model generalization, and interpretability are the three major gaps that need to be solved in the current field of machine learning for ROP prediction.

Data availability

All the data needed to evaluate the conclusions are presented in the paper. Additional data related to this paper may be requested from the corresponding author.

CRediT authorship contribution statement

Qian Li: Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Methodology, Funding acquisition, Data curation, Conceptualization. **Jun-Ping Li:** Writing – review & editing, Validation, Investigation, Formal analysis. **Lan-Lan Xie:** Writing – review & editing, Validation, Methodology, Formal analysis, Data curation.

Declaration of competing interest

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 study was financially supported by CNOOC China Co., Ltd. Zhanjiang Branch (CNOOC-KJ135ZDXM3 8ZJ05ZJ). And the filed data source used in the paper was provided by Professor Zhu Haiyan from State Key Laboratory of Oil and Gas Reservoir Geology and Exploitation (Chengdu University of Technology) and Mr Zhang Chao from CNOOC China Co., Ltd. Zhanjiang Branch. Theoretical support from Chengdu University of Technology and Institute of Exploration Technology, CAGS are also acknowledged.

Nomenclature List

a	wavelet scaling factor
$ARPE$	average percentage relative error
b_j/b_2	biases associated with the hidden and output layer
$b(t)$	basic functions
$\mathrm{cov}(x, y)$	the covariance between x and y
$ED(i)$	error distribution
f_c	cutoff frequency of the filter
$f(x)$	activation function of ANN
$I_{\text{A-base}}/I_{\text{A-compare}}$	the base/comparison indicator to indicate the degree of fit
$I_{\text{E-base}}/I_{\text{E-compare}}$	the base/comparison indicator to indicate the magnitude of error
k	the number used for cross validation in data splitting
m	the smallest integer that satisfies the needed scaling condition
M	number of input parameters
MAE	mean absolute error
$MAE_{\max}$	maximum absolute error
$MAPE$	mean absolute percentage relative error
MSE	mean square error
n	sample number of data set
N	number of neurons in hidden layer
$NER(i)$	normalized error rate
N_RMSE	normalized root mean square error
$P_5/P_{25}/P_{75}/P_{95}$	5th/25th/75th/95th percentile of dataset
PI	performance index
P_R^2	pseudo coefficient of determination
R	correlation coefficient
R^2	coefficient of determination
$RMSE$	root mean square error
$SMAPE$	symmetric mean absolute percentage error
SSE	sum of square error
VAF	variance account
w	frequency
w_1/w_2	weights between input/hidden and hidden/output layers
x_c	measurement correction value
x_1/x_2	measured/estimated value
x_i	initial value
$\bar{x}$	mean value of the dataset
$X_{\min}/X_{\max}$	the minimum/maximum value in the initial dataset
Y_i	true value in prediction
$\overline{Y}$	the average value of the true values
Y_p	predicted value in prediction
ΔD	minimum depth of the sampling interval of the dataset
τ	wavelet translation factor
$\Psi(t)$	basic wavelet functions
σ	standard error of the initial dataset
σ_1/σ_2	standard error of the measured/estimated value

Acronyms List

AD	average deviation
ANFIS	adaptive network-based fuzzy inference system
ANN	artificial neural network
AZI	azimuth angle
C	traditional statistical regression algorithm
CO	optimized statistical regression algorithm
DT	decision tree
ECD	equivalent circulating density
ELM	extreme learning machine
FA	Factor analysis
FL	filter loss
GA	genetic algorithm
GR	gamma ray
HL	hook load
INC	incline angle
IQR	interquartile range
KNN	k-nearest neighbor
LWD	logging while drilling

MLe	ensemble algorithm machine learning model
MLs	single algorithm machine learning model
MLO	optimized machine learning algorithm
MLP	multilayer perceptron
MW	mud weight
MV	mud viscosity
NSGA-II	Non-dominated sorting genetic algorithm II
OD	original distribution
PCA	principal component analysis
PPG	por pressure gradient
Q	mud flowrate
R	Reynold number
RBFNN	radial basis function neural network
RBF	radial basis function
RF	random forest
ROP	rate of penetration
RPM	rotation per minutes, rotation speed
RQD	rock quality designation
SC	solid content
SG	Savitzky-Golay filter
SPP	stand pipe pressure
SVR	support vector regression
T	torque
TEMP	mud temperature
UCS	uniaxial compressive strength
WOB	weight on bit
YP	yield point

References

Abbas, A.K., Rushdi, S., Alsaba, M., et al., 2019. Drilling rate of penetration prediction of high-angled wells using artificial neural networks. J. Energy Resour. Technol. 141 (11), 112904. https://doi.org/10.1115/1.4043699.

