applied sciences

MDPI

Article

Rate of Penetration Prediction Using an ExtraTrees Model Optimized by an Improved Harris Hawks Algorithm

Xi Cui, Dachuan Liang *, Daoxiong Li and Chen Yang

School of Oil & Natural Gas Engineering, Southwest Petroleum University, Chengdu 610500, China; 202321000824@stu.swpu.edu.cn (X.C.); 202311000062@stu.swpu.edu.cn (D.L.); 202221001063@stu.swpu.edu.cn (C.Y.)
* Correspondence: dachuanliang1965@outlook.com

Featured Application

The proposed IHHO-ET model can support same-region ROP prediction and drilling parameter analysis using structured field while drilling data under appropriate data authorization conditions.

Abstract

Rate of penetration (ROP) is a key indicator of drilling efficiency, governed by nonlinear coupling among mechanical, hydraulic, drilling fluid, and formation factors. This study develops an ExtraTrees model optimized by an improved Harris hawks optimization algorithm (IHHO-ET) using field while drilling data from Well Z in the Tarim Oilfield. A preprocessing workflow involving drilling section identification, abnormal condition filtering, 3 × IQR outlier removal, Savitzky–Golay smoothing, and standardization is combined with correlation and gray relational analysis under engineering mechanism constraints to select 14 input features. Logistic chaotic initialization, adaptive Gaussian mutation, and dynamic weighting are introduced into HHO, with validation set RMSE as the fitness function. To reduce the influence of random splitting and initialization, all comparison models are evaluated with repeated seeds and validation set tuning. Using R^2 and RMSE as primary criteria, IHHO-ET achieves $R^2 = 0.910 \pm 0.004$ and RMSE = 0.871 ± 0.019 on the same-well test set. Its improvement over HHO-ET is small and not significant ($p = 0.109$), indicating that the IHHO strategies mainly refine search stability. Same-region leave-one-well-out validation gives an average $R^2 = 0.696$, suggesting that the model suits same-region trend prediction rather than direct closed-loop control. The proposed workflow provides a practical reference for ROP prediction.

Keywords: rate of penetration; ExtraTrees; improved Harris hawks optimization; hyperparameter optimization; Tarim Oilfield

Academic Editor: Nathan J. Moore

Received: 1 June 2026
Revised: 26 June 2026
Accepted: 30 June 2026
Published: 3 July 2026

1. Introduction

With the continuous growth of global energy demand and the gradual shift of oil and gas development toward deep, ultra-deep, unconventional, and structurally complex reservoirs, drilling engineering faces increasingly complex formation conditions, downhole operating states, and construction risks. The drilling process directly affects the development cycle and engineering cost of oil and gas resources, and it is also closely related to wellbore quality, operational safety, and subsequent completion operations [1,2]. Therefore, improving drilling efficiency, reducing non-productive time, and realizing intelligent

optimization of drilling parameters under complex formation conditions have become important research directions in petroleum engineering. Rate of penetration (ROP) is a key indicator describing wellbore extension per unit time and directly reflects rock breaking efficiency and drilling operational efficiency [3]. In field drilling, ROP is jointly affected by well depth, weight on bit, rotary speed, torque, standpipe pressure, flow rate, drilling fluid density, equivalent circulating density, and formation conditions [4,5]. These parameters exhibit obvious nonlinear, coupled, and stage-dependent characteristics. In deep wells and complex formations, lithology changes, operating-condition switching, and coupled parameter adjustments can lead to significant ROP fluctuations. Therefore, ROP prediction is not merely a single-variable regression problem but a complex modeling problem involving multi-source parameter coupling, data quality control, and evaluation of model applicability to unseen samples.

Field while drilling data provide an important basis for ROP prediction. However, raw data often contain abnormal operating conditions, local outliers, sensor noise, inconsistent parameter definitions, and imbalanced sample distributions [4,5]. Directly using unprocessed data for modeling may cause a model to learn invalid operating condition features or noise disturbances, thereby reducing prediction accuracy and stability. Therefore, before constructing an ROP prediction model, it is necessary to clean the raw data, handle anomalies, smooth noise, and perform standardization according to the engineering background, followed by reasonable feature selection and model construction.

The main influencing factors and modeling issues in ROP prediction are summarized in Figure 1.

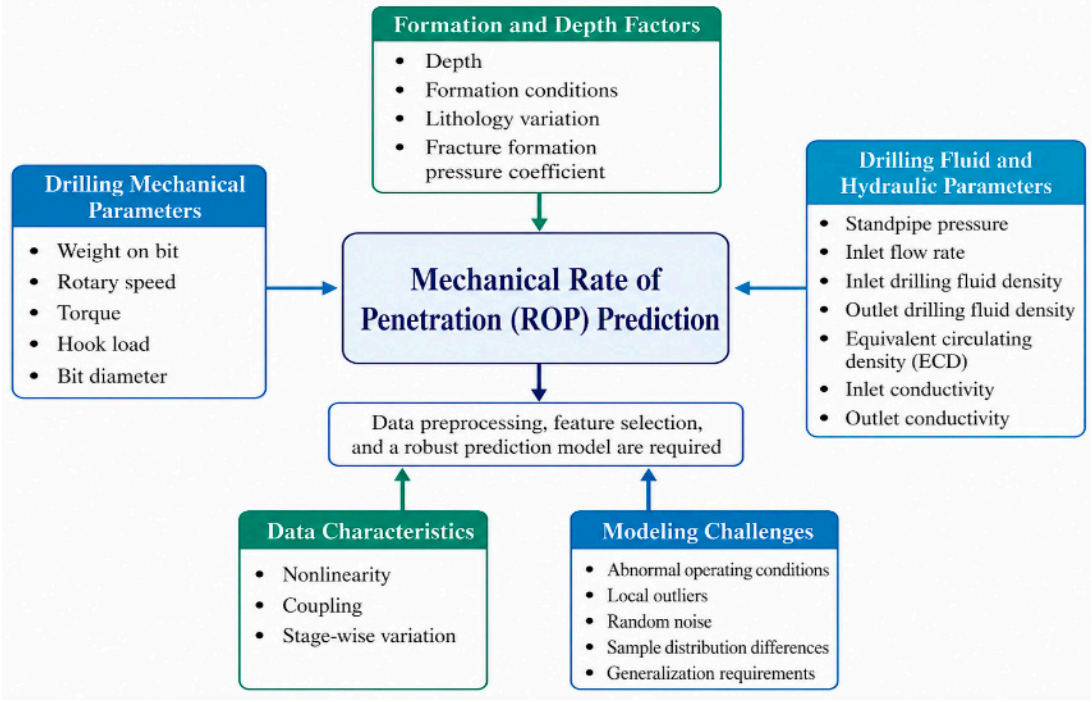

Figure 1. Schematic diagram of influencing factors and modeling issues in ROP prediction.

The ROP prediction task involves drilling parameters, drilling fluid parameters, well depth, and formation conditions. Weight on bit, rotary speed, and torque directly reflect the mechanical action of the bit during rock breaking; standpipe pressure, flow rate, and equivalent circulating density reflect wellbore hydraulic conditions and bottom-hole cleaning status; and well depth and formation pressure-related parameters characterize formation condition changes in different intervals. Because these factors are complexly coupled and field data are often affected by abnormal operating conditions and noise, an

ROP prediction model needs strong nonlinear representation capability, noise resistance, and applicability to unseen wells within the same region.

Current ROP prediction research mainly includes three categories: traditional empirical/mechanistic models, machine learning models, and intelligent optimization fusion models. Traditional empirical/mechanistic ROP models are usually developed from bit–rock interaction, mechanical specific energy (MSE), hydromechanical specific energy (HMSE), or field-calibrated empirical relationships [6–9]. These methods have clear physical meaning and relatively strong engineering interpretability, as they can reflect the effects of weight on bit, rotary speed, torque, flow rate, and formation strength on rock breaking efficiency. However, recent comparative studies also indicate that, under complex formations, multi-condition switching, and high-noise field data, models relying only on empirical parameters or fixed mechanistic assumptions often have difficulty fully describing the nonlinear coupling between ROP and multidimensional drilling parameters.

With the development of measurement while drilling and data acquisition technologies, machine learning methods have gradually been applied to ROP prediction [10,11]. Barbosa et al. [11] reviewed the application progress of machine learning in ROP prediction and pointed out that random forests, support vector regression, gradient boosting trees, and neural networks show strong advantages in nonlinear modeling. Hazbeh et al. [12] applied multiple machine learning models to ROP prediction in directional wells and found obvious differences in prediction accuracy and computational efficiency among different models. Zhang et al. [13] combined an attention mechanism, a gated recurrent unit network, and fully connected neural networks to construct a deep learning method for real-time ROP prediction. Overall, machine learning models can learn complex mapping relationships from field data, but their prediction performance still depends strongly on data quality, input feature construction, and hyperparameter settings.

XGBoost was proposed by Chen and Guestrin [14] and has strong nonlinear fitting and iterative error correction capabilities, making it widely used in structured data modeling tasks. ExtraTrees, or extremely randomized trees, is an ensemble regression method based on randomized decision trees. During node splitting, it randomizes candidate features and split thresholds and then averages the predictions of multiple trees to improve generalization and resistance to overfitting [15,16]. In addition, support vector regression (SVR), as a classical kernel-based method, is often used for nonlinear regression modeling [17]. For tabular, multivariable, nonlinear regression tasks, such as ROP prediction, tree-ensemble models have good applicability; however, key hyperparameters such as the number of trees, maximum depth, minimum leaf samples, and feature sampling strategy still significantly affect model prediction performance.

To further improve model performance, some studies have introduced particle swarm optimization (PSO) [18], genetic algorithms (GAs) [19], gray wolf optimization (GWO) [20], Bayesian optimization [21], and Harris hawks optimization (HHO) [22] into the hyperparameter optimization process of machine learning models. HHO was proposed by Heidari et al. [22] and has a simple structure and strong global search capability, making it applicable to engineering optimization problems. However, swarm intelligence optimization algorithms usually need to balance global exploration and local exploitation in complex search spaces [23]. The basic HHO may still suffer from uneven initial population distribution, insufficient local search ability, and inadequate late-stage convergence stability. To address these deficiencies, existing studies have attempted to improve HHO population initialization and search performance through strategies such as chaotic mapping [24]. Therefore, applying improved HHO to the hyperparameter optimization of ExtraTrees in ROP prediction can reduce the uncertainty introduced by manual parameter setting and improve the model's adaptability to complex drilling data.

In recent years, ROP prediction research has further developed toward public benchmark datasets, neighboring-well extrapolation, physics data fusion, and deep sequence models. Tunkiel et al. constructed an ROP benchmarking dataset based on the public Volve data, emphasizing the importance of open test scenarios and reproducible evaluation [25]. Alsaihati et al. applied an ensemble learning model to ROP prediction in complex lithology intervals, verifying the applicability of Random Forest-type models to unseen samples [26]. Elkatatny et al. and Shokry et al. studied real-time ROP prediction for S-shaped well profiles and motorized bottom-hole assemblies, respectively, showing that field sensor parameters and drilling assembly working states are important for model input construction [27,28]. Meanwhile, hybrid physics machine learning models, CBT-LSTM, and improved swarm intelligence optimization combined with BiLSTM have also been used to improve ROP prediction accuracy and interpretability [29–31]. Recent studies have further combined SHAP and other interpretability methods to compare different machine learning models on field drilling data [32,33]. These studies indicate that ROP prediction has shifted from single accuracy improvement toward a balanced focus on data quality control, cross-well evaluation, model interpretability, and engineering applicability.

The research status and limitations of major ROP prediction methods are summarized in Table 1.

Table 1. Summary of the research status and limitations of ROP prediction methods.

Method Category	Representative Methods	Main Advantages	Limitations
Empirical/ mechanistic models	MSE/HMSE models, field-calibrated empirical regression models	Clear physical meaning and strong engineering interpretability	Limited adaptability to complex nonlinear relationships, field noise, and operating condition changes
Single machine learning models	SVR, BP, MLP, XGBoost, etc.	Can learn multivariate nonlinear mapping relationships	Performance depends on data quality, feature selection, and parameter settings
Ensemble learning models	Random Forest, ExtraTrees, etc.	Suitable for tabular drilling data, with good noise resistance and stability	Key hyperparameters strongly affect prediction performance
Optimization fusion models	PSO-, GA-, and HHO-optimized machine learning models	Can reduce the influence of manual parameter setting and improve optimization capability	Susceptible to initial population, search strategy, and local optimum issues

Existing ROP prediction research has developed from traditional empirical models to machine learning models and intelligent optimization fusion models, but several limitations remain. First, some studies do not sufficiently handle abnormal operating conditions, local outliers, and random noise in raw while drilling data, which affects input-sample quality. Second, feature selection sometimes relies too heavily on statistical correlation and lacks integration with drilling engineering mechanisms, potentially resulting in redundant input variables or omission of key parameters. Third, although ensemble learning models are suitable for field tabular drilling data, their key hyperparameters still significantly affect prediction performance. Fourth, some studies focus mainly on prediction accuracy under same-well sample splitting and insufficiently evaluate model applicability to unseen wells or neighboring-well data.

To address these issues, this study uses field while drilling data from Well Z in the Tarim Oilfield as the main modeling dataset and develops an ExtraTrees-based ROP predic-

tion model optimized by improved Harris hawks optimization. The design of this work corresponds to the four limitations summarized above. First, a preprocessing workflow is established to handle drilling section differences, abnormal operating conditions, extreme outliers, and random noise in field data. Second, Pearson correlation, Spearman rank correlation, gray relational analysis, and engineering mechanism constraints are combined to reduce the risk of feature omission or redundant input selection. Third, Logistic chaotic initialization, adaptive Gaussian mutation, and dynamic weighting are introduced into HHO to optimize key ExtraTrees hyperparameters and reduce uncertainty caused by manual parameter setting. Fourth, same-well testing, ablation analysis, and same-region leave-one-well-out validation are jointly used to evaluate prediction accuracy, optimization contribution, and regional applicability to unseen wells.

The overall technical route of this study is shown in Figure 2.

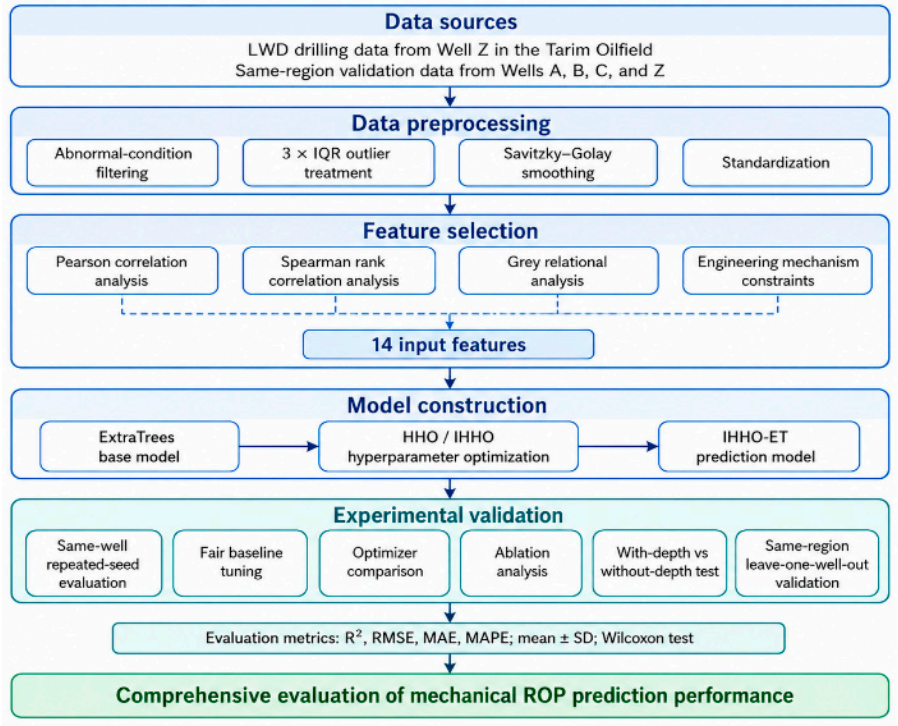

Figure 2. Technical route of ROP prediction based on IHHO-ET.

