Petroleum Science 22 (2025) 787–804

Contents lists available at ScienceDirect

Petroleum Science

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

Original Paper

Predicting the productivity of fractured horizontal wells using few-shot learning

Sen Wang [a, b, *], Wen Ge [b, c], Yu-Long Zhang [a, b], Qi-Hong Feng [a, b, d], Yong Qin [e], Ling-Feng Yue [a, b], Renatus Mahuyu [b], Jing Zhang [f]

[a] State Key Laboratory of Deep Oil and Gas, China University of Petroleum (East China), Qingdao, 266580, Shandong, China
[b] School of Petroleum Engineering, China University of Petroleum (East China), Qingdao, 266580, Shandong, China
[c] Research Institute of Exploration and Development, SINOPEC Jiangsu Oilfield, Yangzhou, 225009, Jiangsu, China
[d] Shandong Institute of Petroleum and Chemical Technology, Dongying, 257061, Shandong, China
[e] Research Institute of Petroleum Exploration and Development, PetroChina, Beijing, 100083, China
[f] Exploration and Development Research Institute, PetroChina Xinjiang Oilfield Company, Karamay, 834000, Xinjiang, China

ARTICLE INFO

Article history:
Received 25 June 2024
Received in revised form
1 November 2024
Accepted 3 November 2024
Available online 5 November 2024

Edited by Yan-Hua Sun

Keywords:
Fractured horizontal well
Machine learning
SMOTE
Few-shot learning
Prediction
Optimization

ABSTRACT

Predicting the productivity of multistage fractured horizontal wells plays an important role in exploiting unconventional resources. In recent years, machine learning (ML) models have emerged as a new approach for such studies. However, the scarcity of sufficient real data for model training often leads to imprecise predictions, even though the models trained with real data better characterize geological and engineering features. To tackle this issue, we propose an ML model that can obtain reliable results even with a small amount of data samples. Our model integrates the synthetic minority oversampling technique (SMOTE) to expand the data volume, the support vector machine (SVM) for model training, and the particle swarm optimization (PSO) algorithm for optimizing hyperparameters. To enhance the model performance, we conduct feature fusion and dimensionality reduction. Additionally, we examine the influences of different sample sizes and ML models for training. The proposed model demonstrates higher prediction accuracy and generalization ability, achieving a predicted R^2 value of up to 0.9 for the test set, compared to the traditional ML techniques with an R^2 of 0.13. This model accurately predicts the production of fractured horizontal wells even with limited samples, supplying an efficient tool for optimizing the production of unconventional resources. Importantly, the model holds the potential applicability to address similar challenges in other fields constrained by scarce data samples.

1. Introduction

Currently, there is a huge demand for oil and gas, while the production rate of conventional resources is declining, leading to a scarcity of these valuable resources. Alternatively, unconventional resources have become pivotal in the global energy landscape. Numerous countries are actively engaged in the exploration and development of unconventional oil and gas to meet the growing energy needs (Manfroni et al., 2022; Akbarabadi et al., 2023; Raza and Lin, 2023). The successful application of multistage hydraulic fracturing in horizontal wells has paved the way for the exploitation of unconventional resources. Predicting the productivity of fractured horizontal wells is crucial for their economic development and production optimization (Manjunath et al., 2023; Hakimi et al., 2023; Tangirala and Sheng, 2019).

Two approaches, numerical simulation, and analytical models, are commonly employed to predict the production of fractured horizontal wells. Numerical simulation involves creating a numerical model of a reservoir, simulating the fluid flow, performing history matching, and obtaining the production variation over time. This technique comprehensively accounts for various influencing factors and mechanisms in oil reservoirs, and it has high prediction accuracy. The numerical simulation models used for unconventional oil reservoirs include the dual-porosity model, discrete fracture model (DFM), embedded discrete fracture model (EDFM), etc. (Jia et al., 2021; Moinfar et al., 2011; Xu et al., 2018; Yu et al., 2018; Azom and Javadpour, 2012). However, numerical simulation always requires a large amount of data, such as geological

* Corresponding author.
E-mail address: fwforest@gmail.com (S. Wang).

https://doi.org/10.1016/j.petsci.2024.11.001

properties, rock and fluid properties, fracture network configurations, etc. It is not trivial to obtain this data, and the data quality has a crucial impact on the simulation results.

The analytical model involves establishing a mathematical model based on the equivalent flow resistance and the superposition theory to recognize the flow patterns and analytically estimate reservoir properties (Deng et al., 2014; Micheal et al., 2021; Afagwu et al., 2023). Among them, the production decline methods, such as the Arps model, the stretched exponential production decline (SEPD) model, and the power law exponential (PLE) decline model, have been widely used due to their simplicity and convenience (Arps, 1945; Ilk et al., 2008; Valkó and Lee, 2010; Blasingame et al., 1989; Fetkovich, 1980; Fraim and Wattenbarger, 1987; Alom et al., 2017). These decline models rely primarily on historical production data and do not explicitly account for physical parameters such as permeability, porosity, and fluid viscosity. Because the production decline method requires historical production data, it is commonly used for wells that have already been put into production. If one wants to use it in a new field, the two fields should be analogous.

In recent years, with the rapid development of artificial intelligence (AI) technologies, ML has been widely applied to complex problems in engineering fields (Cao et al., 2022; Tontiwachwuthikul et al., 2020; Kamrava et al., 2019; Yan et al., 2022; Wu et al., 2023; Liu et al., 2023). ML automatically discovers the rules and correlations through the learning and analysis of a large number of data. In the petroleum industry, ML has been utilized to perform tasks such as working condition diagnosis, reservoir performance prediction, and production optimization (Wang et al., 2023, 2024). Regarding the production forecasting of unconventional reservoirs, ML models are established using static and dynamic data to infer the implicit correlations between production rate and influencing factors. The advantages of ML models in accuracy and efficiency have led to their widespread use.

