

Original Paper

Online residual correction and hybrid modeling framework for cross-regional rate-of-penetration prediction

Xiao-Yang Li ^{a,b,*}, Zong-Jie Mu ^{a,b,**}, Pan-Pan Zhang ^{a,b}, Zhe-Hua Yang ^{a,b}, Yu-Xuan Zhao ^c, Lian-Feng Fu ^d

^a National Key Laboratory of Petroleum Resources and Engineering, China University of Petroleum (Beijing), Beijing, 102249, China

^b Xinjiang Key Laboratory of Intelligent Petroleum Exploration and Engineering, China University of Petroleum (Beijing) at Karamay, Karamay, 834000, Xinjiang, China

^c Research Institute of Oil Production Technology, Xinjiang Petroleum Administration Bureau, Karamay, 834000, Xinjiang, China

^d College of Geosciences, China University of Petroleum (Beijing), Beijing, 102249, China

ARTICLE INFO

Article history:

Received 12 September 2025

Received in revised form

2 February 2026

Accepted 25 May 2026

Available online 29 May 2026

Edited by Jia-Jia Fei

Keywords:

Rate-of-penetration prediction

Online residual correction

Model transferability

Hybrid modeling

Drilling parameter optimization

ABSTRACT

In oil and gas drilling, the mechanical rate of penetration (ROP) is a key indicator of drilling efficiency, and its optimization is crucial for reducing drilling costs and shortening operational cycles. However, owing to the complexity and variability of formation conditions, as well as the diversity of drilling parameters and mechanical tool states, the existing ROP-prediction models often exhibit poor generalization across different regions. To address this issue, a cross-regional adaptive adjustment framework is constructed, which explicitly decouples global stable patterns from local dynamic disturbances to achieve high-precision ROP prediction. Multiregional historical drilling data are used to pretrain deep-learning models (ResNet and FTTransformer), establishing a high-accuracy offline baseline prediction. An online-trained gradient-boosting tree model (LightGBM) then captures regional variations and predicts the discrepancies between the current formation, drilling parameters, and tool conditions, in real time, relative to historical data, for residual correction. The baseline predictions and residual corrections are finally fed into a multilayer perceptron for nonlinear fusion, allowing the magnitudes of residual corrections to adapt dynamically to the real-time operating conditions, thereby enhancing the model's adaptability in new regions. This method is trained on historical data from four ultradeep wells and validated through simulations on an independent deep well. Results show that the offline model achieves a coefficient of determination (R^2) of 0.9585, root mean square error (RMSE) of 5.98 m/h, and mean absolute error (MAE) of 2.90 m/h on the test well. Incorporating online residual correction further improves the performance to an R^2 of 0.9695, RMSE of 5.13 m/h, and MAE of 1.67 m/h. These results demonstrate strong generalization ability under cross-regional and multicondition scenarios, providing effective technical support for real-time ROP prediction and drilling-parameter optimization.

© 2026 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

* Corresponding author.

** Corresponding author.

E-mail addresses: xiaoyang.l@hotmail.com (X.-Y. Li), muzongjie@cupk.edu.cn (Z.-J. Mu).

Peer review under the responsibility of China University of Petroleum (Beijing).

Abbreviations			
RMSE	Root mean square error	CNL	Compensated neutron log
MAE	Mean absolute error	RT	True formation resistivity
R^2	Coefficient of determination	SP	Spontaneous potential
ResNet	Residual network	ABD	Average borehole diameter
TSTransformer	Time-series transformer	SPP	Standpipe pressure
FTTransformer	Feature token transformer	SPM	Strokes per minute
TabTransformer	Tabular transformer	RPM	Revolutions per minute
LightGBM	Light gradient boosting machine	ECD	Equivalent circulating density
XGBoost	Extreme gradient boosting	TG	Total gas
CatBoost	Categorical boosting	C1	Methane
MLP	Multilayer perceptron	C2	Ethane
FFN	Feed-forward network	C3	Propane
STM	Long short-term memory	iC4	Isobutane
CNN	Convolutional neural network	nC4	Normal butane
ROP	Rate of penetration	iC5	Isopentane
WOB	Weight on bit	nC5	Normal pentane
PDC	Polycrystalline diamond compact	DCI	Drilling strength index
GR	Gamma ray	AC	Acoustic log
DEN	Density	CAL	Caliper log
		RXO	Resistivity of flushed zone
		MWD	Measurements while drilling

1. Introduction

In recent years, the optimization of drilling parameters has been widely recognized as a key factor for improving drilling efficiency and reducing operational costs, with the rate of penetration (ROP) serving as a core indicator directly affecting drilling cycles and overall operational performance (Mensa-Wilmot et al., 2010). Field studies have demonstrated that by properly adjusting critical parameters such as weight on bit (WOB) and rotary speed, drilling efficiency can be significantly improved, drilling time shortened, equipment wear reduced, and the stability of the drilling process enhanced (Hegde et al., 2019; Song et al., 2024). Since ROP comprehensively reflects the dynamic characteristics of the drilling process, accurate prediction of ROP is considered a key prerequisite for guiding parameter optimization and achieving high-efficiency drilling operations (Najjarpour et al., 2022; Mazen et al., 2021).

ROP prediction models have undergone long-term development and evolution. Early studies were primarily based on analytical and empirical theoretical frameworks, with Graham and Muench (1959), Maurer (1962), Bingham (1965), and Teale (1965) systematically elucidating the mechanisms of ROP formation from the perspectives of drilling parameter optimization, cuttings removal, rock drillability, and the concept of specific energy, thereby establishing the theoretical foundation for traditional ROP modeling. Subsequently, with the advancement of computational technology, ROP prediction gradually incorporated statistical modeling and multivariate analysis methods. Young (1969) proposed the concept of computerized drilling control, while Bourgoyne and Young (1974) enhanced prediction accuracy by integrating drilling weight, rotational speed, and other parameters through multiple regression models. Warren (1987), Detournay and Defourny (1992), and Hareland and Rampersad (1994) further extended the physical basis of models by considering drill bit types, mechanical mechanisms, and wear effects. In the 21st century, ROP prediction has increasingly been combined with real-time data and field optimization. Eren and Ozbayoglu (2010) introduced a real-time parameter optimization method based on field data, and Soares et al. (2016) systematically evaluated ROP

models for PDC bits, highlighting the significant influence of rock strength on model coefficients and further improving the applicability of traditional models under actual formation conditions.

Recent research on ROP prediction has primarily focused on traditional machine learning models and deep learning models, gradually evolving toward real-time online prediction methods based on these two types of models (Hegde et al., 2015, 2017; Soares and Gray, 2019; Shaygan and Jamshidi, 2023). Across these studies, machine learning methods have consistently played a central role in ROP prediction. Yang et al. (2025) proposed a high-precision prediction framework integrating multiple machine learning algorithms, enhancing model robustness and generalization, and providing support for real-time optimization of drilling parameters. The integration of analytical models with data-driven methods has been shown to significantly improve ROP prediction under complex formation conditions, for example, by dynamically characterizing rock strength parameters through machine learning-based real-time estimation of lithology and P-wave velocity (Sharma et al., 2025). ROP prediction models have gradually been embedded into constraint-aware optimization and rig control frameworks to guide drilling parameter adjustments and improve overall operational performance (Yahia et al., 2025). Furthermore, in deviated wells and complex wellbore conditions, intelligent optimization approaches combining similar-well data selection with ensemble learning models have been demonstrated to further enhance ROP (Karnot et al., 2025).

Deep learning models have gradually shown advantages in ROP prediction, effectively capturing the nonlinear relationships and complex temporal features of drilling parameters (Abbas et al., 2018; Liu et al., 2022; Urdaneta et al., 2024; Bai et al., 2024). However, interpretability and engineering applicability remain major challenges for deep models. To address this, Pei et al. (2024) conducted an interpretable analysis of feedforward neural network-based ROP prediction models by leveraging the piecewise linear characteristics of the ReLU activation function, revealing the influence mechanism of input parameters on prediction outcomes. As model complexity increases, research has focused on prediction performance under heterogeneous formations, constrained inputs, and small-sample scenarios. Fan et al. (2025) proposed a temporal

modeling framework combining signal decomposition and Bayesian causal inference, constructing a model integrating EEMD, Bayesian networks, and TCN to enhance stability and controllability under complex operational conditions. Zhang et al. (2026) further incorporated temporal attention mechanisms and graph neural networks to characterize complex dependencies among multiple variables and their dynamic evolution. Addressing small-sample and heterogeneous formation issues, Peng et al. (2025) applied unsupervised learning for formation clustering and achieved differentiated ROP prediction and parameter optimization; Ma et al. (2025) combined meta-learning with data augmentation to develop a ROP prediction framework for small-sample scenarios; and Ma et al. (2026) systematically compared different sequential models under constrained feature conditions, highlighting that relatively simple architectures exhibit stronger robustness in small-sample scenarios.

Building on these two model categories, real-time online prediction has become a recent research focus. Hao et al. (2025) proposed an online adaptive model combining gradient boosting decision trees with Bayesian optimization, enabling dynamic updating of prediction strategies and improving responsiveness to ROP variations. Ahmednour et al. (2025) developed a stacked generalization ensemble learning framework based on real-time continuous data streams to achieve accurate prediction and optimization of deep well ROP, significantly enhancing model adaptability and accuracy. Bertini and Lavi (2025) designed an online incremental learning algorithm within a data stream framework, specifically addressing high-speed dynamic sensor data during drilling operations, enabling continuous and precise ROP prediction while substantially improving model robustness and intelligence.

The existing ROP prediction methods possess distinct advantages and limitations. Traditional machine-learning models exhibit high computational efficiency and clear structure, but fall short in modeling complex nonlinear relationships and temporal dependencies. Deep-learning models excel in feature representation and temporal modeling, yet entail high training costs, face challenges in real-time application, and exhibit poor generalization across geological regions. Online-prediction models offer advantages in terms of dynamic adaptability and real-time updating; however, they have yet to effectively integrate the superior predictive capabilities of deep-learning methods.

To address the decline in prediction accuracy caused by formation heterogeneity in cross-regional drilling, an adaptive prediction framework integrating an offline deep-learning model with an online tree-based model is developed in this work. The framework leverages historical multiregional drilling data to establish a robust baseline prediction and employs real-time residual correction to enhance adaptability to new operating conditions. Validation using multiformation data demonstrates the method's strong transferability and robustness, providing a feasible approach for real-time ROP prediction and drilling-parameter optimization in complex geological environments.

2. Material and methods

The dataset used in our work comprises multisource drilling and well-logging data from five wells in the Mahu Sag of the Junggar Basin, each drilled to depths exceeding 6000 m. Located on the northwestern margin of the basin, the Mahu Sag features complex geological conditions and has become a key exploration and development area in recent years. The region is dominated by interbedded carbonates, clastic rocks, and mudstones, with Permian Fengcheng and Wuerhe source rocks providing significant hydrocarbon potential. Previous studies have indicated the presence of abundant hydrocarbon resources, mainly hosted in deep

high-quality reservoirs, representing typical deep unconventional resources. Fig. 1 shows the geographical location of the Junggar Basin (Fig. 1(a)), major tectonic units (Fig. 1(b)), tectonic units of Mahu Sag (Fig. 1(c)), and relative locations of the study wells (Fig. 1(d)), reflecting the structural setting and stratigraphic features of the well-related areas.

The five wells (MaHW5101, MaHW5102, MaX5126, MaX5128, and Ma51X) penetrate multiple key Permian formations, targeting carbonate-rich and mixed sedimentary units, yielding a total of 31,132 valid drilling data records, each comprising synchronized measurements of drilling parameters and the corresponding ROP. The well depths and observation counts are summarized in Table 1, showing a broad coverage of drilling and geological features. Although all wells are located within the Mahu Sag, their distribution spans different blocks and structural units, with significant stratigraphic variations, providing a natural scenario for cross-block modeling and prediction.

Data are obtained from three types of engineering files: (1) drilling-parameters file, which contains downhole measurements such as the WOB, pump pressure, pump rate, rotary speed, mechanical ROP, and torque; (2) integrated logging data file containing data on gamma rays (GR), density (DEN), neutron porosity (ϕ_n), resistivity, and spontaneous potential (SP); and (3) well logging data file containing cutting descriptions, gas measurements, and fluorescence data. The parameter categories and representative indicators of the three files are listed in Table 2, which act as multisource inputs for systematic analysis and serve as a basis for data cleaning, feature extraction, and model development.

2.1. Data splitting

The dataset includes data from five wells in the Mahu Depression of the Junggar Basin: MaHW5101, MaHW5102, MaX5126, MaX5128, and Ma51X. Two strategies are used for model evaluation. First, five-fold cross-validation divides the training set into five nonoverlapping subsets for iterative training and validation, to reduce bias (Fig. 2). Second, independent well testing is conducted for temporal model experiments. To prevent future data leakages, the first four wells are used for training, while Ma51X serves as an independent test well to assess cross-well generalization.

2.2. Feature selection

During real-time ROP prediction modeling, certain formation and drilling parameters are excluded for various reasons. Formation-logging parameters, such as acoustic transit time, caliper, ϕ_n , DEN, GR, shallow resistivity, deep resistivity, mud-filtrate resistivity, and SP, rely on specialized logging tools and often suffer from low timeliness and discontinuous measurements, making them unsuitable for real-time modeling. Formation-pressure coefficients, fracture-pressure coefficients, lithology interpretation, mud-logging interpretation, well-testing results, and geochemical interpretations typically require post-processing or manual intervention, and cannot be obtained synchronously during drilling.

Some drilling-fluid parameters, such as mud density and viscosity, are theoretically related to the ROP, but are measured much less frequently than mechanical parameters, and their acquisition is not synchronized with other real-time features, preventing their use as consistent temporal inputs for the model. Volume-related indicators, including average borehole diameter, mud-pit volume changes, and total circulating mud volume, are often incompletely recorded, and retaining them

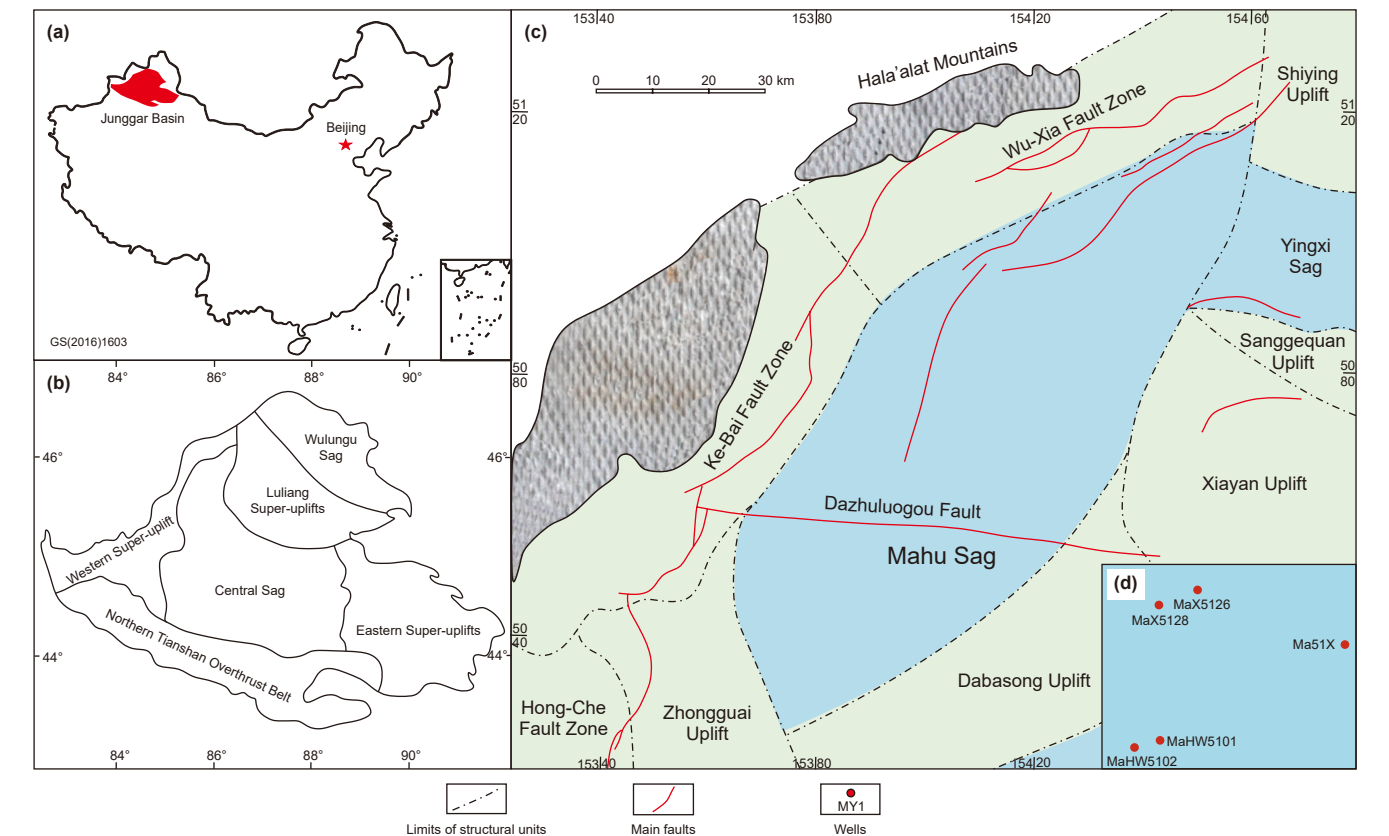


Fig. 1. (a) Location of the Junggar Basin. (b) Main tectonic units of the Junggar Basin. (c) Tectonic units of the Mahu Sag. (d) Relative locations of the study wells (modified from Lv et al., 2025).