Abdelgawad, K.Z., Elzenary, M., Elkatatny, S., et al., 2019. New approach to evaluate the equivalent circulating density (ECD) using artificial intelligence techniques. J. Pet. Explor. Prod. Technol. 9 (2), 1569–1578. https://doi.org/10.1007/s13202-018-0572-y.

Ahmed, A., Elkatatny, S., Ali, A., et al., 2019. New model for pore pressure prediction while drilling using artificial neural networks. Arabian J. Sci. Eng. 44 (6), 6079–6088. https://doi.org/10.1007/s13369-018-3574-7.

Ahmed, O.S., Adeniran, A.A., Samsuri, A., 2019. Computational intelligence based prediction of drilling rate of penetration: a comparative study. J. Petrol. Sci. Eng. 172, 1–12. https://doi.org/10.1016/j.petrol.2018.09.027.

Al-AbdulJabbar, A., Elkatatny, S., Abdulhamid, M.A., et al., 2020. Prediction of the rate of penetration while drilling horizontal carbonate reservoirs using the self-adaptive artificial neural networks technique. Sustainability 12 (4), 1376. https://doi.org/10.3390/su12041376.

Al-AbdulJabbar, A., Elkatatny, S., Mahmoud, M., et al., 2019. A robust rate of penetration model for carbonate formation. J. Energy Resour. Technol. 141 (4), 042903. https://doi.org/10.1115/1.4041840.

Alsaihati, A., Elkatatny, S., Gamal, H., 2022. Rate of penetration prediction while drilling vertical complex lithology using an ensemble learning model. J. Petrol. Sci. Eng. 208, 109335. https://doi.org/10.1016/j.petrol.2021.109335.

Al-AbdulJabbar, A., Mahmoud, A.A., Elkatatny, S., 2021. Artificial neural network model for real-time prediction of the rate of penetration while horizontally drilling natural gas-bearing sandstone formations. Arabian J. Geosci. 14 (2), 117. https://doi.org/10.1007/s12517-021-06457-0.

Ansari, H.R., Sarbaz Hosseini, M.J., Amirpour, M., 2017. Drilling rate of penetration prediction through committee support vector regression based on imperialist competitive algorithm. Carbonates Evaporites 32 (2), 205–213. https://doi.org/10.1007/s13146-016-0291-8.

Ashrafi, S.B., Anemangely, M., Sabah, M., et al., 2019. Application of hybrid artificial neural networks for predicting rate of penetration (ROP): a case study from Marun oil field. J. Petrol. Sci. Eng. 175, 604–623. https://doi.org/10.1016/j.petrol.2018.12.013.

Bani Mustafa, A., Abbas, A.K., Alsaba, M., et al., 2021. Improving drilling performance through optimizing controllable drilling parameters. J. Pet. Explor. Prod. Technol. 11 (3), 1223–1232. https://doi.org/10.1007/s13202-021-01116-2.

Barbosa, L.F.F.M., Nascimento, A., Mathias, M.H., et al., 2019. Machine learning methods applied to drilling rate of penetration prediction and optimization - a review. J. Petrol. Sci. Eng. 183, 106332. https://doi.org/10.1016/j.petrol.2019.106332.

Bezminabadi, S.N., Ramezanzadeh, A., Esmaeil Jalali, S.M., et al., 2017. Effect of rock properties on ROP modeling using statistical and intelligent methods: a case study of an oil well in southwest of Iran. Arch. Min. Sci. 62 (1), 131–144. https://doi.org/10.1515/amsc-2017-0010.

Bodaghi, A., Ansari, H.R., Gholami, M., 2015. Optimized support vector regression for drilling rate of penetration estimation. Open Geosci. 7 (1), 870–879. https://doi.org/10.1515/geo-2015-0054.