Supervised ML-based production prediction methods are categorized into two types: time series model and static model. This study focuses on the latter category; thus, we briefly introduce the current status of time-series studies and concentrate on static models. Time series methods forecast future production performance based on the well dynamic data at the earlier stages. Xue et al. (2023) used a long short-term memory (LSTM) to predict gas well production performance based on the production history of actual wells. Pan et al. (2021) proposed the Laplacian eigenmaps coupled echo-state network method that enables engineers to predict future production performance even with noisy, highly variable history data. Other commonly used techniques include auto-regressive integrated moving average (ARIMA), gated recursive unit (GRU), and prophet (Song et al., 2020; Fan et al., 2021; Ning et al., 2022; Werneck et al., 2022; Li et al., 2022; Zha et al., 2022; Chahar et al., 2022; Zhou et al., 2023; Pan et al., 2023; Jiang et al., 2023).

The static model predicts the production of new wells before fracturing operations based on various types of static parameters such as geological properties and hydraulic fracturing parameters. This kind of model is crucial for optimizing fracturing design because the well does not have to be already put into production. The commonly used methods include random forest (RF), deep neural networks (DNN), and SVR (Zhu et al., 2015; Bhattacharya and Mishra, 2018; Bhattacharya et al., 2019; Luo et al., 2019; Klie and Florez, 2020). Wang et al. (2019) collected 18 characteristic parameters from 2919 wells in Bakken Formation and constructed a DNN model to predict the cumulative oil production in unconventional reservoirs. They also examined the performance of several ML algorithms and concluded that RF is superior to other models, such as adaptive boosting (AdaBoost), SVR, and neural network (NN) (Wang and Chen, 2019). Wang et al. (2021) constructed a database based on numerical simulation and then predicted the production performance of unconventional reservoirs using the deep belief network (DBN) and Bayesian optimization algorithm. Nguyen-Le and Shin (2022) proposed three artificial neural network (ANN) architectures for predicting the peak production and Arps's hyperbolic decline parameters (D_i and b) of a shale gas well in the Montney Formation. Based on the actual data of the Duvernay shale gas field, Hui et al. (2023) constructed a dataset using eight factors from 251 wells to examine the performance of four ML algorithms, including RF, gradient boosting decision tree (GBDT), ANN, and extra tree (ET). Table 1 summarizes the ML-based production prediction models used for unconventional reservoirs in recent years.

A large number of data samples are required for the static model to predict the production of fractured wells. However, creating a sample set using numerical simulations is challenging, and the disparities between simulations and actual fields can lead to low accuracy and poor performance in practical applications. Note that the number of fractured horizontal wells in shale or tight oil reservoirs in China is relatively small, and some essential data are unavailable for these wells. The prevalent data quality issues lead to only limited samples being utilized for dataset preparation and an imbalanced distribution of different data within the available samples. Consequently, the trained model always shows poor generalization capability and low prediction accuracy, impeding the utilization in actual fields. Constructing a production prediction model for fractured horizontal wells using only limited samples is crucial for the exploitation of unconventional resources.

We propose a comprehensive framework for the production prediction of fractured horizontal wells even with a small amount of data samples. Initially, the factors influencing the well productivity are collected based on field data, which includes four types (11 factors): geological factors, fracturing design factors, drilling factors, and well schedules. During well production, abnormal data may arise due to special situations such as temporary shutdowns. Thus, we conduct preprocessing to ensure data quality, such as missing value processing, outlier identification, data normalization, and correlation analysis. To address the issues of limited data samples and imbalanced data ratios, the SMOTE algorithm is employed to synthesize samples and augment the data volume of the minority class. Finally, four ML algorithms are utilized to train the productivity prediction models, from which the optimal model is evaluated. Simultaneously, the hyperparameter configuration is optimized to enhance the model performance. Using this framework, the productivity of fractured horizontal wells is accurately predicted with a small amount of data samples.

2. Data preparation

2.1. Data collection

The data samples are collected from the M reservoir in China. This tight oil reservoir is exploited through the "horizontal drilling + volume fracturing" technology. The burial depth is generally greater than 3000 m, and the minimum horizontal principal stress ranges from 40 to 80 MPa. The main pore types are intragranular dissolution pores and residual intergranular pores, showing typical characteristics of small pores and narrow throats, thus leading to ultra-low porosity (8%–14%) and permeability (0.25×10^{-3}–5.5×10^{-3} μm^2 for gas). The oil density and viscosity under reservoir conditions vary from 0.661 to 0.823 g/cm^3 and 0.55–1.45 mPa·s, respectively. The formation pressure coefficient reaches above 1.6. We collected data from 138 fractured horizontal wells that had been producing for more than 90 days.

Table 1
Summary of ML models for predicting unconventional resources production.

