Petroleum Science 23 (2026) 5706–5722

Contents lists available at ScienceDirect

Petroleum Science

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

Original Paper

Machine learning–driven prediction of profile control effect in oil reservoir based on relative production difference

Kun Xie [a,b], Zhan-Qi Wu [a,b], Xuan-Shuo Tian [a,c], Wei-Jia Cao [a,b,*], Xin Song [d], Er-Long Yang [a,b], Kun Yan [a,b], Xiang-Guo Lu [a,b]

[a] *Enhanced Oil and Gas Recovery of Ministry of Education, Northeast Petroleum University, Daqing, 163318, Heilongjiang, China*
[b] *State Key Laboratory of Continental Shale Oil, Daqing, 163712, Heilongjiang, China*
[c] *College of Petroleum Engineering, China University of Petroleum (Beijing), Beijing, 102249, China*
[d] *CNOOC (China) Limited Tianjin Branch, Tianjin, 300459, China*

ARTICLE INFO

Article history:
Received 20 November 2025
Received in revised form
14 January 2026
Accepted 14 April 2026
Available online 21 April 2026

Edited by Meng-Jiao Zhou

Keywords:
Prediction of profile control effect
Production relative difference
Machine learning
Long Short-term memory network
Random forest

ABSTRACT

Traditional methods for predicting profile control effect have limited accuracy, with deviations of varying degrees between predicted and actual oil production after operation. In order to solve this problem, this study introduces a prediction method of profile control effect based on relative production difference (PRD). Initially, Long Short-Term Memory (LSTM) network, Gated Recurrent Unit (GRU) network, and Transformer neural network were used to predict the benchmark production of water flooding. Subsequently, the study identified factors influencing profile control effect as input features and the relative difference between benchmark production of water flooding and post-treatment actual production as the target output. Then three machine learning models were constructed for prediction of profile control effect, namely Adaptive Boosting (AdaBoost), Extreme Gradient Boosting (XGBoost), and Random Forest (RF). Hyperparameters of all models were optimized using the Optuna framework during training process. Finally, the study applied the optimal profile control effect prediction model to predict the profile control effect in S reservoir. The results show that LSTM performs the best in predicting the benchmark production of water flooding, with R^2 values of 0.9874 and 0.9508 for the training and test sets, respectively. RF performs the best in predicting the profile control effect, with R^2 values of 0.9988 and 0.9995 for the test and cross validation sets, respectively. In oilfield applications, the test set R^2 of RF exceeds 0.9, which proves that the predicted production after profile control is highly consistent with the actual production, and the model has strong generalization ability.

1. Introduction

China's dependence on foreign oil resources has exceeded 70%, which brings severe challenges to the country's energy security. At present, offshore oilfields have become the main source of new oil production in China, which is of great significance for increasing domestic oil production and ensuring national energy security (Sun, 2023). After years of development, some oil reservoirs have gradually entered the high water cut stage. This has led to increased heterogeneity of the reservoir (Wang et al., 2021a), increased difficulty in development, and a significant decrease in crude oil recovery rate. To address these problems, series of chemical profile control measures have been implemented (Cao et al., 2019; Hou et al., 2023; Song et al., 2022; Zhou et al., 2006). The profile control measures inject plugging agents into the formation through injection wells to plug high permeability layers, thereby expanding the swept volume of injected water in medium and low permeability layers, and improving the water flooding effect (Elsharafi and Bai, 2012; Xie et al., 2025). Accurate prediction of profile control effect serves as a foundation for guiding decision-making and optimization design of on-site profile control measures in oilfields, especially in optimizing process parameters, selecting optimal control schemes and reducing operational costs.

Traditional methods for predicting the effect of profile control can be mainly divided into three categories: expert experience

* Corresponding author.
E-mail address: caoweijia131466@163.com (W.-J. Cao).
Peer review under the responsibility of China University of Petroleum (Beijing).

https://doi.org/10.1016/j.petsci.2026.04.021

method, statistical method, and numerical simulation method. Expert experience method requires researchers to summarize the physical laws of the profile control process and on-site profile control practice experience based on reservoir engineering knowledge, and derive a relationship between profile control effect and its influencing factors. This relationship can then be used to evaluate the profile control effect under specific conditions. Wu et al. (2007) summarized empirical formulas for liquid production, oil production, water cut, and polymer concentration before and after profile control, and successfully applied these formulas to predict the profile control effect in Lamadian Oilfield. Chen et al. (2014) introduced the dimensionless oil increment multiplier to evaluate the profile control effect of the overall gel beads. They proposed a set of calculation formula for dimensionless oil increment multiplier and achieved the evaluation of profile control effect. Expert experience method requires researchers to conduct extensive experience practice and summarization, which takes a long time. Meanwhile, there are many factors that affect the profile control effect. Simple relationships are difficult to consider the effects of various factors on the profile control effect, and the prediction accuracy is limited. Statistical methods utilize statistical knowledge to establish an evaluation method for profile control effectiveness. For example, Wang et al. (2021b) established a set of profile control effect evaluation methods using the variance rescaled (V/S) analysis method in statistics. They first constructed a comprehensive evaluation index for water plugging and profile control effect using the apparent water absorption index and the oil change rate per ton of water and regarded it as a nonlinear time series. Then, they used the V/S analysis method to calculate the Hurst index of the comprehensive evaluation index time series, and evaluated the profile control effect through the range of Hurst index values. However, the V/S analysis method is sensitive to data noise (Sun et al., 2014), and on-site data often generates data noise due to equipment failures and other reasons, which affects prediction accuracy. Numerical simulation method mainly uses numerical simulation software to establish geological models for simulating the profile control process. Researchers can evaluate the effect of different profile control measures based on indicators such as oil production increment or water cut, and then complete the optimization design of the profile control scheme. Liu and Zhou (1994) used numerical simulation method to predict the profile control effect. They combined the profile control mechanism with indoor experiment results, established a mathematical model, and solved it using finite difference method through their self-developed simulator. They evaluated the profile control effect based on water cut, oil production and cumulative water production before and after the profile control treatment. Feng et al. (1997) used numerical simulation method to study the dimensionless increment effect of overall profile control in different types of reservoirs, providing a basis for predicting the effect of overall profile control. Numerical simulation method requires researchers with complete professional knowledge to perform the full workflow operation, which demands high level of personal expertise and consequently leads to a longer simulation cycle (Wang, 2024).

Machine learning (ML) is an important branch of artificial intelligence (AI), which has advantages such as high efficiency, intelligence, and low cost in data analysis (Xie et al., 2023, 2024). Essentially, it is the process where computers use mathematical principles and algorithms to identify patterns, extract knowledge and make judgments on new data (Min et al., 2020). In recent years, ML has been widely applied in the field of oil and gas development at home and abroad, showing excellent potential in aspects such as dynamic production parameter prediction (Tariq et al., 2019) and production time series predicting (Song et al., 2020; Kocoglu et al., 2024). In terms of profile control effect prediction, ML has shown potential to outperform traditional prediction methods. Feng et al. (2003) established a predictive model for profile control effect using ANN. They take the wellhead pressure index, water injection pressure, water injection volume, corresponding average water cut of the oil well and plugging agent as inputs, and measures to increase oil production during the effective period of profile control as target variables. After application in the field, it is believed that the reliability of the predicted results of this model is higher than that of traditional methods. Liu et al. (2014) used the oil production after profile control as the evaluation index for profile control effect and constructed a profile control effect prediction model based on BP neural network. Xu et al. (2007) believe that traditional methods are difficult to establish a definite relationship between numerous influencing factors and profile control effect. Therefore, they established a prediction model for profile control effect using support vector machine (SVM) with the target variable of increasing oil or precipitation through profile control measures. However, existing ML methods for profile control effect prediction have following limitations: (1) By directly adopting or simply modifying unconventional oil and gas production prediction algorithms for new applications, these methods rely only on historical profile control production data and absolute production for time series modeling, leading to failure in effectively distinguish the impact of natural decline and profile control measures intervention. Meanwhile, the oil content varies in different layers, so the increase in production after well group profile control may differ greatly. Directly comparing absolute production cannot accurately reflect the effect of profile control. (2) Theses methods neglect the comprehensive effect of multiple factors such as geological parameters, construction parameters, and production dynamic parameters on the profile control effect. (3) The limited generalization ability and engineering applicability of existing models make them difficult to effectively guide practical applications in oilfield sites. Given these challenges, there is an urgent need to develop a more accurate and practically applicable method for profile control effect prediction. At present, ML has developed a variety of algorithm libraries and models, but different models may exhibit significant differences when trained for different tasks, requiring the selection of the most suitable model.

This study comprehensively considers the impacts of reservoir geological conditions, operation parameters, and production characteristics on profile control effect. It further proposes an innovative profile control effect prediction method based on a hybrid approach integrating data-driven and physics-driven methodologies. The method uses the relative difference between the predicted benchmark production of water flooding and the actual production of profile control as core metric for evaluating the oil increment effect of profile control measures. In the study, a high-precision, strongly generalized for water flooding production prediction model and a profile control effect prediction model were constructed. The Optuna algorithm optimization framework has been used for hyperparameter optimization, improving the model's prediction accuracy and generalization ability. The flowchart is shown in Fig. 1.

2. Methodology

2.1. Data cleaning

The oil production after profile control is comprehensively influenced by various factors such as reservoir geological parameters, profile control process parameters and production conditions. This study comprehensively collected the parameters of the on-site profile control well group database from S reservoir (as shown in Table 1).

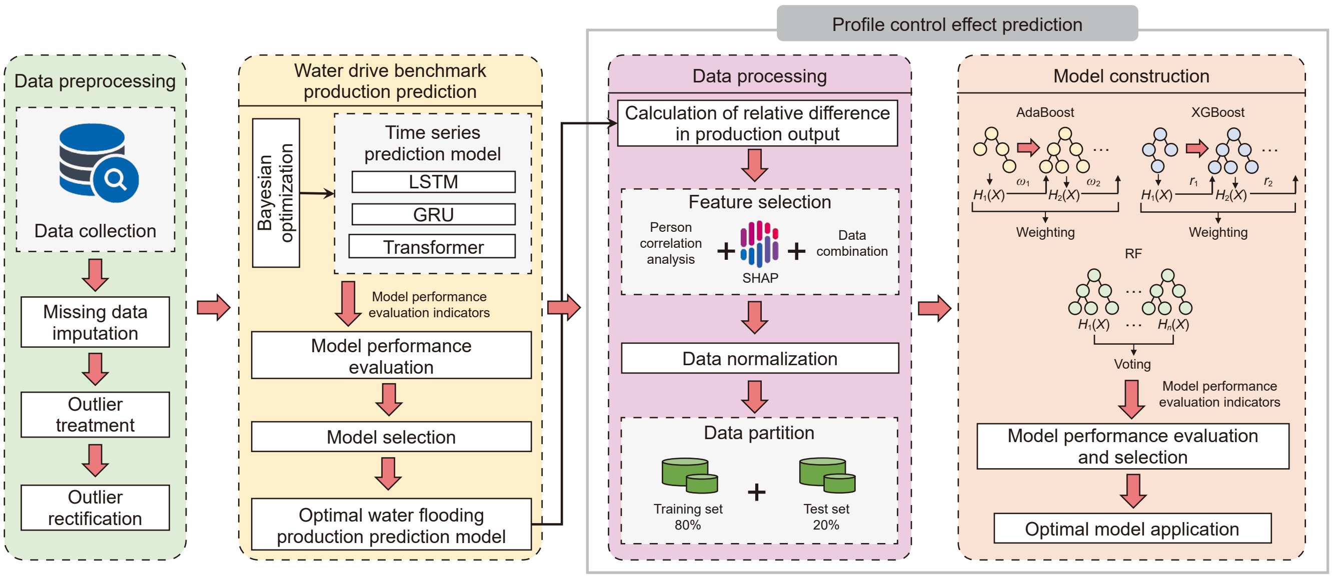

Fig. 1. Roadmap for predicting the effect of profile control.

According to Table 1, there are many marked fields in the collected on-site data, such as well type, injection method, profile control injection fluid type and date. These fields do not contribute significantly to improving the accuracy of oil well production prediction, so they will be deleted. Statistical analysis of data integrity revealed that data of key features such as permeability, porosity, oil production, injection volume, and injection pressure were relatively complete, with missing rates generally below 10%. Given the high availability of these data, the mean imputation method was used to fill in the missing values to ensure that the overall distribution and statistical features of the data did not deviate significantly. However, the data missing rates of features such as formation water salinity, oil pressure, and casing pressure exceeded 20%, and they will be deleted.