Brenjkar, E., Biniaz Delijani, E., Karroubi, K., 2021. Prediction of penetration rate in drilling operations: a comparative study of three neural network forecast methods. J. Pet. Explor. Prod. Technol. 11 (2), 805–818. https://doi.org/10.1007/s13202-020-01066-1.

Brenjkar, E., Biniaz Delijani, E., 2022. Computational prediction of the drilling rate of penetration (ROP): a comparison of various machine learning approaches and traditional models. J. Petrol. Sci. Eng. 210, 110033. https://doi.org/10.1016/j.petrol.2021.110033.

Conradie, W., Craig, A., Palmigiano, A., et al., 2019. Modelling Informational Entropy, pp. 140–160. https://doi.org/10.1007/978-3-662-59533-6_9.

Darbor, M., Faramarzi, L., Shaifzadeh, M., 2019. Performance assessment of rotary drilling using non-linear multiple regression analysis and multilayer perceptron neural network. Bull. Eng. Geol. Environ. 78 (3), 1501–1513. https://doi.org/10.1007/s10064-017-1192-3.

Delavar, M.R., Ramezanzadeh, A., Tokhmechi, B., 2021. An investigation into the effect of geomechanical properties of reservoir rock on drilling parameters—a case study. Arabian J. Geosci. 14 (17), 1763. https://doi.org/10.1007/s12517-021-08168-y.

Deng, S., Wei, M., Xu, M., et al., 2021. Prediction of the rate of penetration using logistic regression algorithm of machine learning model. Arabian J. Geosci. 14 (21), 2230. https://doi.org/10.1007/s12517-021-08452-x.

Diaz, M.B., Kim, K.Y., Kang, T., et al., 2018. Drilling data from an enhanced geothermal project and its pre-processing for ROP forecasting improvement. Geothermics 72, 348–357. https://doi.org/10.1016/j.geothermics.2017.12.007.

Diaz, M.B., Kim, K.Y., Shin, H., et al., 2019. 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. 67, 225–232. https://doi.org/10.1016/j.jngse.2019.05.004.

Diaz, M.B., Kim, K.Y., 2020. Improving rate of penetration prediction by combining data from an adjacent well in a geothermal project. Renew. Energy 155, 1394–1400. https://doi.org/10.1016/j.renene.2020.04.029.

Elkatatny, S., 2018. New approach to optimize the rate of penetration using artificial neural network. Arabian J. Sci. Eng. 43 (11), 6297–6304. https://doi.org/10.1007/s13369-017-3022-0.

Elkatatny, S., 2019. Development of a new rate of penetration model using self-adaptive differential evolution-artificial neural network. Arabian J. Geosci. 12 (2), 19. https://doi.org/10.1007/s12517-018-4185-z.

Elkatatny, S., 2020. Real-time prediction of rate of penetration in s-shape well profile using artificial intelligence models. Sensors 20 (12), 3506. https://doi.org/10.3390/s20123506.

Elkatatny, S., 2021. Real-time prediction of rate of penetration while drilling complex lithologies using artificial intelligence techniques. Ain Shams Eng. J. 12 (1), 917–926. https://doi.org/10.1016/j.asej.2020.05.014.

Encinas, M.A., Tunkiel, A.T., Sui, D., 2022. Downhole data correction for data-driven rate of penetration prediction modeling. J. Petrol. Sci. Eng. 210, 109904. https://doi.org/10.1016/j.petrol.2021.109904.

Eskandarian, S., Bahrami, P., Kazemi, P., 2017. A comprehensive data mining approach to estimate the rate of penetration: application of neural network, rule based models and feature ranking. J. Petrol. Sci. Eng. 156, 605–615. https://doi.org/10.1016/j.petrol.2017.06.039.

Fan, H., Wu, M., Cao, W., et al., 2021. An operating performance assessment strategy with multiple modes based on least squares support vector machines for drilling process. Comput. Ind. Eng. 159, 107492. https://doi.org/10.1016/j.cie.2021.107492.

Gan, C., 2019. Intelligent Modeling of Formation Drillability Field and Drilling Rate of Penetration Optimization in Complex Conditions. PhD Thesis.. China University of Geosciences (in Chinese).

Gan, C., Cao, W., Wu, M., et al., 2019a. Prediction of drilling rate of penetration (ROP) using hybrid support vector regression: a case study on the Shennongjia area, Central China. J. Petrol. Sci. Eng. 181, 106200. https://doi.org/10.1016/j.petrol.2019.106200.