Table 1
Parameter categories in the raw data files.

Raw data file	Parameter categories (abbreviations)
Integrated well logging file	Formation; Lithology; Bit size; Average borehole diameter; WOB; SPP; SPM; Mud density; Mud viscosity; RPM; ROP; Total pump rate; Torque; Hook load; Sigma index; Formation pressure coefficient; Formation fracture pressure coefficient; Inlet mud density; Outlet mud density; ECD; Inlet mud temperature; Outlet mud temperature; Inlet mud conductivity; Outlet mud conductivity; Inlet mud flow rate; Outlet mud flow rate; Mud pit volume change; Total circulating mud volume
Mud logging data file	TG; C1; C2; C3; iC4; iC5; nC4; nC5; Drilling time; DCI; Drilling coring; Sidewall coring; Mud density; Mud viscosity; Dry fluorescence exposure; Logging interpretation; Gas logging interpretation; Geochemical interpretation; Drilling data interpretation; Oil testing conclusion
Well logging data file	AC; CALI; CNL; DEN; GR; RT; RXO; SP

Table 2
Overview of five wells.

Well name	Start depth, m	End depth, m	Number of observations
MaHW5101	11.00	6500.00	6490
MaHW5102	11.00	6667.00	6657
MaX5126	1.00	6355.00	6355
MaX5128	12.00	6278.00	6267
Ma51X	539.00	5900.00	5362

would significantly reduce the sample effectiveness and model robustness.

Although the total pump rate is an important control parameter, it is highly correlated with the pump stroke, resulting in redundant information. The drill time is also highly correlated with the ROP, with substantial information overlap that can cause data leakage, and is therefore excluded. Overall, feature selection considers the real-time availability, sampling consistency, missing data, redundancy, and data-leakage risk, establishing a set of model inputs that are both real-time and practically applicable.

2.2.1. Feature-importance analysis

In the drilling-ROP prediction, accurately identifying the key features that exert a significant influence on the ROP is a fundamental prerequisite for improving the model's predictive performance. To systematically evaluate the relationships between the input variables and ROP, this study integrates a gain-based feature-importance analysis with a mutual information (MI) analysis, thereby quantitatively assessing and ranking the features from two complementary perspectives: model contribution and information dependency (Figs. 3 and 4).

In the gradient-boosting-tree (GBT)-based analysis, feature importance is employed to quantify the contribution of each input parameter to the reduction of prediction error during model training. Specifically, a feature is regarded as more influential for ROP prediction if its use in node splitting leads to a substantial decrease in the model loss, with the quantitative formulation given in Eq. (1). The node gain is calculated from the first- and second-order gradient information of the left and right child

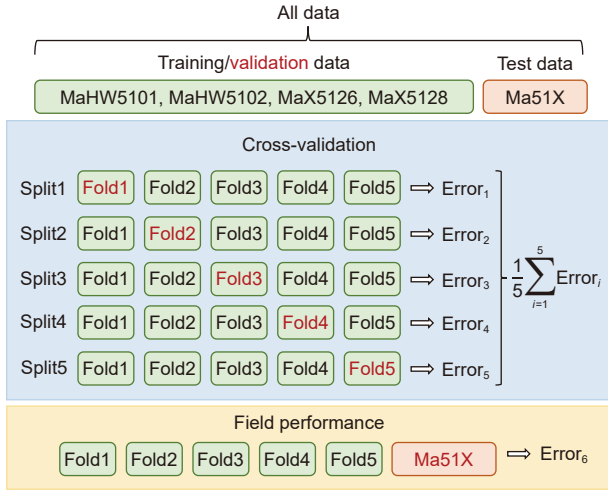


Fig. 2. Five-fold cross-validation and independent well-testing workflow.

nodes, as expressed in Eq. (2). To mitigate the influences of differences in splitting frequency and numerical scale among features, the resulting feature-importance scores are further normalized, as described in Eq. (3).

To further characterize the potential nonlinear dependencies between the features and ROP, an MI-based analysis is introduced. MI measures the amount of shared information between two random variables, and is defined in Eq. (4). Based on this formulation, the pairwise MI between features (Eq. (5)) as well as that between individual features and the target-variable ROP (Eq. (6)) can be computed. By jointly considering the results from the gain-based importance analysis and MI assessment, 26 key parameters that exhibit strong relevance to ROP are ultimately identified, providing a robust data foundation for the subsequent model construction and training.

$$G(\mathbf{F}_i) = \sum_{t \in T_i} \frac{N_t}{N} \Delta \text{Loss}_t \quad (1)$$

$$G_t = \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} - \gamma \quad (2)$$

$$I_i = \frac{G(\mathbf{F}_i)}{\sum_{j=1}^m G(\mathbf{F}_j)} \quad (3)$$

where $\mathbf{F}_i$ denotes the i -th feature and T_i represents the set of tree nodes for which $\mathbf{F}_i$ is used for splitting during model construction. The number of samples at node t is denoted by N_t , where N represents the total number of training samples. ΔLoss_t indicates the reduction in the objective-function loss after the split at node t , which quantifies the performance improvement brought by this split. The cumulative gain $G(\mathbf{F}_i)$ of feature $\mathbf{F}_i$ measures its overall contribution to the reduction in the objective-function loss throughout the training process and serves as the primary indicator of feature importance. Specifically, the node-split gain G_t is computed from the first-order gradients, G_L and G_R , and second-order gradients, H_L and H_R , of the left and right child nodes, together with the L_2 regularization coefficient λ and minimum split-gain threshold γ , which are introduced to determine whether a split is beneficial or not; the gradient terms characterize the direction and magnitude of the objective-function change induced by the split, whereas the regularization coefficient and gain threshold are used to control the model complexity and mitigate overfitting. The normalized importance score I_i of the i -th feature represents its relative contribution among all features, facilitating a fair comparison of feature importance.

$$I(\mathbf{X}; \mathbf{Y}) = \sum_{x \in \mathbf{X}} \sum_{y \in \mathbf{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \quad (4)$$

$$\mathbf{M}_{ij} = I(\mathbf{F}_i; \mathbf{F}_j) \quad (5)$$

$$I(\mathbf{F}_i; \mathbf{Y}) = \sum_{f \in \mathbf{F}_i} \sum_{y \in \mathbf{Y}} p(f, y) \log \frac{p(f, y)}{p(f)p(y)} \quad (6)$$

Mutual information is employed to quantify the amount of information shared between random variables: for two variables $\mathbf{X}$ and $\mathbf{Y}$, where $x \in \mathbf{X}$ and $y \in \mathbf{Y}$ denote their possible values, $p(x)$ and $p(y)$ are the marginal probability distributions, $p(x, y)$ is the joint probability distribution, and $I(\mathbf{X}; \mathbf{Y})$ denotes the mutual information between $\mathbf{X}$ and $\mathbf{Y}$. In the context of feature relevance analysis, the mutual information $\mathbf{M}_{ij}$ characterizes the dependency between the i -th and j -th features, whereas the mutual information $I(\mathbf{F}_i; \mathbf{Y})$ between feature $\mathbf{F}_i$ and the target variable $\mathbf{Y}$ quantifies the contribution of that feature to target prediction, with marginal distributions $p(f)$ and $p(y)$ and joint distribution $p(f, y)$, respectively.

2.2.2. Model input-parameter selection

Based on feature-importance analysis and systematic screening, the 26 selected parameters can be grouped into five categories:

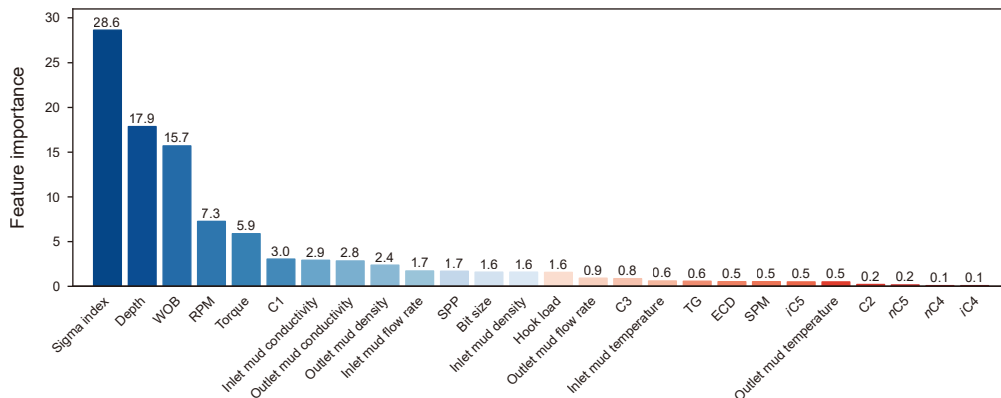


Fig. 3. Gain-based feature importance for ROP prediction.

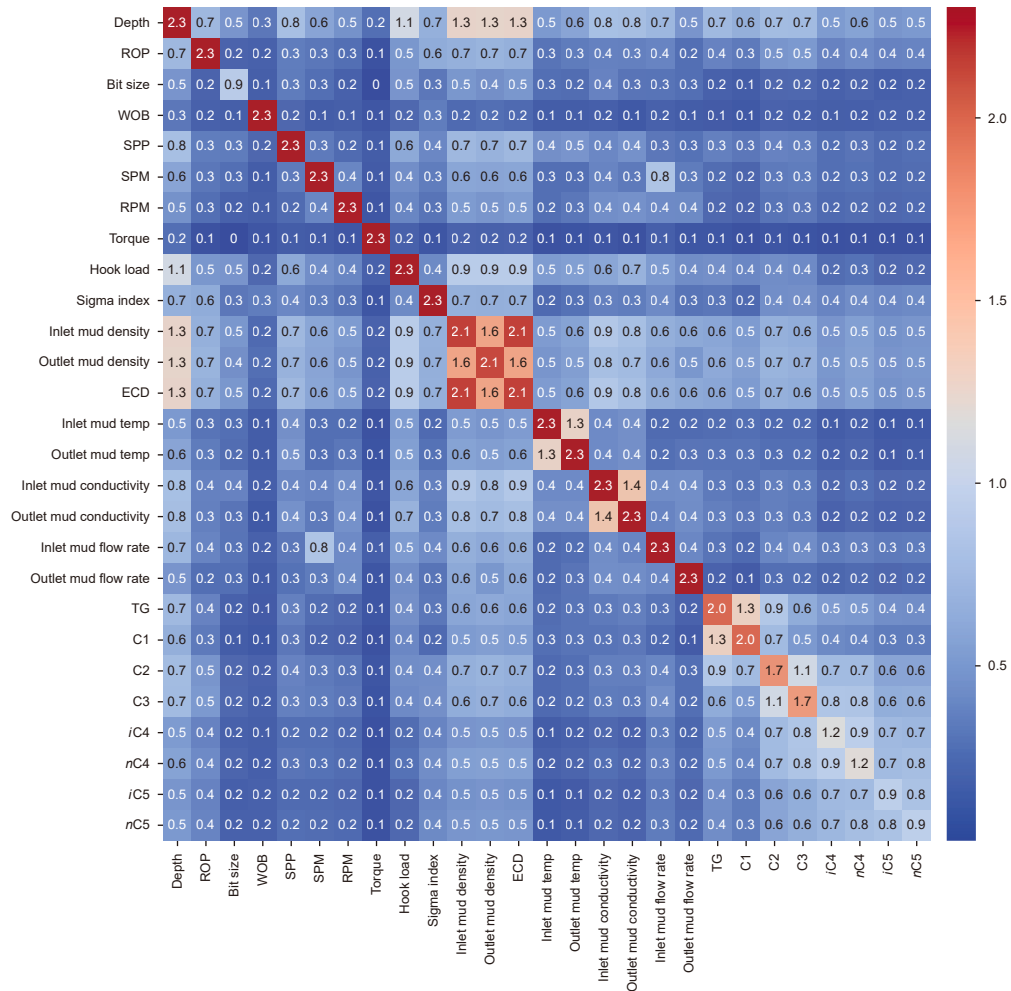


Fig. 4. Mutual information heatmap of drilling parameters.

- (1) Structural parameters: such as the drill-bit diameter, which is determined by the bit type that reflects the wellbore size and affects the upper limit of ROP. This parameter remains constant during a single bit cycle and, although it does not change in real time, serves as an important initial input for modeling.
- (2) Formation drillability indicators: primarily represented by the sigma index, which characterizes the ease of formation breakage. The sigma index is derived from downhole mechanical conditions and can be calculated in real time using measurement-while-drilling measurements and physical models, providing a timely reflection of the geological control on the ROP.
- (3) Mechanical operation parameters: including well depth, WOB, standpipe pressure (SPP), pump strokes, rotary speed, torque, and hook load, which directly describe the drilling system under various operating conditions. These parameters are collected in real time via surface or wellhead sensors, with high update frequencies and dynamic variations, making them highly responsive to ROP changes, and are core features of real-time ROP-prediction models.
- (4) Drilling-fluid circulation parameters: covering inlet and outlet mud density, equivalent circulating density (ECD),

mud temperature, conductivity, and flow rate. These are continuously monitored by the drilling-fluid system, and provide high real-time fidelity. ECD is calculated from pressure and DEN measurements and may exhibit slight delays, while the flow rate, conductivity, and temperature can respond rapidly, reflecting downhole cleaning efficiency and wellbore stability, thereby indirectly influencing the ROP.

- (5) Gaseous components: including total hydrocarbons, methane, ethane, propane, isobutane, *n*-butane, isopentane, and *n*-pentane. These gases are continuously monitored on-site using gas-measurement devices to track hydrocarbon release from the drilling fluid. Although a delay of several seconds to minutes exists, these measurements can assist in identifying potential geological interfaces or anomalous layers, enhancing the geological responsiveness of the drilling process.