No.	Method	Reference	Forecast target	Algorithm
1	Time series models	Song et al. (2020)	Production of fractured horizontal wells in a volcanic reservoir	LSTM
2		Fan et al. (2021)	Production rate and daily production time of gas wells	ARIMA, LSTM
3		Ning et al. (2022)	Production rate of a future time sequence in shale reservoir	ARIMA, LSTM, Prophet
4		Werneck et al. (2022)	Fluid rates and bottom-hole pressures in oil and gas reservoirs for 30 days	RNN
5		Li et al. (2022)	Daily production data of a shale oil well	CNN, PSO, LSTM
6		Zha et al. (2022)	Monthly gas field production	CNN, LSTM
7		Chahar et al. (2022)	Daily oil production	ANN, RF, GB regressor
8		Zhou et al. (2023)	Shale oil production performance	CNN, BiGRU, AM
9		Pan et al. (2023)	Monthly production performance of oil wells	CNN, LSTM, AM
10		Jiang et al. (2023)	Oil production in real-time	LSTM, AFSA
11	Static models	Wang et al. (2019)	6-month and 18-month production of multi-stage fractured horizontal wells	DNN
12		Wang et al. (2021)	Production performance of unconventional reservoirs	DBN, Bayesian optimization
13		Xue et al. (2021)	Shale gas production	MORF
14		Liu et al. (2021)	EUR of shale gas wells	DFNN
15		Wang et al. (2022a)	Absolute open flow potential	SVR
16		Lu et al. (2022)	Shale oil production	DNN, PSO
17		Niu et al. (2022)	EUR of shale gas wells	RSM, MLFNN, SVR
18		Nguyen-Le and Shin (2022)	Peak production and Arps's hyperbolic decline parameters of a shale gas well	ANN
19		He et al. (2023)	6 years of cumulative production in shale gas reservoirs	RF
20		Hui et al. (2023)	Productivity of shale wells	RF, GBDT, ANN, ET
21		This study	Productivity of fractured horizontal wells with limited samples	SMOTE, SVR, PSO

Notes: BiGRU = Bidirectional gated recurrent unit, AM = Attention mechanism, AFSA = Artificial fish swarming algorithm, EUR = Estimated ultimate recovery, MORF = Multi-objective random forest, DFNN = Deep feedforward neural network, SVR = Support vector regression, RSM = Response surface method, MLFNN = Multi-layer feedforward neural network.

During migration and accumulation, tight oil has been impacted by geological processes such as diagenesis and tectonism, resulting in variations in the oil occurrence. Meanwhile, large-scale volume fracturing and diverse production techniques create highly complex flow mechanisms. Thus, the factors influencing production in fractured wells are intricate.

Geological factors directly impact the production of fractured horizontal wells by determining the hydrocarbon reserves and the fluid transport capability. Zou et al. (2015) suggested that the evaluation of "sweet area" in tight oil reservoirs should take into account the porosity, brittleness, and oil saturation. Using information theory, grey relational analysis, and experimental design, Liang et al. (2013) identified permeability and porosity as the important controlling factors of the fractured horizontal wells' productivity in the Bakken tight oil reservoir. Wu et al. (2024) evaluated the feasibility of volumetric fracturing technology in unconventional reservoirs by considering factors such as the rock brittleness index, natural fracture distribution, and rock mechanics properties, highlighting that a higher brittleness index is favorable for the fracture network propagation. Here, we use porosity, permeability, oil saturation, and brittleness index, to characterize the influences of geological factors on production, because the information on these four geological factors is available for the wells in the target area, which avoids missing data during the analyses.

Fracturing parameters determine the hydraulic fracturing network properties, which affect the stimulated reservoir volume and fluid flow behavior near the wellbore. We propose two parameters, namely fracturing fluid injection per unit length (FIPL) and proppant injection per unit length (PIPL), to eliminate the influence of horizontal well length. These parameters along with the number of fracturing sections and clusters are used to characterize the effects of hydraulic fracturing on productivity.

As suggested by Baihly et al. (2015), the total lateral length and the lateral length of the well (LLOW) in the target layer, which tremendously affect the horizontal well production, were selected to account for the drilling effect. During the initial stages, most oil wells in the tight and shale reservoirs operate as flowing wells; thus, we use the average nozzle size to characterize the influence of the well schedule. In summary, we have extracted a total of 11 influencing factors and 1 prediction target (initial productivity) for each horizontal well (Table 2). The data on the influencing factors and production of all 138 wells are shown in Fig. 1.

2.2. Data preprocessing

During the actual production of oil reservoirs, activities such as changing oil nozzles, repairing and inspecting pumps, and inter-well interference can cause typical issues such as missing, repetitive, singular, and non-standard data. As indicated by Fig. 1, abnormal data points among these influencing factors bring complexity and unnecessary difficulty to data analysis and productivity prediction. Therefore, it is essential to preprocess the data to ensure data integrity and standardization, thereby improving data quality. The preprocessing workflow includes missing value handling, outlier identification, data normalization, and correlation analysis.

To deal with missing values along the data stream in time series studies, Pan et al. (2019) proposed a physics-based deep learning method that reconstructs the data and generates the missing production history, which enables the entire history to be applicable during production analysis. However, for models that use static data to predict productivity at a certain time, missing values of selected influencing factors render the well unavailable as a sample for model training. A large amount of missing data leads to model underfitting and impedes the prediction accuracy. The typical methods for addressing missing data include imputation and deletion. Different methods are chosen based on the number of missing values. For fewer missing values within the same influencing factor, typically less than 5%, the imputation method is used to supplement the missing data. Currently, the mean value, the mode value, etc. are commonly used to fill in. However, if there are numerous missing values, the imputation method causes significant deviation in the dataset, impeding the practical application performance of the model. Despite potentially losing numerous training samples, the deletion method is utilized to remove oil wells with missing values to prevent data bias interference on the model. In addition, one of the innovations of this study lies in sample expansion, whereby the deletion method can be directly

Table 2
Influencing factors and prediction target in our model.