Gan, C., Cao, W., Wu, M., et al., 2019b. Two-level intelligent modeling method for the rate of penetration in complex geological drilling process. Appl. Soft Comput. 80, 592–602. https://doi.org/10.1016/j.asoc.2019.04.020.

Gan, C., Cao, W., Lu, K., et al., 2020. A new hybrid bat algorithm and its application to the ROP optimization in drilling processes. IEEE Trans. Ind. Inf. 16 (12), 7338–7348. https://doi.org/10.1109/TII.2019.2943165.

Geekiyanage, S.C.H., Sui, D., Aadnoy, B.S., et al., 2018. Drilling data quality management: case study with a laboratory scale drilling rig. In: ASME 2018 37th International Conference on Ocean. Offshore and Arctic Engineering, Madrid, Spain.

Geng, L., 2021. Application status and development suggestions of big data technology in petroleum engineering. Petroleum Drilling Techniques 49, 72–78. https://doi.org/10.11911/syztjs.2020134 (in Chinese).

Gowida, A., Elkatatny, S., Gamal, H., 2021. Unconfined compressive strength (UCS) prediction in real-time while drilling using artificial intelligence tools. Neural Comput. Appl. 33 (13), 8043–8054. https://doi.org/10.1007/s00521-020-05546-7.

Hassan, A., Elkatatny, S., Al-Majed, A., 2020. Coupling rate of penetration and mechanical specific energy to Improve the efficiency of drilling gas wells. J. Nat. Gas Sci. Eng. 83, 103558. https://doi.org/10.1016/j.jngse.2020.103558.

Hegde, C., Daigle, H., Millwater, H., et al., 2017. Analysis of rate of penetration (ROP) prediction in drilling using physics-based and data-driven models. J. Petrol. Sci. Eng. 159, 295–306. https://doi.org/10.1016/j.petrol.2017.09.020.

Kor, K., Altun, G., 2020. Is support vector regression method suitable for predicting rate of penetration? J. Petrol. Sci. Eng. 194, 107542. https://doi.org/10.1016/j.petrol.2020.107542.

Kor, K., Ertekin, S., Yamanlar, S., et al., 2021. Penetration rate prediction in heterogeneous formations: a geomechanical approach through machine learning. J. Petrol. Sci. Eng. 207, 109138. https://doi.org/10.1016/j.petrol.2021.109138.

Law, K., Stuart, A., Zygalakis, K., 2015. Data Assimilation: A Mathematical Introduction. Springer Link. https://doi.org/10.1007/978-3-319-20325-6.

Leng, S., Lin, J., Hu, Z., et al., 2020. A hybrid data mining method for tunnel engineering based on real-time monitoring data from tunnel boring machines. IEEE Access 8, 90430–90449. https://doi.org/10.1109/ACCESS.2020.2994115.

Li, L., Zhang, X., Xue, M., 2013. Explaining information gain and information gain ratio in information theory. ICIC Express Lett 7 (8), 2385–2391.

Li, Q., Cao, Y., Zhu, H., 2021a. Discussion on the lower limit of data validity for ROP prediction based on artificial intelligence. Drilling Engineering 48, 21–30. https://doi.org/10.12143/j.ztgc.2021.03.003 (in Chinese).

Li, Q., Qu, F., He, J., et al., 2021b. Prediction model of mechanical ROP during drilling based on BAS-BP. Journal of Xi'an Shiyou University (Natural Science Edition) 36, 89–95. https://doi.org/10.3936/j.issn.173-064X.2021.06.014 (in Chinese).

Li, Q., Qu, F., He, J., et al., 2021c. Rate of penetration for drilling prediction model based on PSO-BP. Sci. Technol. Eng. 21, 7984–7990 (in Chinese).

Li, Q., Zhou, C., Zhu, H., et al., 2021d. Application of integration and preliminary analysis of production data in drilling. In: 21st National Mining Engineering (Geotechnical Drilling and Excavation Engineering) Academic Exchange Annual Conference. Shanxi, China (in Chinese).

Li, Y., 2020. Research on Parameters Optimization Using Machine Learning Algorithm in Deep Sea Oil Drilling Field. Master Thesis.. Beijing University of Posts and Telecommunications (in Chinese).