These parameters not only cover the core features of the mechanical and fluid systems but also integrate formation responses and geological backgrounds, and can be monitored and acquired in real time. They provide high-dimensional, time-sensitive inputs for machine-learning-based ROP modeling and prediction. Finally,

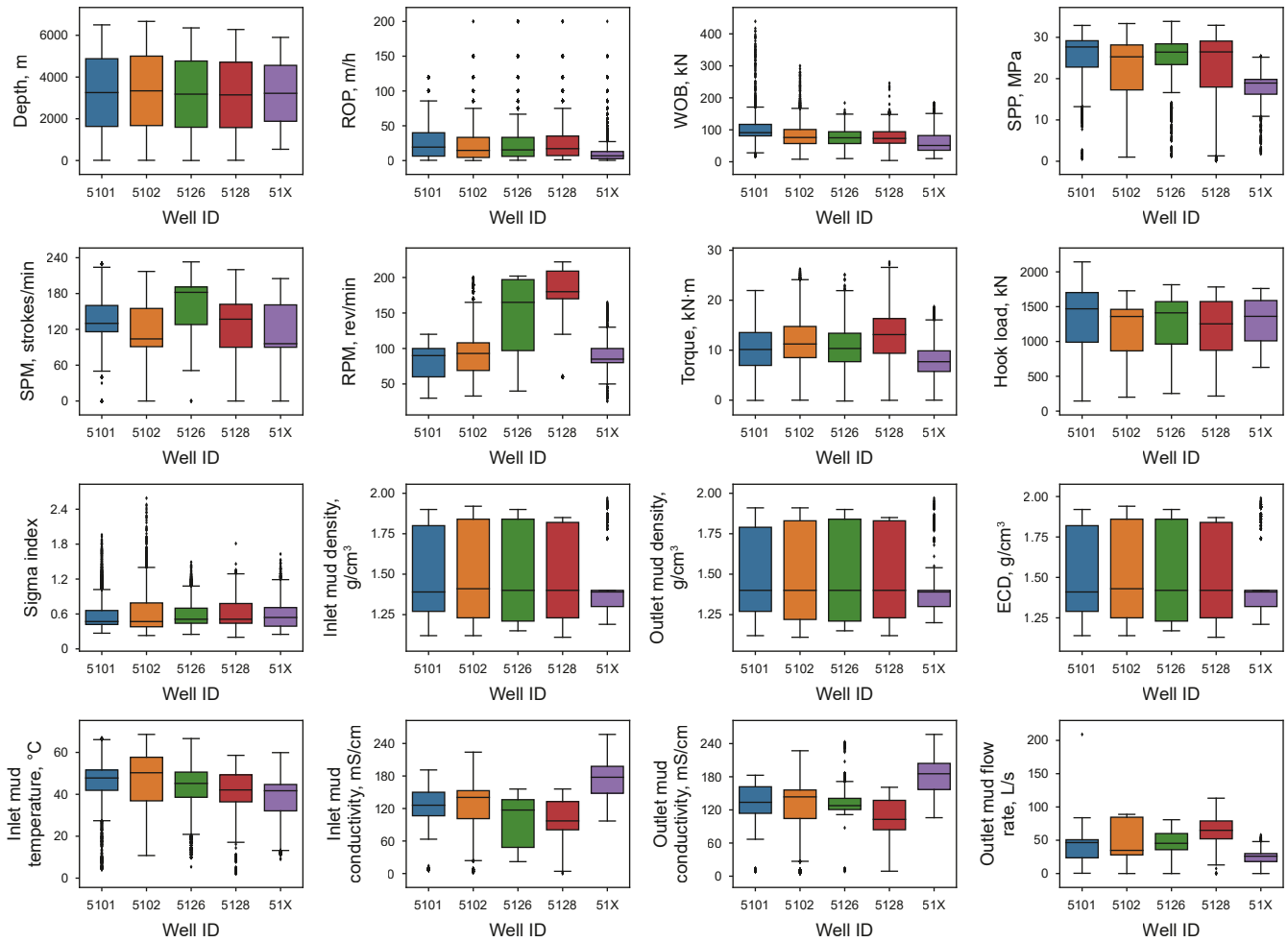


Fig. 5. Boxplots of statistical information for 16 input parameters.

the statistical plots of some input parameters (Fig. 5) and the well log profiles (Fig. 6) are presented to visually illustrate the parameter distributions and characteristics of the well sections.

2.2.3. Selection of optimizable drilling parameters

Based on the analyses presented in the preceding sections, the downhole WOB, rotary speed, and inlet mud flow rate have been identified as critical drilling parameters with real-time controllability, and are therefore considered suitable optimization variables. Although several other parameters exhibit relatively high importance scores and are theoretically adjustable, they have been excluded from real-time optimization for the following reasons.

- (1) Mud resistivity: This parameter primarily reflects the electrochemical properties of the drilling fluid. Its influence on the ROP is generally indirect and delayed, as it does not act directly on the rock-breaking process at the bit. Instead, its effects tend to manifest gradually following changes in the downhole environment. Moreover, mud resistivity is difficult to regulate accurately over short time scales, rendering it unsuitable as a real-time optimization variable.
- (2) Sigma index: The sigma index is essentially a derived indicator used to characterize drilling state features. Its value is jointly determined by multiple fundamental parameters, rather than an independently controllable variable. Directly incorporating such composite indices into the optimization process can introduce strong correlations among

optimization variables, thereby reducing the physical interpretability of the Bayesian optimization framework and impairing its convergence stability.

- (3) Torque: Torque is influenced by a combination of factors, including formation heterogeneity, bit wear, well trajectory, and borehole friction. Variations in torque primarily reflect the system-level response of the drilling process to formation and operational conditions, rather than an independently controllable input parameter. In real-time operations, torque is typically adjusted indirectly through changes in the WOB or rotary speed. Therefore, treating it as a direct optimization variable may introduce strong coupling effects and potential conflicts among optimization objectives.
- (4) SPP: SPP represents the overall hydraulic state of the drilling circulation system and is simultaneously affected by the flow rate, drilling-fluid rheology, well depth, and borehole conditions. As such, it is difficult to achieve targeted and independent control of the ROP through direct manipulation of this parameter. Including SPP in real-time optimization may introduce redundant coupling relationships and compromise optimization robustness.

2.3. Data cleaning

Drilling-parameter data are relatively complete, with no significant missing values, and measurement units are standardized, requiring no additional conversion. Common missing-value and outlier-handling

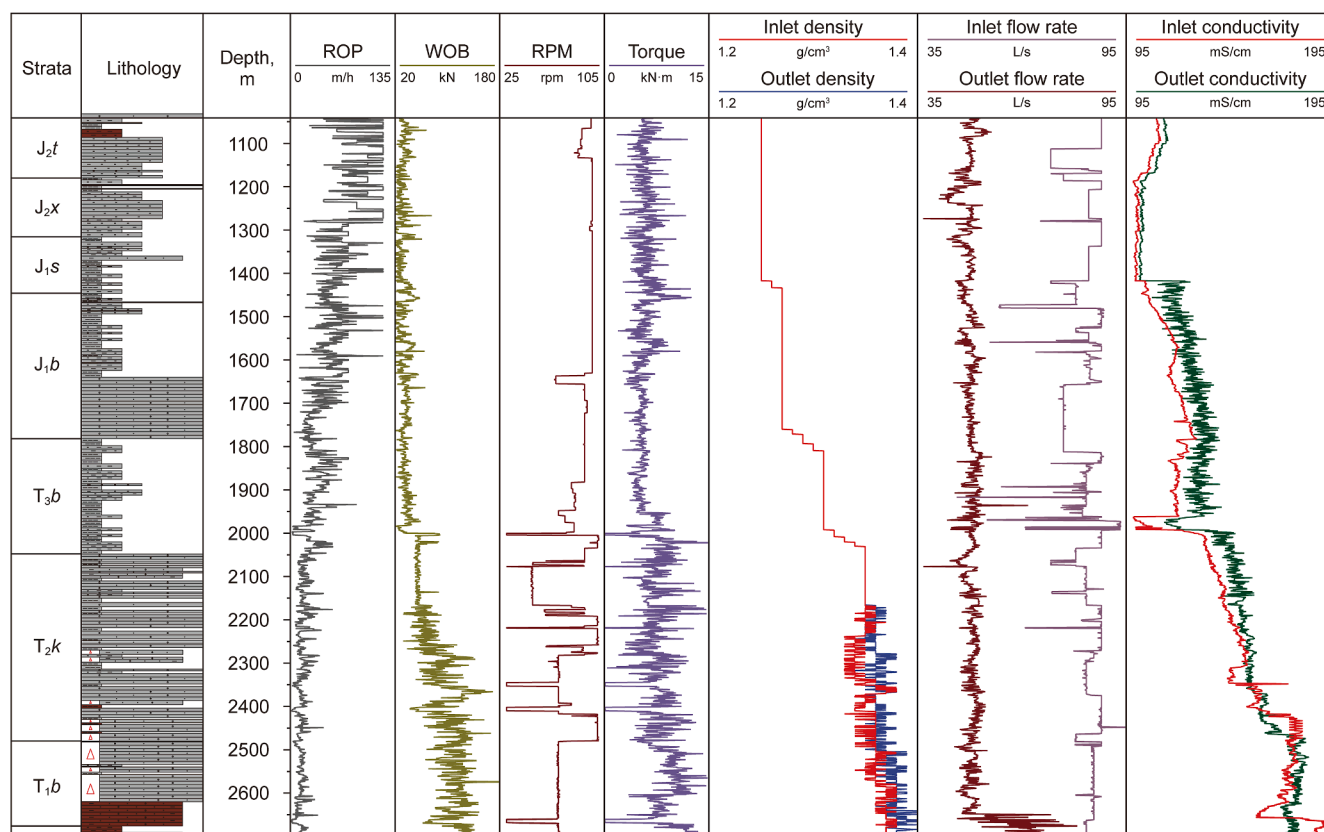


Fig. 6. Well-logging profiles of partial input parameters for several intervals, from Well MaHW5101.

methods such as mean/median imputation, time-series interpolation, sliding window, and interquartile-range method have been considered to evaluate the impact of data denoising on the model performance. Results show that removing outliers can reduce the training error but may weaken model generalization under extreme conditions, potentially inducing overfitting (Fig. 7). Some “outlier” data may actually reflect extreme drilling conditions. Removing them can shift the data distribution and reduce model adaptability in edge cases. Additionally, outlier detection may misclassify legitimate but highly fluctuating data, affecting the model performance.

To further quantify this effect, CatBoost is used for modeling. This algorithm has strong nonlinear modeling capability, built-in handling of missing values, and regularization strategies, making it suitable for prediction tasks under complex conditions. An early stopping mechanism is applied during training, which terminates the process if the validation performance did not improve over several rounds. Notably, after applying denoising, the training error decreases faster and validation performance converges earlier, triggering early stopping and resulting in fewer training iterations than that in the case without denoising. This indicates that the model tends to fit the smoothed features of the training data more easily, leading to overfitting. In summary, all original data are retained during modeling to ensure data completeness and representativeness.

2.4. Dataset limitations

The offline dataset used in this study is relatively limited in size, and comprises data from only five wells, primarily collected from the Fengcheng Formation in the Mahu Sag. These data reflect the typical alkaline lacustrine depositional environment and multi-stage tectonic evolution of the study area, which may constrain the

model's generalization capability when applied to other basins or formations with different geological settings. During data acquisition, potential recording errors or measurement uncertainties may exist, and discrepancies in measurement methods across wells can further affect the model's ability to capture global trends and statistical characteristics. In addition, drilling practices and equipment may evolve across wells or over time, including changes to bit types, drilling tools, and operational strategies; however, the dataset does not encompass all such variations, which may limit the model's adaptability to newly introduced tools or drilling practices.

For online prediction, recent sampling points have been used to train a short-term residual correction model. If the training window is too short, local trends may not be adequately captured, whereas an excessively long window may lead to the accumulation of measurement errors, thereby reducing the model's responsiveness to local variations. Consequently, the choice of window length plays a critical role in balancing model stability and sensitivity. Moreover, real-time drilling data are often affected by high-frequency noise, transmission delays, or short-term data loss, which can adversely impact the accuracy and robustness of online residual correction. Therefore, the limitations related to data volume, data quality, and operational-condition coverage should be fully considered when interpreting the model's prediction results and its applicability.

3. Theory

This section introduces the theoretical foundations and modeling methods involved. It first reviews tree-based model series, attention-based Transformer series, sequence models, and

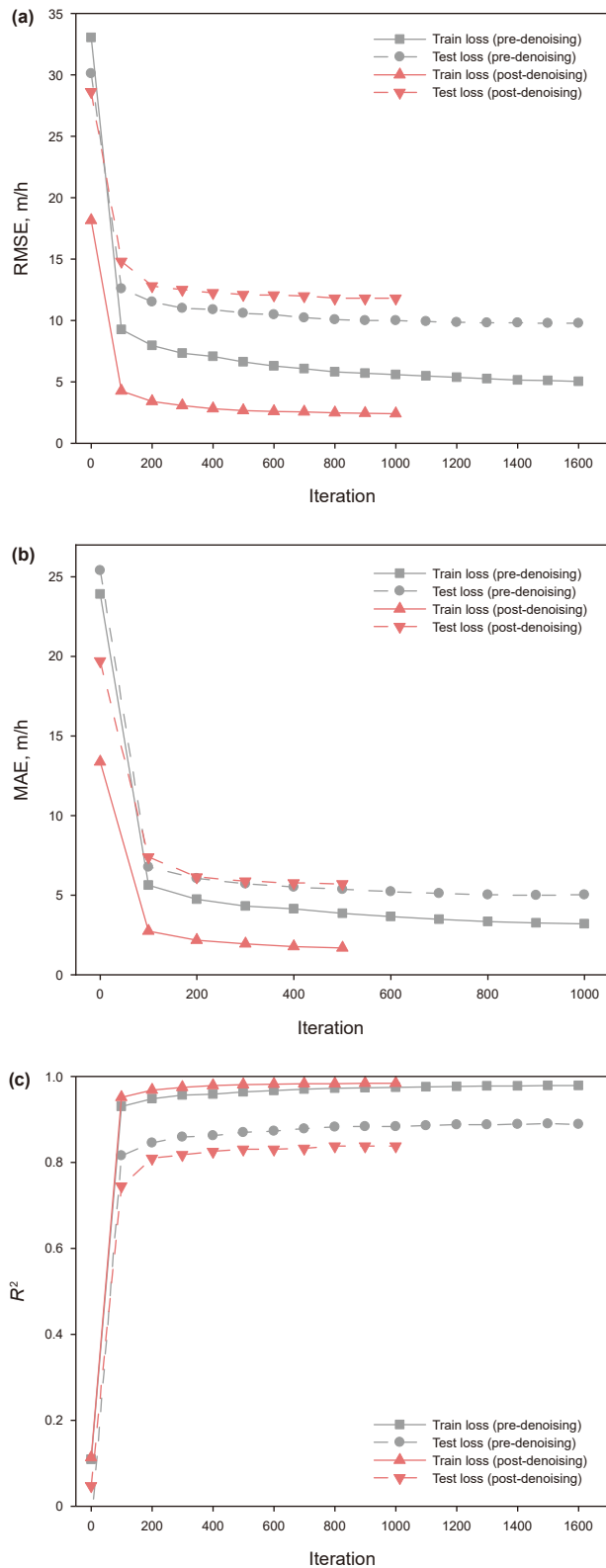


Fig. 7. Training processes of the model before and after denoising: (a) with RMSE minimization as the objective; (b) with MAE minimization as the objective; (c) with R^2 maximization as the objective.

convolutional networks, followed by the hybrid model framework proposed herein. Residual definitions and their advantages are then discussed, alongside Bayesian approaches and model-

evaluation strategies, providing a theoretical basis for subsequent experiments.