Type	Parameter	Unit
Geological factor	Permeability	10^{-3} μm^2
	Porosity	%
	Oil saturation	%
	Brittleness index	%
Fracture design factor	Fracturing fluid injection per unit length (FIPL)	m^3/m
	Proppant injection per unit length (PIPL)	m^3/m
	Number of fracturing sections	–
	Number of fracturing clusters	–
Drilling factor	Lateral length	m
	Lateral length of the well in the target layer (LLOW)	m
Well schedule	90-day average nozzle size	mm
Target production	90-day average daily oil production	m^3

used to remove oil wells with missing values.

Outliers, which are also referred to as anomalies, are data points in a dataset that exhibit values that are unreasonable in comparison to the other points (Hawkins, 1980). For small sample datasets, outliers may distort the data distribution and impact model training and performance evaluation. Identifying and removing outliers manually is a challenging task that is labor extensive and prone to inaccuracies, which potentially introduces identification bias and results in the removal of meaningful observations. Outlier detection technologies include several types (Yehia et al., 2022). In this study, a classical outlier detection algorithm, isolated forest (IF) was employed (Liu et al., 2008). Unlike approaches that directly measure differences between abnormal samples and other samples using indicators such as distance and density, this algorithm directly characterizes sparsity between samples. Thus, this simple and efficient method is capable of handling large multidimensional data and is widely used in industry. The IF algorithm divides data points by randomly selecting m features and selecting a random value between the maximum and minimum values of the selected features. The observations are recursively partitioned until each observation is separated into its own cluster, and the number of times each data point is partitioned can be recorded. The fewer number of partitions corresponds to the abnormal data points.

Data normalization refers to converting data from 0 to 1 through certain transformations, such as maximum-minimum normalization (Eq. (1)) and Z-score normalization. This ensures not only that all factors are of the same order of magnitude, eliminating the effect of different dimensions on the correlations, but also reduces the data dimensionality, improving the training efficiency and predictive performance of the model.

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

where x is the original data of a feature; $x_{\min}$ is the minimum value in the feature; $x_{\max}$ is the maximum value in the feature; x^* is the normalized feature data.

Correlation analysis uses ML algorithms to select features, from available feature arrays, that contribute significantly to the prediction target. When multiple features impact the target, this method fuses and reduces the dimensionality of the features, greatly reducing training time, improving prediction accuracy, and avoiding overfitting. Correlation analysis is a crucial step in data mining. However, many relevant studies fail to account for the impact of correlations between distinct features.

Therefore, the correlation analysis in this study consists of two steps: an analysis of correlations between diverse features, and an analysis of correlations between features and targets. The 1st step involves feature fusion, which is accomplished by analyzing the correlation between features using the Pearson correlation coefficient (PCC). Then the features having high correlations with other features are removed to avoid duplication. The PCC assesses the relationship between two features by calculating the correlation coefficient using Eq. (2). The correlation coefficient ranges from −1 to 1. The degree of correlation increases as the correlation coefficient approaches its endpoint values.

In the 2nd step, we use several feature selection techniques to examine the correlations between features and targets, to identify and eliminate the less important features. Feature selection techniques are categorized into three types: filtering, wrapper, and embedded methods (Jović et al., 2015). Filtering methods, such as linear correlation coefficient (LCC) and PCC, rank features by measuring the non-linear strength between features and targets. It mainly analyzes the influence of a single feature on the targets, with fast computational efficiency and a simple model; however, it cannot analyze the correlation between multiple variables. Both wrapper and embedding methods process data by training ML models. Wrapper methods include forward, backward, bidirectional, and recursive feature elimination (RFE_lr). Embedding methods include penalty term-based methods such as linear regression (LR), and tree-based methods such as random forest regression (RFR). Both the wrapper and embedding methods can perform multivariate and big data computations, but the computational process is relatively complex, and the results depend on the ML accuracy.

To balance the advantages and disadvantages of each method, and obtain reasonable influencing factors, we select 8 methods from the 3 categories to build a comprehensive feature selection model. These methods include LCC, PCC, and maximum information coefficient (MIC) in the filtering category, RFE_lr in the wrapper category, LR, L1 regularization (Lasso), L2 regularization (Ridge), and RFR in the embedded category. After calculating the scores of each feature by different methods, we normalize the scores and sum the scores of the same feature (Eq. (3)) to obtain the influence coefficient P_{ex}. Then the contributions of the features on the well productivity are sorted and the main controlling factors are identified.

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n} \frac{(X_i - E(X))}{\sigma_X} \frac{(Y_i - E(Y))}{\sigma_Y}}{n} \tag{2}$$

where $\mathrm{Cov}(X, Y)$ is the population covariance between two features; X is the feature data; Y is another feature data; i is the ordinal number of the sample in the feature data; n is the total number of samples for each feature; X_i is the i-th sample in the feature X; Y_i is the i-th sample in the feature Y; $E(X)$ and $E(Y)$ are the mathematical expectations of X and Y, respectively; σ_X and σ_Y are the standard