Liao, X., Khandelwal, M., Yang, H., et al., 2020. Effects of a proper feature selection on prediction and optimization of drilling rate using intelligent techniques. Eng. Comput. 36 (2), 499–510. https://doi.org/10.1007/s00366-019-00711-6.

Liu, N., Gao, H., Zhao, Z., et al., 2021. A stacked generalization ensemble model for optimization and prediction of the gas well rate of penetration: a case study in Xinjiang. J. Pet. Explor. Prod. Technol. 12 (6), 1595–1608. https://doi.org/10.1007/s13202-021-01402-z.

Liu, R., 2021. Research on Drilling Parameters Optimization under Big Data Environment. Master Thesis.. Shiyou University, Xi'an (in Chinese).

Liu, S., Sun, J., Gao, X., et al., 2019. Analysis and establishment of drilling speed prediction model for drilling machinery based on artificial neural networks. Computer Science 46, 605–608 (in Chinese).

Mariani, M., Tweneboah, O., Beccar Varela, M., 2021. Principal component analysis. In: Data Science in Theory and Practice. Wiley, pp. 151–163. https://doi.org/10.1002/9781119674757.ch11.

McLaughlin, D., 2014. Data assimilation. In: Encyclopedia of Remote Sensing. Springer, New York, pp. 131–134. https://doi.org/10.1007/978-0-387-36699-9_33.

Mehrad, M., Bajolvand, M., Ramezanzadeh, A., et al., 2020. Developing a new rigorous drilling rate prediction model using a machine learning technique. J. Petrol. Sci. Eng. 192, 107338. https://doi.org/10.1016/j.petrol.2020.107338.

Najjarpour, M., Jalalifar, H., Norouzi-Apourvari, S., 2020. The effect of formation thickness on the performance of deterministic and machine learning models for rate of penetration management in inclined and horizontal wells. J. Petrol. Sci. Eng. 191, 107160. https://doi.org/10.1016/j.petrol.2020.107160.

Najjarpour, M., Jalalifar, H., Norouzi-Apourvari, S., 2022. Half a century experience in rate of penetration management: application of machine learning methods and optimization algorithms - a review. J. Petrol. Sci. Eng. 208, 109575. https://doi.org/10.1016/j.petrol.2021.109575.

Oyedere, M., Gray, K., 2020. ROP and TOB optimization using machine learning classification algorithms. J. Nat. Gas Sci. Eng. 77, 103230. https://doi.org/10.1016/j.jngse.2020.103230.

Qi, W., 2020. Data Preparation and Feature Engineering. Publishing House of Electronics Industry, Beijing (in Chinese).

Qu, F., 2021. Research on Establishment and Application of Drilling Parameter Optimization Model Based on Big Data and Intelligent Algorithms. Master Thesis. Xi'an Shiyou University (in Chinese).

Reshef, D.N., Reshef, Y.A., Finucane, H.K., et al., 2011. Detecting novel associations in large data sets. Science 334 (6062), 1518–1524. https://doi.org/10.1126/science.1205438.

Sabah, M., Talebkeikhah, M., Wood, D.A., et al., 2019. A machine learning approach to predict drilling rate using petrophysical and mud logging data. Earth Science Informatics 12 (3), 319–339. https://doi.org/10.1007/s12145-019-00381-4.

Samaei, M., Ranjbarnia, M., Nourani, V., et al., 2020. Performance prediction of tunnel boring machine through developing high accuracy equations: a case study in adverse geological condition. Measurement 152, 107244. https://doi.org/10.1016/j.measurement.2019.107244.

Savitzky, A., Golay, M., 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36, 1627–1639. https://doi.org/10.1021/ac60214a047.

Soares, C., Gray, K., 2019. Real-time predictive capabilities of analytical and machine learning rate of penetration (ROP) models. J. Petrol. Sci. Eng. 172, 934–959. https://doi.org/10.1016/j.petrol.2018.08.083.

Soares, C., Armenta, M., Panchal, N., 2020. Enhancing reamer drilling performance in deepwater Gulf of Mexico wells. SPE Drill. Complet. 35 (3), 329–356.

Sun, X., 2006. Research on Economic Analysis and Evaluation of Drilling Engineering in Petroleum Enterprises. Master Thesis. Southwest Petroleum University (in Chinese).