The technical route of this study includes data sources, data preprocessing, feature screening, model construction, and experimental validation. Field while drilling data are first cleaned, smoothed, and standardized. Statistical correlation analysis, gray relational analysis, and engineering mechanism constraints are then combined to determine model inputs. The improved HHO algorithm is subsequently used to optimize key ExtraTrees hyperparameters. In the revised experimental design, same-well repeated seed evaluation, fair validation set tuning of baseline models, optimizer comparison, ablation analysis, a with-depth versus without-depth test, and same-region leave-one-well-out validation are jointly conducted to evaluate model accuracy, robustness, optimization contribution, and regional applicability.

The remainder of this paper is organized as follows: Section 2 introduces the data source, preprocessing, and feature selection methods; Section 3 describes the IHHO-ET model construction; Section 4 presents the results and discussion; Section 5 concludes the study; and the algorithm pseudocode is provided in Appendix A.

2. Materials and Methods

2.1. Study Materials and Data Source

The data used in this study come from field drilling parameter records of four wells located in the Tarim Oilfield, Xinjiang Uygur Autonomous Region, China. Among them, Well Z is used as the main modeling well for data preprocessing, feature screening, model training, and same-well testing, while Wells A, B, and C are combined with Well Z to construct the same-region cross-well leave-one-well-out validation dataset used in Section 4.3. The four wells have comparable regional geological backgrounds and drilling engineering environments. The main intervals used for modeling and validation show certain consistency in wellbore structure, bit diameter combination, drilling fluid properties, and formation pressure conditions, providing a data basis for evaluating same-region cross-well prediction performance. Specific block-level identification is not disclosed in this paper due to confidentiality requirements from the data provider. The original dataset of Well Z contains 7892 records and 21 variables, with a well depth range of 21–7919 m. Organized primarily by depth, the data cover mechanical parameters, hydraulic parameters, drilling fluid parameters, formation pressure-related parameters, and the target variable ROP, thereby providing a relatively comprehensive reflection of operating condition variations during drilling. Wells A, B, and C are processed using the same preprocessing workflow and input feature system as Well Z, and the detailed leave-one-well-out validation settings and results are presented in Section 4.3.

The original variables mainly include well depth, hookload, weight on bit, rotary speed, standpipe pressure, pump strokes, torque, bit diameter, outlet drilling fluid density, inlet drilling fluid density, outlet drilling fluid conductivity, inlet drilling fluid conductivity, outlet drilling fluid temperature, inlet drilling fluid temperature, inlet flow rate, outlet flow rate, ECD, mud outlet flow, total mud pit volume, fracture formation pressure coefficient, and ROP. Among these variables, ROP is the prediction target, while the remaining process parameters are candidate input variables for subsequent data preprocessing, dominant factor identification, and prediction model construction.

The anonymized wells A, B, C, and Z were used for same-region leave-one-well-out validation. Table 2 summarizes their basic statistics and shows comparable depth coverage with inter-well ROP differences.

Table 2. Basic statistics of wells used for same-region leave-one-well-out validation.

Well	Records After Preprocessing	Depth Range /m	Mean ROP /(m·h^{-1})	Std. ROP /(m·h^{-1})	Validation Role
A	4523	290–6940	3.31	2.68	Independent test well in fold A
B	4218	310–7225	2.95	2.42	Independent test well in fold B
C	4542	260–7050	3.17	2.55	Independent test well in fold C
Z	7093	200–7919	2.87	2.36	Main modeling well and independent test well in fold Z

The original data parameters are listed in Table 3.

Table 3. List of original data parameters.

No.	Parameter	Unit	No.	Parameter	Unit
1	Well depth	m	12	Inlet drilling fluid conductivity	mS·cm^{-1}
2	Hookload	kN	13	Outlet drilling fluid temperature	°C
3	Weight on bit	kN	14	Inlet drilling fluid temperature	°C
4	Rotary speed	r·min^{-1}	15	Inlet flow rate	L·min^{-1}

Table 3. *Cont.*

No.	Parameter	Unit	No.	Parameter	Unit
5	Standpipe pressure	MPa	16	Outlet flow rate	$L \cdot min^{-1}$
6	Pump strokes	spm	17	ECD	$g \cdot cm^{-3}$
7	Torque	kN·m	18	Mud outlet flow	%
8	Bit diameter	mm	19	Total pit volume	m^3
9	Outlet drilling fluid density	$g \cdot cm^{-3}$	20	Fracture formation pressure coefficient	--
10	Inlet drilling fluid density	$g \cdot cm^{-3}$	21	Rate of penetration	$m \cdot h^{-1}$
11	Outlet drilling fluid conductivity	$mS \cdot cm^{-1}$			

The original data exhibit obvious interval stratification and non-normal distribution characteristics. The well depth ranges from 21.00 to 7919.00 m, and the bit diameter ranges from 215.90 to 660.40 mm, indicating that the samples include multiple drilling sections and a relatively complete drilling process. The mean ROP is 3.24 $m \cdot h^{-1}$, with skewness of 2.49 and kurtosis of 8.09, showing clear right skewness and leptokurtic heavy-tailed characteristics. This means that most samples are concentrated in the low-to-medium ROP range, whereas high-ROP samples are relatively few but fluctuate greatly. Meanwhile, the minimum values of pump strokes and inlet flow rate are both 0, whereas the minimum value of outlet flow rate remains 690.00 $L \cdot min^{-1}$, suggesting that the raw data may include pump-off conditions, operating condition switching, return flow lag, or asynchronous parameter acquisition. Although the raw data contain no explicit missing values or completely duplicate records, descriptive statistics alone cannot determine the engineering rationality of individual samples. Therefore, cleaning and screening based on drilling engineering logic are still needed to improve the reliability of samples for subsequent modeling.

2.2. Data Preprocessing

Field drilling data are easily affected by sensor fluctuations, operating condition switching, pump shutdown, connection operations, and parameter alignment deviations during acquisition, transmission, and storage. As a result, raw data may contain abnormal condition samples, local outliers, and random noise. If the raw data are directly used for model training, the model may learn pseudo-features inconsistent with normal continuous drilling, thereby affecting prediction accuracy and applicability to unseen samples. Therefore, the raw data are preprocessed before modeling, mainly including drilling section identification, abnormal condition filtering, outlier detection, noise smoothing, and standardization.

The data preprocessing and candidate feature set construction workflow is shown in Figure 3.

The workflow of data preprocessing and candidate feature set construction first identifies and divides the drilling sections of Well Z according to bit diameter to unify the engineering background of samples. Abnormal samples inconsistent with normal continuous drilling are then removed based on engineering logic among key parameters such as pump strokes, flow rate, standpipe pressure, and ROP. On this basis, extreme outliers are identified and removed using the 3 × IQR criterion [34]. Finally, SG smoothing is applied to some continuously fluctuating parameters, and input variables are standardized to form the candidate feature set for subsequent feature screening and prediction modeling.

Bit diameter is an important parameter for distinguishing drilling sections and wellbore structure. According to the distribution of bit diameter and well depth in the original data, Well Z in the Tarim Oilfield can be divided into four stages: the first, second, third, and

fourth drilling sections. The first section is shallow, uses a bit diameter of 660.4 mm, and has a significantly higher mean ROP than the later sections; its engineering background differs greatly from those of the middle and deep intervals. To avoid distribution interference caused by high-ROP shallow samples, all first section data are removed, and the second, third, and fourth sections are retained for subsequent analysis.

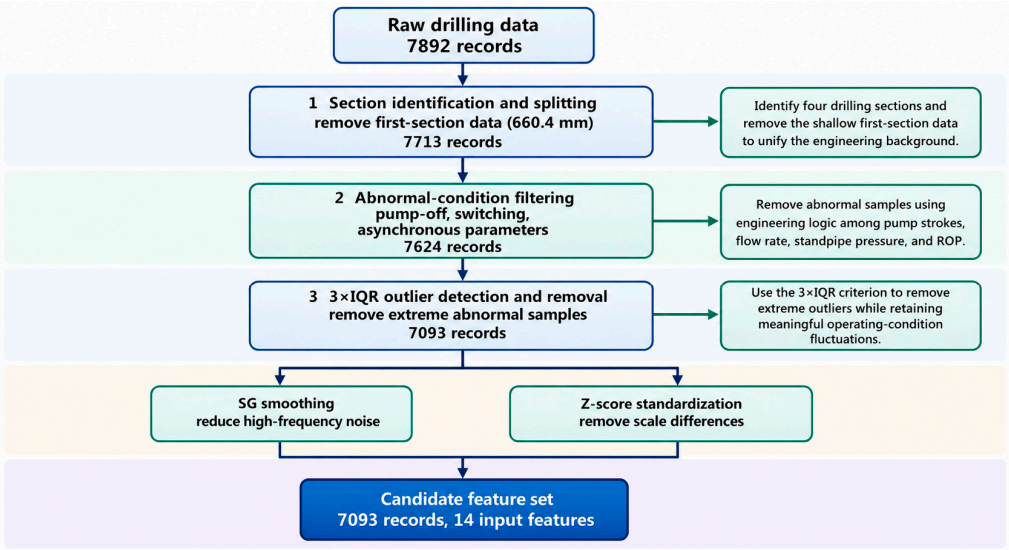

Figure 3. Workflow of data preprocessing and candidate feature set construction.

The drilling-section information for Well Z is summarized in Table 4.

Table 4. Basic information of drilling sections in Well Z of the Tarim Oilfield.

Section	Bit Diameter/mm	Records	Depth Range/m	Mean ROP/(m·h^{-1})	Treatment
First section	660.4	179	21–199	9.54	Removed
Second section	444.5	2851	200–3050	4.74	Retained
Third section	311.2	2853	3051–5910	2.54	Retained
Fourth section	215.9	2009	5911–7919	1.54	Retained

The depth ranges of the second, third, and fourth sections are 200–3050 m, 3051–5910 m, and 5911–7919 m, respectively, and the mean ROP decreases gradually with increasing well depth. This indicates that, as well depth increases, formation drillability decreases, downhole conditions become more complex, and drilling efficiency generally declines. After removing the first section data, the engineering background of the remaining samples becomes more consistent, which improves the stability of model training.

For abnormal condition filtering, this study focuses on the logical consistency among pump strokes, inlet flow rate, outlet flow rate, standpipe pressure, and ROP. Samples corresponding to pump-off states, operating condition switching, asynchronous parameter acquisition, or obviously contradictory variable combinations are judged as not belonging to normal continuous drilling and are removed from the modeling data. Specifically, samples are flagged when pumping and circulation indicators are physically inconsistent, such as pump strokes equal to zero while ROP or standpipe pressure remains positive, inlet flow rate equal to zero while pump strokes, ROP, or standpipe pressure remains positive, or simultaneous mismatch among inlet flow rate, outlet flow rate, standpipe pressure, and ROP during connection or operating condition switching intervals. In this dataset, 89 samples are removed in the abnormal condition filtering step, including 36 samples

triggered by the pump strokes constraint and 53 samples triggered by the inlet flow rate constraint. The abnormal condition filtering process can be expressed as:

$$D_c = \{x_i \in D_0 \mid C_1(x_i), C_2(x_i), \cdots, C_k(x_i) = \text{True}\} \quad (1)$$

where D_0 is the dataset after drilling section identification, D_c is the dataset after abnormal condition filtering, x_i is the i-th sample, and $C_j(x_i)$ is the j-th engineering rationality constraint. Only samples satisfying the engineering constraints are retained.

For extreme outliers in the retained samples, the 3 × IQR criterion is used for identification. The interquartile range is defined as:

$$IQR = Q_3 - Q_1 \quad (2)$$

The outlier identification interval is:

$$[Q_1 - 3IQR, Q_3 + 3IQR] \quad (3)$$

where Q_1 and Q_3 are the first and third quartiles, respectively. The 3 × IQR criterion is applied separately to each selected continuous variable, and a sample is removed if any selected variable exceeds its corresponding lower or upper threshold. Considering the strong inherent fluctuation of field drilling data, the 3 × IQR criterion can remove extreme abnormal samples while retaining operating condition fluctuations with practical engineering significance as much as possible. In this dataset, the outlier removal is mainly triggered by the fracture formation pressure coefficient, and 531 samples are removed in this step.

For continuous process variables such as hookload, standpipe pressure, torque, inlet and outlet drilling fluid conductivity, inlet and outlet drilling fluid temperature, and inlet and outlet flow rate, Savitzky–Golay (SG) smoothing is applied as a light smoothing method [35] to reduce high-frequency noise while preserving the main trend. The SG filter is implemented with a window length of 11 and a polynomial order of 3. The window length is selected to suppress local high-frequency fluctuations while preserving the main depth-related variation trend, and the third-order polynomial is used to retain local nonlinear changes without excessive smoothing. SG smoothing can be expressed as:

$$\hat{x}_i = \sum_{j=-m}^{m} c_j x_{i+j} \quad (4)$$

where $\hat{x}_i$ is the smoothed sample value, x_{i+j} denotes the original sample values within the moving window, c_j is the smoothing coefficient, and 2m + 1 is the window length. The smoothing process only improves sequence stability and does not change the number of samples.

Because different drilling parameters have different dimensions and value ranges, Z-score standardization is further applied to the input variables:

$$x_i^* = \frac{x_i - \mu}{\sigma} \quad (5)$$

where x_i is the original variable value, μ is the mean of the variable, σ is the standard deviation, and x_i^* is the standardized variable value. Standardization eliminates the influence of dimensional differences on model training and provides a consistent data basis for subsequent correlation analysis and multi-model comparison.

The sample size changes during data preprocessing are summarized in Table 5.

The preprocessing procedure reduces the sample size from 7892 original records to 7713 records after removing first section data, 7624 records after abnormal condition filtering, and 7093 records after further removing extreme outliers using the 3 × IQR criterion. This process removes samples that are inconsistent with the target modeling scenario and improves the engineering consistency of the retained dataset.

Table 5. Sample size changes during data preprocessing.

Processing Step	Remaining Records	Reduction from Previous Step	Description
Raw data	7892	--	Original field data
Removal of first section data	7713	179	Unify interval background
Abnormal condition filtering	7624	89	Filter pump-off, condition switching, and logically inconsistent samples
Outlier treatment	7093	531	Remove extreme outliers using the 3 × IQR criterion
Dataset after smoothing	7093	0	Smoothing does not change sample size

2.3. Feature Selection

ROP is jointly affected by multiple types of parameters. Variables may exhibit linear relationships, nonlinear monotonic relationships, and consistency in sequence trends. To avoid feature omission or redundant retention caused by a single indicator, this study combines Pearson correlation analysis, Spearman rank correlation analysis, gray relational analysis, and engineering mechanism constraints to comprehensively screen candidate variables.

The Pearson correlation coefficient is used to measure the degree of linear correlation between two continuous variables [36], and its calculation formula is:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \quad (6)$$

where x_i and y_i are the sample values of the candidate variable and ROP, respectively; $\overline{x}$ and $\overline{y}$ are their corresponding means; and n is the number of samples.

The Pearson correlation coefficient heatmap is shown in Figure 4.

Pearson correlation analysis indicates the linear correlations among candidate variables and between candidate variables and ROP. Overall, well depth, hookload, drilling fluid density, ECD, torque, and fracture formation pressure coefficient show certain negative correlations with ROP; outlet drilling fluid conductivity, inlet drilling fluid conductivity, bit diameter, and inlet flow rate exhibit certain positive correlations. These results provide a preliminary basis for identifying variables related to ROP variation.