To reduce the negative impact of outliers and data noise on prediction results, this study uses an efficient data compression algorithm, Autoencoder (AE), for unsupervised outlier detection of oilfield production data. Its core mechanism is to use ANN to reduce the dimensionality and reconstruct input feature vectors, calculate the minimum reconstruction error, capture key information of data and achieve noise and outlier filtering. Minimizing reconstruction error is the minimum value of error between the original data and the reconstructed data. Abnormal data points often exhibit significantly high reconstruction errors due to deviation from the mainstream data distribution. Therefore, AE can effectively identify outliers in the original data. The reconstruction error is defined as the mean square error (MSE) between the input vector x_i and the reconstructed vector $\widehat{x}_i$:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N} (x_i - \widehat{x}_i)^2 \tag{1}$$

where N is the number of elements in vectors. This study calculates the anomaly detection threshold using the following equation:

$$\tau = \mu + k \times \sigma \tag{2}$$

where μ is reconstruction error, σ is the standard deviation of reconstruction error, and k is a hyperparameter. Based on the existing research on the value of k (Luo et al., 2026) and the strictness of anomaly detection in this study, we set $k = 3$.

2.2. Construction of water flooding benchmark production prediction model

The benchmark production of water flooding refers to the natural production trend exhibited by oil wells during normal water flooding development without intervention of profile control measures. It is an important prerequisite for quantitatively evaluating the effect of profile control and helps to accurately clarify the relationship between natural decline and intervention measures. In order to enable neural networks to better capture and learn the latent temporal features in data, it is necessary to transform continuous time series data into samples that are compatible with neural network architectures. This study uses the temporal sliding window method for processing, which can effectively extract the dynamic changes of temporal data and form a supervised learning data format.

The schematic diagram of the sliding window method is shown in Fig. 2. The core idea of the sliding window method is to set a fixed size window and gradually slide it over the original time series, generating a new training sample for each step of sliding. Specifically, assuming the original time series is $y_1, y_2, \ldots, y_T$ and the sliding window size is defined as W. For each time step t ($W \le t < T$), a training sample can be generated, where the input sequence is data from W consecutive time steps in the past, i.e. $X_t = [y_t - W+1, y_t - W+2, \ldots, y_t]$, and the corresponding output is the yield data y_t+1 for the next time step. By traversing the entire original sequence through the sliding window process, the original time series $y_1, y_2, \ldots, y_T$ can be transformed into $N = T - W$ training samples, where each sample contains data from W input time steps and corresponds to a prediction target. This processing method not only ensures the temporal continuity of data, but also meets the input requirements of neural networks.

Table 1
Production database parameters of profile control well group in S reservoir.

Parameter type	Features
Well information	Block, well number, well type, injection layer, injection method, recovery
Geological parameters	Average permeability of the target layer, maximum permeability of the target layer, breakthrough coefficient, thickness of the target layer, effective porosity, pore throat radius of the reservoir, viscosity of underground crude oil, salinity of formation water, concentration of calcium and magnesium ions, reservoir temperature
Engineering parameters	Profile control injection fluid type, injection concentration of profile control agents, injection speed of profile control agents, profile control agent dosage, gel strength
Static production parameters	Average daily injection volume before operation, average daily injection volume after operation, average pressure before operation, average pressure after operation, difference between maximum pressure and trial injection pressure during operation, pressure difference before and after one month of measures, PI before and after measures, average apparent water absorption index before operation, average apparent water absorption index after operation, apparent resistance coefficient, apparent residual resistance coefficient, water cut of the well group before measures, cumulative oil production of the well group
Dynamic production parameters	Date, injection time, wellhead pressure, casing pressure, daily water injection volume, daily injection volume of polymer, total volume of water and polymer, cumulative water injection, cumulative polymer injection, production time, daily liquid production, daily gas production, daily oil production, water cut, daily water production, gas oil ratio, gas lift rate dynamic liquid level, pump frequency, pump current, oil pressure, pressure difference, flowing pressure, backpressure, pump inlet pressure, pump outlet pressure, wellhead temperature, flowing temperature, cumulative oil production, cumulative liquid production, cumulative gas production

Based on the historical daily oil production data of reservoir well groups in S reservoir, this study selects LSTM, GRU, and Transformer, which are commonly used in time series prediction and have demonstrated good performance (Alzakari et al., 2025; Elshewey et al., 2023; Su et al., 2025; El-Rashidy et al., 2025), to construct univariate time series production prediction models. The hyperparameters were optimized by Optuna framework. Subsequently, the stability and generalization ability of the three models in engineering applications are compared, and the optimal model is selected.

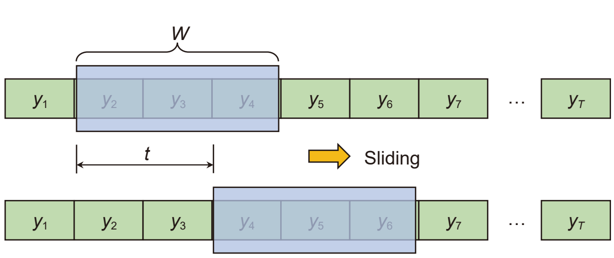

Fig. 2. Schematic diagram of sliding window method.

LSTM (Hochreiter and Schmidhuber, 1997) is one of the most prevalent gated recurrent neural networks, excelling in the processing of both time series and nonlinear data. In contrast to conventional neural networks, the hidden layer of an LSTM is constructed from multiple interconnected memory units that are sequentially arranged. Each memory unit is equipped with three gate structures (Chen et al., 2020; Lee et al., 2019): forget gate, input gate and output gate (Fig. 3(a)). These gates use activation functions such as Sigmoid or Tanh to regulate the information flow within the sequence.

GRU is a simplified variant of LSTM. Compared to LSTM, GRU only comprises two gated units: an update gate and a reset gate (Fig. 3(b)). Reset gate replaces the role of forget gate in LSTM, deciding whether to discard information from previous hidden states, while the update gate determines how much information from previous hidden states needs to be retained in the current state. This structure is beneficial to reduce model complexity and improve computational efficiency (Yu et al., 2021; Chen et al., 2024; Tarek et al., 2023).

Transformer is a deep learning model based entirely on self-attention mechanism. It is composed of two modules: encoder and decoder (Yuan et al., 2024) (Fig. 3(c)). Compared to LSTM and GRU, Transformer breaks away from the conventional iterative approach of recurrent neural networks (RNN) and adopts attention mechanism to process all elements in the sequence in parallel, enabling the model to more effectively capture long-range dependencies and improve computing efficiency.

After data cleaning, the historical water flooding production in chronological order was extracted from data samples. Firstly, LSTM, GRU and Transformer were used for time series prediction respectively, and the Optuna framework was used to perform 200 Bayesian optimizations. Subsequently, the optimal hyperparameter configurations were used to reconstruct the water flooding production prediction benchmark models. Finally, the model with superior predictive accuracy and generalization performance was selected for deployment.

2.3. Construction of profile control effect prediction model

2.3.1. Quantification of the effect of profile control measures

To achieve the prediction of the profile control effect, it is necessary to quantify this effect to uniformly evaluate the production changes of each well group under the effect of profile control measures. This study takes the predicted production of water flooding in natural conditions as the benchmark, and uses the relative difference between the actual production of profile control and the benchmark production to characterize the increase in production after profile control. This approach effectively quantifies the effect of profile control measures and reduces discrepancies in production scale among different well groups, thereby facilitating a comparative analysis of profile control effects across well groups. The specific quantification method is as follows.

(1) Based on historical production data of water flooding prior to profile control implementation, time-series forecasting models were developed to predict benchmark daily oil production for each well group under natural depletion conditions. The prediction cycle was established as a 12-month period commencing immediately after profile control initiation, consistent with the field-reported average duration of measure effectiveness. To incorporate the temporal accumulation effects of water injection, daily predictions were

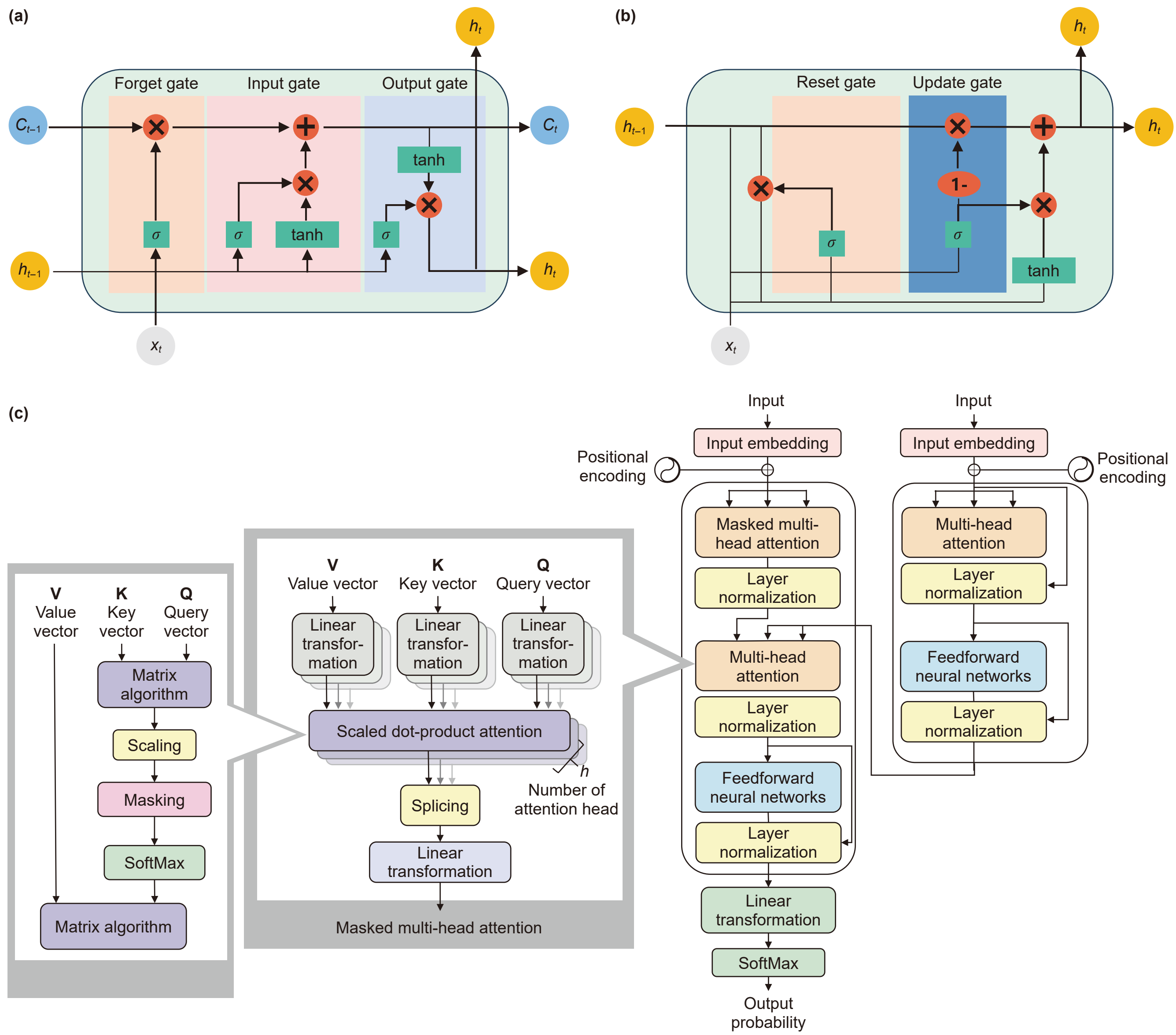

Fig. 3. Schematic diagram of LSTM, GRU, and Transformer structures. **(a)** Schematic diagram of LSTM gating structure. **(b)** Schematic diagram of GRU gating structure. **(c)** Schematic diagram of Transformer structure.

sequentially aggregated to derive the total annual oil production that would occur in the absence of profile control interventions;

(2) Daily oil production data for each well group were extracted from actual field records over one consecutive year after profile control implementation. Considering the cumulative impact of the injected mass of profile control agents on oil production, the actual daily cumulative oil production after profile control within one year is obtained through cumulative calculation;

(3) Calculate the relative difference between the actual production after profile control and the predicted water flooding benchmark production in chronological order:

$$\mathrm{PRD} = \frac{Q_\mathrm{P}(t) - Q_\mathrm{W}(t)}{Q_\mathrm{W}(t)} \quad (t = 1, 2, \ldots, 360) \tag{3}$$

where $Q_\mathrm{W}(t)$ is the predicted water flooding benchmark production, $Q_\mathrm{P}(t)$ is the actual production after profile control, PRD is relative production difference. Fig. 4 shows the calculation method for the difference between the actual profile control production and the predicted waterflooding baseline production. Dividing the absolute difference by $Q_\mathrm{W}(t)$ yields PRD.