Tan, Y., 2019. Research on Application and Optimization of Machine Learning Algorithm in Oil Drilling Field. Master Thesis.. Beijing University of Post and Telecommunications (in Chinese).

Tan, Y., Yin, Z., Xu, L., et al., 2019. Research on outlier marking method of drilling data in machine learning. In: 6th International Academic Conference on Digital Oilfield, Shaanxi, China (in Chinese).

Wang, W., Liu, X., Dou, P., et al., 2018. A ROP prediction method based on neutral network for the deep layers. Oil Drilling and Production Technology 40, 121–124. https://doi.org/10.13639/j.odpt.2018.S0.034 (in Chinese).

Xiong, H., Li, Q., 2018. ROP prediction model based on formation composition and drilling parameters. Explor. Eng. 45, 195–201 (in Chinese).

Youcefi, M.R., Hadjadj, A., Bentriou, A., et al., 2020. Rate of penetration modeling using hybridization extreme learning machine and whale optimization algorithm. Earth Science Informatics 13 (4), 1351–1368. https://doi.org/10.1007/s12145-020-00524-y.

Yu, Y., Huang, K., Li, H., 2021. Research on ROP prediction method based on machine learning and multi-source data preprocessing technology. China Petroleum and Chemical Standards and Quality 41, 133–136 (in Chinese).

Yuswandari, A., Prayoga, A., Purba, D., 2019. Rate of penetration (ROP) prediction using artificial neural network to predict rop for nearby well in a geothermal field. In: 44th Workshop on Geothermal Reservoir Engineering. Stanford University, Stanford, California.

Zhang, C., Song, X., Su, Y., et al., 2022. Real-time prediction of rate of penetration by combining attention-based gated recurrent unit network and fully connected neural networks. J. Petrol. Sci. Eng. 213, 110396. https://doi.org/10.1016/j.petrol.2022.110396.

Zhang, C., Song, X., Liu, Z., et al., 2023. Real-time and multi-objective optimization of rate-of-penetration using machine learning methods. Geoengy Science and Engineering 223, 211568. https://doi.org/10.1016/j.geoen.2023.211568.

Zhang, W., Tang, L., Chen, F., et al., 2021. Prediction for TBM penetration rate using four hyperparameter optimization methods and random forest model. J. Basic Sci. Eng. 29, 1186–1200. https://doi.org/10.16058/j.issn.1005-0930.2021.05.009 (in Chinese).

Zhang, Z., 2020. Generalized mutual information. Stats 3. https://doi.org/10.3390/stats3020013.

Zhao, Y., Noorbakhsh, A., Koopialipoor, M., et al., 2020. A new methodology for optimization and prediction of rate of penetration during drilling operations. Eng. Comput. 36 (2), 587–595. https://doi.org/10.1007/s00366-019-00715-2.

Zhao, Y., Sun, T., Yang, J., et al., 2019. Extreme learning machine-based offshore drilling ROP monitoring and real-time optimization. China Offshore Oil Gas 31, 138–142. https://doi.org/10.11935/j.issn.1673-1506.2019.06.018 (in Chinese).

Zhou, Z., 2016. Machine Learning. Tsinghua University Press, Beijing.

Zhou, Y., Chen, X., Zhao, H., et al., 2021a. A novel rate of penetration prediction model with identified condition for the complex geological drilling process. J. Process Control 100, 30–40. https://doi.org/10.1016/j.jprocont.2021.02.001.

Zhou, Y., Chen, X., Wu, M., et al., 2021b. Modeling and coordinated optimization method featuring coupling relationship among subsystems for improving safety and efficiency of drilling process. Appl. Soft Comput. 99, 106899. https://doi.org/10.1016/j.asoc.2020.106899.

Zhu, Y., 2021. Research on Deepwater Drilling Rate Optimization for Intelligent Application. Master Thesis. Beijing University of Post and Telecommunications (in Chinese).

Zielinski, T., 2021. Fast fourier transform. In: Starting Digital Signal Processing in Telecommunication Engineering. Springer, pp. 93–114.

Zuo, D., 2018. Research on ROP Increasing in Keshen Block Based on Big Data Analysis. Master Thesis. China University of Petroleum (in Chinese).