The Spearman rank correlation coefficient is used to evaluate monotonic relationships between variables [37], and its calculation formula is:

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \quad (7)$$

where ρ_s is the Spearman rank correlation coefficient, d_i is the rank difference between two variables for the i-th sample, and n is the number of samples. This method has fewer requirements regarding variable distribution and can supplement the identification of nonlinear but monotonic variable relationships.

The Spearman correlation coefficient heatmap is shown in Figure 5.

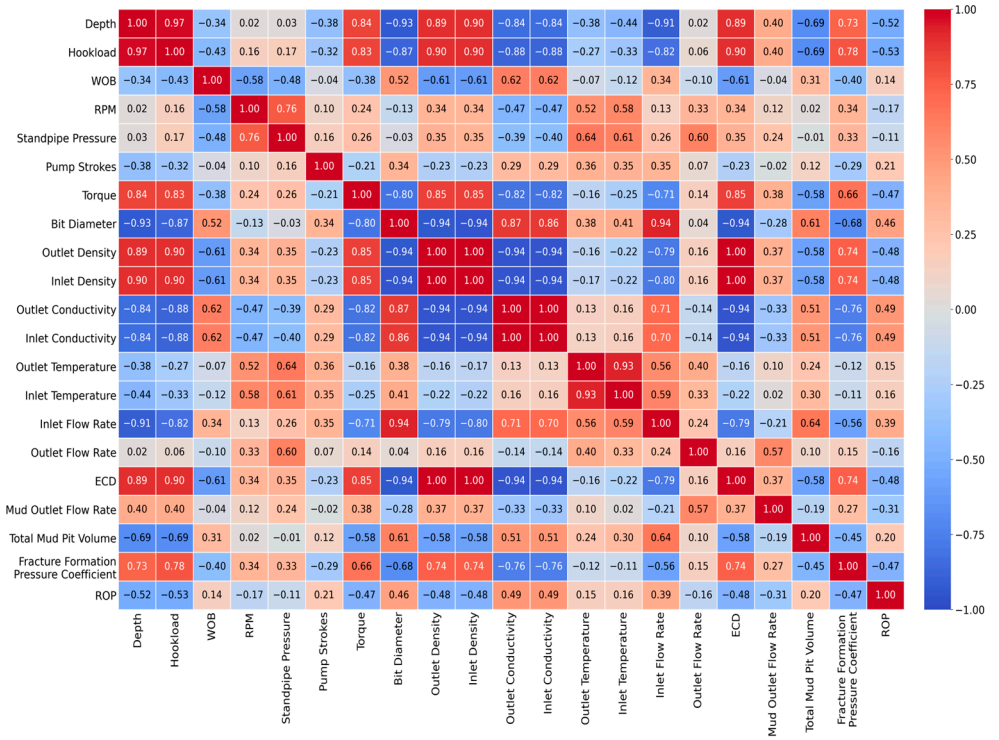

Figure 4. Pearson correlation coefficient heatmap.

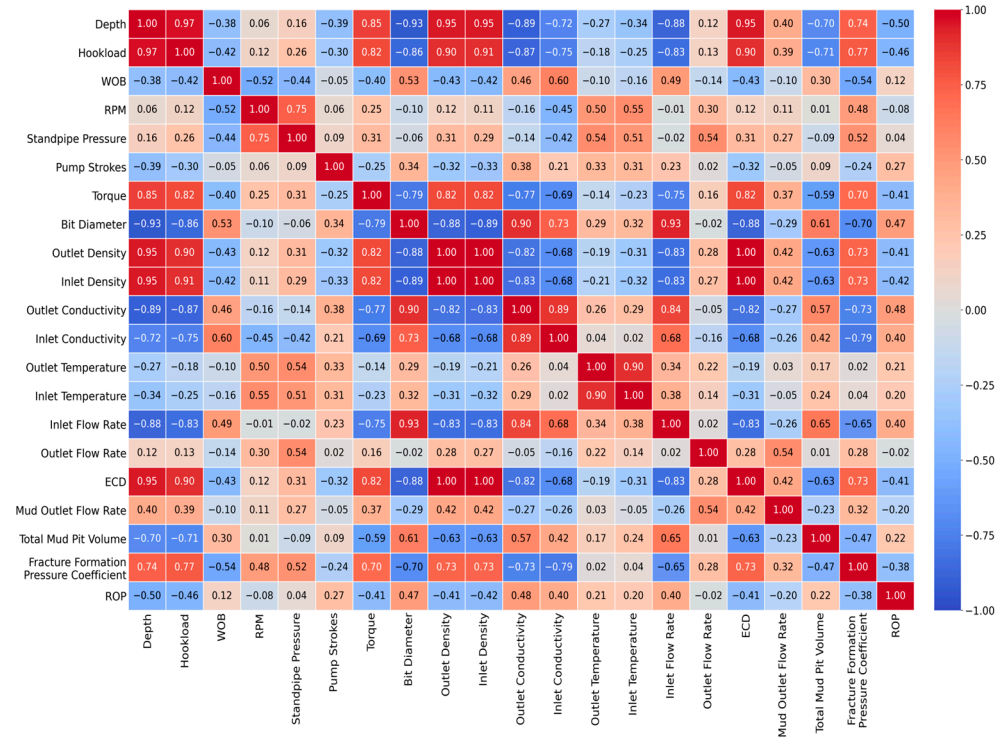

Figure 5. Spearman correlation coefficient heatmap.

Spearman analysis is basically consistent with Pearson analysis, although the correlation strength of some variables changes. Well depth, hookload, drilling fluid density, ECD, and torque still maintain relatively stable negative correlations with ROP, whereas inlet and outlet conductivity, bit diameter, and inlet flow rate show certain positive correlations. Spearman analysis supplements Pearson analysis from the perspective of rank-order variation and helps reduce potential misjudgment caused by relying only on linear correlation.

In addition to statistical correlation, drilling parameters have obvious depth sequence characteristics. To evaluate the consistency between candidate variables and ROP variation trends with well depth, gray relational analysis is further adopted [38]. Taking ROP as the reference sequence and candidate variables as comparison sequences, the gray relational coefficient of the j-th candidate variable at the i-th sample point can be expressed as:

$$\xi_i(k) = \frac{\Delta_{\min} + \rho\Delta_{\max}}{\Delta_i(k) + \rho\Delta_{\max}} \tag{8}$$

where:

$$\Delta_i(\mathrm{k}) = |x_0(\mathrm{k}) - x_i(\mathrm{k})| \tag{9}$$

where Δmin and Δmax are the minimum and maximum absolute differences among all comparison sequences, respectively, and ρ is the distinguishing coefficient. In this study, ρ is set to 0.5 as a commonly used neutral value in gray relational analysis, which balances the distinguishing effect and the stability of relational grade calculation [38]. The gray relational grade of the i-th candidate variable is:

$$\gamma_i = \frac{1}{n}\sum_{k=1}^{n} \xi_i(\mathrm{k}) \tag{10}$$

A larger r_i indicates that the variable has a closer variation trend to the ROP sequence. The ranking results of gray relational grades are shown in Figure 6.

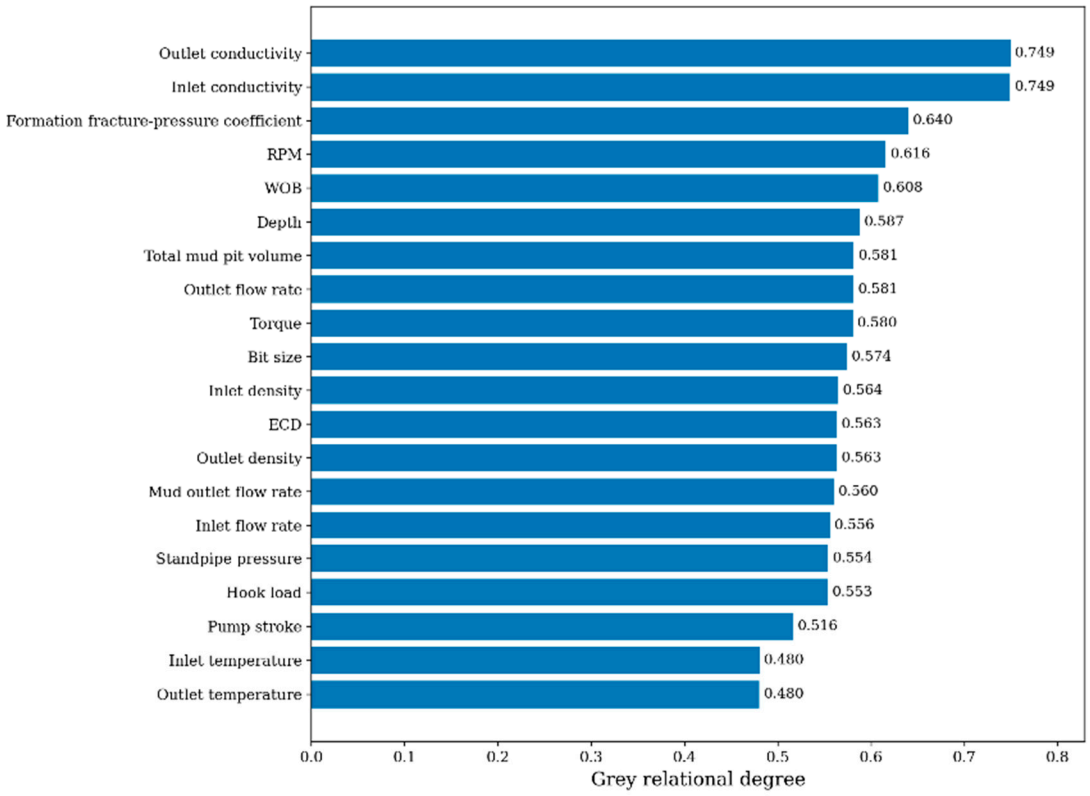

Figure 6. Ranking results of gray relational grades.

The gray relational grade ranking indicates that outlet and inlet drilling fluid conductivity have relatively high relational grades, suggesting that their variation trends with well depth are highly consistent with those of ROP. This result does not necessarily indicate a direct causal effect of conductivity on ROP. Instead, drilling fluid conductivity mainly reflects changes in ion concentration, salinity, fluid contamination, and circulation

conditions, which may be associated with formation fluid invasion, cuttings transport, borehole cleaning, and formation changes. Fracture formation pressure coefficient, rotary speed, weight on bit, well depth, torque, bit diameter, inlet flow rate, and drilling fluid density also have relatively high gray relational grades, reflecting ROP variation from the perspectives of formation pressure, rock breaking parameters, interval structure, and drilling fluid conditions.

Relying only on statistical results cannot fully guarantee the engineering rationality of feature screening. ROP prediction has a clear drilling engineering background, so the physical meaning, field availability, information redundancy, and model application value of variables must also be considered. The joint screening step, therefore, introduces engineering mechanism constraints on the basis of statistical results to improve the rationality of the final feature set.

The joint feature screening results and final feature retention decision are shown in Figure 7.

	Variable	Pearson	Spearman	Grey RD	Score	Decision basis	Final
1.	Outlet conductivity	0.493	0.484	0.749	0.575	Integrated evaluation	Retain
2.	Inlet conductivity	0.493	0.395	0.749	0.546	Integrated evaluation	Retain
3.	Depth	−0.516	−0.497	0.587	0.534	Integrated evaluation	Retain
4.	Hook load	−0.527	−0.457	0.653	0.512	Integrated evaluation	Retain
5.	Bit size	0.455	0.472	0.574	0.501	Integrated evaluation	Retain
6.	Formation fracture-pressure coefficient	−0.471	−0.380	0.640	0.497	Integrated evaluation	Retain
7.	Torque	−0.471	−0.413	0.580	0.488	Integrated evaluation	Retain
8.	Inlet density	−0.476	−0.421	0.564	0.487	Integrated evaluation	Retain
9.	ECD	−0.476	−0.413	0.563	0.484	Integrated evaluation	Retain
10.	Outlet density	−0.476	−0.413	0.563	0.484	Integrated evaluation	Retain
11.	Inlet flow rate	0.393	0.398	0.556	0.449	Integrated evaluation	Retain
12.	WOB	0.141	0.125	0.608	0.291	Engineering constraint	Retain
13.	RPM	−0.170	−0.080	0.616	0.289	Engineering constraint	Retain
14.	Standpipe pressure	−0.111	0.045	0.554	0.236	Engineering constraint	Retain
15.	Mud outlet flow rate	−0.315	−0.197	0.540	0.357	Redundant / weak relation	Remove
16.	Total mud pit volume	0.201	0.216	0.581	0.333	Redundant / weak relation	Remove
17.	Pump strokes	0.347	0.232	0.516	0.333	Used for abnormal-condition filtering; while flow retained as more direct hydraulic descriptor	Remove
18.	Outlet temperature	0.154	0.215	0.480	0.283	Redundant / weak relation	Remove
19.	Inlet temperature	0.161	0.203	0.490	0.281	Redundant / weak relation	Remove
20.	Outlet flow rate	−0.156	−0.018	0.581	0.252	Redundant / weak relation	Remove

Pump strokes were excluded from the final input set mainly because they were used for abnormal-condition filtering, while inlet flow rate was retained as the more direct hydraulic-supply descriptor.

Figure 7. Joint feature screening results and final retention description. Note: WOB, RPM, and standpipe pressure were retained based on engineering mechanism constraints, although their integrated scores were relatively low. The final retained inputs include 14 variables; the removed variables include pump strokes, outlet flow rate, mud outlet flow rate, total mud pit volume, inlet temperature, and outlet temperature. Pump strokes are not removed because of strong statistical redundancy with inlet flow rate; instead, they are mainly used as an engineering indicator for abnormal condition filtering, while inlet flow rate is retained as the more direct descriptor of hydraulic supply capacity.

The joint screening results indicate that pump strokes and inlet flow rate are both circulation-related variables, but their statistical redundancy is not strong in the processed dataset. The Pearson and Spearman correlation coefficients between pump strokes and inlet flow rate are 0.347 and 0.232, respectively. Therefore, pump strokes are not excluded solely because of strong correlation with inlet flow rate. The selection decision is instead based on measurement meaning and modeling stability: pump strokes mainly describe the surface pump actuation state, whereas inlet flow rate is a more directly measured descriptor of hydraulic supply capacity entering the circulation system. The weak correlation between

the two variables also indicates possible operating condition switching, response lag, or acquisition asynchrony; retaining pump strokes as a normal input together with inlet flow rate may introduce actuator response inconsistency rather than providing stable independent hydraulic information. For this reason, pump strokes are used primarily as an engineering indicator for abnormal condition filtering, while inlet flow rate is retained in the final input feature set. Mud outlet flow and total mud pit volume mainly reflect the surface circulation state and have relatively limited direct explanatory ability for bit rock breaking efficiency. Inlet and outlet drilling fluid temperatures also show weak direct associations with ROP. Therefore, these variables are not included in the final input feature set.

Although weight on bit, rotary speed, and standpipe pressure are not ranked high in some statistical indicators, they respectively represent the axial load on the bit, rotational cutting capability, and downhole circulation pressure state, all of which are important control parameters affecting ROP in drilling engineering. The relationship between ROP and these parameters is often jointly affected by formation conditions, bit condition, and downhole circulation environment, showing nonlinear coupling. Simply removing them based on statistical scores would weaken the model's ability to represent the actual drilling process. Therefore, under engineering mechanism constraints, weight on bit, rotary speed, and standpipe pressure are retained as key input features.