2.3.2. Feature selecting

Given the high dimensionality of the input feature space, this study implements a two-step feature selection to avoid the adverse effects of irrelevant features on model accuracy and generalization performance. Firstly, Pearson correlation analysis was conducted to quantify relationships among features, followed by domain expert knowledge to remove redundant features. Subsequently, the Shapley Additive Explanations (SHAP) method was used for feature importance assessment.

(1) Correlation analysis

Pearson correlation analysis is a statistical measure of the strength of linear correlation between two variables. Its core idea

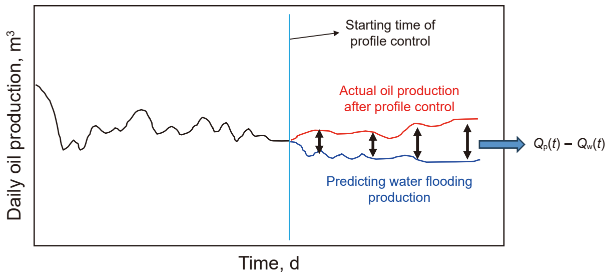

Fig. 4. Schematic diagram of quantification method for profile control effect.

is to evaluate the correlation between two variables by calculating the ratio of their covariance to their standard deviation:

$$r = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n} (X_i - \overline{Y})^2}} \tag{4}$$

where r represents the Pearson correlation coefficient, n denotes the total number of samples, X_i and Y_i are random variables, and $\overline{X}$ and $\overline{Y}$ represent the means. Based on the research findings of Wang and Fu (2018) and the actual situation of this study, the linear correlation strength corresponding to absolute value range is summarized in Table 2.

(2) Feature importance analysis

SHAP method is an interpretable machine learning method based on game theory, which can provide individual sample level feature contribution explanations for the prediction results of complex models. Its core idea originates from the Shapley value in game theory (Li, 2022). The SHAP method calculates the sum of marginal contributions of each feature in the feature subset and then obtains the degree of influence of each feature on the prediction results. Meanwhile, it can demonstrate the positive or negative impact of the feature on the target variable. In this method, the Shapley value of a certain feature is represented as the weighted average of its marginal contribution across all feature subsets, while the final prediction value of the model for a given sample can be expressed as the weighted sum of the Shapley values of all features in that sample.

(3) Data combination

To fully consider the comprehensive impact of static geological parameters and dynamic engineering parameters on yield during the production process, this study constructed a multidimensional input feature feature set by “static-dynamic parameter coupling”. Specifically, for each well group, its static parameters (such as average permeability, effective porosity, etc.) are treated as time-invariant features and embedded into the feature vector of each dynamic sample. Then, the dynamic parameters at corresponding time points are concatenated and combined in chronological order to form a time-varying feature sequence. By combining data, the consistency of data dimensions can be effectively maintained, ensuring that the model can capture the synergistic effects of static reservoir conditions and dynamic production parameters on production.

Table 2
Pearson correlation coefficient and linear correlation strength.

Pearson correlation coefficient absolute value	Linear correlation strength
[0.6, 1]	Strength
[0.2, 0.6)	Medium
[0, 0.2)	Weak

2.3.3. Data transformation and partition

Owing to the scale inconsistency among input features in the dataset, the predictive accuracy of subsequent profile control effect may be significantly reduced. Hence, it is imperative to normalize the original data to a unified numerical range, ensuring comparability across feature values. This study used the Max-Min normalization method to normalize the data:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{5}$$

where x' represents the normalized data, x denotes the specific value in the data, $\min(x)$ is the minimum value of the feature data, and $\max(x)$ is the maximum value of the feature data.

Subsequently, the original dataset was divided into a training set and a test set in an 8:2 ratio. In order to fully utilize the sample data, this study used K-fold cross validation to comprehensively evaluate the robustness and generalization ability of the profile control effect prediction model. This method divides the dataset into K parts, and each iteration selects a subset as the reserved test set, while aggregating the remaining $K-1$ subsets into the training set. This process will continue until each subset happens to serve as a test set once to ensure comprehensive model evaluation. The mean prediction error of the model on K test sets is used as the final model error (Fig. 5).

2.3.4. Model construction

Currently, three types of ensemble learning algorithms, namely AdaBoost, XGBoost and RF, have been applied in numerous disciplines such as petroleum science (Li et al., 2025), agronomy (Elshewey, 2025) and medicine (Yeşilkanat, 2020), demonstrating significant advantages in efficiency, accuracy, and stability. Therefore, this study takes the main controlling factors of profile control effect as input features and the production relative difference as output features, and utilizes the above three algorithms to construct profile control effect prediction models. The Optuna framework is used for hyperparameter optimization, and then the optimal model is selected for further profile control effect prediction. The predicted relative difference can be used to back-calculate, in accordance with Eq. (3), the predicted cumulative oil production after profile modification, and the predicted daily oil production after profile modification can be obtained through discrete difference calculation. The discrete difference calculation formula is shown in Eq. (6).

$$Q_d = Q_i - Q_{i-1} \tag{6}$$

where Q_d represents the daily production after profile control. Q_i is the cumulative production from the start time of profile control to the i-th day ($i > 1$). Q_{i-1} is the cumulative production from the start time of profile control to the ($i-1$)-th day.

AdaBoost is an ensemble learning method based on Boosting strategy, which gradually trains model in a forward stagewise additive manner (Lahiri et al., 2016). During the training process, the weights of data points are automatically adjusted based on the error rate in the previous sub prediction model. Then, new weighted data are used to train the next sub model. When the number of iterations or error rate meets the set value, the

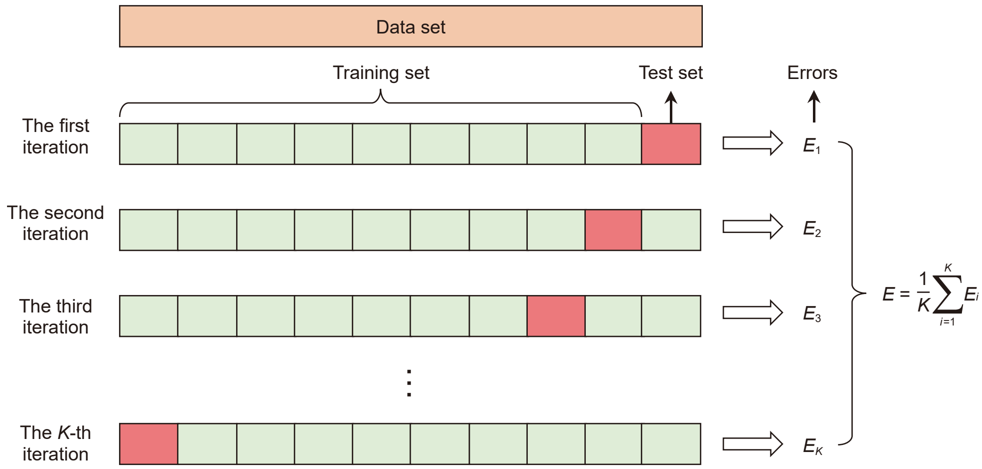

Fig. 5. *K*-fold cross validation schematic.

algorithm automatically stops running and saves all sub prediction models trained in a certain weight combination as final model (Wang et al., 2024).

XGBoost is an improvement of the Gradient Boosting Decision Tree (GBDT) algorithm, which is also based on the Boosting strategy. It uses a second-order Taylor expansion to approximate the loss function, thereby optimizing the modeling process and improving model accuracy. At the same time, it adds a regularization term to the objective function. The regularization term controls the complexity of the model to a certain extent, thereby helping reduce the risk of overfitting and effectively improves the generalization performance of the model (Chen et al., 2022).

RF was officially proposed by Breiman (2001), which is based on the Bagging strategy. Firstly, the training samples are sampled using the Bootstrap sampling method to form several sub samples. Based on these sub sample data, independent weak learners are trained. Finally, the arithmetic mean of all weak learner results is taken to obtain the final result. This modeling approach enables RF to effectively reduce variance, prevent overfitting and improve prediction accuracy (He et al., 2022; Jackins et al., 2021).

2.4. Data driven optimization algorithm

Conventional hyperparameter optimization algorithms, such as grid search and random search, exhibit limitations such as high computational costs, low efficiency and vulnerability to local optima. Specifically, grid search necessitates a priori specification of discrete hyperparameter values, whereas random search offers no convergence guarantees to global optima. In contrast, Bayesian optimization, a probabilistic model-based approach, addresses these challenges through data-driven sequential decision-making. The fundamental principle of Bayesian optimization is to construct surrogate models such as Gaussian Process (GP) (Jiang et al., 2025) or Tree-structured Parzen Estimator (TPE) (Song et al., 2023) to approximate the objective function, thereby dynamically constructing the mapping relationship between the parameter space and the objective function. At the same time, it uses acquisition functions such as expected improvement (EI) or probability improvement (PI) to assist proxy models in efficiently searching the parameter space, in order to find the optimal parameter combination, gradually approaching the global optimal solution, and adept at handling continuous or mixed parameter spaces. Optuna is an open-source hyperparameter optimization framework tailored for ML algorithms. Its core objective is to automate the search for optimal hyperparameters by adaptively adjusting the search strategy and range, thereby enhancing model performance and eliminating the cumbersome manual tuning steps associated with traditional methods (Akiba et al., 2019; Rodriguez and Salazar, 2022; Yang et al., 2025). Furthermore, Optuna

Table 3
Evaluation indicators and calculation formulas for ML models.

Evaluation indicators	Computational formulations
MAE	$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} \lvert y_i - \hat{y}_i \rvert$
MAPE	$\mathrm{MAPE} = \frac{100\%}{N}\sum_{i=1}^{N} \frac{\lvert y_i - \hat{y}_i \rvert}{y_i}$
RMSE	$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}$
R^2	$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}$

Table 4
Optimal hyperparameter combination for time series prediction model.

Model hyperparameters	LSTM	GRU	Transformer
Num_layers	4	1	1
Hidden_units	64	64	/
Learning_rate	0.001	0.001	0.001
Dropout	0.1	/	0.1
Batch_size	16	32	32
Activation	ReLU	Tanh	ReLU
Optimizer	Adam	Adam	Adam
Epochs	300	300	300
*d*_model	/	/	64
Num_heads	/	/	4
*d*_ff	/	/	128

Notes: Num_layers refers to the number of neural networks, *d*_model refers to the characteristic dimension of each time step, num_heads refers to the number of heads with multi head attention, and *d*_ff refers to the middle layer dimension of feedforward neural network.