3.1. Tree-based models

Mechanical ROP is influenced by the coupling of drilling parameters and formation characteristics, exhibiting significant nonlinearity and uncertainty. This renders the traditional empirical and linear models insufficient for providing stable and reliable predictions under complex geological conditions. Tree-based models, which partition the feature space for nonlinear modeling, can automatically capture high-order interactions among input parameters. They are robust to small-scale datasets and missing values and are relatively interpretable, as the contributions of different drilling parameters to the ROP can be assessed through feature-importance analysis. Among tree-based models, GBT algorithms, including XGBoost, LightGBM, and CatBoost, are widely used for predicting ROP from drilling parameters and formation features. XGBoost (Chen and Guestrin, 2016) introduces regularization terms in the objective function to control model complexity, supports parallel computation and sparse data processing, and demonstrates strong robustness and interpretability for numerical data scenarios. LightGBM (Ke et al., 2017) employs a histogram-based leaf-wise growth strategy, significantly improving the training efficiency, while maintaining prediction accuracy, making it suitable for engineering applications with high computational demands. CatBoost (Ostroumova et al., 2017) effectively reduces target leakage through an ordered boosting mechanism, and provides native support for categorical features, achieving good stability and generalization on small- to medium-scale datasets with numerous discrete features. Nevertheless, tree-based models have limited ability to capture complex temporal dependencies and respond to rapidly changing drilling conditions, and are therefore, not suitable directly for modeling time-dependent ROP features.

3.2. Attention-based models

Drilling parameters not only exhibit complex nonlinear coupling relationships, but their interaction effects also evolve dynamically with operating conditions, leading to pronounced high-order interactions and global dependency characteristics. Such characteristics render models based on fixed weights or local-dependency assumptions inadequate for capturing the full interrelationships among parameters. To address this, the Transformer architecture (Vaswani et al., 2017) is employed, which leverages an adaptive weighting mechanism based on feature correlations to model the dependencies among input features, thereby dynamically adjusting the relative contributions of different drilling parameters to ROP prediction under various drilling states.

In ROP prediction, the input feature sequence is first linearly mapped to generate Query, Key, and Value representations. Attention weights are computed from similarities between each feature's Query and other features' Keys, and the corresponding Values are aggregated according to Eq. (7), thereby capturing global dependencies among drilling parameters. Here, Query represents the feature pattern the model focuses on under the current sample drilling state, while the attention information represents the matchable features provided by each input parameter under different states. The degree of matching between the two determines the relative importance of each parameter for that sample. For example, for samples with high WOB and stable drilling, Query attends to features such as lithology and rotational speed (RPM); for samples under high vibration states, Query

attends to features such as torque and rotational fluctuations. This attention mechanism is learned through multisample training, and the specific attention results are dynamically generated for individual samples rather than manually defined.

To enhance the model's ability to capture complex feature interactions, a multihead attention structure is introduced to learn dependencies across different feature subspaces in parallel, with the parallel computation and fusion processes presented in Eqs. (8) and (9). Subsequently, residual connections and normalization operations are applied to stabilize the training of deep networks, as shown in Eq. (10), followed by position-wise nonlinear transformations to further improve the modeling of higher-order nonlinear relationships, as expressed in Eq. (11).

$$\mathbf{Z} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \quad (7)$$

$$\mathbf{Z} = [\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \dots, \mathbf{Z}^{(h)}] \mathbf{W}_O \quad (8)$$

$$\mathbf{Z}^{(i)} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{W}_Q^{(i)}(\mathbf{K}\mathbf{W}_K^{(i)})^T}{\sqrt{d_k}}\right)\mathbf{V}\mathbf{W}_V^{(i)} \quad (9)$$

$$\mathbf{X}' = N(\mathbf{X} + F(\mathbf{X})) \quad (10)$$

$$F(\mathbf{x}) = \max(0, \mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2 \quad (11)$$

In Eq. (7), $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ represent the three feature representations obtained from the input feature sequence through linear mappings, d_k denotes the dimensionality of the feature mapping space used to scale the vector dot product results, and $\mathbf{Z}$ indicates the output feature matrix obtained by weighting based on feature similarity. In Eq. (8), $\mathbf{Z}^{(i)}$ represents the weighted mapping result in the i -th feature subspace, h denotes the number of parallel feature subspaces, $[\cdot]$ indicates concatenation along the feature dimension, and $\mathbf{W}_O$ is the linear transformation matrix used to fuse the mapping results from each subspace. In Eq. (9), $\mathbf{W}_Q^{(i)}$, $\mathbf{W}_K^{(i)}$, and $\mathbf{W}_V^{(i)}$ represent the linear mapping matrices in the i -th feature subspace, while $\mathbf{Z}^{(i)}$ indicates the weighted feature representations calculated based on feature similarity within that subspace. In Eq. (10), $\mathbf{X}$ denotes the input feature of the sublayer, $F(\cdot)$ is the mapping function applied to the sublayer input, $N(\cdot)$ is the normalization operator used to normalize the feature distribution after residual addition, and $\mathbf{X}'$ is the normalized output feature. In Eq. (11), $\mathbf{x}$ represents the input feature vector, $\mathbf{W}_1$, $\mathbf{W}_2$, and $\mathbf{b}_1$, $\mathbf{b}_2$ are the weight matrices and bias terms in the nonlinear mapping, and $\max(0, \cdot)$ is the element-wise nonlinear activation function used to enhance the model's capacity to represent complex nonlinear feature relationships.

3.2.1. Tabular Transformer model

Both numerical and categorical features play a crucial role in predicting the mechanical ROP. To effectively capture the high-order interactions between numerical and categorical features, the Tabular Transformer (TabTransformer) model (Huang et al., 2020) is introduced. This model maps categorical features into low-dimensional embedding vectors, and employs a self-attention mechanism to model global dependencies among different features, thereby extracting highly discriminative high-order feature combinations. Subsequently, the embedding vectors are fused with normalized numerical features and fed into a multilayer perceptron (MLP) for final prediction, with the overall model architecture illustrated in Fig. 8.

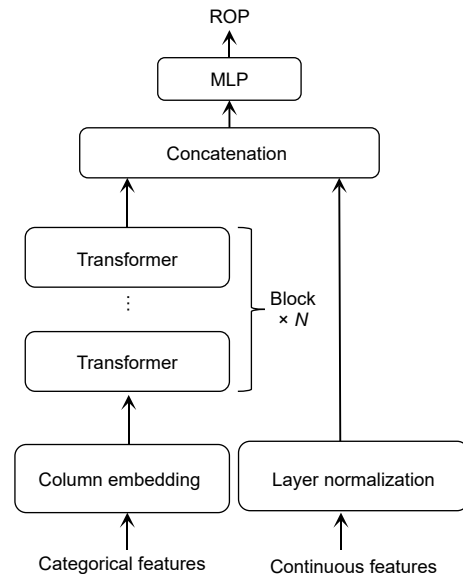


Fig. 8. TabTransformer architecture.

In the model input, categorical features refer to variables with discrete and limited values that lack inherent numerical magnitude or physical order, such as the formation type or drill bit type. Numerical features are continuous variables with physical units and magnitudes, such as WOB and RPM. Although the model is effective in capturing categorical features and their interactions, its capacity for directly modeling numerical features is relatively limited.

3.2.2. Feature Token Transformer model

Variations in the ROP are primarily driven by numerical features, while categorical features play an auxiliary modulating role under specific drilling conditions; therefore, model design should place greater emphasis on the effective modeling of continuous features. Accordingly, the Feature Token Transformer (FTTransformer, Gorishniy et al., 2021) is adopted, which emphasizes efficient modeling of numerical features in tabular data and typically outperforms TabTransformer in tasks dominated by numerical inputs. FTTransformer applies linear projections to map continuous numerical features and discrete categorical features into feature vectors of a unified dimensionality, after which the fused representations are fed into a self-attention module. This mechanism enables the model to capture global dependencies among input features in a high-dimensional feature space and automatically extract discriminative interaction features, thereby enhancing its modeling capability for ROP prediction (Fig. 9). It should be noted that the effectiveness of categorical-feature embeddings remains dependent on the sample size and may become unstable when category samples are scarce. In addition, FTTransformer has limited capability in modeling strict temporal dependencies, and its ability to capture both long- and short-term patterns under dynamic drilling conditions remains constrained.

3.2.3. Time-series transformer model

The drilling ROP varies dynamically with drilling conditions, exhibiting pronounced temporal dependencies and condition-dependent variability; therefore, accurate modeling of temporal contextual information is crucial. The Time-series Transformer (TSTransformer, Zerveas et al., 2021) model processes multiple time-series samples simultaneously and introduces a time-step masking mechanism, which encourages the model to rely on the

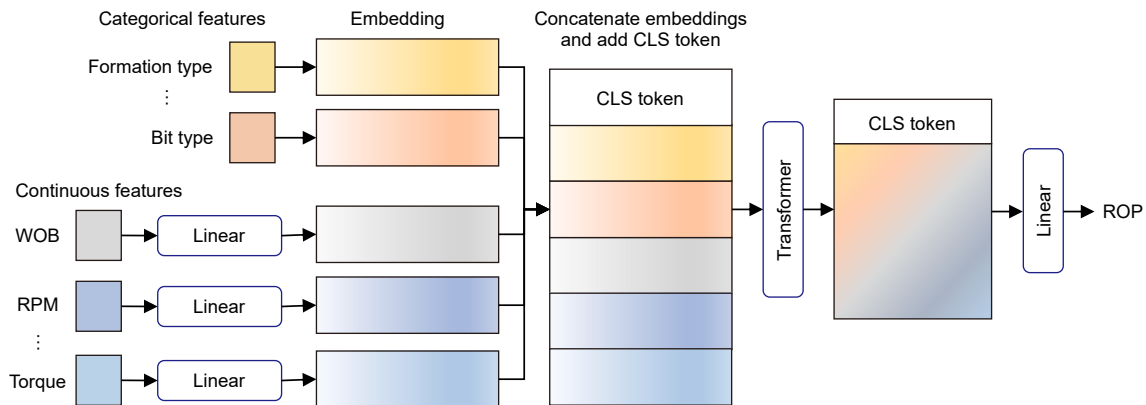


Fig. 9. FITransformer architecture (CLS: Class Token, a special token that aggregates all input features into a single vector for ROP prediction).

joint information from neighboring time steps when reconstructing the masked values, thereby effectively learning sequential contextual relationships. The masked sequences are then fed into a Transformer encoder, where complex dependencies between input variables and the temporal dimension are modeled in a high-dimensional space, facilitating unified representation of both local and global temporal patterns. It should be noted that under small-sample conditions, the generalization capability of TSTransformer may be inferior to that of FITransformer, and its ability to model categorical features remains relatively limited.

3.3. Sequence model

Given that the ROP and related drilling parameters exhibit pronounced temporal dependencies, a classical long short-term memory (LSTM) network is introduced for comparison with other time-series modeling approaches. The LSTM captures both long-term dependencies and short-term dynamic variations through its memory cells and gating mechanisms, making it well suited for modeling continuous time-series data.

To capture the temporal evolution of drilling parameters, sequential modeling plays a key role in ROP prediction. As shown in Fig. 10, the LSTM network alleviates the vanishing-gradient problem in long sequences by introducing a memory cell, where the cell C_t stores long-term information and hidden state H_{t-1} carries short-term memory. At each time step, the current input X_t and previous hidden state H_{t-1} are processed through the forget, input, and output gates to update the cell and hidden states, with the Sigmoid activation σ compressing the gate values to [0,1]. The updated hidden state H_t and cell C_t jointly determine the current

output, which is also passed to the next time step, facilitating continuous contextual learning and improving the ROP prediction accuracy. Compared to Transformers, LSTM exhibits limited parallel-computing capability and struggles to simultaneously capture global dependencies and high-order interaction features.

3.4. Convolutional neural network model

ROP data can be represented as a multivariate time series or feature matrix, encompassing both local patterns (e.g., short-term fluctuations in drilling pressure and instantaneous changes in drill-rotation speed) and global trends (e.g., overall increase in penetration rate and gradual changes in drilling-fluid concentration). To effectively extract these features, convolutional neural networks (CNNs) have been introduced for tabular data modeling tasks. CNNs leverage local receptive fields and hierarchical feature-learning mechanisms to sensitively capture local patterns.

To improve the training stability of deep networks, a ResNet structure (He et al., 2016) is incorporated on top of the CNN backbone (see Fig. 11), using residual connections to alleviate the vanishing-gradient problem in deep networks. Residual learning introduces the input features directly into subsequent layers via identity mappings, with the forward-propagation process described in Eq. (12). The residual mapping is composed of convolutional operations, batch normalization, and nonlinear activation functions (Eq. (13)). This architecture maintains network depth, while allowing stable training and extraction of multiscale local features. Compared with LSTM and Transformer, ResNet is better suited for capturing local patterns, but exhibits limited capability for modeling global temporal dependencies.

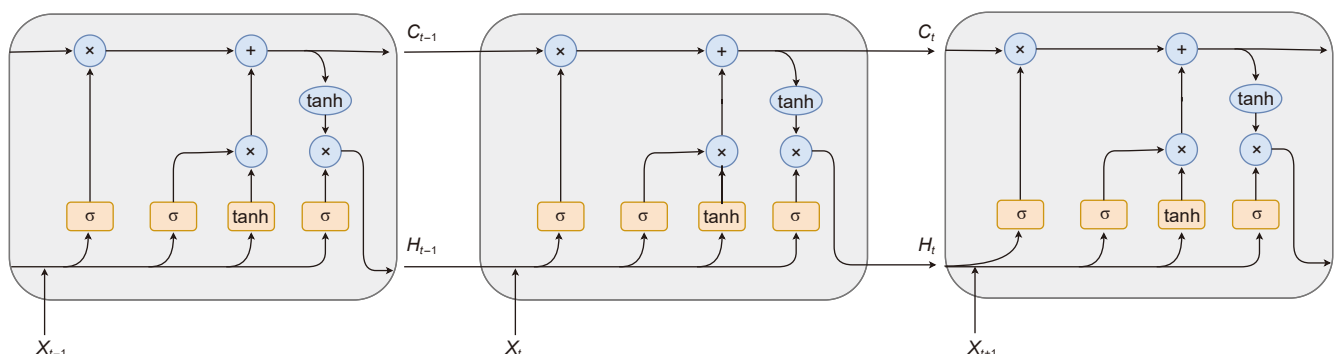


Fig. 10. LSTM architecture.

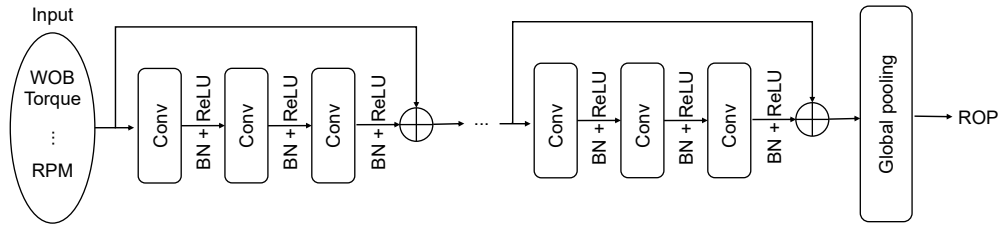


Fig. 11. ResNet architecture with residual blocks.

$$\mathbf{z}_{l+1} = \mathbf{z}_l + F(\mathbf{z}_l, \mathbf{W}_l) \quad (12)$$

$$F(\mathbf{z}_l, \mathbf{W}_l) = \sigma(\text{BN}(\mathbf{W}_l * \mathbf{z}_l)) \quad (13)$$

In the equations, $\mathbf{z}_l$ denotes the input feature vector of the l -th residual block, while $\mathbf{z}_{l+1}$ represents the output feature vector of the corresponding residual block. The dimensions of both vectors are kept consistent to enable identity mapping. $F(\cdot)$ is the residual mapping function, which learns the nonlinear difference between the input and output features. $\mathbf{W}_l$ denotes the learnable weights of the convolutional layer in the l -th residual block. The symbol $*$ indicates the convolution operation. BN refers to batch normalization, which accelerates network convergence and improves training stability. Finally, $\sigma(\cdot)$ represents a nonlinear activation function (e.g., ReLU) that enhances the model's nonlinear representation capability.