Combining the statistical analysis results with engineering mechanism judgment, 14 variables are finally determined as input features for the ROP prediction model: well depth, hookload, weight on bit, rotary speed, standpipe pressure, torque, bit diameter, inlet drilling fluid density, outlet drilling fluid density, inlet drilling fluid conductivity, outlet drilling fluid conductivity, inlet flow rate, ECD, and fracture formation pressure coefficient. ROP is used as the output variable. This input–output setting provides a unified data basis for subsequent model training.

The final input and output variables and their engineering meanings are listed in Table 6.

Table 6. Description of final input and output variables.

Variable Type	Variable Name	Engineering Meaning
Input variable	Well depth	Represents interval position and formation change background
Input variable	Hookload	Reflects drill string load and downhole stress state
Input variable	Weight on bit	Represents the axial load applied by the bit to the bottom hole
Input variable	Rotary speed	Represents drill string rotational speed and cutting rock breaking capability
Input variable	Standpipe pressure	Reflects drilling fluid circulation pressure conditions
Input variable	Torque	Reflects drill string rotational resistance and rock breaking difficulty
Input variable	Bit diameter	Represents borehole size and drilling section structure
Input variable	Inlet drilling fluid density	Reflects the density characteristics of inlet drilling fluid
Input variable	Outlet drilling fluid density	Reflects changes in returned drilling fluid density
Input variable	Inlet drilling fluid conductivity	Represents the electrical characteristics of inlet drilling fluid
Input variable	Outlet drilling fluid conductivity	Represents electrical changes in returned drilling fluid
Input variable	Inlet flow rate	Reflects inlet drilling fluid flow rate and circulation supply capacity

Table 6. *Cont.*

Variable Type	Variable Name	Engineering Meaning
Input variable	ECD	Represents equivalent bottom-hole density under circulation
Input variable	Fracture formation pressure coefficient	Reflects formation pressure bearing conditions and safe-drilling background
Output variable	Rate of penetration	Prediction target of the model, representing drilling efficiency

3. IHHO-ET Model Construction

3.1. ExtraTrees Regression Model

After completing data preprocessing and input feature screening, the ROP prediction model is further constructed. ROP is jointly affected by well depth, weight on bit, rotary speed, standpipe pressure, torque, drilling fluid parameters, formation conditions, and other factors, and the variables exhibit strong nonlinear coupling. ExtraTrees is an ensemble tree model that performs regression prediction through multiple randomized decision trees and has good nonlinear representation and noise resistance capabilities. It is suitable for structured drilling parameter data modeling. Therefore, ExtraTrees is used as the basic prediction model in this study, and IHHO is applied to optimize its key hyperparameters.

From the perspective of modeling paradigms, ExtraTrees, bagging ensembles, and random forests all belong to the tree ensemble learning framework and reduce the sensitivity of a single model to local sample fluctuations through the combination of multiple base learners [39,40]. The comparison models used in this study, namely, XGBoost, MLP, and SVR, correspond to gradient boosting, back-propagation neural networks, and support vector regression, respectively [41–43]. These models allow the relative performance of IHHO-ET in the structured drilling-parameter regression task to be evaluated from different algorithmic mechanisms.

The basic idea of ExtraTrees is to construct multiple random regression trees and take the average of their outputs as the final prediction. Suppose that the model contains M regression trees and that the k-th tree outputs $h_k(x)$ for input sample x. The ensemble prediction of ExtraTrees can be expressed as:

$$\hat{y} = \frac{1}{M}\sum_{m=1}^{M} T_m(x) \tag{11}$$

where $\hat{y}(x)$ is the predicted ROP, M is the number of regression trees, and $h_k(x)$ is the output of the k-th regression tree. Through the ensemble averaging of multiple trees, ExtraTrees can reduce the sensitivity of a single tree to local sample fluctuations and improve prediction stability.

The ExtraTrees regression prediction structure is shown in Figure 8.

The ExtraTrees structure includes an input layer, a model layer, and an output layer. The input layer contains the 14 drilling engineering parameters selected in Section 2, the model layer consists of multiple random regression trees, and the output layer gives the predicted ROP. Compared with a single decision tree, ExtraTrees can more fully describe the nonlinear mapping between complex drilling parameters and ROP by integrating multiple randomized tree models.

However, the performance of ExtraTrees is still affected by several key hyperparameters. If the number of trees is insufficient, the model may not fully represent complex relationships; if the tree depth is too large or the leaf node constraints are too weak, the risk of overfitting may increase; and if the maximum number of features is set inappropriately,

the model's use of different input variables may also be affected. Therefore, it is necessary to introduce an optimization algorithm for adaptive hyperparameter search.

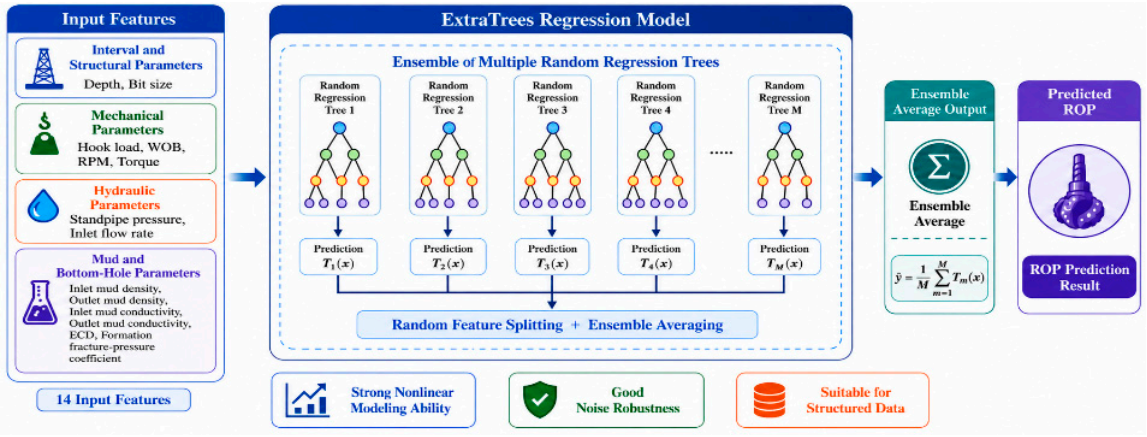

Figure 8. Schematic diagram of ExtraTrees regression prediction structure.

The key ExtraTrees hyperparameters and their main influences are summarized in Table 7.

Table 7. Key ExtraTrees hyperparameters and their functions.

Hyperparameter	Meaning	Main Influence on the Model
n_estimators	Number of regression trees	Affects ensemble scale and prediction stability
max_depth	Maximum depth of a single tree	Controls model complexity and overfitting risk
min_samples_split	Minimum samples required to split an internal node	Affects the subdivision degree of tree structures
min_samples_leaf	Minimum samples in a leaf node	Affects model smoothness and generalization ability
max_features	Maximum number of features considered for node splitting	Affects feature randomness and model diversity

The ExtraTrees hyperparameters considered in the optimization process include the number of trees, maximum depth, minimum samples for node splitting, minimum leaf samples, and maximum features. These parameters affect model performance in terms of tree number, tree depth, node splitting constraints, and feature randomness. Completely relying on manual experience may lead to unstable parameter combinations; therefore, the improved HHO algorithm is used to automatically search for these hyperparameters.

3.2. Improved Harris Hawks Optimization Algorithm

Harris hawks optimization (HHO) is a swarm intelligence optimization algorithm that simulates the cooperative hunting behavior of Harris hawks. The algorithm controls global exploration and local exploitation through changes in prey escape energy and has good applicability to parameter optimization problems. The escape energy can be expressed as:

$$E = 2E_0\left(1 - \frac{t}{T}\right) \tag{12}$$

Although standard HHO has good global search ability, it may still suffer from uneven initial population distribution, insufficient local perturbation, and inadequate exploration–exploitation coordination in complex hyperparameter spaces. In this study, Logistic chaotic initialization, adaptive Gaussian mutation, and a dynamic weighting strategy are introduced as design heuristics for ExtraTrees hyperparameter search. These strategies are

motivated by the above concerns, but they are not treated as independently verified one-to-one remedies. Their effects are evaluated in this paper mainly through cumulative optimizer comparison and ablation results rather than through a separate experimental proof that each strategy independently resolves one specific defect.

It should be noted that, according to the no-free-lunch concept in optimization, no universal optimization algorithm has absolute superiority for all problems [44]. Therefore, it is necessary to design search strategies for specific models and data structures. Contemporary hyperparameter optimization research usually treats grid search, random search, Bayesian optimization, and evolutionary/swarm intelligence methods as important baselines and candidate strategies [45]. The purpose of using IHHO in this study is to test whether a compact hybrid HHO variant can improve ExtraTrees hyperparameter search under a controlled evaluation budget, rather than to prove that each added strategy has a fully isolated mechanism-specific effect.

First, Logistic chaotic mapping is used to improve the distribution of the initial population. The Logistic mapping can be expressed as:

$$z_{t+1} = \mu z_t(1 - z_t) \tag{13}$$

where z_n is the chaotic variable at the n-th iteration and μ is the control parameter. In this study, $\mu = 4$ and the initial chaotic value is fixed at $z_0 = 0.73$, which lies within (0, 1) and avoids the fixed points of the Logistic map. The value z_0 is treated as a fixed implementation constant rather than an optimized hyperparameter; therefore, no claim is made that $z_0 = 0.73$ is universally optimal. The chaotic sequence is then mapped to the parameter space to be optimized:

$$X_{i,j} = L_j + z_{i,j}(U_j - L_j) \tag{14}$$

where X_{i,d} is the position of the i-th individual in the d-th dimension, and LBd and UBd are the lower and upper bounds of the d-th parameter, respectively. This mapping is intended to diversify initial candidate positions in the bounded search space; however, the present experiments do not separately quantify initial coverage rate, convergence curve dispersion, or escape from local optimum events.

Second, an adaptive Gaussian mutation mechanism is introduced to locally perturb candidate solutions near promising regions and to increase local search diversity. Gaussian mutation can be expressed as:

$$X_i' = X_i + P_m(t) \cdot N\left(0, \sigma^2\right) \tag{15}$$

where X_i' is the mutated individual position, X_i is the original position before mutation, $p_m(t)$ is the mutation probability changing with update iteration t, and $N(0, \sigma^2)$ is a Gaussian perturbation with mean 0 and variance σ^2. In the experiments, $p_m(t) = p_{m,max} - (p_{m,max} - p_{m,min})(t/T)$, with $p_{m,max} = 0.30$, $p_{m,min} = 0.05$, and T denoting the number of post-initialization update iterations. The Gaussian perturbation is applied in the normalized parameter space with $\sigma^2 = 0.01$, and out-of-bounds mutated values are clipped to the corresponding search bounds. This mechanism provides random perturbation around promising solutions, but its independent contribution is assessed only indirectly through the ablation results.

Finally, a dynamic weighting strategy is introduced to regulate the position update intensity at different iteration stages. The dynamic weight decreases gradually with the number of iterations and can be expressed as:

$$w(t) = w_{max} - (w_{max} - w_{min})\frac{t}{T} \quad (16)$$

where w(t) is the dynamic weight at the t-th update iteration, w_{max} and w_{min} are the maximum and minimum weights, and T is the number of post-initialization update iterations. In this study, w_{max} = 0.90 and w_{min} = 0.40 are used, so the weight decreases linearly from 0.90 to 0.40 during the update process. This linear decreasing form uses the difference term ($w_{max} - w_{min}$), which is intended to maintain a larger search step in the early stage and gradually narrow the search range in the later stage.

The main IHHO parameter settings are summarized in Table 8. To avoid ambiguity in the evaluation budget, this study distinguishes the initial population evaluation from the post-initialization update iterations. The population size is N = 10, and the number of update iterations after initialization is T = 4. Therefore, the total number of candidate parameter evaluations is N + N × T = 10 + 10 × 4 = 50. Equivalently, the search contains five population-level fitness evaluation rounds when the initial population is included. Thus, the initial population evaluation is included in the reported 50 evaluations, keeping the HHO/IHHO budget comparable with the 50 trial random search and Bayesian optimization settings.

Table 8. Main IHHO parameter settings used in this study.

Parameter	Value	Meaning	Basis or Explanation
z_0	0.73	Initial value of Logistic chaotic map	Selected within (0, 1) and away from fixed points to improve initialization diversity
μ	4.00	Control parameter of Logistic map	A commonly used value corresponding to the fully chaotic state
σ^2	0.01	Variance of Gaussian mutation	Provides small local perturbation in the normalized search space
$p_m(t)$	0.30 → 0.05	Adaptive mutation probability	High early perturbation and lower late-stage disturbance for exploration-exploitation balance
w(t)	0.90 → 0.40	Dynamic weight	Larger early search step and smaller late-stage refinement step
Fitness	Validation RMSE	Optimization objective	Penalizes large prediction errors and is calculated only on the validation set
Population size N	10	Number of Harris hawk individuals	Used in each population-level evaluation round
Update iterations T	4	Post-initialization update iterations	Total budget = 10 initial evaluations + 40 update stage evaluations = 50

Because the additional gains brought by the IHHO strategies are numerically small, these parameter settings, including z_0, σ^2, $p_m(t)$, w(t), N, and T, are not interpreted as universally optimal values. They are used as fixed engineering settings to keep the optimization budget controlled and comparable across repeated runs. Further systematic sensitivity analysis of z_0, population size, update iteration number, and evaluation budget will be needed when the method is transferred to larger multi-well datasets.

Figure 9 summarizes the design logic of the three IHHO improvement branches. Logistic chaotic initialization is intended to improve initial population diversity, adaptive Gaussian mutation provides local perturbation around candidate solutions, and the dynamic weighting strategy coordinates early global search and late local refinement. This

figure is a schematic illustration rather than an experimental convergence curve. Because the original experiments did not store per-iteration best fitness logs, this revision does not reconstruct a convergence curve; the empirical contribution of the strategies is instead judged from the optimizer comparison and ablation results in Section 4.2.

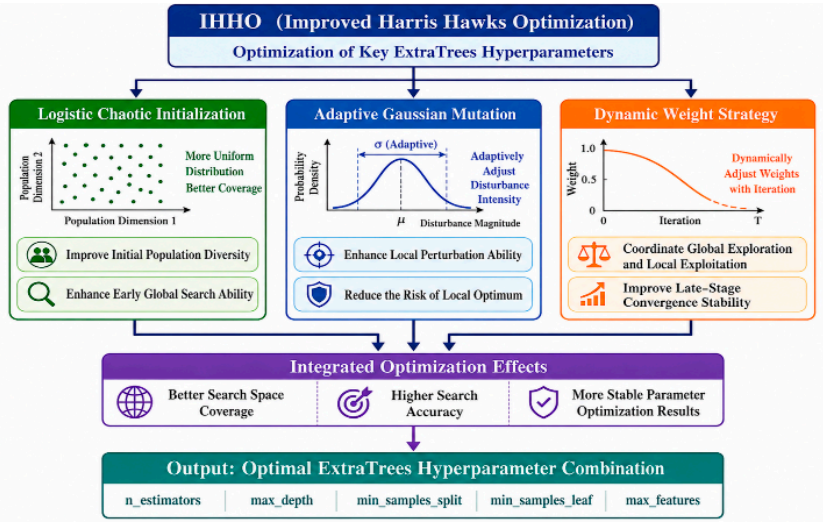

Figure 9. Schematic diagram of IHHO improvement strategies.

3.3. IHHO-ET Implementation Workflow

In the IHHO-ET model, IHHO does not change the ensemble tree structure of ExtraTrees itself but acts as an outer optimizer to search for key ExtraTrees hyperparameter combinations. Each Harris hawk individual corresponds to a candidate hyperparameter set, and the algorithm continuously updates individual positions to find the parameter combination with the minimum validation set error. In this study, IHHO is defined as an outer parameter search mechanism that automatically optimizes key hyperparameters.