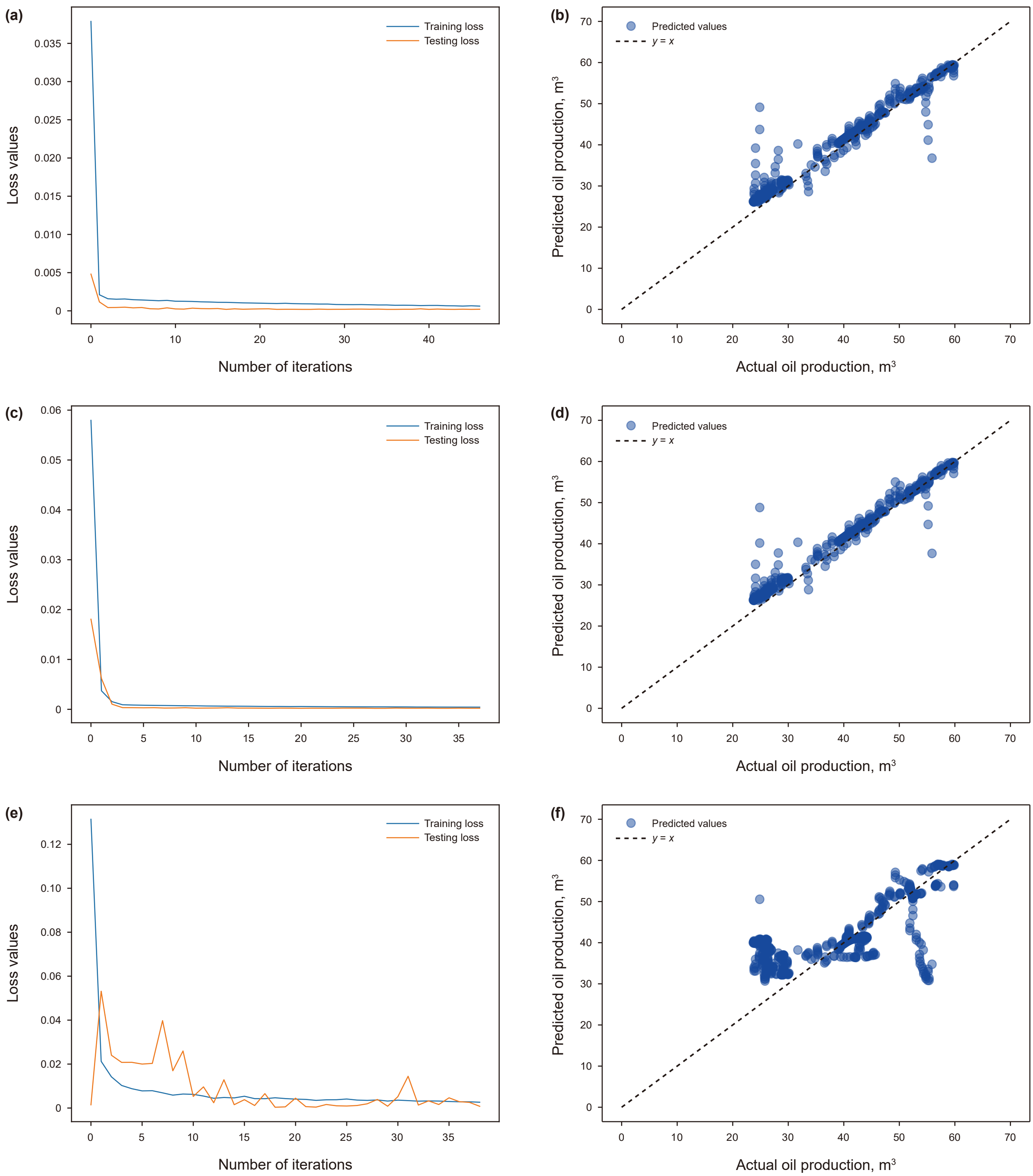

Fig. 6. Comparison chart of production forecast results. **(a)** Trend of loss value changes in LSTM model. **(b)** Comparison between LSTM predicted oil production and actual oil production. **(c)** Trend of loss value changes in GRU model. **(d)** Comparison between GRU predicted oil production and actual oil production. **(e)** Trend of loss value changes in Transformer model. **(f)** Comparison between transformer predicted oil production and actual oil production.

incorporates an intelligent built-in early stopping mechanism that intervenes immediately and terminates model training when the initial results are unsatisfactory, significantly improving computational efficiency. This framework supports various optimization algorithms such as Bayesian optimization, random search and CMA-ES, and provides parallel experiment management, dynamic pruning, and visual analysis tools. This study employs Bayesian optimization principles, using the Optuna framework to construct a TPE surrogate model. This model performs automated hyperparameter optimizing for predictive models during the prediction of water flooding benchmark production and profile control effects. TPE divides

Table 5
Model performance evaluation.

Models	Data set	MAE	MAPE	RMSE	R^2
LSTM	Training set	1.45	1.64%	3.49	0.9874
	Test set	1.53	4.74%	2.62	0.9508
GRU	Training set	1.82	2.00%	4.01	0.9834
	Test set	2.09	6.43%	2.94	0.9379
Transformer	Training set	8	7.79%	10.24	0.8918
	Test set	5.65	16.17%	8.21	0.5170

the posterior probability $P(x|y)$ of traditional Bayesian algorithms into the following two categories based on the quantile threshold y^*:

$$P(x|y) = \begin{cases} l(x), \text{ if } \ y < y^* \\ g(x), \text{ if } \ y \geq y^* \end{cases} \tag{7}$$

where $l(x)$ is the distribution of parameters with objective function values better than the threshold, and $g(x)$ is the distribution of parameters with objective function values worse than the threshold. Subsequently, TPE adopts the Expected Improvement (EI) criterion to select new hyperparameters, striving to maximize the EI value of the new parameters as much as possible:

$$\mathrm{EI}(x) = \frac{l(x)}{\gamma l(x) + (1-\gamma)g(x)} \tag{8}$$

where γ is the probability of $y < y^*$.

2.5. Evaluation of model prediction performance

The accuracy and robustness of the previously developed water flooding benchmark production prediction model and profile control effect prediction model were assessed using four standardized evaluation indicators: Mean absolute error (MAE), Mean absolute percentage error (MAPE), Root mean squared error (RMSE) and Coefficient of determination (R^2). The smaller the three types of errors and the closer the coefficient of determination is to 1, the better the model stability and accuracy. Given the total sample size N, actual target values y_i, predicted values $\hat{y}$ and mean actual value $\bar{y}$, the computational formulations for these indicators are presented in Table 3.

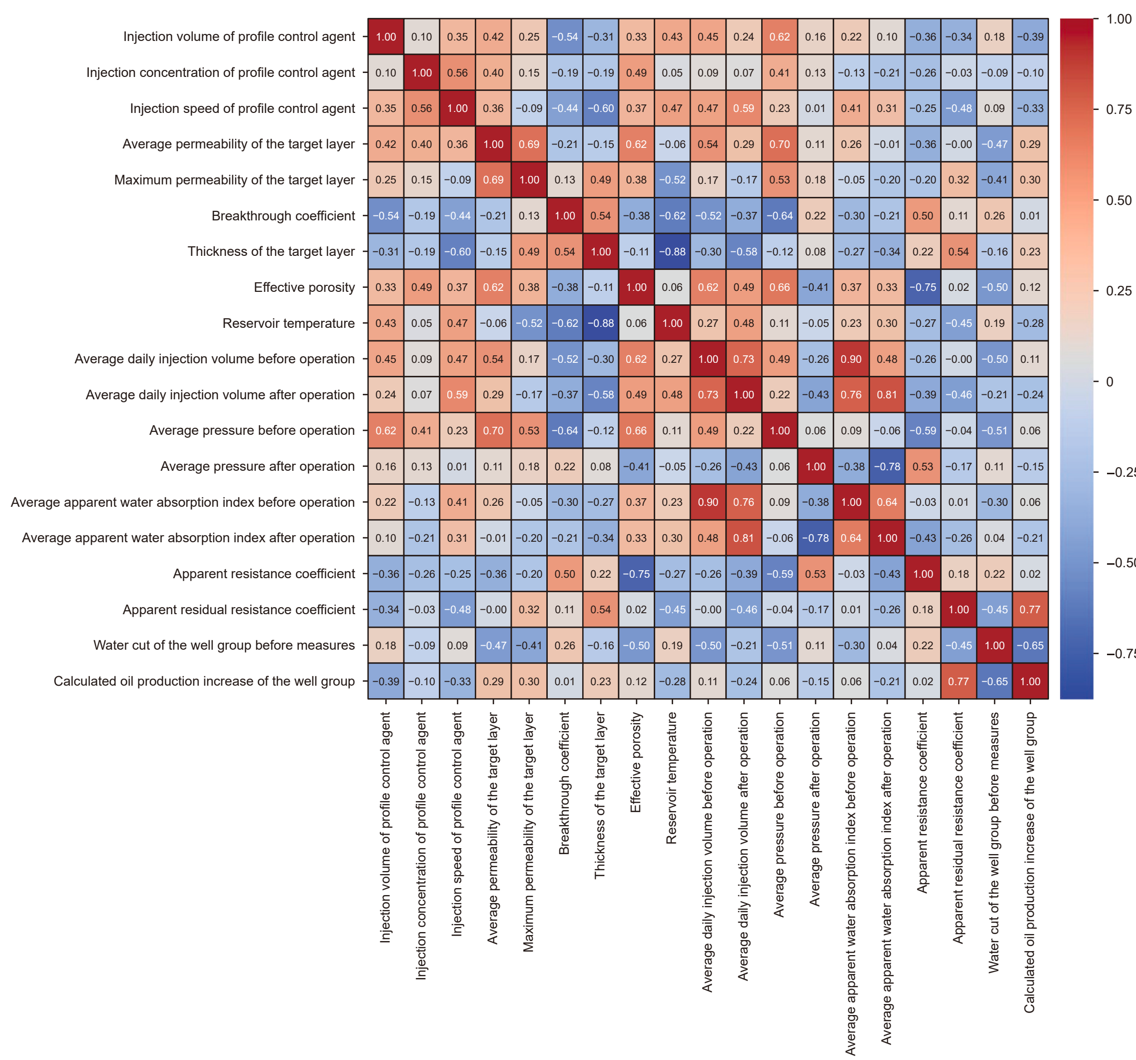

Fig. 7. Pearson correlation coefficient heatmap of each input feature.

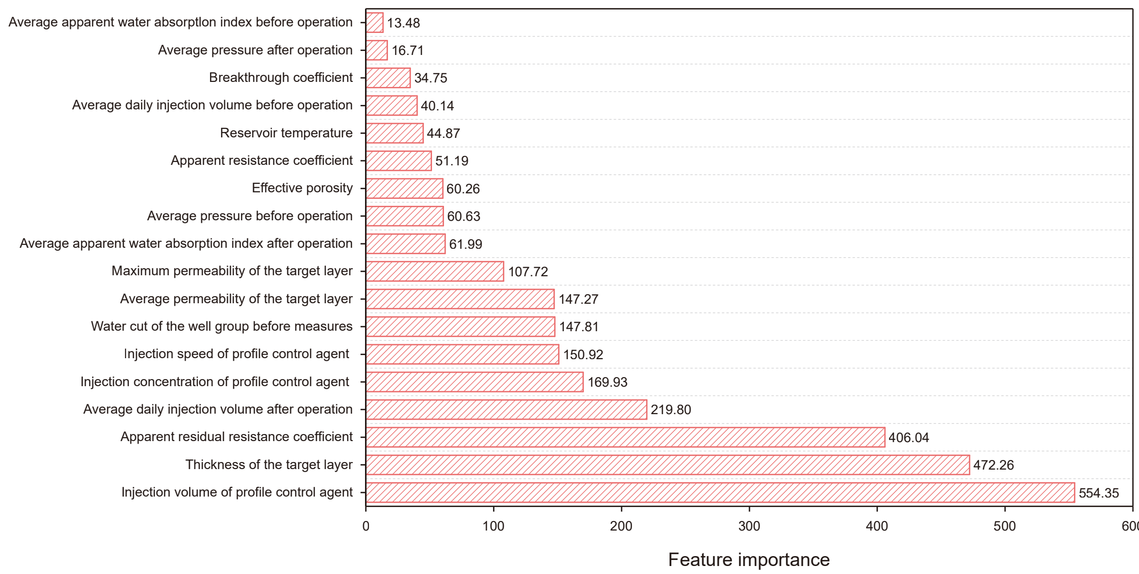

Fig. 8. SHAP value feature importance.

3. Results and discussions

3.1. Prediction of water flooding benchmark production

Table 4 presents the optimized hyperparameter configurations for each time-series prediction model. Utilizing the optimal hyperparameters from Table 4, the water flooding benchmark production model was trained.

The comparative prediction results are illustrated in Fig. 6 and the model performance metrics are detailed in Table 5. As shown in Fig. 6(a) and (b), LSTM exhibits rapid initial loss reduction followed by stable convergence, indicating robust model training without overfitting. The predicted and actual oil production data points predominantly cluster near the $y = x$, demonstrating strong agreement between predictions and actual values. Table 5 reveals that LSTM achieves low prediction errors, with R^2 values of 0.9874 (training set) and 0.9508 (test set), confirming high predictive accuracy and the absence of significant overfitting.