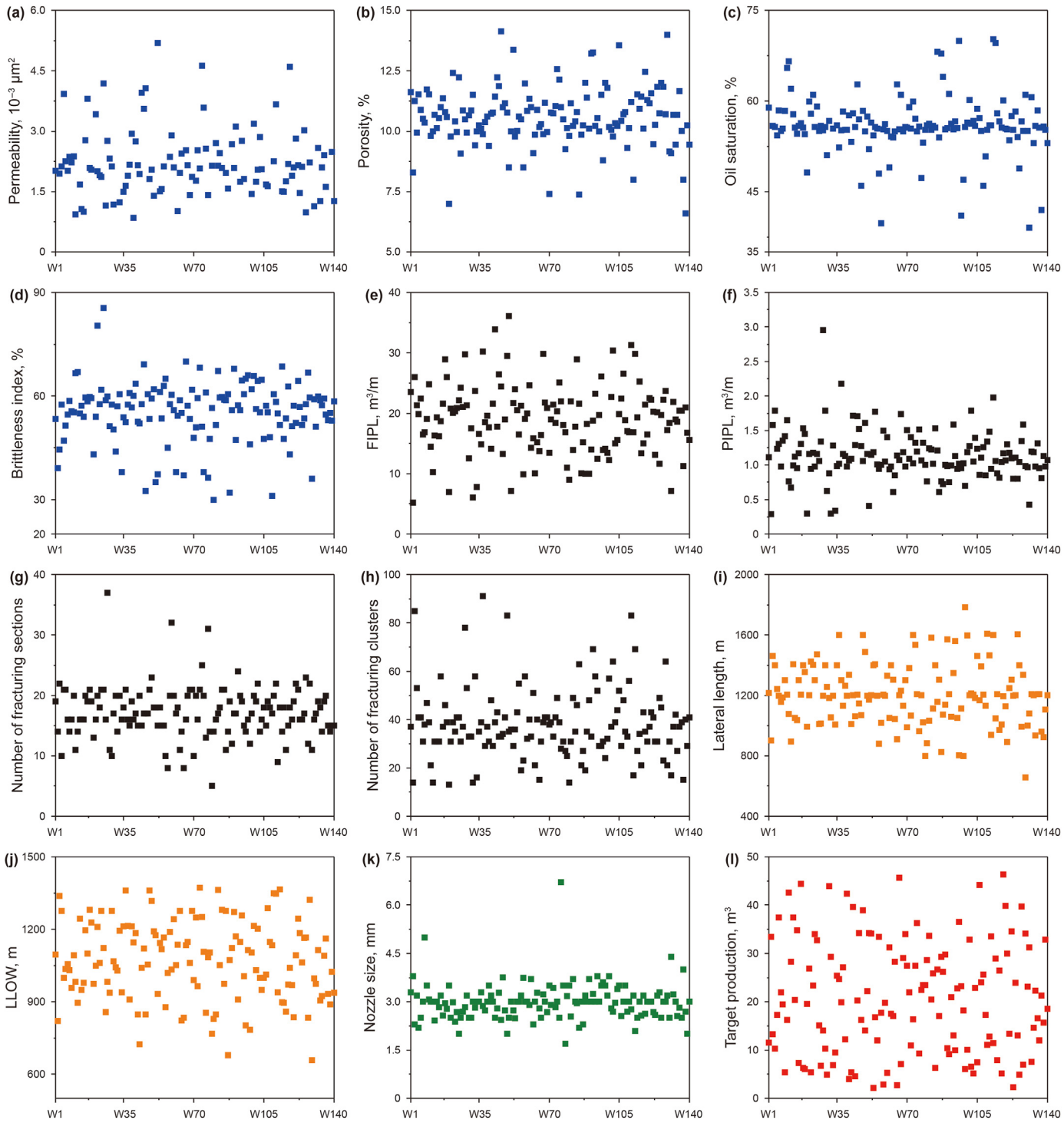

Fig. 1. Scatter plot of data from 138 fractured horizontal wells: (**a**) permeability, (**b**) porosity, (**c**) oil saturation, (**d**) brittleness index, (**e**) FIPL, (**f**) PIPL, (**g**) number of fracturing sections, (**h**) number of fracturing clusters, (**i**) lateral length, (**j**) LLOW, (**k**) 90-day average nozzle size, and (**l**) 90-day average daily oil production.

deviation of X and Y, respectively.

$$P_{\mathrm{ex}} = \sum_{i=1}^{8} P_{ix} \tag{3}$$

where P_{ix} is the normalized evaluation score of the feature x by the i-th feature selection model; P_{ex} is the influence coefficient of feature x on the target.

3. Method

We establish a comprehensive framework for predicting the production of fractured horizontal wells with limited samples based on machine learning (Fig. 2). Firstly, we collect field data including the influencing factors and the target production. Subsequently, the data undergo preprocessing to meet the requirements of model training and prediction. Then we expand the