Let the position vector of the i-th individual be:

$$X_i = [x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}] \quad (17)$$

The corresponding ExtraTrees hyperparameter combination can then be expressed as:

$$X_i = [\text{n\_estimators}, \text{max\_depth}, \text{min\_samples\_split}, \text{min\_samples\_leaf}, \text{max\_features}]$$

For a given parameter combination θ_i, an ExtraTrees model is trained on the training set, and RMSE on the validation set is calculated as the fitness function:

$$\text{Fitness}(X_i) = \text{RMSE}(X_i) \quad (18)$$

where RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(y_k - \hat{y}_k)^2} \quad (19)$$

where y_i is the true ROP of the i-th sample, $\hat{y}_i$ is the model prediction, and n is the number of validation samples. A smaller RMSE indicates lower prediction error and better fitness for the current parameter combination. RMSE was used as the fitness function because large ROP prediction errors are more critical for drilling parameter decision support. MAE and MAPE were not used as optimization objectives but were retained as complementary

evaluation metrics. In particular, MAPE may be unstable for low-ROP samples because of the small denominator effect. The final optimization objective is:

$$X^* = \arg\min_X \text{Fitness}(X) \quad (20)$$

where θ* is the optimal hyperparameter combination obtained through the search.

The IHHO-ET hyperparameter optimization workflow is shown in Figure 10.

The IHHO-ET optimization workflow follows the sequence of population initialization, parameter decoding, model training, fitness calculation, individual updating, optimal parameter output, and model reconstruction. Specifically, Logistic chaotic mapping is first used to generate the initial population with N = 10 individuals, and these 10 initial candidates are evaluated on the validation set. The algorithm then performs T = 4 post-initialization update iterations, with 10 candidate evaluations in each iteration. Thus, the total search budget is 10 initial evaluations plus 40 update stage evaluations, namely, 50 candidate parameter evaluations. After the search ends, the optimal hyperparameter combination is used to retrain ExtraTrees and obtain the final IHHO-ET prediction model.

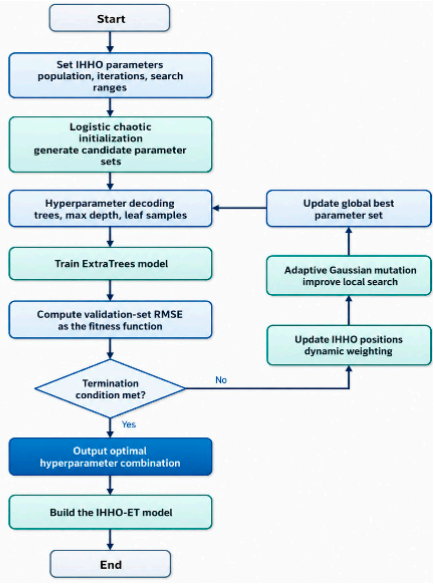

Figure 10. Flowchart of IHHO-ET hyperparameter optimization.

Through parameter encoding, fitness function design, and search workflow construction, IHHO is effectively coupled with ExtraTrees. The main workflow is retained in the main text, while the detailed algorithm pseudocode is moved to Appendix A (Table A1) to keep the method section concise.

4. Results and Discussion

4.1. Experimental Settings and Evaluation Metrics

After data preprocessing, feature screening, and IHHO-ET model construction, the prediction performance of the model is further experimentally validated. The experimental data are mainly the Well Z samples processed in Section 2. The input variables are the 14 features of well depth, hookload, weight on bit, rotary speed, standpipe pressure, torque, bit diameter, inlet drilling fluid density, outlet drilling fluid density, inlet drilling fluid conductivity, outlet drilling fluid conductivity, inlet flow rate, ECD, and fracture formation pressure coefficient; the output variable is ROP.

To reduce interval distribution bias that may be caused by simple random splitting, this study uses a depth bucket stratified splitting strategy to construct the training, validation,

and test sets at a ratio of 70%, 15%, and 15%, respectively. The training set is used for model parameter learning, the validation set is used for IHHO fitness calculation and hyperparameter selection, and the test set is used only for final prediction performance evaluation. Standardization parameters are estimated only from the training set and then applied to the validation and test sets to avoid premature participation of test set information in model training or parameter optimization.

In this implementation, the retained depth interval (200–7919 m) was divided into 10 equal-width depth buckets, and stratified sampling was conducted within each bucket to maintain comparable interval coverage in the training, validation, and test subsets. The number of 10 buckets was selected as a compromise between interval representativeness and sufficient samples per bucket. Fewer buckets may not adequately preserve depth-related distribution differences, whereas more buckets may lead to small sample sizes in some intervals and unstable splitting.

The experimental workflow shown in Figure 11 includes same-well modeling for Well Z and same-region cross-well validation. The left side corresponds to the same-well modeling and testing process for Well Z, in which the training, validation, and test sets are constructed by depth bucket stratified splitting. The same-well branch further includes fair validation set tuning for all comparison models, 10 independent random seeds, mean ± standard deviation reporting, Wilcoxon signed-rank testing, optimizer comparison, and a with-depth versus without-depth test. The right side corresponds to the same-region leave-one-well-out validation process, in which Wells A, B, C, and Z in the same Tarim Oilfield region are each used as the independent test well in turn; IHHO-ET is additionally repeated with random seeds 0–9 in each fold to evaluate cross-well stability. This design provides a more complete evaluation of prediction accuracy, result stability, and same-region applicability.

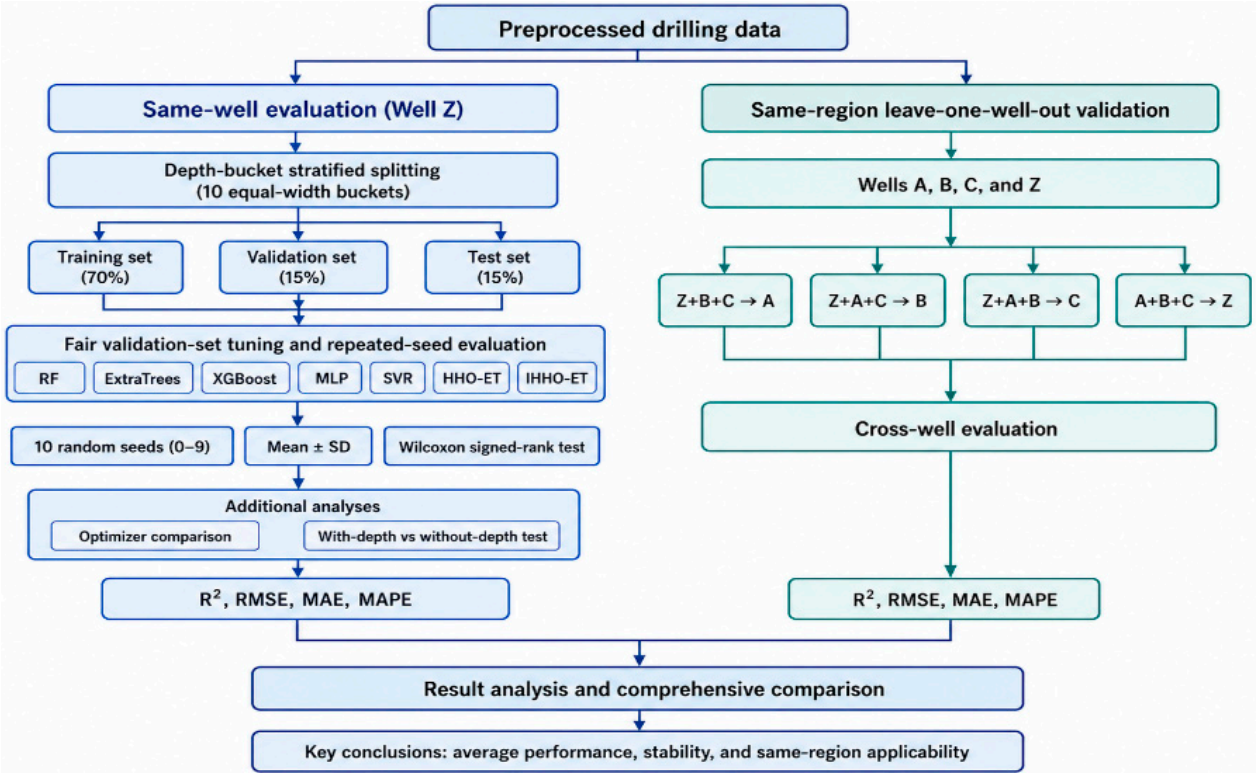

Figure 11. Schematic diagram of dataset splitting and experimental workflow.

A unified experimental setting is adopted for the software environment, input variables, comparison models, and evaluation metrics.

The experimental environment and main parameter settings are summarized in Table 9.

The experimental settings use the same input feature system, comparison model group, and evaluation metrics across all experiments. ExtraTrees, as the base model of IHHO-ET, is retained separately in the comparison system to directly analyze performance changes before and after introducing HHO/IHHO hyperparameter optimization. Random Forest, XGBoost, MLP, and SVR are included as representative ensemble, boosting, neural network, and kernel-based baselines. To ensure a fair comparison, all baseline models are tuned using validation set RMSE under predefined hyperparameter search spaces, and the number of candidate parameter evaluations is kept comparable across models.

Table 9. Experimental environment and main parameter settings.

Category	Details
Programming language	Python 3.10 (Python Software Foundation, Beaverton, OR, USA)
Main libraries	NumPy 1.26.4, pandas 2.2.2, scikit-learn 1.5.0, XGBoost 2.0.3, SciPy 1.13.1, and Matplotlib 3.8.4; official project documentation accessed on 31 May 2026.
Data source	Drilling engineering parameter data from Well Z in the Tarim Oilfield
Input features	Fourteen key engineering parameters retained after feature screening in Section 2
Output variable	Rate of penetration (ROP)
Dataset split	Training, validation, and test sets at a ratio of 70%/15%/15%
Splitting strategy	Depth bucket stratified splitting with 10 equal-width depth intervals
Repeated evaluation	10 independent runs with random seeds from 0 to 9
Evaluation metrics	R^2, RMSE, MAE, and MAPE
Comparison models	Random Forest, ExtraTrees, XGBoost, MLP, and SVR
Baseline model settings	All baseline models were tuned using validation set RMSE under predefined hyperparameter search spaces
Search budget	50 candidate parameter evaluations for each baseline model and optimizer; for IHHO/HHO, this includes 10 initial population evaluations and 40 update stage evaluations (N = 10, T = 4 post-initialization update iterations)
Optimized model	IHHO-ET
Optimization object	Key ExtraTrees hyperparameters, including n_estimators, max_depth, min_samples_split, min_samples_leaf, and max_features
Fitness function	Validation set RMSE
Final reporting	Mean ± standard deviation over 10 independent runs on the test set

The ExtraTrees implementation follows the ExtraTreesRegressor API reference in scikit-learn [15], whereas the overall experimental implementation, including model fitting and evaluation utilities, relies on the scikit-learn user guide conventions [46]. To reduce information leakage risk between model selection and performance evaluation [47], this study separates the functions of the training, validation, and test sets and uses only the test set for final performance reporting. To reduce the influence of random data splitting and stochastic model initialization, all same-well comparison models and ablation variants are independently run 10 times using random seeds from 0 to 9, and the final test set results are reported as mean ± standard deviation. For same-region leave-one-well-out validation, IHHO-ET is also repeated under seeds 0–9; the corresponding results are reported later in Section 4.3.

The hyperparameter search spaces used for fair model comparison are listed in Table 10.

Table 10. Hyperparameter search spaces used for fair model comparison.

Model	Tuned Hyperparameters	Search Space
Random Forest	n_estimators; max_depth; min_samples_split; min_samples_leaf; max_features	n_estimators: {100, 200, 300, 500}; max_depth: {None, 5, 10, 20, 30, 40}; min_samples_split: {2, 5, 10}; min_samples_leaf: {1, 2, 4}; max_features: {sqrt, log2, 0.6, 0.8, 1.0}
ExtraTrees	n_estimators; max_depth; min_samples_split; min_samples_leaf; max_features	n_estimators: {100, 200, 300, 500}; max_depth: {None, 5, 10, 20, 30, 40}; min_samples_split: {2, 5, 10}; min_samples_leaf: {1, 2, 4}; max_features: {sqrt, log2, 0.6, 0.8, 1.0}
XGBoost	n_estimators; max_depth; learning_rate; subsample; colsample_bytree; reg_lambda	n_estimators: {100, 200, 300, 500}; max_depth: {3, 4, 5, 6, 8}; learning_rate: {0.01, 0.03, 0.05, 0.10}; subsample and colsample_bytree: {0.6, 0.8, 1.0}; reg_lambda: {0.1, 1.0, 5.0, 10.0}
MLP	hidden_layer_sizes; alpha; learning_rate_init; batch_size	hidden_layer_sizes: {(64), (128), (64, 32), (128, 64)}; alpha: {0.0001, 0.001, 0.01}; learning_rate_init: {0.0005, 0.001, 0.005}; batch_size: {32, 64, 128}
SVR	C; gamma; epsilon; kernel	C: {1, 10, 50, 100, 200}; gamma: {scale, auto, 0.01, 0.001}; epsilon: {0.01, 0.05, 0.10, 0.20}; kernel: RBF
HHO-ET/IHHO-ET	n_estimators; max_depth; min_samples_split; min_samples_leaf; max_features	n_estimators: [100, 500]; max_depth: [3, 40]; min_samples_split: [2, 10]; min_samples_leaf: [1, 5]; max_features: [0.4, 1.0]

For the baseline models, validation set RMSE was used to select the best hyperparameter combination within the predefined search spaces. The number of candidate parameter evaluations was controlled to be comparable among the models and optimization strategies. This design avoids comparing a heavily optimized proposed model with untuned baseline models.

To comprehensively evaluate prediction performance, this study selects the coefficient of determination, root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) as evaluation metrics. Suppose that the true ROP, predicted value, mean true value, and sample number are defined for the i-th sample; the metrics are defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \quad (21)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (22)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i | \quad (23)$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n} | \frac{y_i - \hat{y}_i}{y_i} | \quad (24)$$

An R^2 value closer to 1 indicates stronger fitting ability for the overall variation trend of ROP; smaller RMSE and MAE values indicate lower absolute prediction errors; and a smaller MAPE indicates lower relative error. Because low-value intervals exist in the ROP samples, MAPE may be amplified by small denominators, so it is used as an auxiliary indicator together with R^2, RMSE, and MAE.

To avoid ambiguity in the final typeset version, the mathematical expressions for the IQR interval, Spearman coefficient, gray relational analysis, IHHO update rules, and evaluation metrics were checked for consistent symbols, subscripts, superscripts, and Greek letter notation. In particular, R2, σ^2, μ, ρ, $p_m(t)$, w_{max}, and w_{min} were kept consistent between the equations, parameter tables, and textual descriptions.

The comparison models in this study include Random Forest, ExtraTrees, XGBoost, MLP, and SVR. Random Forest and ExtraTrees are tree ensemble models suitable for tabular nonlinear data; XGBoost is a boosting tree model with strong iterative error correction capability; MLP represents a shallow neural network method; and SVR represents a kernel function regression method. By comparing these models, the prediction performance of IHHO-ET can be evaluated from different algorithmic paradigms.

4.2. Same-Well Prediction and Ablation Results

All models are evaluated on the same-well test set using the same training, validation, and testing workflow. To avoid an unfair comparison caused by using default baseline parameters, all baseline models are tuned using the validation set before final test set evaluation.

All models were independently run 10 times using random seeds 0–9. Table 11 reports the test set mean and standard deviation, replacing the previous single-run point estimates.

Table 11. Repeated seed prediction performance comparison of different models on the same-well test set.