Fig. 6(c) and (d) demonstrate that GRU exhibits rapid initial loss reduction followed by stable convergence, indicating robust training dynamics without overfitting. Although the predicted and actual oil production data points predominantly cluster near $y = x$, the dispersion of data points is marginally greater compared to LSTM. Table 5 reveals that GRU has relatively small errors in the training and test sets, with R^2 values of 0.9834 (training set) and 0.9379 (test set), which are slightly lower than LSTM, indicating that the performance of the two models is similar.

Fig. 6(e) demonstrates that the loss value of Transformer has a certain degree of volatility, indicating poor model stability. Fig. 6(f) demonstrates that the overall deviation of Transformer's predicted oil production and actual oil production data from line $y = x$ is greater, indicating that the model has lower prediction accuracy. Table 5 reveals that Transformer has higher errors than LSTM and GRU, with R^2 values of 0.5170 (training set) and 0.8918 (test set), indicating that the model has poor generalization ability for unknown data. Overall, LSTM performs better than GRU and Transformer, making it suitable as a benchmark production prediction model for water flooding to participate in subsequent production relative difference calculations. This may be due to the fact that the number of parameters in the Transformer model is greater than that in the LSTM, and the limited sample size of the training set in this study is insufficient to support the large number of parameters in the Transformer (Yang and Wu, 2025). At the same time, the high computational complexity of the Transformer's full attention mechanism makes it difficult to handle noise interference in long sequences, resulting in the model being sensitive to data noise and prone to overfitting.

3.2. Profile control effect prediction

Using the Pandas tool in Python, Pearson correlation coefficient method was used to analyze the linear correlation between various features, and the results are shown in Fig. 7. And Fig. 8 shows the feature importance data calculated by SHAP algorithm,

Table 6
Well group parameters after data combination.

Average permeability, 1.0×10^{-3} μm^2	Breakthrough coefficient	Target layer thickness, m	Porosity, %	Water cut of the well group before measures, %	Average apparent water absorption index	Cumulative mass of profile control agents, kg	Accumulated water injection volume, m^3
3725	1.65	56.4	33.2	75	80.77	0.95	0
3725	1.65	56.4	33.2	75	80.77	1.91	0
…							
2941	2.06	48.6	31.0	89	104.14	19.80	123279
2941	2.06	48.6	31.0	89	104.14	19.80	123737

Table 7
The optimal hyperparameter combination for the prediction model of profile control effect.

Model hyperparameter	AdaBoost	XGBoost	RF
*n*_estimators	200	300	300
Max_depth	/	6	5
Min_samples_split	/	/	4
Min_samples_leaf	/	/	2
Max_features	/	/	0.8
Learning_rate	0.1	0.05	/
Subsample	/	0.8	/
Colsample_bytree	/	0.7	/
Reg_alpha	/	0.1	/
Reg_lambda	/	1	/

where the horizontal axis represents the factors affecting production and the vertical axis represents the calculated feature importance score. The higher the feature importance score, the greater the impact of the factor on oil production.

Fig. 7 reveals the that the correlation coefficient between the average permeability and the highest permeability stands at 0.69, suggesting a strong linear correlation. According to reservoir engineering knowledge and expert knowledge, it is evident that the average permeability offers a more accurate representation of the reservoir's physical properties. Consequently, the highest permeability is disregarded. There is a strong linear correlation between the injection speed of profile control agents and both the injection volume and concentration of profile control injection fluid. According to reservoir engineering knowledge and expert knowledge, it is observed that under standard production conditions, an accelerated injection rate corresponds to an increased injection volume per unit of time, demonstrating a clear positive correlation between the two variables. Therefore, the injection speed can be indirectly reflected by the injection volume. In order to reduce redundant features, the injection speed for profile control is excluded. There is a strong linear correlation between the average visual water absorption index before operation and the average daily injection volume before and after operation. Based on knowledge and expert experience in reservoir engineering, tit is recognized that the average apparent water absorption index before operation effectively characterizes the reservoir's water absorption capacity. Consequently, the average daily injection volumes before and after operation are excluded. There is a negative correlation between water cut and cumulative oil production in the well group. From the perspective of reservoir engineering and expert experience, water cut reflects the degree of water drive development in the reservoir. The higher the water cut, the higher the degree of oil recovery in the reservoir. The dominant flow channels for water drive gradually form, leading to an imbalance in oil-water distribution and ineffective water

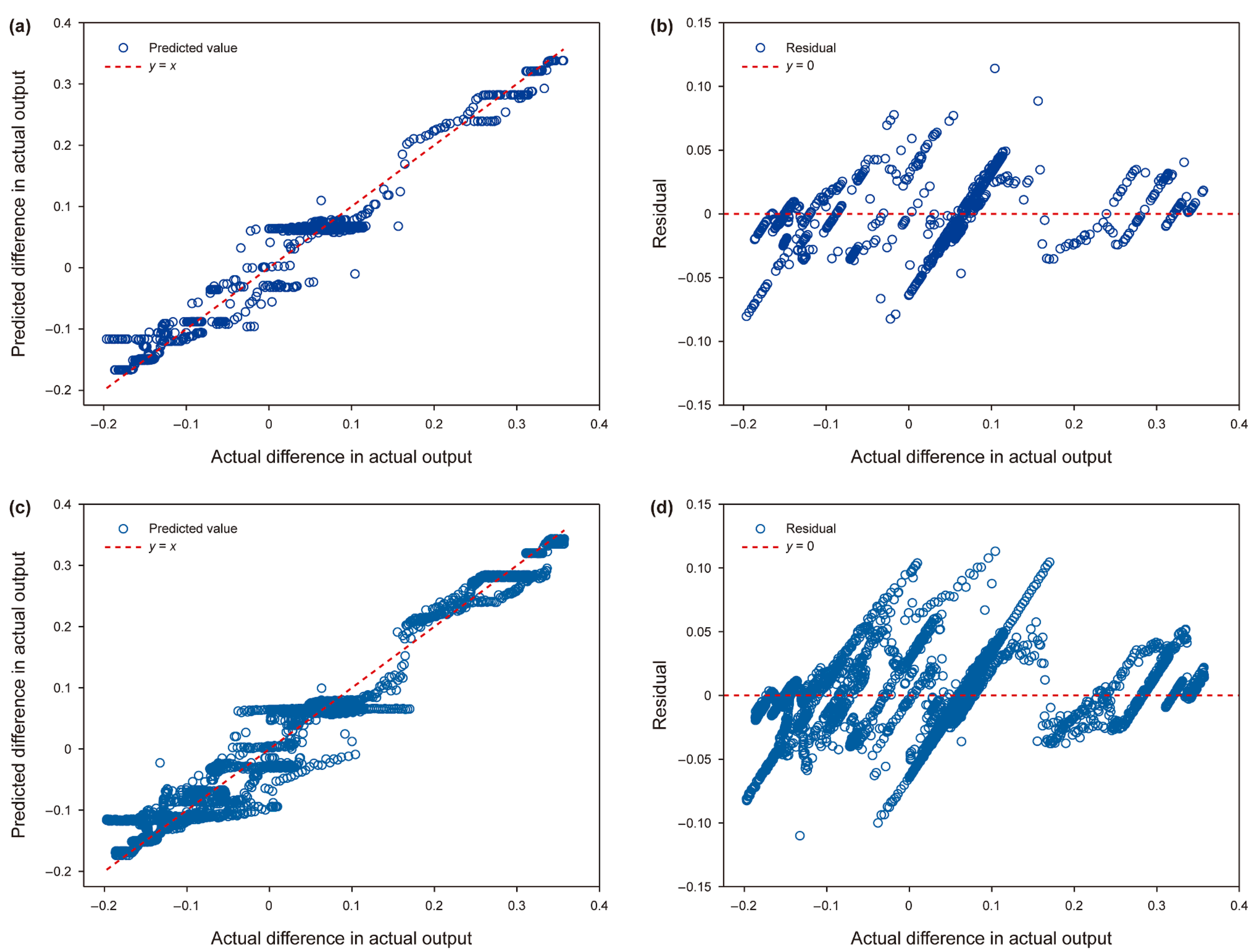

Fig. 9. AdaBoost model prediction results. **(a)** Comparison between the relative difference between the actual production and the predicted production in the test set. **(b)** Test set prediction residuals. **(c)** Comparison between the relative difference between the actual production and the predicted production in cross validation test set. **(d)** Cross validation test set prediction residuals.

circulation, ultimately resulting in a decrease in oil production. Therefore, water cut of the well group before measures should be the main controlling factor for profile control. In addition, from the perspective of reservoir engineering and expert experience, the injection volume, injection concentration, and injection speed jointly determine the actual effect of profile control measures. Higher injection volume and concentration can effectively seal high permeability layers and improve water drive efficiency, but excessive injection volume and concentration may cause blockages in the near wellbore area, affecting subsequent injections. Furthermore, considering that the true factor affecting oil production is the mass actually injected into the formation, this study multiplies the injection volume by the injection concentration to convert them into a comprehensive parameter representing the total mass of the injected agent. This parameter will be used in subsequent research and analysis to replace the injection amount and injection concentration of the agent. The correlation coefficient between reservoir temperature and calculated oil production increase of the well group is only −0.28. Considering that the temperature range of S reservoir is 64–65 °C, the impact of reservoir temperature on profile control effect is not significant. Thus, this characteristic is excluded.

According to Fig. 8, the reservoir physical parameters that have a significant impact on oil production include target layer thickness, average permeability, porosity and post operation average apparent water absorption index. The construction parameters that have a significant impact on increasing oil production include injection volume, injection concentration, injection speed and average daily injection volume after operation of profile control agents. The production parameters that have a significant impact on increasing oil production include the well group water cut before operation and the apparent residual resistance coefficient. However, parameters such as apparent residual resistance coefficient need to be calculated through dynamic data after profile control operations and cannot be used for predicting the production of new well groups. Therefore, these parameters are excluded when predicting the effect of profile control on increasing production. Breakthrough coefficient is the ratio of the maximum permeability to the average permeability of the target reservoir, reflecting the degree of reservoir heterogeneity. Research has shown that the magnitude of the breakthrough coefficient has a significant impact on the profile control effect (Qi et al., 2018). Combining expert experience, this study incorporates this parameter as a key controlling factor for profile control effectiveness. Based on the results of feature importance analysis and expert experience, the remaining parameters have limited impact on oil production increase and are not considered.

Through systematic evaluation of feature correlations and importance metrics, static and dynamic parameters were