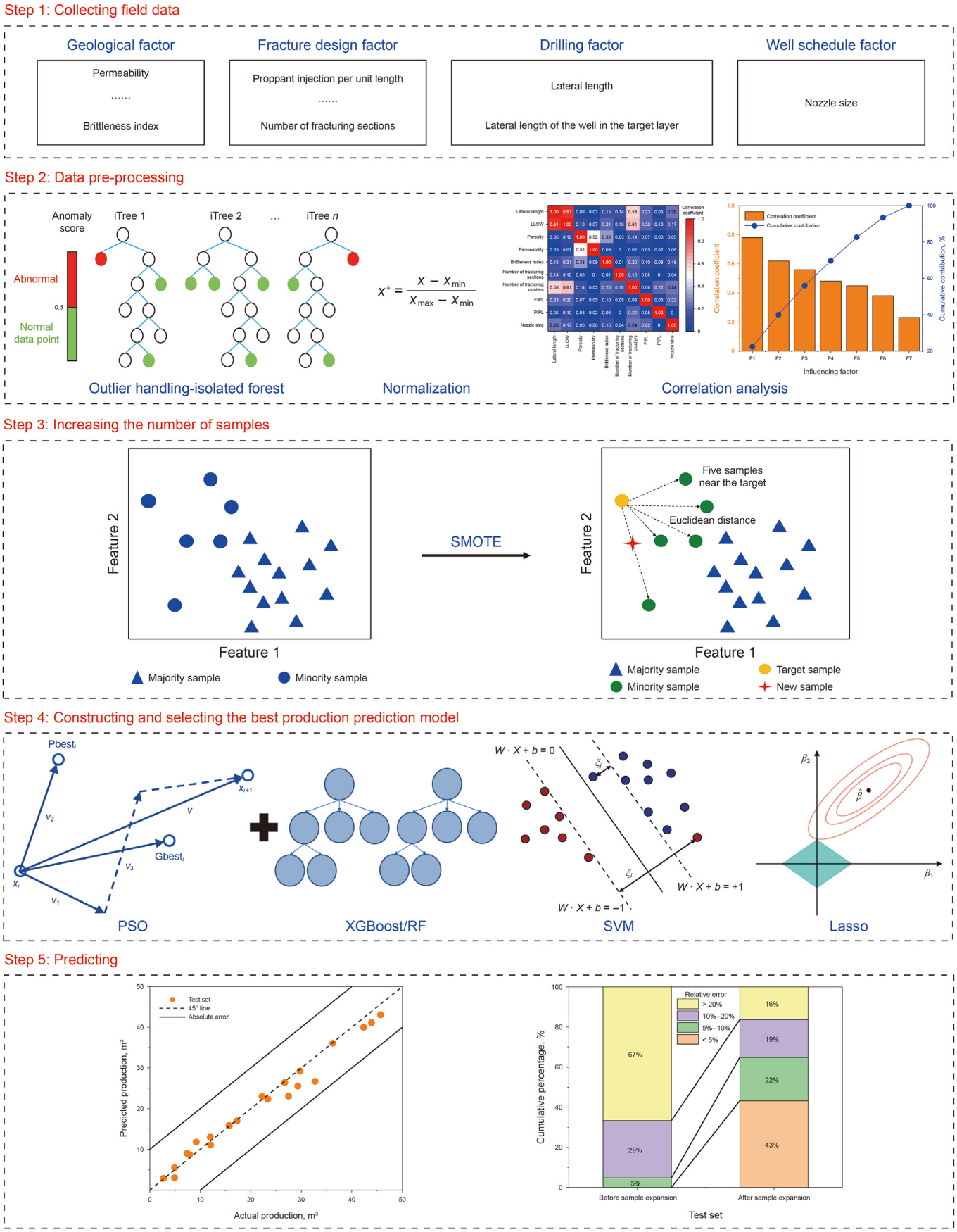

Fig. 2. Framework for predicting the productivity of fractured horizontal wells with limited samples.

preprocessed data and divide the samples into training datasets and validation datasets. After examining the training performance, the optimal ML algorithm is selected from several ML models, and the hyperparameters are optimized to meet the accuracy requirements. Finally, the production of fractured horizontal wells is predicted and the model performance is examined through the comparison with real data. Our objective is to accurately predict the productivity of fractured horizontal wells using limited data samples. The novelty lies in the utilization of machine learning techniques to enhance the sample data. Below is a brief overview of data enhancement and prediction modeling (Fig. 2).

3.1. Data expansion

The dataset in this study is obtained from the real field, exhibiting imbalanced data ratios and a limited sample size. Imbalanced datasets show a substantial disparity in the proportion of majority and minority samples, usually with a majority sample proportion exceeding 75% of the total samples. ML algorithms favor majority samples while ignoring or incorrectly discarding minority samples as noise or outliers (Díez-Pastor et al., 2015), resulting in poor model performance.

Oversampling and undersampling algorithms are often used to deal with the influence of unbalanced datasets on the prediction. The undersampling algorithm involves discarding a subset of samples from the majority class. However, this method is unsuitable for datasets with limited samples, because it leads to the loss of critical information (Lin et al., 2017). To overcome the influence of imbalanced datasets and a small amount of data samples on the prediction model accuracy, we use the SMOTE (Chawla et al., 2002) in the oversampling process to generate a dataset with a more balanced distribution.

Samples are generated by interpolation methods that merge the features of neighboring class samples. Figs. 3 and 4 show the specific process and the basic principle of SMOTE, respectively. First, determine whether the data set is balanced by the clustering algorithm. If yes, then all samples are constructed; if it is imbalanced, then only minority samples are constructed. Second, the Euclidean distance between each sample and other samples is estimated separately to determine its k-nearest neighbors. Then, arbitrarily select one of the minority samples as the target sample and several samples from the k-nearest neighbors of the target sample, and conduct linear interpolation using Eq. (4) to create new samples. Finally, the created and original samples are integrated to obtain an expanded sample set. The SMOTE algorithm not only solves the problem of an imbalanced dataset but also increases the number of data samples.

$$x_{\text{construct}} = x + rand(0,1) * \left(x_{\text{nearby}} - x\right) \tag{4}$$

where $x_{\text{construct}}$ is new samples created; x is a selected sample; x_{nearby} is a randomly selected sample close to x; $rand(0,1)$ is a random number between 0 and 1, but not including 0 and 1 themselves.

3.2. Machine learning models

We evaluate four ML algorithms to guarantee the prediction model performance, including SVR, RF, extreme gradient boosting (XGBoost), and least absolute shrinkage and selection operator (Lasso). We haven't used a neural network (NN) for the comparison because the performance of the NN model has been examined in previous studies. For example, to optimize the fracturing parameters, Li et al. (2024) used five ML models, such as RF, SVR, and multilayer perceptron (MLP)—a typical feedforward neural network, to predict the production rate of shale oil based on the numerical simulation data. They concluded that NN performs worse than the other models like RF. Similar conclusions have also been reported by Bassey et al. (2024). Given that the small sample size of our data set may lead to serious overfitting issues for NN models, we have not used NN for the comparison.