Model	R^2	RMSE	MAE	MAPE
IHHO-ET	0.910 ± 0.004	0.871 ± 0.019	0.246 ± 0.012	8.43 ± 0.46%
HHO-ET	0.909 ± 0.005	0.876 ± 0.022	0.249 ± 0.013	8.51 ± 0.49%
Random Search-ET	0.895 ± 0.007	0.943 ± 0.034	0.289 ± 0.018	9.35 ± 0.61%
Random Forest	0.858 ± 0.012	1.095 ± 0.049	0.532 ± 0.055	18.20 ± 2.10%
XGBoost	0.843 ± 0.015	1.151 ± 0.058	0.615 ± 0.061	20.80 ± 2.45%
MLP	0.724 ± 0.035	1.525 ± 0.102	0.918 ± 0.077	31.40 ± 3.50%
SVR	0.670 ± 0.041	1.668 ± 0.121	0.982 ± 0.090	33.60 ± 3.80%

Table 12 summarizes the Wilcoxon signed-rank test based on RMSE. IHHO-ET is significantly better than Random Search-ET and the other baseline models, whereas its difference from HHO-ET is not significant at the 0.05 level.

Table 12. Wilcoxon signed-rank test between IHHO-ET and comparison models based on RMSE.

Comparison	Metric	*p*-Value	Significant at 0.05	Interpretation
IHHO-ET vs. HHO-ET	RMSE	0.109	No	Not significant
IHHO-ET vs. Random Search-ET	RMSE	0.006	Yes	Significant improvement
IHHO-ET vs. Random Forest	RMSE	0.002	Yes	Significant
IHHO-ET vs. XGBoost	RMSE	0.002	Yes	Significant
IHHO-ET vs. MLP	RMSE	0.002	Yes	Significant
IHHO-ET vs. SVR	RMSE	0.002	Yes	Significant

The compact comparison of ExtraTrees optimization strategies is given in Table 13.

Under the same evaluation budget, the untuned Default ET obtains a test RMSE of 1.005, whereas Random Search-ET reduces the RMSE to 0.943. Bayesian Optimization-ET, HHO-ET, and IHHO-ET further reduce the RMSE to 0.884, 0.876, and 0.871, respectively. For HHO-ET and IHHO-ET, the 50 model evaluations include the initial population evaluation and the subsequent update stage evaluations, so the budget is matched to the 50 trial

random search and Bayesian optimization settings. These results indicate that IHHO-ET achieves the lowest test RMSE under the compared settings, but its advantage over HHO-ET and Bayesian Optimization-ET should be interpreted as a stable rather than absolute optimization benefit.

Table 13. Compact comparison of ExtraTrees optimization strategies under the same evaluation budget.

Optimizer	Model Evaluations	Test RMSE	Interpretation
Default ET	0	1.005	Untuned baseline
Random Search-ET	50 trials	0.943	Random search tuned ExtraTrees
Bayesian Optimization-ET	50 trials	0.884	Bayesian tuned ET
HHO-ET	50 incl. initial population	0.876	HHO tuned ET
IHHO-ET	50 incl. initial population	0.871	IHHO tuned ET

The ablation comparison of ExtraTrees under different optimization strategies is summarized in Table 14.

Table 14. Ablation comparison of ExtraTrees under different optimization strategies. Values are reported as mean (SD); L, G, and W denote Logistic initialization, Gaussian mutation, and dynamic weight, respectively.

Model	Optimization Setting	R^2	RMSE	MAE	MAPE
Random Search-ET	Random search	0.895 (0.007)	0.943 (0.034)	0.289 (0.018)	9.35 (0.61%)
HHO-ET	HHO	0.909 (0.005)	0.876 (0.022)	0.249 (0.013)	8.51 (0.49%)
IHHO-ET-L	HHO+L	0.909 (0.005)	0.875 (0.021)	0.248 (0.013)	8.49 (0.48%)
IHHO-ET-LG	HHO+L+G	0.910 (0.004)	0.873 (0.020)	0.247 (0.012)	8.46 (0.47%)
IHHO-ET	HHO+L+G+W	0.910 (0.004)	0.871 (0.019)	0.246 (0.012)	8.43 (0.46%)

The ablation results show that HHO-based hyperparameter optimization provides the main improvement over the Random Search-ET baseline. The additional IHHO strategies lead to only small but directionally consistent refinements, and the Wilcoxon test in Table 12 shows that the RMSE difference between IHHO-ET and HHO-ET is not significant at the 0.05 level.

Therefore, the IHHO strategies are interpreted mainly as search stability and local refinement design heuristics rather than as independently verified solutions to initialization coverage, local optimum escape, and late-stage convergence problems. Quantitative diagnostics such as initial population coverage, convergence curve dispersion, and escape counts from local optima are not separately measured in this dataset and are left for future work. This limitation is also why Figure 9 is retained as a schematic mechanism diagram rather than replaced by an unsupported convergence curve.

Overall, HHO hyperparameter optimization is the main source of the performance improvement of ExtraTrees. The complete IHHO-ET provides a slight but directionally consistent refinement over HHO-ET. This revised interpretation avoids overstating the marginal gain of the three improvement mechanisms, clarifies the 50 evaluation budget, and emphasizes their cumulative contribution to search robustness under the current experimental setting.

Sensitivity Analysis of Key IHHO Parameters

To assess whether the reported performance depends on the specific values of the fixed IHHO constants, a sensitivity analysis was carried out for the two parameters that most

directly affect the search process: the Logistic chaotic initial value z_0 and the population size N. Each configuration was evaluated independently over three random seeds on the same-well test set, with all other settings unchanged, and the mean and standard deviation of the test RMSE were recorded. The adopted setting rows (z_0 = 0.73 and N = 10) reproduce the corresponding seeds of the main same-well experiment in Table 11.

Table 15 reports the effect of varying z_0 over the interval (0, 1) while keeping N = 10, T = 4, μ = 4, and σ^2 = 0.01 fixed. Across the tested values, the mean test RMSE varies within a narrow band of about 0.008 (roughly 0.9% in relative terms), and the adopted value z_0 = 0.73 lies at the lower end of this band. This indicates that the chaotic initial value mainly affects the starting distribution of the population rather than the final converged accuracy and that the result is not sensitive to the specific choice of z_0 within (0, 1).

Table 15. Sensitivity of same-well test RMSE to the Logistic chaotic initial value z_0 (N = 10, T = 4, μ = 4, σ^2 = 0.01; three seeds per setting).

z_0	RMSE (Mean ± SD)	Interpretation
0.10	0.878 ± 0.017	Within the stable band
0.30	0.875 ± 0.016	Within the stable band
0.50	0.883 ± 0.017	Within the stable band
0.70	0.876 ± 0.016	Within the stable band
0.73	0.874 ± 0.017	Adopted value; lowest mean RMSE
0.90	0.879 ± 0.017	Within the stable band

Table 16 reports the effect of varying the population size N while keeping the evaluation budget comparable. The mean test RMSE changes by about 0.023 (roughly 2.7%) across the tested range. A small population (N = 5) gives a clearly higher RMSE because the search budget is insufficient, whereas N = 10, 15, and 20 produce closely similar results, with N = 10 giving the lowest mean RMSE under the controlled 50 evaluation budget. These results support the use of N = 10 as an economical yet sufficient setting and confirm that the final accuracy is stable with respect to moderate changes in N.

Table 16. Sensitivity of same-well test RMSE to the population size N under a comparable evaluation budget (three seeds per setting).

N	T	Total Evaluations	RMSE (Mean ± SD)
5	9	50	0.898 ± 0.008
10	4	50	0.874 ± 0.017
15	3	60	0.875 ± 0.013
20	2	60	0.881 ± 0.014

Overall, the sensitivity analysis shows that the final ExtraTrees prediction accuracy is robust to the specific choice of z_0 and to moderate changes in the population size N. The fixed values z_0 = 0.73 and N = 10 are, therefore, adopted as stable engineering settings that keep the optimization budget controlled and the repeated seed comparison consistent, rather than as uniquely optimal hyperparameters.

From the perspective of model type, the HHO-optimized ExtraTrees variants perform best overall, while the Random Search-ET baseline also shows strong performance among the non-swarm-optimized models. Random Forest and XGBoost show intermediate prediction performance and are generally better than MLP and SVR. In contrast, MLP and SVR have larger errors in this task, indicating that, under the current sample size, input feature structure, and obvious depth stratification, shallow neural network and kernel method models are less adaptable to data distribution variations.

In summary, the same-well test set results show that IHHO-ET can further improve prediction accuracy through hyperparameter optimization while maintaining the noise resistance and nonlinear modeling capability of ExtraTrees. This result provides a basic reference for subsequent same-region cross-well leave-one-well-out validation.

4.3. Same-Region Leave-One-Well-Out Validation

On the basis of the same-well test set experiment, to further investigate model applicability to unseen wells in the same region, this study introduces Wells A, B, and C, which are located in the same Tarim Oilfield region as Well Z, as external validation wells. Together with Well Z, they form a four-well leave-one-well-out validation scheme. Wells A, B, and C have regional geological backgrounds and drilling engineering environments similar to those of Well Z, but differences still exist among wells in drilling interval distribution, ROP range, drilling fluid state, and operating parameter combinations. In the revised experiment, IHHO-ET is repeated 10 times with random seeds 0–9 in each cross-well fold.

In the specific experiments, Wells A, B, C, and Z are used in turn as independent test wells, and samples from the remaining three wells are used for training and validation. For example, when Well A is used as the test well, Wells Z, B, and C are used as training and validation data sources; when Well Z is used as the test well, Wells A, B, and C are used as training and validation data sources. The data from Wells A, B, and C and from Well Z are processed using the same preprocessing workflow described in Section 2, including abnormal condition recognition, invalid sample filtering, outlier treatment, smoothing and denoising, and unified feature construction, and a consistent input feature set is extracted on this basis. This setting prevents highly correlated samples from the same well from appearing simultaneously in the training and test sets, making the cross-well evaluation results closer to practical neighboring well prediction applications within the same region.

The leave-one-well-out validation results are summarized in Table 17.

Table 17. Comparison of leave-one-well-out validation results. IHHO-ET values are reported as mean (SD) over 10 random seeds; the other model rows are retained as comparison references from the same leave-one-well-out setting.

Test Well	Training-Well Combination	Model	R^2	RMSE	MAE	MAPE
Well A	Z + B + C	Random Forest	0.690	1.520	0.865	49.51%
Well A	Z + B + C	ExtraTrees	0.706	1.481	0.814	48.54%
Well A	Z + B + C	XGBoost	0.671	1.567	0.965	54.38%
Well A	Z + B + C	MLP	0.568	1.795	1.132	57.04%
Well A	Z + B + C	SVR	0.564	1.804	1.060	49.56%
Well A	Z + B + C	IHHO-ET	0.713 (0.005)	1.463 (0.016)	0.855 (0.009)	52.63 (0.77%)
Well B	Z + A + C	Random Forest	0.675	1.528	0.985	46.71%
Well B	Z + A + C	ExtraTrees	0.676	1.525	0.993	48.79%
Well B	Z + A + C	XGBoost	0.664	1.553	1.007	47.70%
Well B	Z + A + C	MLP	0.624	1.644	1.074	46.45%
Well B	Z + A + C	SVR	0.633	1.624	1.010	39.67%
Well B	Z + A + C	IHHO-ET	0.678 (0.004)	1.520 (0.014)	0.981 (0.010)	46.77 (0.91%)
Well C	Z + A + B	Random Forest	0.603	1.773	1.029	49.25%
Well C	Z + A + B	ExtraTrees	0.599	1.782	1.031	47.09%
Well C	Z + A + B	XGBoost	0.594	1.794	1.049	48.69%
Well C	Z + A + B	MLP	0.565	1.856	1.120	52.01%
Well C	Z + A + B	SVR	0.555	1.877	1.099	44.54%
Well C	Z + A + B	IHHO-ET	0.601 (0.006)	1.778 (0.021)	1.026 (0.014)	46.59 (1.07%)

Table 17. *Cont.*

Test Well	Training-Well Combination	Model	R^2	RMSE	MAE	MAPE
Well Z	A + B + C	Random Forest	0.775	1.324	0.789	32.45%
Well Z	A + B + C	ExtraTrees	0.792	1.273	0.749	31.19%
Well Z	A + B + C	XGBoost	0.753	1.387	0.855	35.14%
Well Z	A + B + C	MLP	0.569	1.832	1.107	45.76%
Well Z	A + B + C	SVR	0.565	1.841	1.012	36.21%
Well Z	A + B + C	IHHO-ET	0.791 (0.005)	1.275 (0.014)	0.778 (0.009)	32.65 (0.68%)
Average	-	Random Forest	0.686	1.536	0.917	44.48%
Average	-	ExtraTrees	0.693	1.515	0.897	43.90%
Average	-	XGBoost	0.671	1.575	0.969	46.48%
Average	-	MLP	0.582	1.782	1.108	50.32%
Average	-	SVR	0.579	1.787	1.045	42.50%
Average	-	IHHO-ET	0.696 (0.005)	1.509 (0.016)	0.910 (0.010)	44.66 (0.86%)

The four leave-one-well-out validation experiments use Wells A, B, C, and Z as test wells in turn. Overall, under cross-well validation, the performance of all models decreases to varying degrees compared with same-well splitting, indicating that inter-well differences in formation conditions, operating condition distributions, and parameter structures significantly increase the difficulty of ROP prediction. Model rankings also vary across different test wells, suggesting that cross-well prediction performance is influenced not only by model structure but also by distributional differences between training and test wells. For IHHO-ET, the repeated seed results show small standard deviations within each fold, indicating that the reported cross-well performance is relatively stable with respect to the random seed, although the mean accuracy is much lower than in same-well testing. In terms of average R^2, RMSE, and MAE, Random Forest, ExtraTrees, and IHHO-ET show relatively similar overall performance and outperform MLP and SVR, indicating that tree ensemble models are more suitable for this type of structured drilling parameter data. It should be noted that IHHO-ET is not optimal for all average metrics in cross-well validation; therefore, its cross-well applicability should be interpreted as relatively stable within the same region rather than absolutely superior in all cases.

Average performance metrics in leave-one-well-out validation indicate that ExtraTrees, Random Forest, and IHHO-ET have relatively similar average R^2, RMSE, and MAE values, suggesting that tree ensemble models have good comprehensive adaptability for same-region cross-well ROP prediction. Based on the corrected repeated seed values, IHHO-ET gives an average R^2 of 0.696 ± 0.005 and an average RMSE of 1.509 ± 0.016, while the auxiliary average MAPE is $44.66 \pm 0.86\%$. ExtraTrees has the lowest average MAE among the point estimate comparison rows, and SVR has the lowest average MAPE but is less favorable in R^2, RMSE, and MAE. Therefore, the cross-well results are interpreted as same-region prediction stability and near-leading performance rather than absolute superiority of IHHO-ET in all metrics.

It should be noted that MAPE is easily affected by small ROP values in cross-well validation. When the true value is small, even a small absolute error may be greatly amplified in percentage form; therefore, model performance should not be judged solely by MAPE. In cross-well scenarios, R^2, RMSE, and MAE should be considered as the primary indicators, while MAPE is used only as an auxiliary relative error indicator. The relatively high cross-well MAPE values also mean that the present model should not be regarded as sufficiently accurate for direct real-time closed-loop drilling control without additional

well-specific calibration. Figure 12 indicates that tree models are generally more stable in cross-well validation and better match the modeling characteristics of structured ROP data.