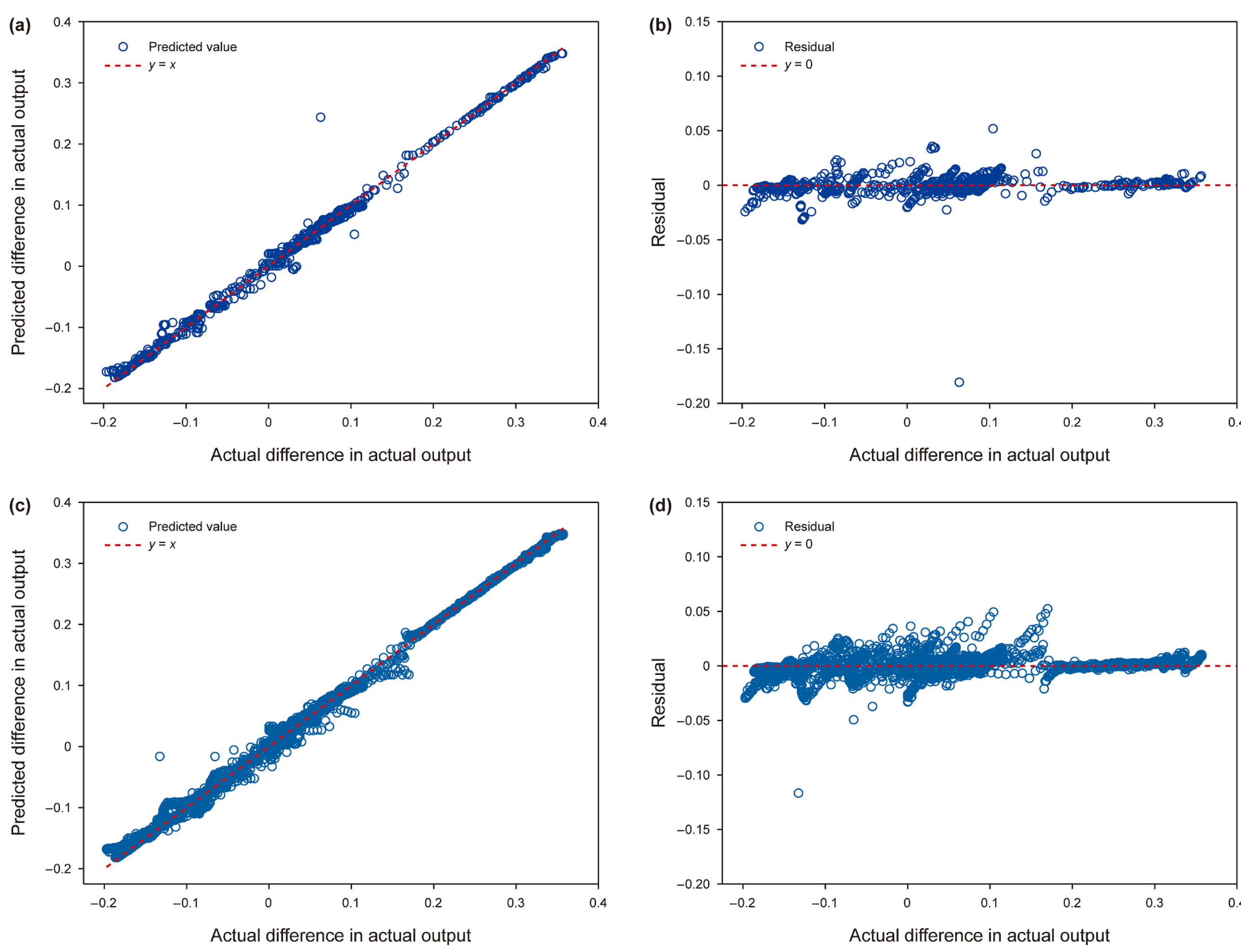

Fig. 10. XGBoost model prediction results. **(a)** Comparison between the relative difference between the actual production and the predicted production in the test set. **(b)** Test set prediction residuals. **(c)** Comparison between the relative difference between the actual production and the predicted production in cross validation test set. **(d)** Cross validation test set prediction residuals.

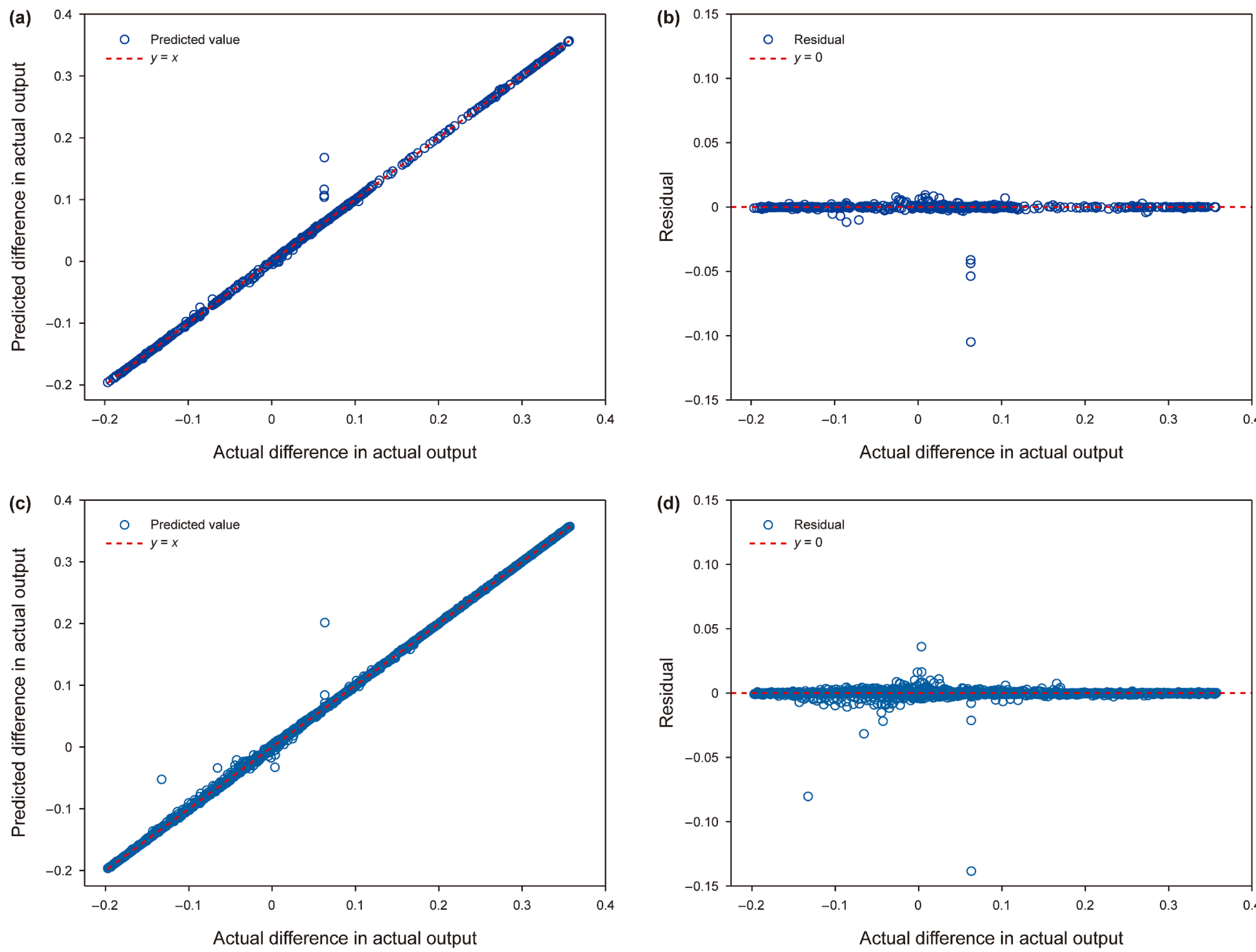

Fig. 11. RF model prediction results. **(a)** Comparison between the relative difference between the actual production and the predicted production in the test set. **(b)** Test set prediction residuals. **(c)** Comparison between the relative difference between the actual production and the predicted production in cross validation test set. **(d)** Cross validation test set prediction residuals.

combined to construct the final input features of profile control effect prediction models (Table 6).

After feature selecting, AdaBoost, XGBoost and RF were used to construct predictive models for profile control effect. Subsequently, 200 Bayesian optimization were conducted using the Optuna framework. The optimal parameter combination was obtained as shown in Table 7.

The study allocated 80% of the dataset to training set and 20% to test set, and trained profile control effect prediction models using the optimal hyperparameter configurations from Table 7. To fully utilize the dataset, the K-fold cross validation method ($K = 10$) was used to further verify the model's fitting ability and generalization performance. Figs. 9–11 respectively present the comparison and residual values of the relative difference between actual and predicted yields in three prediction models, respectively.

Fig. 9 shows that in both the test set and cross validation evaluation, most data points deviate from the baseline, indicating significantly poor prediction accuracy and limited generalization performance of AdaBoost. Fig. 10 shows that in both the test set and cross-validation evaluations, most data points slightly deviate from the baseline. Notably, cross validation exacerbates these deviations. This indicates that while XGBoost achieves satisfactory prediction accuracy, it exhibits limited generalization capability and compromised stability. Fig. 11 shows that in both the test set and the overall evaluation of cross validation, most of the data points are very close to the baseline, with only a few data points deviating. This indicates that predicted values of RF are close to the true values, and model has high prediction accuracy and strong generalization ability in different data ranges.

Further utilizing performance evaluation indicators to evaluate the RF model, the results are shown in Table 8. It can be seen that the MAE and MAPE values are low, and the R^2 value is close to 1. The consistency of indicator values in the test set before and after cross-validation demonstrates that RF model achieves high prediction accuracy without overfitting. Meanwhile, the low RMSE further indicates its excellent robustness against extreme outliers, suggesting superior generalization capability compared to both AdaBoost and XGBoost. Consequently, RF was selected for subsequent on-site profile control effect prediction application.

Table 8
RF model performance evaluation.

Data set	MAE	RMSE	MAPE	R^2
Test set	0.0008	3.63%	0.0047	0.9988
Cross validation test set	0.0006	4.24%	0.0029	0.9995

4. Model application

This study takes the production and development data of well group A in S reservoir as experimental data. The well group has one injection well and three production wells. The reservoir

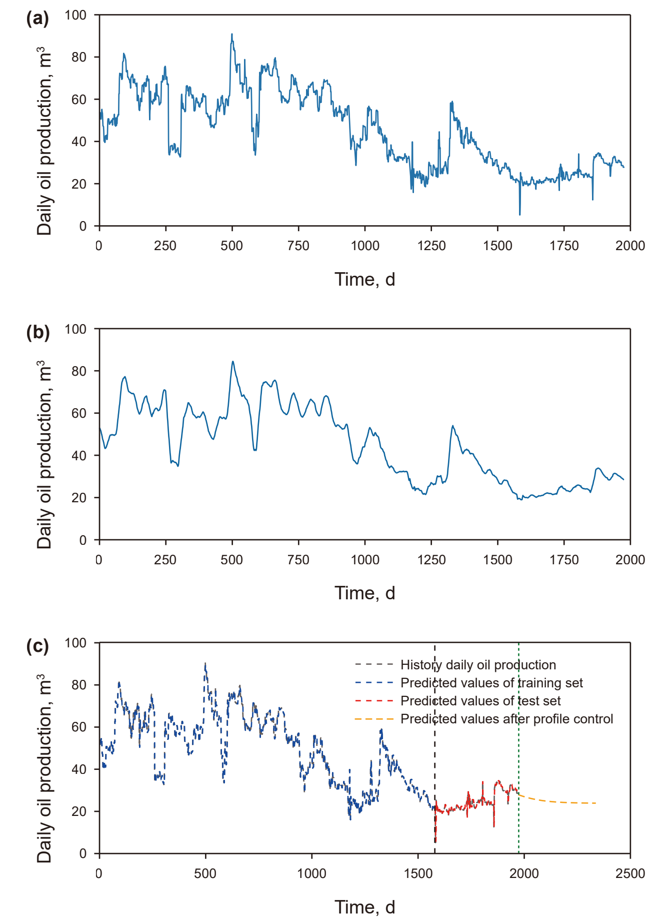

Fig. 12. Oil production curve of well group A. **(a)** Actual oil production curve of well group A. **(b)** Oil production curve of well group A after data preprocessing. **(c)** Prediction results of water drive benchmark production in well group A.

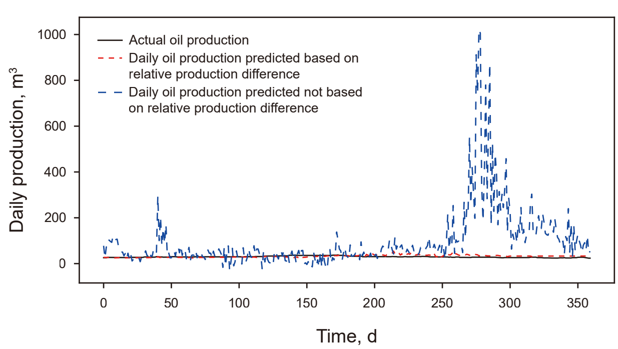

Fig. 13. Comparison between predicted daily production and actual daily production after profile control of well group A.

Table 9
Performance evaluation of the prediction model for profile control effect in well group A.

Model	Predicting cumulative production after profile control, m^3	MAE	MAPE	RMSE	R^2
RF	11222.35	5.44	18.89%	6.79	0.91
Actual cumulative production, m^3	10627.37	–	–	–	–

temperature is 64 °C, the average permeability is 3848 × 10^{-3} μm^2, the breakthrough coefficient of 2.82, the target layer thickness is 29.3 m, the effective porosity is 31%, the average apparent water absorption index is 29.43. The total water cut of this well group reaches 93.6%, which indicates that it has entered the high water cut stage. The oil production data of the well group before profile control measures was extracted, and a trend chart of oil production over time was plotted (Fig. 12). Fig. 12(a) shows that the water flooding curve of the well group conforms to the natural decline law and can be used for subsequent research.

After outlier treatment and missing values imputation on the on-site data, the daily oil production curve of water flooding was obtain as shown in Fig. 12(b). Use the established LSTM water flooding benchmark production prediction model to predict the water drive production of well group A for the next year, as shown by the yellow curve in Fig. 12(c).