3.5. Proposed hybrid model

To address the insufficient transferability of artificial intelligence models in cross-regional geological applications, a cross-regional adaptive adjustment framework is developed (Fig. 12), where global stable patterns and local dynamic disturbances are explicitly decoupled, facilitating accurate ROP prediction. This framework is built upon multi-regional historical drilling data, which are used for training both the offline baseline model and fusion module. Specifically, the offline model adopts deep-learning architectures with relatively high structural complexity and training cost to learn the global stable patterns of drilling processes across different geological regions and provide baseline ROP predictions.

In the online stage, a short-term temporal data window is introduced as the main information source for model adaptation.

The supervision label is defined as the residual between the ROP predicted by the offline model and actual measured ROP, which is used to train a lightweight online model (e.g., tree-based models), allowing it to learn and capture local dynamic disturbances during drilling. During real-time prediction, the drilling-related parameters at the current time step are first fed into the offline model to generate baseline ROP predictions. These predictions can optionally be used as high-level features for the online model or joint inputs along with the original real-time parameters to estimate the residual corrections, thereby reducing the learning difficulty of the online model and improving prediction stability. If offline prediction is introduced as an additional feature input, the feature dimensionality of the short-term data increases by m , compared with that of the original setting. Finally, the fusion module nonlinearly fuses the residual compensation with the baseline prediction and adaptively modulates the residual correction magnitude by exploiting the characteristics of different ensemble models and high-level features of the operational state, thereby yielding ROP predictions that adapt to the current drilling conditions.

In the implementation and validation processes, all measurement data from four wells (MaHW5101, MaHW5102, MaX5126, and MaX5128) are selected as the training set, and a simulated field-condition validation is conducted on the Ma51X well. Fig. 13 presents an online prediction example of the proposed method using real drilling data within the depth interval of 1000–2000 m. In this case, the offline component adopts the optimal model combination of ResNet and FTTransformer determined through comprehensive evaluation. The online component employs a lightweight and computationally efficient LightGBM model, and rapid adaptive updating is achieved by constructing a short-term temporal window. The short-term data consist of the most recent window measurements, prior to the current time t (with a

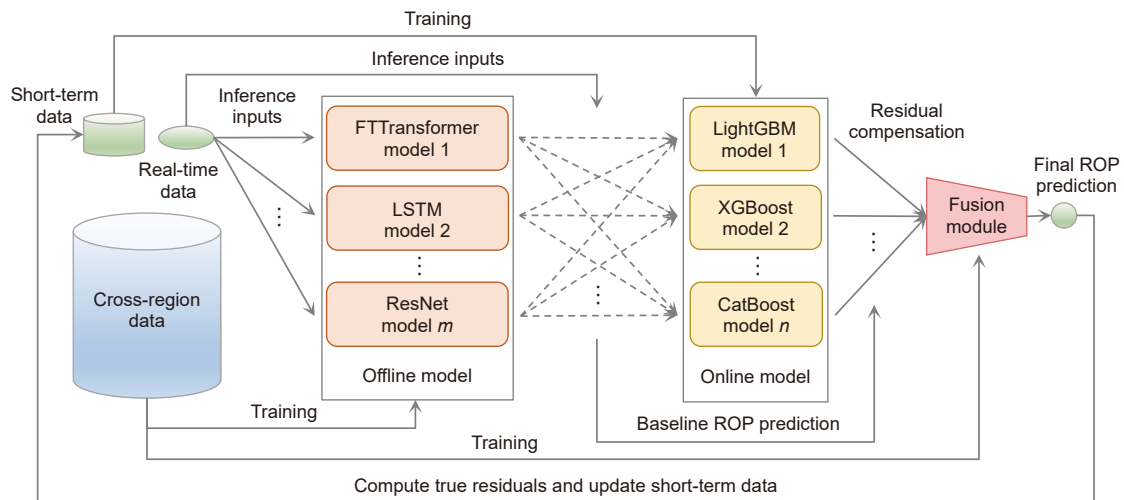


Fig. 12. Data flow of the hybrid online-offline ROP-prediction framework.

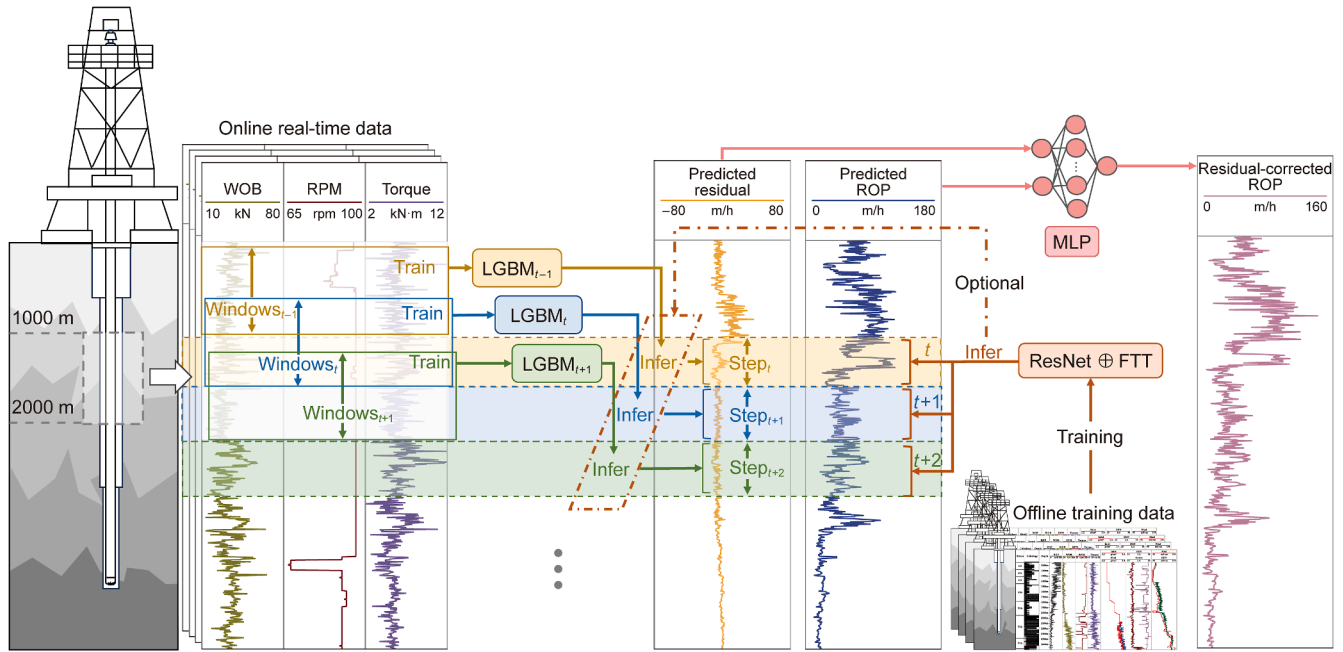


Fig. 13. Hybrid online-offline ROP prediction framework example (FTT denotes FTTransformer; LGBM denotes LightGBM).

default value of 200). The online model is trained in real time based on these short-term data to predict the ROP deviation caused by short-term changes in drilling states, from time t to $t + \text{step}$ (with a default value of 1), under different formation constraints. The final fusion module adopts an MLP to further enhance the robustness and field adaptability of ROP prediction.

This method effectively integrates deep feature extraction from historical data with real-time response to on-site variations, significantly improving the model's generalization capability and prediction accuracy in cross-regional applications. The integration strategy of the online and offline models not only addresses the challenge of model transferability, but also provides a new technical pathway for real-time drilling optimization under complex geological conditions. Section 4 will present a detailed discussion of the criteria for model selection and performance comparison.

3.6. Residual definition and online-offline fusion mechanism

The prediction objective of the online model is explicitly defined as the dynamic estimation of the prediction error (residual) of the offline model, rather than the direct prediction of the formation petrophysical parameters. This modeling strategy decouples relatively stable drilling patterns under cross-regional conditions from local operational disturbances, thereby ensuring overall prediction stability while effectively reducing the dependence of online modeling on information that cannot be obtained in real time.

At time step t , the offline model produces a baseline prediction of the ROP, based on the real-time input-feature vector, as expressed in Eq. (14). The offline model is trained using cross-regional historical drilling data to learn relatively stable drilling patterns under different geological conditions. On this basis, the deviation between the true ROP and offline prediction is defined as the residual, as shown in Eq. (15). This residual jointly reflects the effects of short-term factors such as abrupt changes in the local formation, drilling-parameter adjustments, tool-state variations, and on-site stochastic disturbances, and essentially represents the

systematic prediction bias of the offline model under the current operating conditions.

During the online stage, the ROP is not modeled directly; instead, a short-term time-series data window composed of the most recent w sampling points is used as the input to perform real-time residual prediction, as described in Eqs. (16) and (17). To avoid the excessive correction or noise amplification caused by simple linear superposition, an MLP is further introduced to nonlinearly fuse the offline prediction and online residual prediction, yielding the final ROP prediction, as given in Eq. (18). From an overall modeling perspective, this online-offline collaborative residual-correction mechanism can be regarded as a decomposition and local approximation of the original high-dimensional nonlinear mapping relationship, with its mathematical formulation expressed in Eqs. (19) and (20), where the offset term is approximated by the online model within the short-term time-series window, thereby achieving dynamic correction of local operating-condition variations, while maintaining global prediction stability.

$$\hat{y}_t^{\text{off}} = f_{\text{off}}(\mathbf{x}_t) \quad (14)$$

$$r_t = y_t - \hat{y}_t^{\text{off}} \quad (15)$$

$$\mathbf{X}_t^{\text{win}} = \{\mathbf{x}_{t-w+1}, \dots, \mathbf{x}_t\} \quad (16)$$

$$\hat{r}_t = g_{\text{on}}(\mathbf{X}_t^{\text{win}}) \quad (17)$$

$$\hat{y}_t = f_{\text{MLP}}(\hat{y}_t^{\text{off}}, \hat{r}_t) \quad (18)$$

$$y_t = f_{\text{off}}(\mathbf{x}_t) + \Delta(\mathbf{x}_t) \quad (19)$$

$$\Delta(\mathbf{x}_t) \approx g_{\text{on}}(\mathbf{X}_t^{\text{win}}) = \hat{r}_t \quad (20)$$

Here, $\mathbf{x}_t$ denotes the real-time drilling input feature vector at time step t , y_t represents the corresponding true ROP, and $\hat{y}_t^{\text{off}}$ is

the baseline prediction generated by the offline model $f_{\text{off}}(\cdot)$; r_t and $\hat{r}_t$ denote the true residual and its online predicted value, respectively, $\mathbf{X}_t^{\text{win}}$ represents a short-term temporal window composed of the most recent w sampling points, $g_{\text{on}}(\cdot)$ denotes the online residual prediction model, and $f_{\text{MLP}}(\cdot)$ is an MLP used to nonlinearly fuse the offline prediction with the residual compensation term; while $\Delta(\mathbf{x}_t)$ represents the prediction deviation induced by local operating-condition disturbances or data-distribution drift.

3.6.1. Online residual modeling guided by offline predictions

The core task of the online model is to perform short-term dynamic modeling of the prediction errors generated by the offline model during real-time drilling operations. Specifically, the online model is continuously trained or updated using drilling-related parameters, collected within a recent time window, and their corresponding residual labels. In the inference stage, the residual is estimated based on the drilling-related parameters at the current time step t , as formulated in Eq. (21). This scheme has a relatively simple structure and does not explicitly rely on the output of the offline model; however, under conditions of limited short-term samples or rapidly changing operating conditions, it exhibits a relatively strong dependence on the quality and quantity of online data.

To enhance the predictive stability of the online model under complex conditions, the offline model prediction at the current time step can be incorporated as an auxiliary input into the online residual-prediction process, as expressed in Eq. (22). The offline model, trained on cross-well and cross-field historical data, captures globally stable patterns of drilling processes and serves as a high-level semantic feature during the online stage, providing a stable reference for the current operation. From a modeling perspective, this strategy realizes a form of model-level information transfer, enabling the online model to avoid relearning the complete nonlinear mapping and instead, focus on locally and dynamically correcting the systematic prediction deviations generated by the offline model under the current operating conditions, thereby reducing the learning complexity and improving predictive stability and robustness.

$$\hat{r}_t = g_{\text{on}}(\mathbf{x}_t) \quad (21)$$

$$\hat{r}_t = g_{\text{on}}(\mathbf{x}_t, \hat{y}_t^{\text{off}}) \quad (22)$$

Here, $\hat{r}_t$ denotes the residual predicted by the online model at time step t , $\hat{y}_t^{\text{off}}$ denotes the prediction generated by the offline model at time step t .

3.6.2. Comparison of residual compensation and model update strategies

Model fine-tuning and incremental training typically require continuous updates within the original model-parameter space, and their predictive performance is highly dependent on the scale of the newly added data, update frequency, and appropriateness of parameter freezing or unfreezing strategies. In application scenarios involving significant changes across geological regions or operating conditions, such methods often struggle to effectively distinguish globally stable patterns learned from historical data from local perturbations arising under current operational conditions, which may lead to catastrophic forgetting or unstable model predictions. In contrast, the proposed approach keeps the offline model parameters fixed and introduces a lightweight online residual-correction mechanism only at the prediction-error level,

thereby avoiding interference with the overall structure and previously learned knowledge of the original model.

As the online model only performs local modeling of the offline prediction errors, its training sample size and model complexity are significantly lower than those required for full model retraining or incremental updates. Consequently, the model does not need to repeatedly revisit large-scale historical data; nor does it rely on complex training schedules or parameter synchronization processes, enabling rapid convergence and immediate effects within short-term temporal windows under computationally constrained and real-time field conditions. In cross-region or abrupt operational-change scenarios, this residual modeling strategy structurally decouples the globally stable mapping from local dynamic deviations, allowing the model to adapt to new operating conditions while maintaining the stability of the original predictive baseline, thereby demonstrating clear advantages over conventional model fine-tuning and incremental training strategies in terms of the generalization ability, predictive stability, and operational controllability.

3.6.3. Comparison of the residual-decoupling framework and temporal models

End-to-end temporal models such as LSTM typically perform holistic predictions of future mechanical ROPs directly, based on historical time series, and their modeling performance heavily depends on long-term continuous data with relatively regular time intervals, consistent distributions, and sufficient annotations. However, in actual drilling operations, on-site data are often affected by sensor malfunctions, communication delays, and frequent operational changes, resulting in missing values, high noise levels, and irregular sampling. These issues generally require additional interpolation, resampling, or masking, which in turn introduce new sources of uncertainty. Moreover, in scenarios involving abrupt changes across geological regions or operating conditions, the historical memory mechanism of LSTM is easily influenced by data-distribution shifts and the accumulation of irrelevant historical information, leading to a significant decline in the predictive performance.