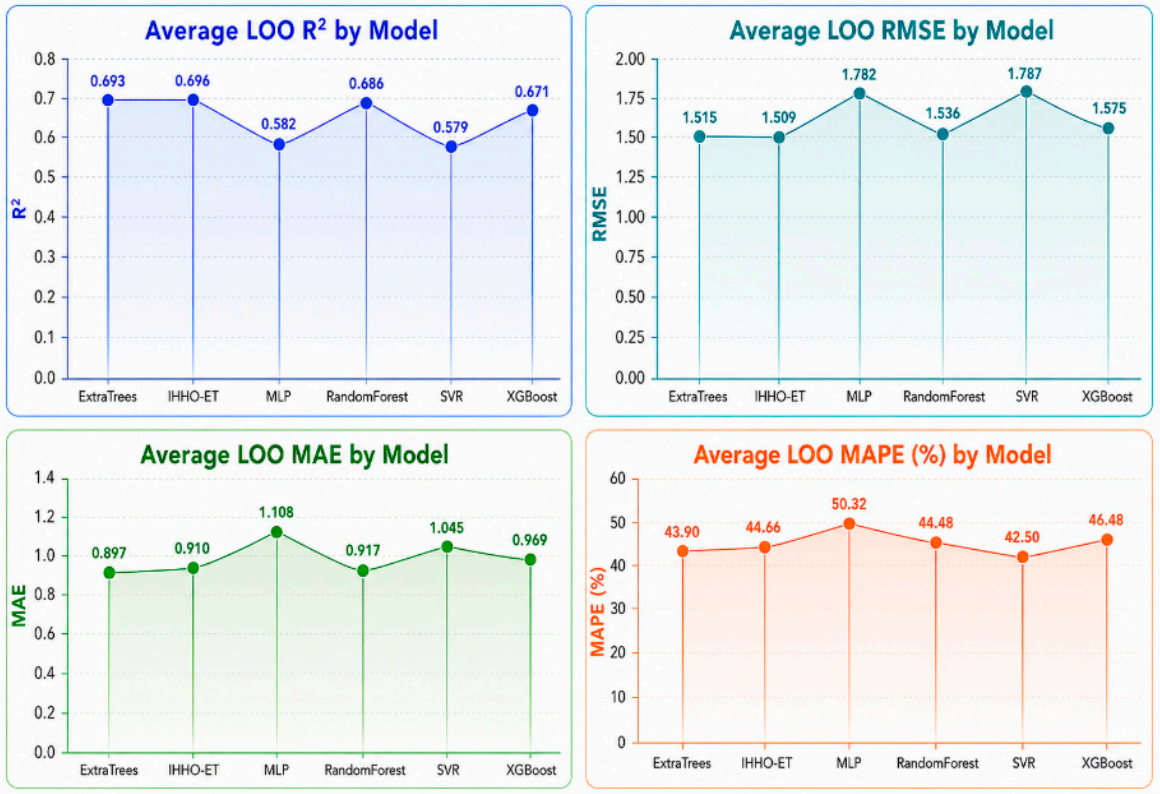

Figure 12. Comparison of average performance metrics in same-region leave-one-well-out validation.

A special case appears in the Well B fold: SVR obtains a lower MAPE than ExtraTrees and Random Forest, although its R^2, RMSE, and MAE are still worse than those of the tree ensemble models. This does not indicate better overall predictive ability of SVR. Instead, MAPE emphasizes relative error and is sensitive to the denominator distribution; when the test well contains more low-to-medium and relatively concentrated ROP samples, a smoother SVR prediction may reduce some percentage errors while still producing larger absolute errors and weaker trend explanation. Therefore, the Well B SVR result is interpreted as a metric-specific phenomenon rather than as evidence that SVR is more suitable for cross-well ROP prediction.

The true and predicted ROP values of IHHO-ET on each test well are compared in Figure 13.

To further examine the fitting effect of IHHO-ET in same-region cross-well validation, the true and predicted ROP values for Wells A, B, C, and Z are compared. The results show that IHHO-ET can follow the overall variation trend of ROP on all four test wells, especially in the low-to-medium ROP range, where the predicted and true curves have good synchronization. This indicates that IHHO-ET can still capture the overall depth-related variation pattern of ROP under inter-well distribution differences.

However, the model still shows some lag or smoothing when responding to local peaks and abrupt change intervals, manifested as a smaller fluctuation amplitude in the predicted curve than in the true curve. This indicates that under cross-well prediction conditions, the model is better at fitting the overall trend but is relatively limited in describing a small number of sudden high-value samples and strongly fluctuating local intervals. This may be caused partly by distribution shift between training and test wells and partly by the relatively small number of extreme condition samples and insufficient learning of high-

fluctuation intervals. Across different test wells, IHHO-ET shows better trend tracking performance on Wells Z and A, whereas it still reflects the overall trend on Wells B and C but has more obvious local fluctuation errors. This suggests that inter-well differences in operating conditions and sample structures directly affect prediction accuracy and further confirms that cross-well ROP prediction is more challenging than same-well sample splitting experiments.

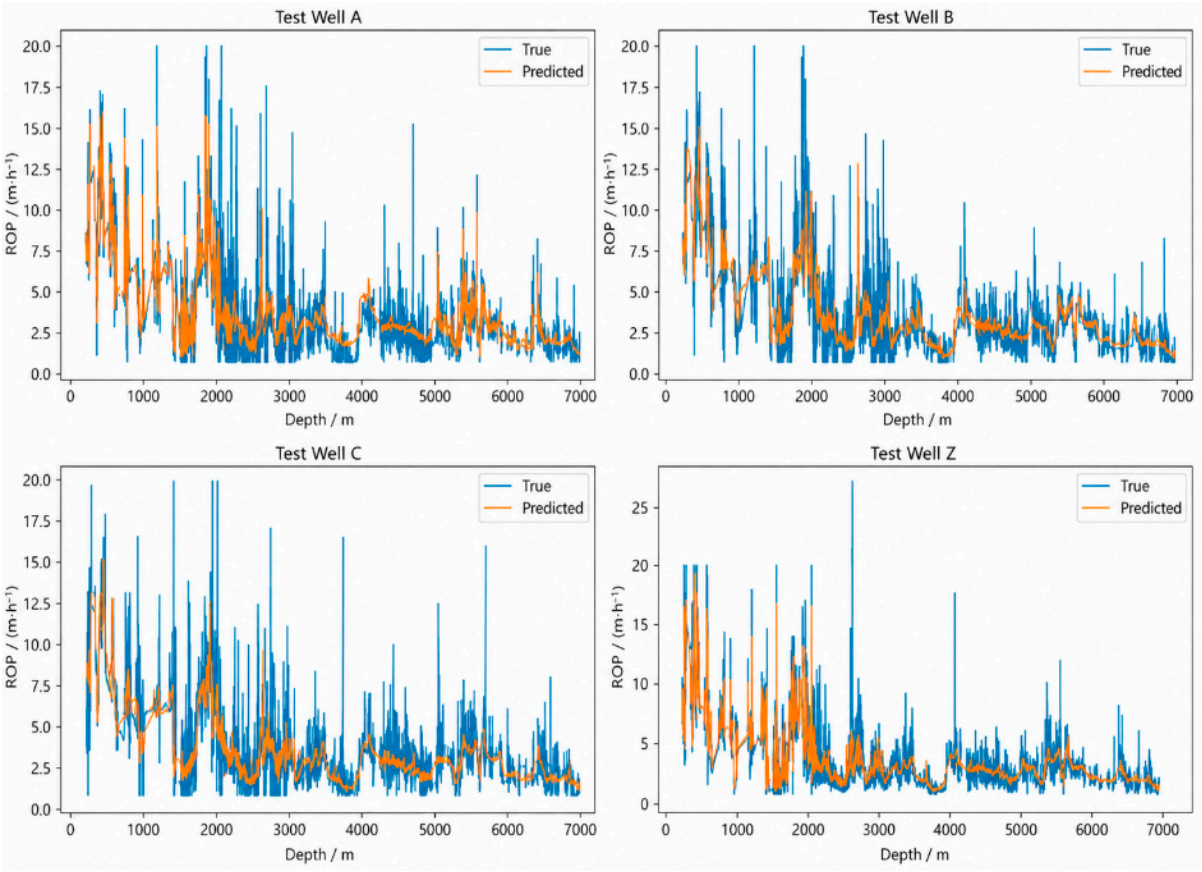

Figure 13. Comparison of true and predicted values of IHHO-ET on each test well.

The residual histograms and KDE curves of IHHO-ET on the test wells are shown in Figure 14.

The residual histograms and KDE curves of IHHO-ET on the four test wells show that most residuals are concentrated near zero and exhibit a clear unimodal pattern. The vertical zero residual reference line indicates that the main error mass remains close to zero, suggesting that the model maintains reasonable error control capability in same-region leave-one-well-out validation. This result is consistent with the preceding R^2, RMSE, and MAE analysis. Meanwhile, the density curves show varying degrees of right-skewed tails, which means that the model still underestimates true ROP for some high-value or abrupt change samples. This finding is also consistent with the limited response to local peaks observed in the true-predicted comparisons.

From the perspective of well differences, the residual shapes of Wells A, B, and C are generally similar, with a concentrated central region and gradual attenuation on both sides. The residuals of Well Z are also centered near zero, but its tail range is relatively wider under the influence of local high-value samples. These differences indicate that the error distribution is affected not only by average error magnitude but also by inter-well variations in operating conditions and sample structures. Overall, IHHO-ET maintains a relatively stable residual distribution pattern across same-region test wells,

although its ability to describe extreme samples and local abrupt fluctuations still needs further improvement.

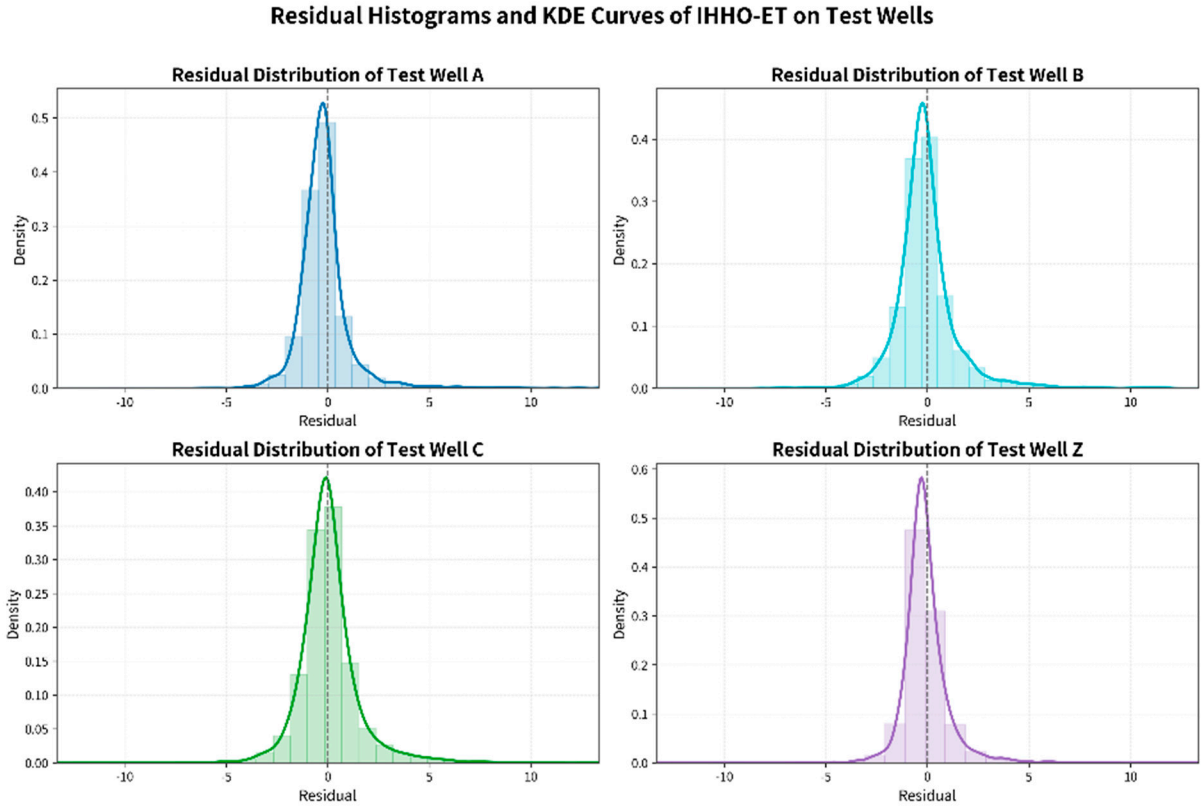

Figure 14. Residual histograms and KDE curves of IHHO-ET on test wells.

4.4. Discussion and Limitations

The combined same-well testing and same-region leave-one-well-out validation results show a clear gap between interpolation-like same-well testing and unseen-well prediction. IHHO-ET achieves $R^2 = 0.910 \pm 0.004$ and RMSE = 0.871 ± 0.019 on the same-well test set, whereas the repeated seed leave-one-well-out average decreases to $R^2 = 0.696 \pm 0.005$ and RMSE = 1.509 ± 0.016. This decline confirms that cross-well ROP prediction is affected not only by model accuracy but also by distribution shift among wells. The shift may arise from differences in depth intervals, bit diameter proportions, drilling fluid properties, hydraulic settings, formation pressure background, and the frequency of local high-ROP or low-ROP intervals.

From the engineering perspective, the high cross-well MAPE of IHHO-ET ($44.66 \pm 0.86\%$) should be interpreted cautiously. It is not acceptable as a standalone criterion for automated parameter control because low-ROP samples can amplify percentage error and because local peaks are still smoothed in unseen-well prediction. Therefore, the current cross-well results are more suitable for offline trend analysis, preliminary drilling parameter screening, and same-region decision support after local calibration. Future applications should incorporate domain adaptation strategies, such as per-well distribution alignment, sample reweighting, transfer calibration with a small number of target-well samples, feature-invariant learning, or multi-well hierarchical modeling, to reduce inter-well distribution shift.

Meanwhile, using well depth as an input variable involves an explicit accuracy–transferability trade-off. The with-depth versus without-depth test shows that removing well depth decreases the average R^2 from 0.910 to 0.823 and increases the average RMSE from 0.871 to 1.260. This indicates that depth contains important information about forma-

tion transition, drilling section, and operational stage, and, therefore, improves same-well prediction accuracy. However, depth is also a positional proxy rather than a direct controllable drilling parameter. If the formation sequence or drilling program differs among wells, the model may partially learn well-specific depth patterns, which can weaken transfer performance. For this reason, depth is retained in this study for same-well and same-region prediction, but cross-well deployment should use depth together with formation/section consistency checks, target-well calibration, or a parallel depth-free model as a robustness reference. MAPE is easily affected by the small-denominator effect in low-ROP samples, so cross-well validation should jointly consider R^2, RMSE, and MAE when judging model performance. In addition, the cross-well validation in this study remains neighboring-well validation within the same region and cannot be directly equated with broad generalization under cross-block or cross-basin conditions.

The compact effect of removing well depth from the input feature set is shown in Table 18.

Table 18. Compact effect of removing well depth from the input feature set.

Input Setting	R^2	RMSE	MAE	Interpretation
With depth	0.910 ± 0.004	0.871 ± 0.019	0.246 ± 0.012	Best same-well performance
Without depth	0.823 ± 0.018	1.260 ± 0.065	0.482 ± 0.040	Performance decreases but remains usable

The depth removal check indicates that well depth contains important formation-transition and drilling stage information, but it should be interpreted as a proxy feature rather than a causal operational control variable. The model still retains predictive capability after depth is removed, suggesting that prediction is not solely determined by the depth trend; nevertheless, the accuracy loss shows why depth is retained in the final feature set for the defined same-region application scenario.

5. Conclusions

This study addresses the insufficient accuracy and stability of ROP prediction under complex drilling parameter conditions and constructs an ROP prediction model based on ExtraTrees optimized by improved HHO. The main conclusions are as follows:

(1) Through drilling section identification, abnormal condition filtering, 3 × IQR outlier treatment, Savitzky–Golay smoothing, and standardization, 7093 valid samples are retained from the original 7892 drilling records, improving the engineering consistency and modeling usability of field while drilling data.