The RF model is utilized to predict the relative difference in production within one year after the start of profile control measures. In order to verify the effectiveness of the method proposed in this study to characterize the profile control effect based on the relative difference in production, the RF model established in the previous section was used to directly predict the oil production after profile control. By employing back-calculation based on Eq. (3), we obtain the predicted cumulative oil production after profile modification for that year. Subsequently, a discrete difference calculation was performed to obtain the predicted daily oil production after profile control. The results are shown in Fig. 13, and the model evaluation metrics are presented in Table 9. To assess the uncertainty of the model's predictions and better support on-site decision-making, the 95% prediction interval (PI) of the model was quantified, and prediction interval coverage probability (PICP) was calculated, as illustrated in Fig. 14.

Based on Fig. 13 and Table 9 and it can be observed that the RF exhibits good predictive performance on the new well group. The overall trend of the predicted daily oil production (red line) is consistent with the actual daily oil production after profile control (black line), with small error values. The R^2 value exceeds 0.9, and the predicted cumulative production after profile control closely matches the actual cumulative production after profile control. As clearly illustrated in Fig. 14, the PICP of PI is 95.3%, indicating that the probability of the confidence interval of this model covering the actual value is relatively high, and the model has good reliability. Therefore, RF for predicting profile control effects constructed in this study also demonstrates good performance when applied to the new well group. However, if the oil production after profile control is directly predicted, the prediction curve exhibits strong fluctuations and significant deviations (as shown by the blue broken line in Fig. 13). Moreover, the predicted total oil production is 34908.62 m^3, which significantly differs from the actual

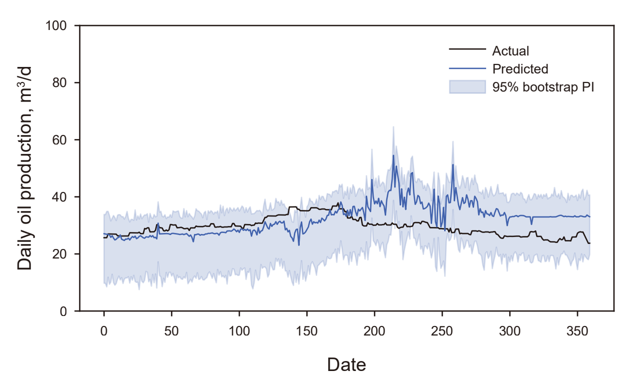

Fig. 14. Production interval of daily oil production.

oil production of 10627.37 m^3. It can be seen that the profile control effect prediction method proposed in this study is scientific and more accurate than the conventional method of directly predicting absolute production. This is because the method proposed in this study adopts the relative difference in production as an evaluation index for profile control effect. As mentioned earlier, this approach enables direct comparison of profile control effects between different well groups. At the same time, the method comprehensively considers the synergistic effects of geological parameters, engineering parameters, and production parameters on profile control effect, effectively enhancing the model's interpretation ability for profile control patterns and prediction accuracy for profile control effects.

5. Conclusions

This study focuses on the complexity and specificity of profile control effect prediction (e.g., being influenced by multiple factors like production dynamic parameters, and the inability to directly compare the production increment from profile control in different well groups). It innovatively constructs a profile control effect prediction framework of “water drive benchmark production + relative difference”, achieving efficient and intelligent prediction of profile control and production increase effect in oil reservoirs. This framework considers the joint influence of geological parameters, construction parameters, and production dynamic parameters on the effect of profile control, and can effectively distinguish the impact of natural decline and profile control measures on oil production. It has high prediction accuracy and strong engineering applicability. The main research results are as follows.

(1) This study constructed time-series prediction models for water flooding production based on LSTM, GRU, and Transformer algorithms. Meanwhile, the Optuna framework was used for hyperparameter optimization. After comparison, it was found that the R^2 of the LSTM training set and test set both exceeded 0.9, indicating that LSTM has the highest prediction accuracy and strongest robustness among the three models. Therefore, LSTM is selected to establish a water flooding benchmark production prediction model, which provides a high-quality data foundation for further of profile control effect prediction.

(2) Innovatively proposing the relative difference between the actual production after profile control and the benchmark production of water flooding to evaluate the effect of profile control measures. By combining expert knowledge, Pearson correlation coefficient method and SHAP analysis method, the main controlling factors of profile control effect were screened. These factors for the selected profile control effect include average permeability, breakthrough coefficient, target layer thickness, porosity, water cut, average apparent water absorption index, cumulative amount of substance of injected chemicals, and cumulative water injection volume.

(3) AdaBoost, XGBoost, and RF algorithms were used to establish profile control effect prediction models based on the relative difference in production, and the Optuna framework was used to achieve hyperparameter optimization. After comparison, it was found that both the R^2 of the RF model test set and the cross-validation test set exceeded 0.99, indicating that RF has the highest prediction accuracy and strongest robustness among the three models. Therefore, RF is utilized to establish a predictive model for profile control effect. Based on the relative difference calculation formula, the predicted oil production for one year after profile control can be inferred, and further discrete difference processing can be used to obtain the predicted daily oil production for that year.

(4) The on-site application results prove that the profile control effect prediction method established in this study has strong engineering practicality, which can accurately reflect the actual production increase contribution of profile control measures, effectively distinguish the effects of natural decline and profile control intervention, and significantly improve the reliability of profile control effect prediction. Compared with existing prediction methods, the method proposed in this study has higher prediction accuracy and can provide effective theoretical guidance for chemical profile control in high water cut oil reservoir, with broad application prospects.

(5) It should be noted that the model training data in this study originates from oilfield sites, recorded either through sensors or manually. There may be some deviations from the true values, and a significant number of feature parameter values are missing. Despite rigorous data preprocessing, this may still affect the model's prediction accuracy to some extent. Currently, this study has only applied the proposed profile control effect prediction model to certain oil reservoirs in offshore oilfields and has not yet been used in onshore oilfields. In the future, the application scope will be expanded, and the model will be continuously optimized to enhance its generalization ability and robustness. Furthermore, hybrid ML models and transfer learning models have been successfully applied in numerous fields (Gu et al., 2025; Li et al., 2026; Zhu et al., 2026; Mohammad et al., 2026). In the future, the prediction framework proposed in this study can be combined with these technologies to achieve more efficient and accurate calculation of the effect of reservoir profile control.

CRediT authorship contribution statement

Kun Xie: Writing – original draft, Methodology, Funding acquisition, Data curation. **Zhan-Qi Wu:** Writing – original draft, Validation, Investigation, Data curation. **Xuan-Shuo Tian:** Writing – review & editing, Visualization, Investigation. **Wei-Jia Cao:** Writing – review & editing, Methodology, Conceptualization. **Xin Song:** Supervision, Formal analysis. **Er-Long Yang:** Validation, Investigation. **Kun Yan:** Writing – review & editing, Investigation. **Xiang-Guo Lu:** Writing – review & editing, Supervision.

Declaration of competing interest

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

Acknowledgements

The following projects are acknowledged to support our work: the National Natural Science Foundation Project (Grant No. 52204037) and the Key Research & Development Program of Heilongjiang Province (Grant No. JD24A008).

Nomenclature

MSE	Mean square error
x_i	Input vector

$\widehat{x}_i$	Reconstructed vector
N	The number of elements in the vector
τ	Anomaly detection threshold
μ	Mean of reconstruction error
k	Hyperparameter
σ	Standard deviation of reconstruction error
PRD	Relative difference between the actual production after profile control and the predicted water flooding benchmark production
$Q_p(t)$	Actual daily cumulative oil production after profile control
$Q_w(t)$	Rredicted water flooding benchmark production
Q_d	Daily production after profile control
Q_i	Cumulative production from the start of profile control to the i-th day ($i > 1$)
Q_{i-1}	Cumulative production from the start of profile control to the ($i-1$)th day
$P(x\|y)$	The posterior probability of traditional Bayesian algorithms
$l(x)$	The distribution of parameters with objective function values better than the threshold
$g(x)$	The distribution of parameters with objective function values worse than the threshold
EI(x)	The EI value of the new parameters
r	Pearson correlation coefficient
X_i, Y_i	Random variable
$\overline{X}$, $\overline{Y}$	Mean value of random variable
n	Total sample size
x'	Normalized data
x	Specific values in the data
max(x)	The maximum value of feature data
min(x)	The minimum value of feature data
MAE	Mean absolute error
MAPE	Mean absolute percentage error
RMSE	Root mean squared error
R^2	Coefficient of determination

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M., 2019. Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. https://doi.org/10.1145/3292500.3330701.

Alzakari, S.A., Alhussan, A.A., Qenawy, A.S.T., Elshewey, A.M., Eed, M., 2025. An enhanced long short-term memory recurrent neural network deep learning model for potato price prediction. Potato Res. 68 (1), 621–639. https://doi.org/10.1007/s11540-024-09744-x.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Cao, W.J., Xie, K., Lu, X.G., Liu, Y.G., Zhang, Y.B., 2019. Effect of profile-control oil-displacement agent on increasing oil recovery and its mechanism. Fuel 237, 1151–1160. https://doi.org/10.1016/j.fuel.2018.10.089.

Chen, X.C., Feng, Q.H., Zhang, A.G., Shi, S.B., 2014. Prediction and evaluation of gel beads profile control performance after polymer flooding. Oil Drill. Prod. Technol. 36, 82–86. https://doi.org/10.13639/j.odpt.2014.03.021 (in Chinese).

Chen, W., Yang, L.Q., Zha, B., Zhang, M., Chen, Y.K., 2020. Deep learning reservoir porosity prediction based on multilayer long short-term memory network. Geophys 85, WA213–WA225. https://doi.org/10.1190/geo2019-0261.1.

Chen, M.J., Meng, X.M., Kan, G.M., Wang, J.Q., Li, G.B., Liu, B.H., Liu, C.G., Liu, Y.G., Liu, Y.X., Lu, J.J., 2022. Predicting the sound speed of seafloor sediments in the east China sea based on an XGBoost algorithm. J. Mar. Sci. Eng. 10, 1366. https://doi.org/10.3390/jmse10101366.

Chen, G.X., Tian, H.L., Xiao, T., Xu, T.F., Lei, H.W., 2024. Time series forecasting of oil production in enhanced oil recovery system based on a novel CNN-GRU neural network. Geoenergy Sci. Eng. 233, 212528. https://doi.org/10.1016/j.geoen.2023.212528.

El-Rashidy, N., Tarek, Z., Elshewey, A.M., Shams, M.Y., 2025. Multitask multilayer-prediction model for predicting mechanical ventilation and the associated mortality rate. Neural Comput. Appl. 37 (3), 1321–1343. https://doi.org/10.1007/s00521-024-10468-9.

Elsharafi, M.O., Bai, B., 2012. Effect of weak preformed particle gel on unswept oil zones/areas during conformance control treatments. Ind. Eng. Chem. Res. 51 (35), 11547–11554. https://doi.org/10.1021/ie3007227.

Elshewey, A.M., 2025. Enhancing crop yield prediction based on dove optimization algorithm and gradient boosting model. Signal, Image and Video Process 19 (11), 951. https://doi.org/10.1007/s11760-025-04545-2.

Elshewey, A.M., Shams, M.Y., Tarek, Z., Megahed, M., El-Kenawy, E.S.M., El-dosuky, M.A., 2023. Weight prediction using the hybrid Stacked-LSTM food selection model. Comput. Syst. Sci. Eng. 46 (1), 765–781. https://doi.org/10.32604/csse.2023.034324.

Feng, Q.H., Chen, Y.W., Jiang, H.Q., Zhang, J.J., Wang, S.Y., 1997. Prediction of the overall profile control effect of the block. J. China Univ. Pet. Nat. Sci. 21 (4), 32–34+116–117 (in Chinese).

Feng, Q.H., Chen, Y.W., Guo, J.L., 2003. Research on the Statistical Method for Predicting the Profile Control Effect of Water Injection Wells. Oil Drill. Prod. Technol. 25 (6), 63–66+87. https://doi.org/10.13639/j.odpt.2003.06.020 (in Chinese).

Gu, X.Q., Li, Y., Xiao, X., Wu, L.F., Yu, H.F., 2025. Optimizing methane flux prediction and key feature identification based on a novel hybrid machine learning model. iScience 28 (12), 114011. https://doi.org/10.1016/j.isci.2025.114011.