To address these challenges, the proposed method explicitly decouples globally stable patterns from local dynamic perturbations at the model-structure level. Within this framework, even when short-term temporal features vary significantly or data exhibit missing values and irregular sampling, the offline predictions can still provide a stable reference for overall forecasting, while the online residual-correction mechanism can quickly respond to local perturbations, effectively suppressing severe oscillations in the predictions. Consequently, compared with traditional end-to-end temporal models such as LSTM, the proposed approach demonstrates superior stability, robustness, and practical applicability under cross-region operations, complex working conditions, and scenarios with limited data quality.

3.7. Role and contribution of the fusion module

In a multimodel collaborative prediction framework, the fusion module is responsible for integrating the outputs of different models into a unified prediction. This module achieves adaptive combination and error correction in the model output space through a learnable nonlinear mapping, thereby capturing the complex interactions among the predictions of different models. As each base model focuses on distinct feature aspects and statistical patterns of the drilling process during training, their outputs often exhibit significant differences in the numerical scale, error distribution, and response characteristics. Directly using fixed weights or heuristic averaging fails to fully exploit the complementary information among models and lacks adaptability

under changing operating conditions or data distributions. By minimizing the overall prediction error on the validation and test sets, the fusion module can automatically learn the relative importance of different model outputs under various conditions, facilitating dynamic adjustment of each model's contribution.

From the perspective of model design and engineering applications, this approach avoids the high computational cost and potential overfitting risk associated with jointly retuning the hyperparameters of all base models, instead concentrating the learning of “combinatorial optimality” within the fusion layer. Accordingly, the fusion module can be regarded as a model-level adaptive regulator, enhancing the overall predictive performance without compromising the generalization capability of individual submodels. This structured fusion strategy demonstrates strong stability and scalability in engineering applications, particularly in drilling prediction scenarios involving heterogeneous model sources and significant changes in data distribution under various operating conditions.

3.8. Bayesian inference algorithms

Bayesian optimization is a probabilistic black-box method for efficiently finding optima when the objective is nondifferentiable or expensive to evaluate. It builds a surrogate model, typically a Gaussian process, and uses an acquisition function to explore promising regions.

In ROP prediction, the trained model serves as a black box: given inputs such as WOB or RPM, it outputs the ROP but cannot directly provide optimal settings. Bayesian optimization approximates the ROP–parameter relationship, iteratively selects points in promising regions, and balances exploration and exploitation to find optimal drilling inputs. It is also used for automatic hyperparameter tuning of AI models.

3.9. Model evaluation strategies

To comprehensively evaluate the performances of the developed models in the ROP-prediction task, multiple commonly used regression metrics are employed to quantitatively assess the prediction results from different perspectives. Among them, the root mean square error (RMSE) is used to measure the overall deviation between predicted and observed values, with its formulation given in Eq. (23). Owing to the squared error term, the RMSE is more sensitive to large prediction errors and can effectively reflect the overall predictive accuracy across the entire dataset. Meanwhile, the coefficient of determination (R^2) is adopted to evaluate the model's capability to explain the variance of the observed data, as defined in Eq. (24), where the mean of the observed values is given in Eq. (25); R^2 ranges from 0 to 1, with higher values indicating better model fits. In addition, the mean absolute error (MAE) quantifies the model performance by computing the average magnitude of absolute prediction errors, with its calculation expressed in Eq. (26). Compared with RMSE, MAE is less sensitive to outliers and provides a more intuitive measure of the average prediction deviation. Through the combined analysis of these evaluation metrics, the predictive performance of different models in the ROP-prediction task can be more comprehensively characterized.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (23)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (24)$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad (25)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (26)$$

where y_i and $\hat{y}_i$ denote the true and predicted ROPs, respectively, of the i -th sample; $\bar{y}$ represents the mean of the observed ROP values, and n denotes the total number of samples; RMSE, R^2 , and MAE are employed to evaluate the overall model performance in the ROP prediction task from the perspectives of error magnitude, goodness of fit, and robustness to prediction errors, respectively.

4. Results and discussion

This section presents a systematic evaluation of the model performance, training efficiency, and real-time applicability. Specifically, we first assess the performance of offline models to determine the optimal ensemble configuration. We then analyze the effects of data scale and training time on model applicability and conduct a systematic exploration of the hyperparameters of the online residual model. Finally, Bayesian optimization is applied to allow real-time adjustment of the key drilling parameters, thereby examining the potential of the proposed framework for ROP improvement.

4.1. Offline model experiments

This section systematically evaluates the performances of various models and their ensemble schemes for the ROP-prediction task. The evaluation includes assessments of model generalization, hyperparameter optimization strategies, and effectiveness of multi-model ensembles, aiming to identify offline-trained models with high predictive accuracies as a foundation for subsequent online applications.

4.1.1. Model performance evaluation

The models were trained using all measurement points from four wells, with an additional well selected as an independent test set to evaluate the generalization performance of different models in the ROP-prediction task. Considering that some machine-learning libraries do not support R^2 as a validation metric and cannot track its dynamic variation during training, RMSE and MAE were consistently used as performance-evaluation metrics. Fig. 14 presents the comparison of RMSE and MAE for different models on the test set. To further illustrate the prediction accuracy, Fig. 15 shows the comparison between the predicted and actual ROP for eight models on a subset of the test set.

Owing to the weak sequential correlation between the well depth and ROP in the dataset, the temporal dependencies exhibit significant instabilities under varying operating conditions, making it difficult for LSTM models to learn generalizable temporal patterns. On the one hand, the ROP is influenced by multiple operational parameters and geological factors, resulting in rapid sequence variations without ideal smoothness, which limits the reference values of historical time steps for future predictions and

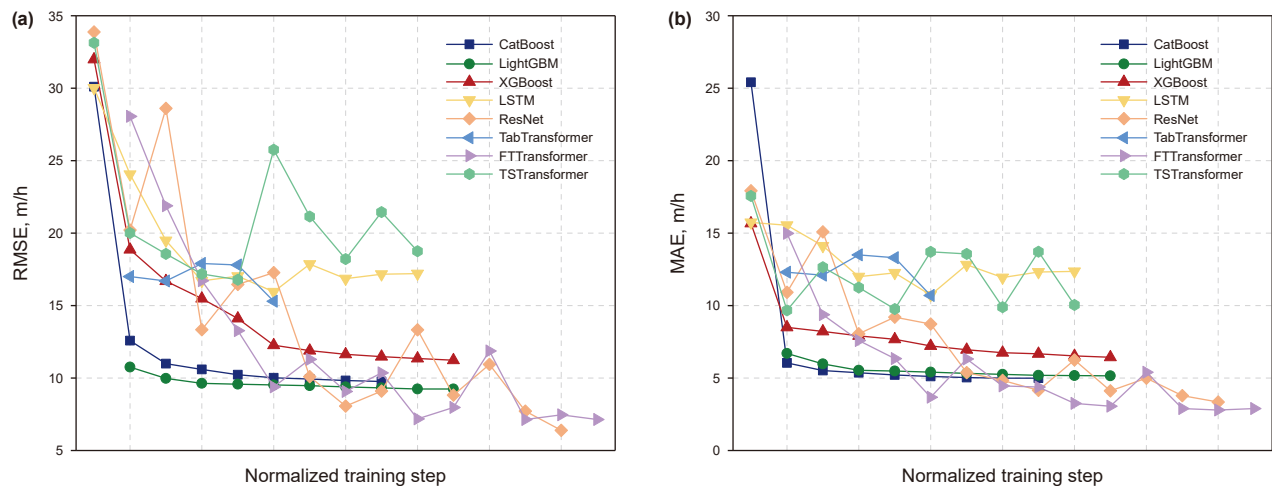


Fig. 14. Comparison of RMSE and MAE for different models in ROP prediction.

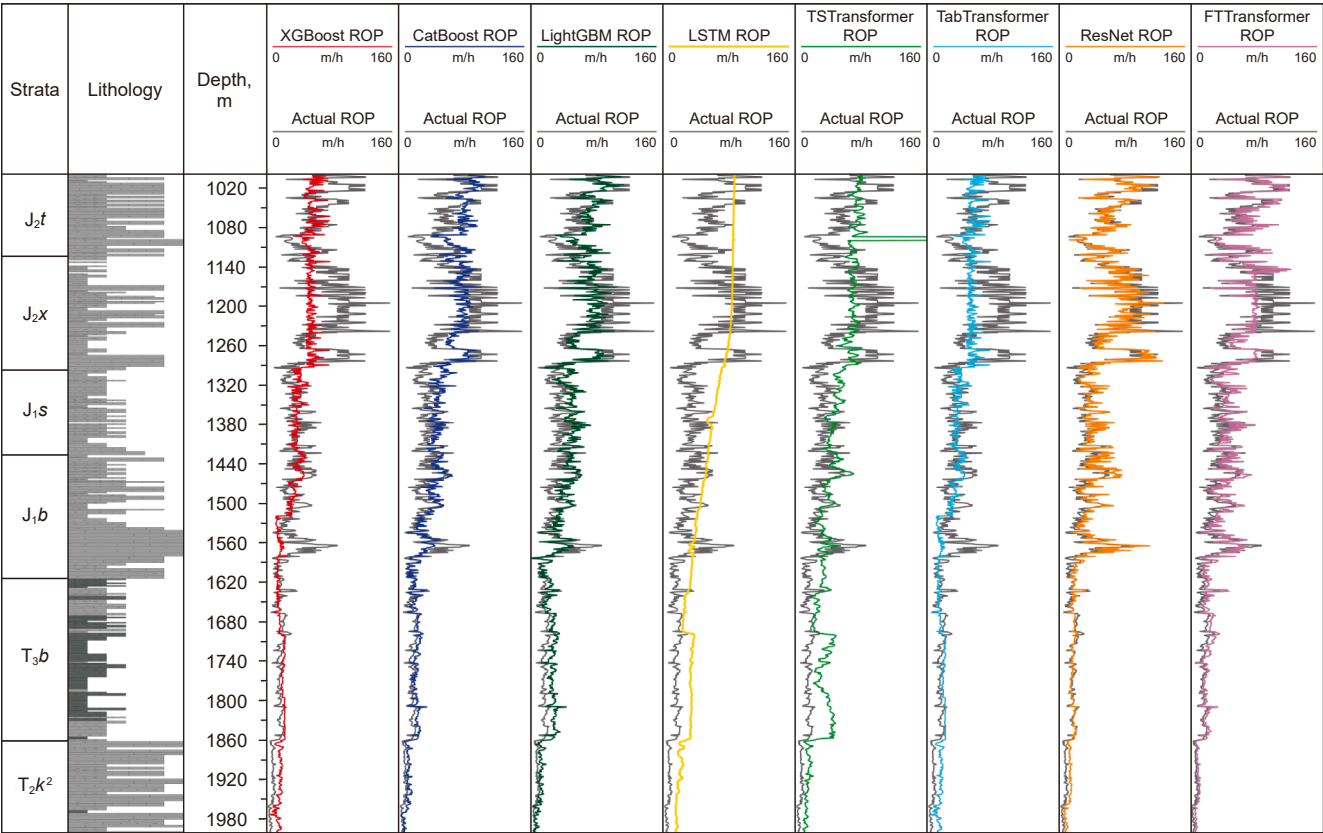


Fig. 15. Predicted vs. actual ROPs of eight models on a subset of the test set.

may lead to underfitting during LSTM training. On the other hand, the drilling data inevitably contain missing values, measurement errors, and extreme operating points, which, if not smoothed, can induce gradient oscillations and negatively affect the stable training of sequence models, thereby constraining the predictive performance of LSTM in this task.

For the TabTransformer model, as shown in Fig. 8, under the data settings used in this study, the limited sample size and scarcity of categorical features preclude explicit modeling of categorical variables. In this context, its self-attention-based feature

interaction module cannot fully leverage its advantages for modeling categorical features, rendering the model functionally close to that of an MLP based solely on continuous features, and thus failing to achieve notable improvements in the prediction accuracy and generalization.

The TSTransformer theoretically possesses the ability to simultaneously model global relationships among features and local temporal dependencies, making it suitable for time-series modeling tasks. However, this model generally assumes that input sequences are relatively continuous and regular, whereas the actual ROP data

are affected by multiple coupled factors, resulting in nonsmooth sequences with pronounced instantaneous fluctuations and poor local-pattern stability, which hinders effective representation learning through masked learning. Moreover, TSTransformer relies on self-supervised reconstruction of masked time steps and is sensitive to input data quality; the presence of intermittent missing values, measurement noise, and extreme operating points can interfere with the training process and cause notable instability.

Therefore, LSTM, TabTransformer, and TSTransformer are excluded from the subsequent model ensemble process. Although tree-based models are generally stable, their optimal prediction accuracies remain inferior to that of neural-network models. Consequently, FTTransformer and ResNet are selected as the baseline models for the offline dataset.

4.1.2. Bayesian optimization of hyperparameters

For tree-based models (XGBoost, LightGBM, and CatBoost), hyperparameters are tuned using Bayesian optimization, which efficiently identifies optimal combinations under limited computational resources and is suitable for models with relatively independent parameters. The final LightGBM parameters are listed in Table 3; hyperparameter settings for all remaining models (i.e., CatBoost, XGBoost, ResNet, FTTransformer, and the residual-correction MLP) are provided in Appendix A (Tables A1–A5).

Neural networks exhibit strong hyperparameter interdependence, and training is sensitive to initial weights and data splits, causing performance variability and high cost, which limits Bayesian optimization. Key parameters (learning rate, hidden-layer size, and dropout rate) were therefore tuned via grid search, selecting the configuration with the best validation performance.

All models were trained using their respective optimization strategies and evaluated on an independent test set. Prediction performance under optimal configurations, including RMSE and MAE, is summarized in Table 4.

4.1.3. Ensemble experiments

This section presents multiple model ensemble experiments to investigate the impact of different model combinations on the ROP prediction performance. The results indicate that the ensembles of FTTransformer and ResNet achieve the best performance across all evaluation metrics, significantly outperforming individual models and other ensemble schemes (see Fig. 16). In contrast, XGBoost shows the poorest performance across all metrics, and its inclusion in ensembles leads to a decline in the overall accuracy. Further analysis reveals that mixed ensembles of tree-based and neural-network models fail to surpass the performance of purely neural-network-based ensembles, as illustrated by the comparison between M3 (neural-network ensemble) and M4 (mixed ensemble of neural networks and tree models). Based on these findings, the FTTransformer and ResNet ensemble models are selected as the offline models, and trained using existing drilling data with a focus on historical data modeling.

Table 3
Optimal hyperparameter settings for the LightGBM model.

Hyperparameter	Search space	Optimal value
Number of estimators	(1000, 10,000)	4480
Maximum depth	(1, 24)	17
Learning rate	(0.005, 0.1)	0.0716
Feature fraction	(0.5, 1)	0.7130
L ₁ regularization	(0.5, 1)	0.7028
L ₂ regularization	(0.5, 1)	0.6394

Table 4
Performance on Ma51X test set with optimal hyperparameters.

Model	RMSE, m/h	MAE, m/h	R ²
CatBoost	9.751	4.889	0.889
LightGBM	9.223	5.139	0.901
XGBoost	11.228	6.229	0.853
ResNet	6.389	3.347	0.952
FTTransformer	7.135	2.895	0.940

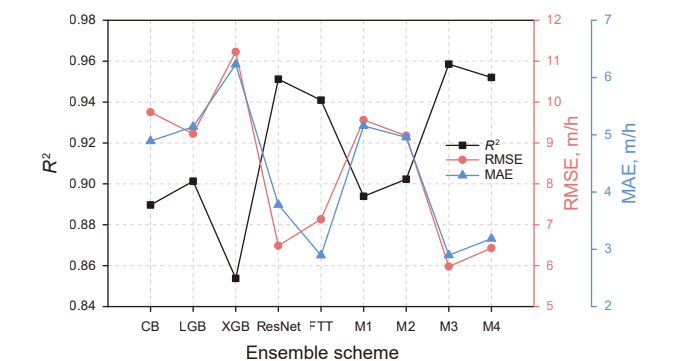


Fig. 16. Performance comparison of different models. CB, CatBoost; LGB, LightGBM; XGB, XGBoost; FTT, FTTransformer; M1, CB + LGB + XGB; M2, CB + LGB; M3, ResNet + FTT; M4, ResNet + FTT + CB + LGB.