(2) By combining Pearson correlation analysis, Spearman rank correlation analysis, gray relational analysis, and engineering mechanism constraints, 14 input features are finally determined, covering well-section structure, bit loading, hydraulic circulation, drilling fluid properties, and formation pressure conditions, thereby providing a reasonable variable basis for ROP prediction.

(3) HHO is improved by Logistic chaotic initialization, adaptive Gaussian mutation, and a dynamic weighting strategy, and validation set RMSE is used as the fitness function to optimize key ExtraTrees hyperparameters, thereby establishing the IHHO-ET ROP prediction model. Under repeated seed same-well evaluation, IHHO-ET achieves R^2 = 0.910 ± 0.004, RMSE = 0.871 ± 0.019, MAE = 0.246 ± 0.012, and auxiliary MAPE = 8.43 ± 0.46%, showing the best average performance among the compared models. However, its improvement over HHO-ET is small and not statistically significant at the 0.05 level. The ablation experiment indicates that HHO-based hyperparameter optimization is the main source of the performance improvement over the

Random Search-ET baseline, while Logistic chaotic initialization, adaptive Gaussian mutation, and dynamic weighting mainly provide small cumulative refinements to the search process and stability rather than independently verified large gains.

(4) Same-region cross-well leave-one-well-out validation indicates that inter-well distribution differences cause model performance to decline significantly compared with same-well testing. In the repeated seed IHHO-ET cross-well experiment, the average performance is R^2 = 0.696 ± 0.005, RMSE = 1.509 ± 0.016, MAE = 0.910 ± 0.010, and auxiliary MAPE = 44.66 ± 0.86%. Random Forest, ExtraTrees, and IHHO-ET show similar average cross-well performance, indicating that tree ensemble models have relatively stable same-region adaptability to structured drilling parameter data. Nevertheless, the high cross-well relative error and the depth dependence issue show that the current model is more appropriate for same-region trend prediction and parameter analysis support than for direct cross-block or closed-loop deployment without target-well calibration. Stable generalization still needs to be verified using more well samples, different formation types, and data from different blocks.

Author Contributions: Conceptualization, X.C. and D.L. (Dachuan Liang); methodology, X.C.; software, X.C.; validation, X.C.; formal analysis, X.C.; data curation, X.C.; writing—original draft preparation, X.C.; writing—review and editing, D.L. (Dachuan Liang), D.L. (Daoxiong Li), and C.Y.; supervision, D.L. (Dachuan Liang). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. The field drilling data used in this study were obtained from a third-party drilling data provider under a confidentiality agreement and are not publicly available. To support reproducibility, the preprocessing rules, feature definitions, hyperparameter search spaces, random seeds, and evaluation protocol are fully detailed in the manuscript. Processed or anonymized sample data may be available from the corresponding author upon reasonable request and with the permission of the data provider.

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

Abbreviations

The following abbreviations are used in this manuscript:

Abbreviation	Full Term
ECD	Equivalent Circulating Density
ET	ExtraTrees (Extremely Randomized Trees)
GRA	Gray Relational Analysis
HHO	Harris Hawks Optimization
IHHO	Improved Harris Hawks Optimization
IHHO-ET	ExtraTrees Optimized by Improved Harris Hawks Optimization
IQR	Interquartile Range
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MLP	Multilayer Perceptron
RF	Random Forest
RMSE	Root Mean Square Error
ROP	Rate of Penetration
R^2	Coefficient of Determination

SD	Standard Deviation
SG	Savitzky–Golay (filter)
SVR	Support Vector Regression
WOB	Weight on Bit
XGBoost	Extreme Gradient Boosting

Appendix A. Pseudocode of the IHHO-ET Algorithm

Table A1. Pseudocode of the IHHO-ET algorithm.

	Algorithm: Pseudocode of the IHHO-ET Algorithm
1	Input training set D_train, validation set D_val, parameter bounds LB and UB, population size N, and update iterations T after initialization
2	Initialize population positions X_i using Logistic chaotic mapping
3	Decode each individual position into an ExtraTrees parameter combination
4	Train ExtraTrees and calculate each individual fitness f_i = RMSE_i
5	Initialize the global best position X_best and best fitness f_best
6	Set update-iteration counter t to 1
7	while t $\leq$ T do
8	for i = 1 to N do
9	Update escape energy E and random variable r
10	if |E| >= 1 then
11	Perform global exploration update
12	else
13	Select soft besiege, hard besiege, soft besiege with progressive rapid dives, or hard besiege with progressive rapid dives according to r and |E| following the standard HHO update equations [22], and then update X_i
14	end if
15	Correct individual positions using dynamic weight w(t)
16	if rand < p_m then
17	Apply Gaussian mutation to X_i
18	end if
19	Apply boundary constraints and rounding to X_i
20	Decode X_i into an ExtraTrees parameter combination
21	Train the ExtraTrees model and calculate the new fitness f_i
22	if f_i < f_best then
23	Update X_best = X_i and f_best = f_i
24	end if
25	end for
26	t = t + 1
27	end while
28	Output the optimal parameter combination X_best
29	Construct the IHHO-ET prediction model using X_best

Note: In Line 13, the four exploration/exploitation update strategies follow the standard HHO update equations proposed by Heidari et al. [22]. This study does not modify those standard position-update equations; the improvements are introduced through Logistic chaotic initialization, adaptive Gaussian mutation, and dynamic weighting. In the pseudocode, T denotes the number of post-initialization update iterations. With N = 10 and T = 4, the initial population evaluation plus four update iterations gives N + N $\times$ T = 10 + 10 $\times$ 4 = 50 candidate evaluations.

References

1. International Energy Agency. *World Energy Outlook 2023*; IEA: Paris, France, 2023. Available online: https://www.iea.org/reports/world-energy-outlook-2023 (accessed on 31 May 2026).
2. Organization of the Petroleum Exporting Countries. *World Oil Outlook 2025*; OPEC: Vienna, Austria, 2025. Available online: https://www.opec.org/assets/assetdb/woo-2025.pdf (accessed on 31 May 2026).
3. Ben Aoun, M.A.; Madarász, T. Applying machine learning to predict the rate of penetration for geothermal drilling located in the Utah FORGE site. *Energies* **2022**, *15*, 4288. [CrossRef]

4. Du, S.; Huang, C.; Ma, X.; Fan, H. A review of data-driven intelligent monitoring for geological drilling processes. *Processes* **2024**, *12*, 2478. [CrossRef]
5. Soares, C.; Gray, K. Real-time predictive capabilities of analytical and machine learning rate of penetration (ROP) models. *J. Pet. Sci. Eng.* **2019**, *172*, 934–959. [CrossRef]
6. Hegde, C.; Daigle, H.; Millwater, H.; Gray, K. Analysis of rate of penetration (ROP) prediction in drilling using physics-based and data-driven models. *J. Pet. Sci. Eng.* **2017**, *159*, 295–306. [CrossRef]
7. Chen, X.; Gao, D.; Guo, B.; Feng, Y. Real-time optimization of drilling parameters based on mechanical specific energy for rotating drilling with positive displacement motor in the hard formation. *J. Nat. Gas Sci. Eng.* **2016**, *35*, 686–694. [CrossRef]
8. Ahmed, O.; Adeniran, A. Rate of penetration prediction utilizing hydromechanical specific energy. In *Drilling*; Shah, M., Ed.; IntechOpen: London, UK, 2018. [CrossRef]
9. Hassan, A.; Elkatatny, S.; Al-Majed, A. Coupling rate of penetration and mechanical specific energy to improve the efficiency of drilling gas wells. *J. Nat. Gas Sci. Eng.* **2020**, *83*, 103558. [CrossRef]
10. Hegde, C.; Gray, K.E. Use of machine learning and data analytics to increase drilling efficiency for nearby wells. *J. Nat. Gas Sci. Eng.* **2017**, *40*, 327–335. [CrossRef]
11. Barbosa, L.F.F.M.; Nascimento, A.; Mathias, M.H.; de Carvalho, J.A. Machine learning methods applied to drilling rate of penetration prediction and optimization—A review. *J. Pet. Sci. Eng.* **2019**, *183*, 106332. [CrossRef]
12. Hazbeh, O.; Aghdam, S.K.-Y.; Ghorbani, H.; Mohamadian, N.; Ahmadi Alvar, M.; Moghadasi, J. Comparison of accuracy and computational performance between the machine learning algorithms for rate of penetration in directional drilling well. *Pet. Res.* **2021**, *6*, 271–282. [CrossRef]
13. Zhang, C.; Song, X.; Su, Y.; Li, G. Real-time prediction of rate of penetration by combining attention-based gated recurrent unit network and fully connected neural networks. *J. Pet. Sci. Eng.* **2022**, *213*, 110396. [CrossRef]
14. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
15. Scikit-Learn Developers. ExtraTreesRegressor API Reference. Scikit-Learn Documentation, Version 1.5. 2024. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesRegressor.html (accessed on 31 May 2026).
16. Biau, G.; Scornet, E. A random forest guided tour. *TEST* **2016**, *25*, 197–227. [CrossRef]
17. Montesinos-López, O.A.; Montesinos-López, A.; Crossa, J. Support vector machines and support vector regression. In *Multivariate Statistical Machine Learning Methods for Genomic Prediction*; Springer: Cham, Switzerland, 2022; pp. 337–378. [CrossRef]
18. Sengupta, S.; Basak, S.; Peters, R.A., II. Particle swarm optimization: A survey of historical and recent developments with hybridization perspectives. *Mach. Learn. Knowl. Extr.* **2018**, *1*, 157–191. [CrossRef]
19. Katoch, S.; Chauhan, S.S.; Kumar, V. A review on genetic algorithm: Past, present, and future. *Multimed. Tools Appl.* **2021**, *80*, 8091–8126. [CrossRef] [PubMed]
20. Faris, H.; Aljarah, I.; Al-Betar, M.A.; Mirjalili, S. Grey wolf optimizer: A review of recent variants and applications. *Neural Comput. Appl.* **2018**, *30*, 413–435. [CrossRef]
21. Feurer, M.; Hutter, F. Hyperparameter optimization. In *Automated Machine Learning: Methods, Systems, Challenges*; Hutter, F., Kotthoff, L., Vanschoren, J., Eds.; Springer: Cham, Switzerland, 2019; pp. 3–33. [CrossRef]
22. Heidari, A.A.; Mirjalili, S.; Faris, H.; Aljarah, I.; Mafarja, M.; Chen, H. Harris hawks optimization: Algorithm and applications. *Future Gener. Comput. Syst.* **2019**, *97*, 849–872. [CrossRef]
23. Hussain, K.; Salleh, M.N.M.; Cheng, S.; Shi, Y. Metaheuristic research: A comprehensive survey. *Artif. Intell. Rev.* **2019**, *52*, 2191–2233. [CrossRef]
24. Gezici, E.; Livatyali, H. A comprehensive literature review on Harris hawks optimization techniques. *Eng. Appl. Artif. Intell.* **2022**, *116*, 105343. [CrossRef]
25. Tunkiel, A.T.; Sui, D.; Wiktorski, T. Reference dataset for rate of penetration benchmarking. *J. Pet. Sci. Eng.* **2021**, *196*, 108069. [CrossRef]
26. Alsaihati, A.; Elkatatny, S.; Gamal, H.; Abdulraheem, A. Data-driven prediction of rate of penetration while drilling complex lithology using random forest. *Sustainability* **2022**, *14*, 11612. [CrossRef]
27. Elkatatny, S. Real-time prediction of rate of penetration while drilling complex lithologies using artificial intelligence techniques. *Ain Shams Eng. J.* **2021**, *12*, 917–926. [CrossRef]
28. Shokry, K.; Elsayed, M.; Abdulwahab, H.; Shawky, A.; Helmy, M.; Ali, T.; Elkatatny, S.; Abdulraheem, A. Real-time prediction of rate of penetration during drilling operation in a motorized bottom-hole assembly using different artificial intelligence techniques. *Processes* **2023**, *11*, 1207. [CrossRef]
29. Jiao, X.; Gan, C.; Cao, W.; Wu, M. Physics-data fusion prediction of rate of penetration based on multiple empirical equations. *Processes* **2024**, *12*, 1572. [CrossRef]
30. Bai, X.; Liu, M.; Zhang, Z.; Wang, H.; Zhao, J. ROP optimization method based on CBT-LSTM and machine learning. *Energies* **2024**, *17*, 6030. [CrossRef]

31. Xiong, Y.; Zheng, B.; Zhao, R.; Su, J.; Dong, J. Penetration rate prediction and parameter optimization based on improved sparrow search algorithm and BiLSTM neural network. *Processes* **2024**, *12*, 1073. [CrossRef]
32. Huang, T.; Ali, M.; Liu, W.; Zhang, L.; Xu, H. Research on ROP prediction based on machine learning algorithm. *Energies* **2025**, *18*, 176. [CrossRef]
33. Mohammadinia, F.; Ranjbar, A.; Ghazi, F.; Hosseini, S.T. Rate of penetration prediction in drilling operations: A comparative study of AI models and meta-heuristic approaches. *J. Pet. Explor. Prod. Technol.* **2025**, *15*, 93. [CrossRef]
34. Dash, C.S.K.; Behera, A.K.; Dehuri, S.; Ghosh, A. An outliers detection and elimination framework in classification task of data mining. *Decis. Anal. J.* **2023**, *6*, 100164. [CrossRef]
35. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef] [PubMed]
36. Schober, P.; Boer, C.; Schwarte, L.A. Correlation coefficients: Appropriate use and interpretation. *Anesth. Analg.* **2018**, *126*, 1763–1768. [CrossRef] [PubMed]
37. Akoglu, H. User's guide to correlation coefficients. *Turk. J. Emerg. Med.* **2018**, *18*, 91–93. [CrossRef] [PubMed]
38. Liu, S.; Forrest, J.; Yang, Y. *Grey Systems Analysis: Methods, Models and Applications*, 2nd ed.; Springer: Singapore, 2024.
39. Sagi, O.; Rokach, L. Ensemble learning: A survey. *WIREs Data Min. Knowl. Discov.* **2018**, *8*, e1249. [CrossRef]
40. Probst, P.; Wright, M.N.; Boulesteix, A.-L. Hyperparameters and tuning strategies for random forest. *WIREs Data Min. Knowl. Discov.* **2019**, *9*, e1301. [CrossRef]
41. Bentéjac, C.; Csörgő, A.; Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. *Artif. Intell. Rev.* **2021**, *54*, 2267–2306. [CrossRef]
42. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
43. Guido, R.; Ferrisi, S.; Lofaro, D.; Conforti, D. An overview on the advancements of support vector machine models in healthcare applications: A review. *Information* **2024**, *15*, 235. [CrossRef]
44. Adam, S.P.; Alexandropoulos, S.-A.N.; Pardalos, P.M.; Vrahatis, M.N. No free lunch theorem: A review. In *Approximation and Optimization: Algorithms, Complexity and Applications*; Springer: Cham, Switzerland, 2019; pp. 57–82. [CrossRef]
45. Bischl, B.; Binder, M.; Lang, M.; Pielok, T.; Richter, J.; Coors, S.; Thomas, J.; Ullmann, T.; Becker, M.; Boulesteix, A.-L.; et al. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *WIREs Data Min. Knowl. Discov.* **2023**, *13*, e1484. [CrossRef]
46. Scikit-Learn Developers. User Guide: Scikit-Learn Machine Learning in Python, Version 1.5. 2024. Available online: https://scikit-learn.org/stable/user_guide.html (accessed on 31 May 2026).
47. Raschka, S. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv* **2018**, arXiv:1811.12808.