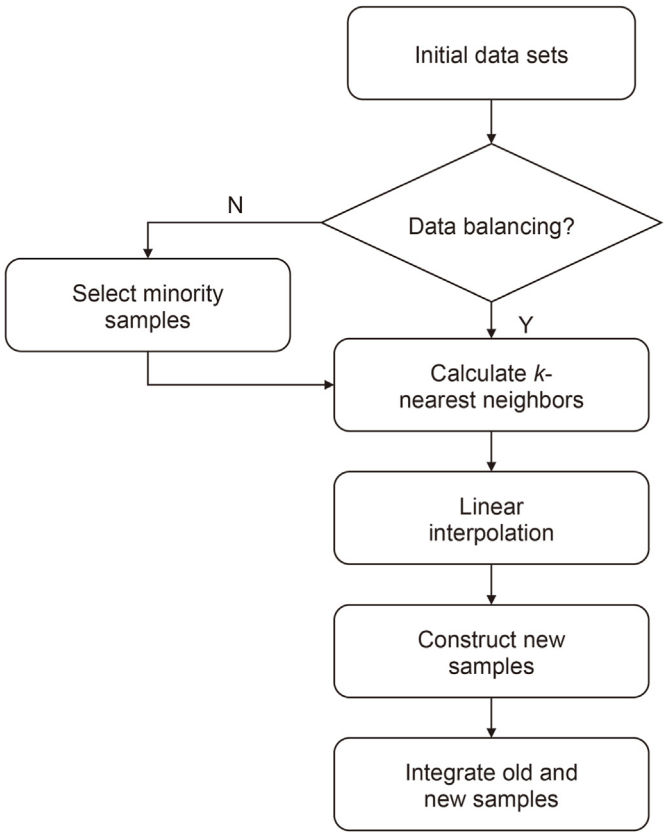

Fig. 3. Specific process of SMOTE algorithm.

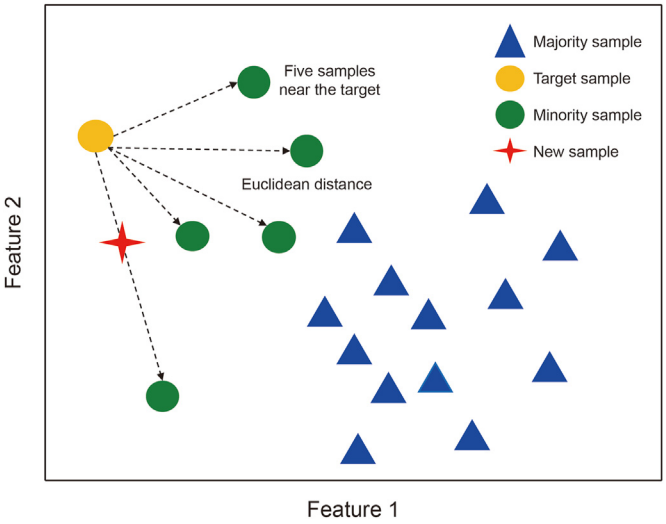

Fig. 4. Basic principle of sample construction using SMOTE.

The support vector machine (SVM) is a supervised ML algorithm proposed by Cortes and Vapnik (1995). Its basic idea is to find the best hyperplane of the original space in a linearly separable space. Relax variables should be added in a linearly inseparable space, and the nonlinear samples should be mapped to the high-dimensional space (feature space). Then the samples become linear, and it is easy to find the optimal hyperplane of the feature space. The SVM can solve regression and classification problems. SVM has important applications in data regression, known as support vector regression (SVR), which finds a regression plane closest to all samples (Smola and Schölkopf, 2004). SVR aims to minimize the error between the model prediction and the observed values while keeping within a given tolerance range. There are two important parameters in SVR: the penalty parameter and the kernel function. A larger penalty parameter leads to more penalized errors. Several kernel functions are available to deal with diverse non-linear problems. The linear kernel function is particularly noteworthy among them, owing to its flexibility and ability to handle large numbers of samples and features. Moreover, the computation is simple; thus, the linear kernel function is favorable for scenarios where the sample and feature sizes are comparable and large. The radial basis function (RBF) kernel performs better when the sample number is moderate, and the feature number is small. The computational complexity of SVM depends on the number of support vectors rather than the dimension of the input space. Meanwhile, the SVM has excellent generalization and high prediction accuracy.

The RF is a supervised machine learning algorithm based on ensemble learning (Breiman, 1996, 2001). It works by training samples through the integration of multiple decision trees into a forest, and then summing the prediction results of each decision tree to derive the average value and obtain final results (Genuer et al., 2017). RF is capable of handling large, high-dimensional datasets. Meanwhile, building independent decision trees in parallel significantly reduces training time (Vyas et al., 2017). Compared to other machine learning algorithms like NN, RF has fewer parameters and can achieve higher prediction accuracy by adjusting fewer parameters (Genuer et al., 2017). In addition, because the RF model is composed of multiple decision trees, it mitigates the variance of the prediction model and improves the accuracy.

XGBoost is proposed based on the gradient boosting decision tree (GBDT) by Chen and Guestrin (2016). It has become increasingly popular in recent years. Unlike traditional GBDT, XGBoost generates an expansion set of weak classification tree models by optimizing the gradient descent of the loss function (Cui et al., 2017). In essence, each subsequent decision tree in XGBoost is trained to fit the residual between the true value and the predicted outcome of the preceding decision tree. Moreover, XGBoost utilizes the second derivative to optimize the objective function and incorporates a regularization term to mitigate overfitting, thereby controlling the model complexity and enhancing training speed.