4.2. Online model experiments

This section focuses on the characteristics of real-time drilling data and presents a series of online training and optimization experiments. It includes sensitivity analysis, with respect to training data volume, comparison of training times across different models, performance evaluation of residual models, and Bayesian optimization of key drilling parameters. Through multifaceted experiments, the adaptability and predictive accuracy of the models in dynamic drilling environments are properly evaluated.

4.2.1. Data sensitivity analysis

Given the strict requirements on real-time performance in online prediction scenarios and the often limited availability of real-time drilling data on site, the direct application of deep-learning models in actual drilling operations is constrained by their reliance on large-scale training data and computational resources. Therefore, it is necessary to systematically evaluate the training efficiencies and predictive performances of lightweight models under small-to-medium sample conditions. Based on this consideration, three typical GBT models—XGBoost, LightGBM, and CatBoost—are selected for sensitivity analysis under various training sample sizes, with an independent well used as a test set to assess the generalization capability.

The experimental results indicate that, as the training-data volume increases, the overall predictive performances of all three model shows stable improvement trends (as illustrated in Fig. 17), although the performance has not yet reached saturation at the current data scale. It should be noted that performance improvement does not always monotonically increase with data volume; in certain ranges, further increasing the training data may lead to slight performance degradation, which is partly related to the accumulated training errors and uneven sample distributions. Hence, merely expanding the dataset does not guarantee continuous performance improvement. The quality and coverage balance of training data are the key factors that affect model performance.

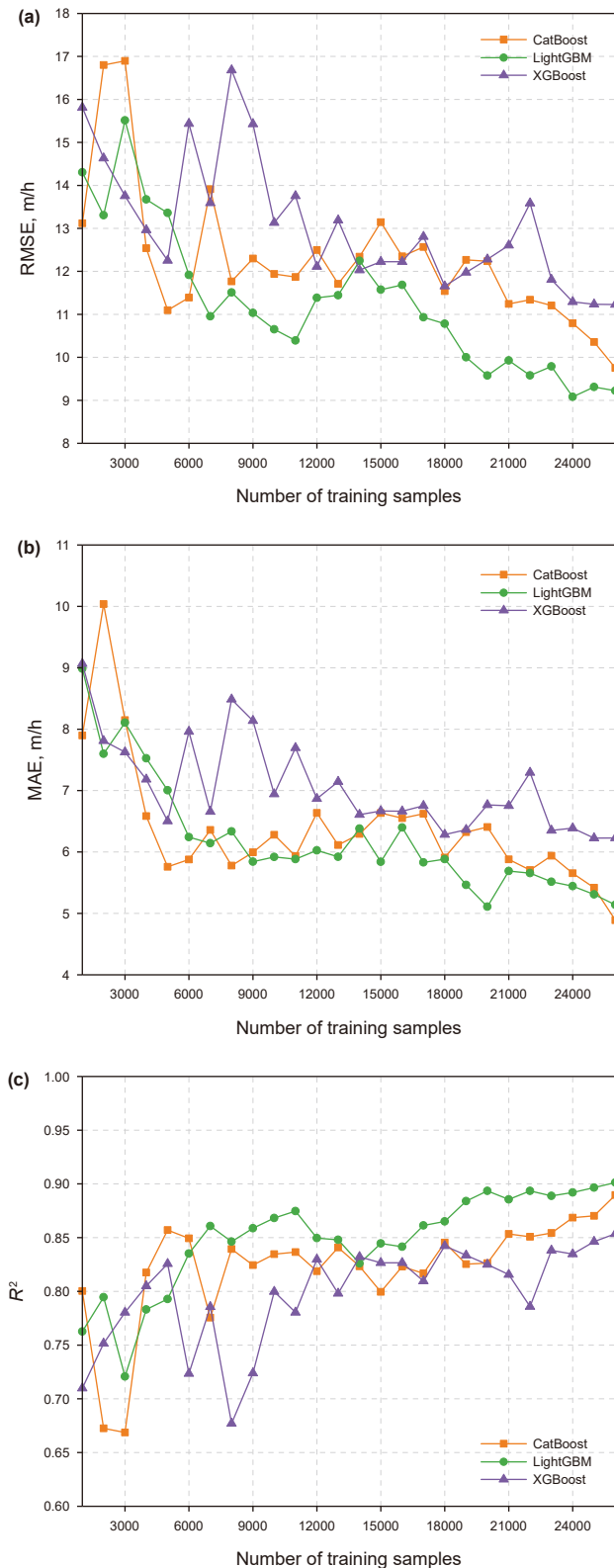


Fig. 17. Prediction performances of XGBoost, LightGBM, and CatBoost, with different training set sizes.

In these comparisons, XGBoost consistently underperforms LightGBM and CatBoost across all evaluation metrics, and is therefore excluded from subsequent analyses. It is noteworthy that

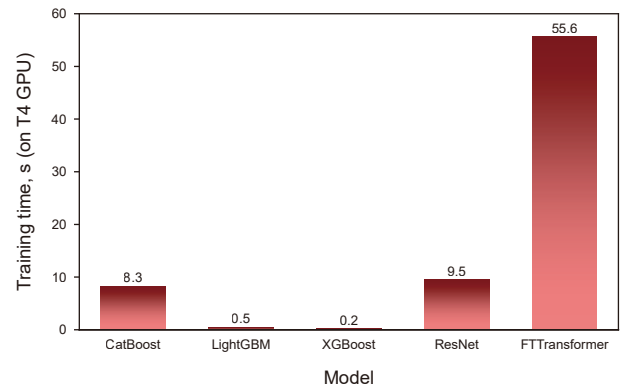


Fig. 18. Training time comparison of different models for online real-time drilling.

LightGBM demonstrates relatively strong robustness under imbalanced sample distributions, which is one of the key reasons for its further adoption in this study. Meanwhile, models supporting incremental training or fine-tuning exhibit limited effectiveness in this task, with some models failing to achieve effective fitting under constrained online data conditions.

To address these limitations, a residual prediction strategy is proposed herein: based on existing offline models, a residual model is trained using limited real-time data, to characterize the systematic deviation between the original model and the current operating conditions. As the residual primarily reflects local perturbations caused by formation differences and operational variations, it has a relatively simple structure and limited magnitude. Thus, its dependency on real-time data volume is significantly lower than that of directly reconstructing the complete ROP mapping. By dynamically adding the predicted residuals to the original predictions, the model's adaptability and predictive accuracy can be enhanced without retraining the main model.

4.2.2. Training time analysis

To meet the requirements of online real-time training, a comparison of training times is conducted to evaluate the suitability of different models in practical drilling environments. Because the number of samples available at each online update is not fixed, all observation points from a single well are used as the training set, representing the maximum amount of data available at that update. Results (see Fig. 18) show that FTTransformer requires nearly 1 min to train, which exceeds the data-acquisition interval of key drilling parameters (e.g., WOB and RPM) collected at the second-level frequency, making it unsuitable for real-time online training. Limited computational resources and complex field conditions make the CatBoost and ResNet training times uncertain and potentially long, thereby excluding them from consideration. Considering both training efficiency and model performance, LightGBM is selected for online residual prediction, as it allows rapid response and dynamic adjustment of drilling parameters and formation states to ensure drilling efficiency and operational safety.

4.2.3. Residual model evaluation

The training length of the residual model is defined as *window*, and the number of forward-predicted samples is defined as *step*. To determine the optimal parameter combination, multiple (*window*, *step*) configurations are designed and evaluated, with experiments labeled in the “*step–window*” format (e.g., 1–200, 50–200). Experiments without online residual prediction are defined as the Baseline.

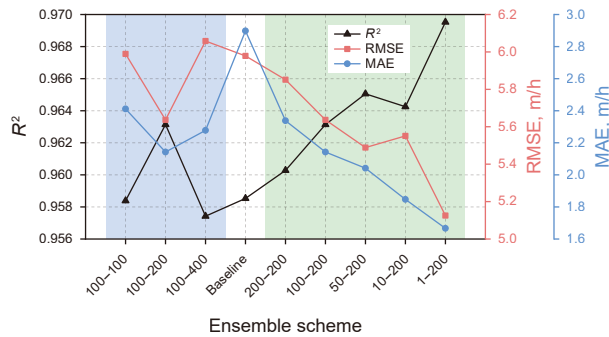


Fig. 19. Effect of different window and step configurations on residual model performance.

As shown in Fig. 19, the model performance is optimal when $window = 200$ (blue region) and $step = 1$ (green region). These results indicate that, during drilling, training the residual model on the most recent 200 observations can effectively capture the prediction errors caused by changes in formation properties, drilling parameters, and tool conditions.

For visualization, the well interval with a relatively stable mechanical ROP between 5000 and 5100 m is selected. Fig. 20(a) shows the residual predictions by the LightGBM model, Fig. 20(b) shows the ROP predicted directly by FTTransformer and ResNet, and Fig. 20(c) shows the final ROP after applying online residual correction to the predictions in Fig. 20(b).

4.2.4. Drilling parameter tuning

To enhance the control and optimization of the actual ROP during real-time drilling, key parameters such as RPM, WOB, and inlet drilling-fluid flow rate are dynamically adjusted using Bayesian optimization. Theoretically, appropriately increasing RPM and WOB improves the bit cutting efficiency, while optimizing fluid flow enhances cutting transport and reduces drilling resistance, thereby increasing ROP.

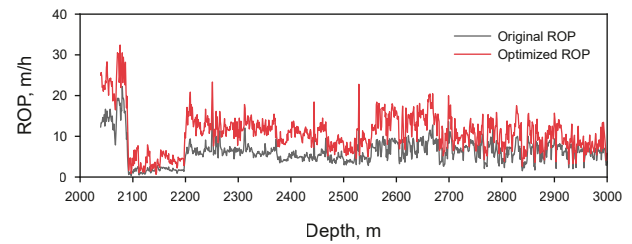


Fig. 21. Online ROP optimization results using Bayesian optimization.

To ensure engineering feasibility, parameter adjustments must strictly adhere to predesign constraints related to well depth and borehole dimensions, to prevent unsafe settings. Adjustments are kept smooth and continuous to accommodate the equipment response and operational safety.

Bayesian optimization efficiently balances complex parameter interactions through iterative sampling and model updates, quickly identifying optimal combinations. Each measurement point undergoes 120 iterations, with an average computation time of approximately 2 s, meeting online real-time requirements. Selected optimization results are presented in Fig. 21. Subsequent cost–risk assessments and field verification are required to ensure both safety and economic efficiency, guiding operational parameter adjustments.

4.3. Field application limitations

Despite the constructed ROP prediction model demonstrating satisfactory performance during field-scale validations, several typical limitations and challenges remain for practical on-site applications:

- (1) Computational resource constraints: The primary computational processes of the model can be executed offline; however, practical online deployment still requires consideration of the limited computational resources. In the

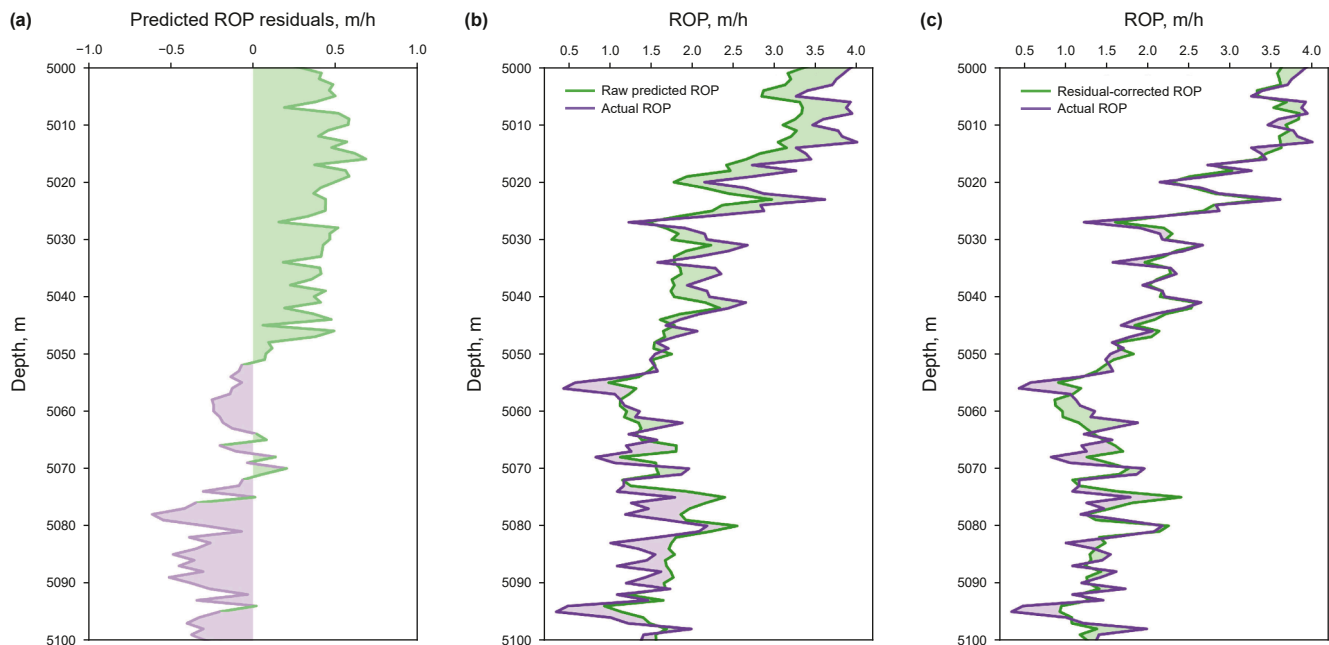


Fig. 20. ROP prediction results of offline models and online residual correction. (a) Online residual prediction. (b) ROP predicted directly by the offline models. (c) ROP after online residual correction.

current framework, intensive computations are concentrated in the offline phase, while the online prediction and residual correction modules are relatively lightweight. Nevertheless, under high-frequency data input or complex operating conditions, the long-term stable operation of the system may be constrained by the performance of field equipment, and warrants further evaluation.

- (2) Data latency limitations: Some input parameters inherently experience unavoidable delays during acquisition, such as gas measurements and logging-while-drilling (LWD) data. Gas measurements typically exhibit a 2–5 min lag. Additionally, because LWD tools are positioned at a certain distance from the drill bit, the real-time downhole data they collect inevitably experiences transmission delays to the surface, thereby limiting the model's ability to respond to rapidly changing downhole conditions in real time.
- (3) Insufficient data volume and limited features: Owing to the limited availability of training data, the model may not adequately learn the feature distributions under certain complex operating conditions, restricting its predictive capability for rare or extreme scenarios. Furthermore, the range of input features is limited, and some critical influencing factors are not incorporated into the model, which not only reduces its adaptability and predictive accuracy across different formations and operating conditions but also constrains its generalization ability and reliability in practical engineering applications.
- (4) Missing data: Real-time field data inevitably contain missing, abnormal, or noisy measurements, which may compromise the completeness of the model inputs and the stability of predictions. Causes of missing data include sensor failures, communication interruptions, equipment limitations, and insufficient acquisition frequencies for certain parameters. Missing data may also reduce the effectiveness of online residual correction, thereby reducing the prediction reliability in rapidly changing formations or complex drilling environments.
- (5) Parameter acquisition discrepancies: Because of cost and equipment limitations, the available drilling-related parameters may vary across wells or well sections, and differences in input feature configurations increase the difficulty of model transfer and reuse across wells. Moreover, variations in measurement precision, sampling frequency, and update intervals for different parameters may further affect the prediction consistency and stability, adding engineering challenges to practical deployment.
- (6) Lack of physical constraints: The proposed method remains primarily data-driven and does not explicitly incorporate physical constraints, such as rock mechanics or drilling mechanics. Consequently, the physical consistency of predictions under extreme conditions still requires verification. The absence of such constraints may lead to unrealistic predictions under unobserved formations or abnormal drilling conditions, limiting the model's reliability for engineering decision support and safety management.
- (7) Limitations of single-objective optimization: The current model optimizes for ROP alone and does not consider other critical engineering objectives such as bit life, wellbore stability, and downhole safety, which imposes certain limitations under complex operating conditions. Moreover, single-objective optimization may result in biased decision-making under anomalous formations, sudden mechanical issues, or extreme scenarios, increasing the operational risk and constraining the model's potential for multiobjective

optimization and comprehensive safety management in practical engineering.