He, J., Wen, X.T., Li, B., Chen, Q.S., Li, L., 2022. The pre-stack fluid identification method based on random forest algorithm. Acta Pet. Sin. 43 (3), 376–385. https://doi.org/10.7623/syxb202203005 (in Chinese).

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Hou, J.R., Song, K.P., Wen, Y.C., 2023. Development trend of enhanced oil recovery technology in old oilfields after polymer flooding. Sci. Technol. Foresight. 2 (2), 47–61. https://doi.org/10.3981/j.issn.2097-0781.2023.02.004 (in Chinese).

Jackins, V., Vimal, S., Kaliappan, M., Lee, M.Y., 2021. AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. J. Supercomput. 77 (5), 5198–5219. https://doi.org/10.1007/s11227-020-03481-x.

Jiang, X.M., Yan, W.C., Xing, H.L., Feng, P., Sun, J.M., Fan, Y.C., 2025. A highly accurate and interpretable gradient boosting machine learning model for predicting permeability in strongly heterogeneous reservoirs using the Optuna framework. J. Geophys. Eng. 22 (3), 810–823. https://doi.org/10.1093/jge/gxaf039.

Kocoglu, Y., Gorell, S.B., Emadi, H., Eyinla, D.S., Bolouri, F., Kocoglu, Y.C., Arora, A., 2024. Improving the accuracy of short-term multiphase production forecasts in unconventional tight oil reservoirs using contextual Bi-directional long short-term memory. Geoenergy Sci. Eng. 235, 212688. https://doi.org/10.1016/j.geoen.2024.212688.

Lahiri, A., Paria, B., Biswas, P.K., 2016. Forward stagewise additive model for collaborative multiview boosting. IEEE Trans. Neural Netw. Learn. Syst. 29, 470–485. https://doi.org/10.1109/TNNLS.2016.2623219.

Lee, K., Lim, J., Yoon, D., Jung, H., 2019. Prediction of shale-gas production at Duvernay formation using deep-learning algorithm. SPE J. 24 (6), 2423–2437. https://doi.org/10.2118/195698-PA.

Li, Z.Q., 2022. Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. Comput. Environ. Urban Syst. 96, 101845. https://doi.org/10.1016/j.compenvurbsys.2022.101845.

Li, J., Fargalla, M.A.M., Yan, W., Zhang, Z.X., Zhang, W., Zou, Z.C., Qing, T., Yang, T., Tan, C.D., Li, G.C., 2025. A novel fusion of interpretable boosting algorithm and feature selection for predicting casing damage. Pet. Sci. 22 (10), 4157–4173. https://doi.org/10.1016/j.petsci.2025.08.011.

Li, S.P., Wen, S., Liu, F.G., Zhang, H.Y., Ge, Y.F., Deng, Z.Y., 2026. Research on real time mapping and transfer learning based 3D semantic segmentation of unstructured orchards. Comput. Electron. Agric. 242, 111323. https://doi.org/10.1016/j.compag.2025.111323.

Liu, X.P., Zhou, K.H., 1994. Numerical simulation of adjusting water injection profile with solid particle type profile control agent. J. Oil Gas Technol. (J. Jianghan Petroleum Inst.) 16 (3), 52–56 (in Chinese).

Liu, N., Liu, S.M., Li, M., 2014. Prediction model of conformance control effect based on Bp neural network. Complex Hydrocarbon Reservoirs 7 (2), 51–53. https://doi.org/10.16181/j.cnki.fzyqc.2014.02.016 (in Chinese).

Luo, Y.P., He, A.L., Yu, K.M., Zhang, K., Liao, F.Y., 2026. Structural damage identification method based on convolutional neural network and autoencoder. Steel Constr. (Chin. Engl.) 1–12. https://link.cnki.net/urlid/10.1609.TF.20251231.1740.002.

Min, C., Dai, B.R., Zhang, X.H., Du, J.P., 2020. A review of the application progress of machine learning in oil and gas industry. J. Southwest Pet Univ. (Sci. Technol. Ed.) 42 (6), 1–15. https://doi.org/10.11885/j.issn.1674-5086.2020.06.05.03 (in Chinese).

Mohammad, S.I., Owida, H.A., Vasudevan, A., Ballal, S., Hasnaawei, S.I., Ray, S., Talniya, N.C., Aashna, S., Vatsal, J., Ahmad, A., 2026. Hybrid machine learning models for predicting density, dynamic viscosity, and speed of sound in aliphatic biogenic polyamines. Results Eng. 29, 108741. https://doi.org/10.1016/j.rineng.2025.108741.

Qi, Y.M., Ding, J.G., Zhang, H.L., Zhou, J., 2018. The choosing decision technique study of injector well for profile control based on geology and production data. J. Oil Gas Technol. (J. Jianghan Petroleum Inst.) 40 (6), 83–89. https://doi.org/10.12677/jogt.2018.406124.

Rodriguez, A.X., Salazar, D.A., 2022. Methodology for the prediction of fluid production in the waterflooding process based on multivariate long–short term memory neural networks. J. Pet. Sci. Eng. 208, 109715. https://doi.org/10.1016/j.petrol.2021.109715.

Song, X., Liu, Y., Xue, L., Wang, J., Zhang, J., Wang, J., Jiang, L., Cheng, Z., 2020. Time-series well performance prediction based on Long Short-Term Memory (LSTM) neural network model. J. Petr. Sci. Eng. 186, 106682. https://doi.org/10.1016/j.petrol.2019.106682.
Song, K.P., Tao, J.W., Lyu, X.Q., Xu, Y., Liu, S.P., Wang, Z.B., Liu, H.F., Zhang, Y.X., Fu, H.T., Meng, E., Liu, M.X., Hu, G., 2022. Recent advances in polymer flooding in China. Molecules 27 (20), 6978. https://doi.org/10.3390/molecules27206978.
Song, S.H., Zhou, Q.J., Zhang, T., Hu, Y.T., 2023. Automatic concrete dam deformation prediction model based on TPE-STL-LSTM. Water 15 (11), 2090. https://doi.org/10.3390/w15112090.
Su, L.L., Zuo, X.M., Li, R.L., Wang, X.W., Zhao, H., Huang, B.D., 2025. A systematic review for transformer-based long-term series forecasting. Artif. Intell. Rev. 58 (3), 80. https://doi.org/10.1007/s10462-024-11044-2.
Sun, F.J., 2023. Status and prospects of efficient development and EOR technologies in China offshore oilfields. China Offshore Oil Gas 35 (5), 91–99. https://doi.org/10.11935/i.issn.1673-1506.2023.05.009 (in Chinese).
Sun, D.Y., Zhang, H.B., Huang, Q., 2014. Application of sliding removal of rescaled variance in dynamic structural mutation testing. Acta Phys. Sin. 63 (20), 434–441. https://doi.org/10.7498/aps.63.209203 (in Chinese).
Tarek, Z., Shams, M.Y., Towfek, S.K., Alkahtani, H.K., Ibrahim, A., Abdelhamid, A.A., Eid, M.M., Khodadadi, N., Abualigah, L., Khafaga, D.S., 2023. An optimized model based on deep learning and gated recurrent unit for COVID-19 death prediction. Biomim 8 (7), 552. https://doi.org/10.3390/biomimetics8070552.
Tariq, Z., Mahmoud, M., Abdulraheem, A., 2019. An artificial intelligence approach to predict the water saturation in carbonate reservoir rocks. In: Proceedings of SPE Annual Technical Conference and Exhibition. https://doi.org/10.2118/195804-MS. SPE-195804-MS.
Wang, X.Y., 2024. Prediction method for deep profile control effect based on PSO-SVR. Comput. Program. Skills Maint. 31 (5), 22–24+84. https://doi.org/10.16184/j.cnki.comprg.2024.05.029 (in Chinese).
Wang, X.R., Cai, H., Zhang, G.H., Liu, D., Wang, Y.P., 2021a. Effect of polymer flooding on oil field developmentunder intra-layer heterogeneity. Nat. Gas Oil 39, 56–61. https://doi.org/10.3969/j.issn.1006-5539.2021.03.009 (in Chinese).
Wang, L., Fu, X.B., 2018. Data-driven analyses of low salinity water flooding in sandstones. Fuel 234, 674–686. https://doi.org/10.1016/j.fuel.2018.07.063.
Wang, Z.L., Shi, S.B., Zhang, X.M., Yin, X.W., Chao, Z., 2021b. Comprehensive evaluation of deep water shutoff and profile controlbased on V/S analysis. Pet. Geol. Recovery Effic. 28 (3), 77–83. https://doi.org/10.13673/j.cnki.cn37-1359/te.2021.03.009 (in Chinese).
Wang, Y.Y., Xu, D.H., Li, X.P., Wang, W., 2024. Prediction model of ammonia nitrogen concentration in aquaculture based on improved AdaBoost and LSTM. Math 12 (5), 627. https://doi.org/10.3390/math12050627.
Wu, N., Jiang, Y.Z., Jiang, W.D., 2007. Theoretical study on optimization of profile control parameters. Spec. Oil Gas Reserv. 14 (3), 91–94+110. https://doi.org/10.3969/j.issn.1006-6535.2007.03.026 (in Chinese).
Xie, K., Wu, Z.Q., Li, Y.Y., Mei, J., Cao, W.J., Zhao, W.Y., 2023. Application of machine learning in oil and gas development field and its prospect. J. Xi'an Shiyou Univ. (Nat Sci. Ed.) 38 (5), 58–67. https://doi.org/10.3969/i.issn.1673-064X.2023.05.008 (in Chinese).
Xie, K., Wu, Z.Q., Li, X.Y., Cao, W.J., Yang, E.L., Mei, J., Gao, M.X., Yuan, S.L., 2025. A review of research progress of in situ polyacrylamide gel for profile control and water plugging. Phys. Fluids 37 (9), 091303. https://doi.org/10.1063/5.0277891.
Xie, K., Wu, Z.Q., Liu, C.L., Mei, J., Cao, W.J., Ding, H.N., Zhang, X.Q., Xu, H.L., Suo, Y., Tian, X.S., Lu, X.G., 2024. A review of mechanisms and ML-based research on factors affecting spontaneous imbibition of surfactant. Geoenergy Sci. Eng. 213071. https://doi.org/10.1016/j.geoen.2024.213071.
Xu, Y.D., Ren, Y.P., Ding, L.C., Tang, X.H., 2007. Prediction of profile control result using support vector machines. Fault-Block Oil Gas Field 14 (2), 50–52+92. https://doi.org/10.3969/j.issn.1005-8907.2007.02.017 (in Chinese).
Yang, Y., Wu, Y.Q., 2025. Applications of vision Transformer in surface defect detection: research progress and challenges. Chin. J. Sci. Instrum. 1–22. https://doi.org/10.19650/j.cnki.cjsi.J2514531.
Yang, Q.C., Rong, D.W., Guo, Q.W., Bao, R.J., Zhang, D.W., 2025. Ensemble learning-driven multi-objective optimization of the co-pyrolysis process of biomass and coal for high economic and environmental performance. Chin. J. Chem. Eng. 84 (8), 23–34. https://doi.org/10.1016/j.cjche.2025.06.001.
Yeşilkanat, C.M., 2020. Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm. Chaos. Solitons Fractals 140, 110210. https://doi.org/10.1016/j.chaos.2020.110210.
Yu, J.X., Zhang, X., Xu, L.L., Dong, J., Zhangzhong, L.L., 2021. A hybrid CNN-GRU model for predicting soil moisture in maize root zone. Agric. Water Manag. 245, 106649. https://doi.org/10.1016/j.agwat.2020.106649.
Yuan, S., Qin, H.L., Yan, X., Akhtar, N., Mian, A., 2024. Sctransnet: Spatial-channel cross transformer network for infrared small target detection. IEEE Trans. Geosci. Remote Sens. 62, 1–15. https://doi.org/10.1109/TGRS.2024.3383649.
Zhou, S.W., Han, M., Xiang, W.T., Zhang, J., Zhou, W., Jiang, W., 2006. Application of EOR technology by means of polymer flooding in Bohai Oilfields. China Offshore Oil Gas 18 (6), 386–389+412. https://doi.org/10.3969/j.issn.1673-1506.2006.06.006 (in Chinese).
Zhu, Z.P., Han, Y.Y., Du, A.M., 2026. Enhanced eco-driving and energy management for heterogeneous hybrid electric vehicles through integrated transfer learning and deep reinforcement learning. Energy 342, 139626. https://doi.org/10.1016/j.energy.2025.139626.