In summary, the practical deployment of the proposed model requires careful consideration of site-specific conditions and the aforementioned limitations to ensure the reliability and engineering applicability of the prediction results.

5. Conclusions

This study first conducted a systematic screening of input features, ultimately identifying 26 key representative parameters for mechanical ROP prediction. The contributions of individual features to the prediction performance were quantitatively evaluated, based on which the WOB, rotary speed, and inlet drilling-fluid flow rate were selected as engineering-feasible real-time optimization variables. The impact of data-denoising strategies on model generalization was investigated. The results indicated that retaining anomalous data could effectively enhance the model's adaptability under extreme operating conditions. In addition, the adaptability of drilling parameters across different model architectures—including tree-based models, attention-based models, sequence models, and CNNs—was comparatively analyzed, providing guidance for subsequent model-structure selection.

To address the limited transferability of artificial intelligence models in cross-regional geological applications, this study proposed a high-accuracy ROP prediction framework that integrated offline modeling with online residual correction. The residual was defined as a dynamic estimate of the prediction error of the offline model, and served as a comprehensive representation of short-term influences such as local formation variations, drilling-parameter adjustments, and stochastic operational disturbances. Compared with model fine-tuning, incremental training, and purely temporal modeling approaches, the proposed residual-modeling strategy achieves structural decoupling between globally stable drilling patterns and locally dynamic deviations, thereby exhibiting superior generalization capability and prediction stability. In the offline stage, eight commonly used artificial-intelligence models were systematically evaluated, and a neural-network ensemble composed of FTTransformer and ResNet was selected. In the online stage, a residual-based dynamic-correction strategy was developed to address generalization under small- and medium-sample conditions. Through comparative analyses of the training efficiency and predictive performance, LightGBM was ultimately chosen as the online residual-prediction model, with the optimal parameter configuration determined as $window = 200$ and $step = 1$.

CRedit authorship contribution statement

Xiao-Yang Li: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Formal analysis. **Zong-Jie Mu:** Supervision, Funding acquisition, Conceptualization. **Pan-Pan Zhang:** Writing – review & editing, Supervision, Project administration. **Zhe-Hua Yang:** Resources, Data curation. **Yu-Xuan Zhao:** Resources, Investigation. **Lian-Feng Fu:** Visualization.

Data availability statements

The datasets, feature engineering code, model files, training and inference scripts, and figure source files are available from the corresponding author upon reasonable request.

Declaration of interests

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

Acknowledgements

The authors gratefully acknowledge the financial support provided by the Science and Technology Program of Xinjiang Uygur Autonomous Region (Grant No. 2025B01014-1) and the National Natural Science Foundation of China (NSFC, Grant No. 52274017).

Appendix A

Table A1

Optimal hyperparameter settings for the CatBoost model.

Hyperparameter	Search space	Optimal value
Number of iterations	(1000, 5000)	2591
Depth of tree	(1, 16)	7
Learning rate	(0.005, 0.1)	0.0752
L ₂ regularization	(1, 20)	3

Table A2

Optimal hyperparameter settings for the XGBoost model.

Hyperparameter	Search space	Optimal value
Learning rate	(0.005, 0.1)	0.0978
Maximum depth	(1, 16)	7
L ₂ regularization	(1, 10)	5.2379
L ₁ regularization	(0.1, 1)	0.3482

Table A3

Optimal hyperparameter settings for the ResNet model.

Hyperparameter	Search space	Optimal value
Number of blocks	{1, 2, 3, 4, 5}	4
Main dimension	{16, 32, 64, 128}	64
Hidden dimension	{32, 64, 128, 256}	128
First dropout rate	{0, 0.1, 0.2, 0.3}	0.1
Second dropout rate	{0, 0.1, 0.2, 0.3}	0.1

Table A4

Optimal hyperparameter settings for the FTTransformer model.

Hyperparameter	Search space	Optimal value
Batch size	{32, 64, 128, 256}	128
Learning rate	{0.1, 0.01, 0.001}	0.001
Attention blocks	{1, 2, 3, 4}	2
Number of heads	{2, 4, 6, 8}	4
Embedding dim	{8, 16, 32, 64}	32

Table A5

Optimal hyperparameters for the residual-correction MLP model.

Hyperparameter	Search space	Optimal value
Input dimension	Fixed	2
Output dimension	Fixed	1
Hidden layers	{1, 2, 3}	1
Hidden dimension	{8, 16, 32, 64, 128}	16
Activation function	{ReLU, LReLU, ELU}	ReLU
Dropout rate	{0, 0.05, 0.1, 0.2, 0.3}	0.1

References

- Abbas, A.K., Rushdi, S., Alsaba, M., 2018. Modeling rate of penetration for deviated wells using artificial neural network. In: Abu Dhabi International Petroleum Exhibition and Conference. <https://doi.org/10.2118/192875-MS>.
- Ahmednour, O., Chen, D., Liu, J., et al., 2025. Comparative analysis of rate of penetration prediction and optimization in deep wells using real-time continuous stacked generalization ensemble learning: A case study in Shunbei. *Geoenergy Sci. Eng.* 247, 213674. <https://doi.org/10.1016/j.geoen.2025.213674>.
- Bai, K., Jin, S., Zhang, Z., et al., 2024. Drilling rate of penetration prediction based on CBT-LSTM neural network. *Sensors* 24 (21), 6966. <https://doi.org/10.3390/s24216966>.
- Bertini, J.R., Lavi, B., 2025. Enhancing rate of penetration prediction in drilling operations: A data stream framework approach. *Eng. Appl. Artif. Intell.* 143, 110034. <https://doi.org/10.1016/j.engappai.2025.110034>.
- Bingham, M.G., 1965. *A New Approach to Interpreting Rock Drillability*. Petroleum Publishing Company, Tulsa.
- Bourgoyne, A., Young, F.S., 1974. A multiple regression approach to optimal drilling and abnormal pressure detection. *Soc. Pet. Eng. J.* 14, 371–384. <https://doi.org/10.2118/4238-PA>.
- Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. <https://doi.org/10.1145/2939672.2939785>.
- Detournay, E., Defourny, P., 1992. A phenomenological model for the drilling action of drag bits. *Int. J. Rock Mech. Min. Sci. Geomech. Abstr.* 29, 13–23. [https://doi.org/10.1016/0148-9062\(92\)91041-3](https://doi.org/10.1016/0148-9062(92)91041-3).
- Eren, T., Ozbayoglu, M.E., 2010. Real time optimization of drilling parameters during drilling operations. *SPE Oil and Gas India Conference and Exhibition*. <https://doi.org/10.2118/129126-MS>.
- Fan, Y.-D., Pang, H.-W., Jin, Y., et al., 2025. A new method for the rate of penetration prediction and control based on signal decomposition and causal inference. *Pet. Sci.* 22 (6), 2414–2437. <https://doi.org/10.1016/j.petsci.2025.04.005>.
- Gorishniy, Y., Rubachev, I., Khrulkov, V., et al., 2021. Revisiting deep learning models for tabular data. *arXiv: 2106.11959*. <https://doi.org/10.48550/arXiv.2106.11959>.
- Graham, J.W., Muench, N.L., 1959. Analytical determination of optimum bit weight and rotary speed combinations. In: *Fall Meeting of the Society of Petroleum Engineers of AIME*. <https://doi.org/10.2118/1349-G>.
- Hao, J., Xu, H., Peng, Z., et al., 2025. An online adaptive ROP prediction model using GBDT and Bayesian optimization algorithm in drilling. *Geoenergy Sci. Eng.* 246, 213596. <https://doi.org/10.1016/j.geoen.2024.213596>.
- Hareland, G., Rampersad, P., 1994. Drag-bit model including wear. In: *SPE Latin America/Caribbean Petroleum Engineering Conference*. <https://doi.org/10.2118/26957-MS>.
- 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>.
- Hegde, C., Millwater, H., Pyrcz, M., et al., 2019. Rate of penetration (ROP) optimization in drilling with vibration control. *J. Nat. Gas Sci. Eng.* 67, 71–81. <https://doi.org/10.1016/j.jngse.2019.04.017>.
- Hegde, C., Wallace, S., Gray, K., 2015. Using trees, bagging, and random forests to predict rate of penetration. *SPE Middle East Intelligent Oil and Gas Conference and Exhibition*. <https://doi.org/10.2118/176792-MS>.
- He, K., Zhang, X., Ren, S., et al., 2016. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. <https://doi.org/10.1109/CVPR.2016.90>.
- Huang, X., Khetan, A., Cvitkovic, M., et al., 2020. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv:2012.06678*. <https://doi.org/10.48550/arXiv.2012.06678>.
- Karnot, M., Talib, M., Alkamil, E.H.K., et al., 2025. Maximizing drilling efficiency in deviated wells using AI based ROP optimization. In: *International Petroleum Technology Conference*. <https://doi.org/10.2523/IPTC-24964-MS>.
- Ke, G., Meng, Q., Finley, T., et al., 2017. LightGBM: A highly efficient gradient boosting decision tree. In: *Neural Information Processing Systems*.
- Liu, H., Jin, Y., Song, X., et al., 2022. Rate of penetration prediction method for ultra-deep wells based on LSTM-FNN. *Appl. Sci.* 12 (15), 7731. <https://doi.org/10.3390/app12157731>.
- Lv, J.-H., Hu, T., Zhang, W., et al., 2025. Microscopic oil occurrence in the Permian alkaline lacustrine shales: Fengcheng Formation, Mahu Sag, Junggar Basin. *Pet. Sci.* 22 (4), 1407–1427. <https://doi.org/10.1016/j.petsci.2025.01.005>.
- Ma, Y., Liu, Y., Liu, Z., et al., 2026. Addressing the paradox of data-driven rate of penetration prediction: Challenges and solutions. *Geoenergy Sci. Eng.* 257 (Pt. B), 214281. <https://doi.org/10.1016/j.geoen.2025.214281>.
- Ma, Z., Weng, J., Zhang, J., et al., 2025. Intelligent prediction of rate of penetration through meta-learning and data augmentation synergy under limited sample. *Geoenergy Sci. Eng.* 250, 213818. <https://doi.org/10.1016/j.geoen.2025.213818>.
- Maurer, W.C., 1962. The “perfect-cleaning” theory of rotary drilling. *J. Petrol. Technol.* 14 (11), 1270–1274. <https://doi.org/10.2118/408-PA>.
- Mazen, A.Z., Rahmanian, N., Mujtaba, I., et al., 2021. Prediction of penetration rate for PDC bits using indices of rock drillability, cuttings removal, and bit wear. *SPE Drill. Complet.* 36 (3), 320–337. <https://doi.org/10.2118/204231-PA>.
- Mensa-Wilmot, G., Harjadi, Y., Langdon, S., et al., 2010. Drilling efficiency and rate of penetration-Definitions, influencing factors, relationships and value (SPE-128288-MS). In: *IADC/SPE Drilling Conference and Exhibition*. <https://doi.org/10.2118/128288-MS>.

- 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 (Part D), 109575. <https://doi.org/10.1016/j.petrol.2021.109575>.
- Ostroumova, L., Gusev, G., Vorobev, A., et al., 2017. CatBoost: Unbiased boosting with categorical features. In: *Neural Information Processing Systems*.
- Pei, Z.-J., Song, X.-Z., Wang, H.-T., et al., 2024. Interpretation and characterization of rate of penetration intelligent prediction model. *Pet. Sci.* 21 (1), 582–596. <https://doi.org/10.1016/j.petsci.2023.10.011>.
- Peng, C., Zhang, H.-L., Fu, J.-H., et al., 2025. A novel drilling parameter optimization method based on big data of drilling. *Pet. Sci.* 22 (4), 1596–1610. <https://doi.org/10.1016/j.petsci.2025.03.002>.
- Sharma, A., Burak, T., Nygaard, R., et al., 2025. Hybrid ROP modeling: combining analytical and data-driven approaches for drilling. *Geoenergy Sci. Eng.* 251, 213877. <https://doi.org/10.1016/j.geoen.2025.213877>.
- Shaygan, K., Jamshidi, S., 2023. Prediction of rate of penetration in directional drilling using data mining techniques. *Geoenergy Sci. Eng.* 221, 111293. <https://doi.org/10.1016/j.petrol.2022.111293>.
- Soares, C., Daigle, H., Gray, K., 2016. Evaluation of PDC bit ROP models and the effect of rock strength on model coefficients. *J. Nat. Gas Sci. Eng.* 34, 1225–1236. <https://doi.org/10.1016/j.jngse.2016.08.012>.
- 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>.
- Song, W., Mu, H., Ji, W., et al., 2024. Evaluation and optimization of drilling efficiency while drilling based on improved rock-breaking specific energy model of bit. *Geomech. Geophys. Geo-Energy Geo-Resour.* 10, 174. <https://doi.org/10.1007/s40948-024-00872-9>.
- Teale, R., 1965. The concept of specific energy in rock drilling. *Int. J. Rock Mech. Min. Sci. Geomech. Abstracts* 2 (1), 57–73. [https://doi.org/10.1016/0148-9062\(65\)90022-7](https://doi.org/10.1016/0148-9062(65)90022-7).
- Urdaneta, C., Jeong, C., Wu, X., et al., 2024. Deep learning method for improving rate of penetration prediction in drilling. *SPE J.* 29, 3440–3448. <https://doi.org/10.2118/219746-PA>.
- Vaswani, A., Shazeer, N., Parmar, N., et al., 2017. Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010.
- Warren, T., 1987. Penetration rate performance of roller cone bits. *SPE Drill. Eng.* 2, 9–18. <https://doi.org/10.2118/13259-PA>.
- Yahia, H., Figliuzzi, B., Gerbaud, L., et al., 2025. A real-time rig control ROP optimization framework using machine learning and predictive vibration modeling. In: *ADIPEC*. <https://doi.org/10.2118/229215-MS>.
- Yang, Y., Cen, X., Ni, H., et al., 2025. A highly accurate and robust prediction framework for drilling rate of penetration based on machine learning ensemble algorithm. *Geoenergy Sci. Eng.* 244, 213423. <https://doi.org/10.1016/j.geoen.2024.213423>.
- Young, F., 1969. Computerized drilling control. *J. Pet. Technol.* 21, 483–496. <https://doi.org/10.2118/2241-PA>.
- Zerveas, G., Jayaraman, S., Patel, D., et al., 2021. A Transformer-based framework for multivariate time series representation learning. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2114–2124. <https://doi.org/10.1145/3447548.3467401>.
- Zhang, R., Zhu, Z., Yan, Z., et al., 2026. EHTGNN: An explainable hybrid temporal graph neural network for robust rate of penetration prediction in drilling operations. *Geoenergy Sci. Eng.* 259, 214347. <https://doi.org/10.1016/j.geoen.2025.214347>.