Lasso is a linear regression method proposed by Tibshirani (1996). It constructs a loss coefficient model with an L1 regularization penalty term. Using this model, the sum of the absolute values of the coefficients is constrained to be less than a constant, thereby compressing the variables (reducing the dimensionality) and mitigating the overfitting issue. The penalty coefficient in Lasso serves to control the model complexity. For numerous variables, a higher penalty coefficient leads to a more stringent penalization, resulting in a model with fewer variables and reduced complexity. The optimal model can be achieved by integrating cross-validation techniques with Lasso (Pedregosa et al., 2011).

3.3. Optimization algorithm

There are two distinct types of parameters in ML models. The 1st type is model parameters, which encapsulate configuration variables within the model that are learned and estimated from the data. Examples of model parameters include the weights in NN, support vectors in SVM, and coefficients in linear regression. The 2nd type is hyperparameters, representing configurations external to the model to aid in estimating model parameters. Examples of hyperparameters include the learning rate in NN, penalty parameters and kernel functions in SVM, and the depth of trees in decision tree models. The primary disparity between these two types of parameters lies in the fact that hyperparameters necessitate manual configuration, whereas parameters are automatically optimized during model training. The setting of hyperparameters significantly impacts model performance. A more judicious selection of hyperparameters enhances the model's generalization capability, augments training efficiency, and improves prediction accuracy (Bergstra et al., 2011).

To find the hyperparameters that can make the prediction model perform best on the sample set, we use the PSO algorithm to optimize the model hyperparameters by minimizing the root mean square error (RMSE) between the actual and prediction results. PSO is a global bionic optimization algorithm inspired by the flight and foraging behavior of birds (Kennedy and Eberhart, 1995). Unlike genetic algorithms, PSO does not involve "crossover" and "mutation"; instead, it searches for the optimal global solution by dynamically adjusting particle positions within the search space while sharing information regarding the current optimal value. PSO requires minimal parameter tuning and demonstrates high accuracy and rapid convergence, rendering it a prominent subject of contemporary optimization studies.

In PSO, there is a group of m particles flying at a given speed in the D-dimensional search space. Each particle i individually has a position represented by x_i, a velocity denoted by v_i, and an optimal value found by p_i. The optimal global solution discovered by all particles is denoted as p_g, and the updated equations for particle velocity and position are given by Eqs. (5)–(7).

$$v_{id}^{k+1} = \omega v_{id}^{k} + c_1 r_1 \left(p_{id}^{k} - x_{id}^{k}\right) + c_2 r_2 \left(p_{gd}^{k} - x_{gd}^{k}\right) \tag{5}$$

$$x_{id}^{k+1} = x_{id}^{k} + v_{id}^{k+1} \tag{6}$$

$$1 \le i \le m, \quad 1 \le d \le D \tag{7}$$

where c_1 and c_2 are typical numerical values known as learning factors, often set to 2; r_1 and r_2 are two random numbers equally distributed between 0 and 1; ω is inertial weight which determines how much it inherits from the current velocity of the particles (Shi and Eberhart, 1998).

3.4. Evaluation indexes

We utilize four types of factors as input variables to train the ML models, including geological factors, fracturing design factors, drilling factors, and well schedules. The initial production rates of fractured horizontal wells are used as the output variable. We use two evaluation indicators to evaluate the performance of different ML models and select the best algorithm for application: the coefficient of determination (R^2) and RMSE. R^2 (Eq. (8)) gauges the degree of fit of the sample data. A higher R^2 value indicates a better fit of the model, with its value typically ranging between 0 and 1.

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{8}$$

RMSE (Eq. (9)) evaluates the average error between the predicted and true values. A smaller RMSE corresponds to a smaller standard deviation of the residuals, indicating a higher accuracy.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2} \tag{9}$$

where $\hat{y}_i$ is the predicted value of the sample data; y_i is the true value of the sample data; $\bar{y}$ is the average of the true data; n is the total number of samples.

4. Results and discussion

Based on actual field data, we use the proposed framework to predict the productivity of fractured horizontal wells.

4.1. Sample set preparation

The data of 138 fractured horizontal wells is obtained from the M reservoir in China. The average daily oil production over the 1st 90 days is recorded as the target predicted by the model. A total of 11 factors of four distinct types are identified based on field data availability. The 11 factors, alongside the 90-day average daily oil production data, constitute the fundamental sample set for the ML model.

4.1.1. Missing values and outliers handling

Data preprocessing is essential for improving data quality. We first address missing data. A manual inspection suggests that 23 out of 138 fractured horizontal wells do not have brittleness index information. The missing values account for approximately 15% of the total samples. This proportion is deemed too substantial. Using the mean-filling method to address the missing values would introduce significant inaccuracies in the dataset. Therefore, we remove the 23 samples from the dataset. Subsequently, the IF algorithm is used to detect outliers among the samples. Fig. 5 shows a 2D schematic of the detection results, with red points indicating the identified outliers. Horizontal wells containing abnormal points were directly removed, leading to the elimination of 9 samples from the dataset. Following the handling of missing values and outliers, data from 106 fractured horizontal wells remain in the sample set. Then we utilize the maximum-minimum normalization method to standardize the data and mitigate the impact of varying magnitudes on the results.

4.1.2. Analysis of influencing factors

The analysis not only examines the linear relationship between production and influencing factors but also assesses the correlation among different features. For feature fusion and data dimensionality reduction, we employ both the data interpretation method and PCC method to analyze the correlation between various factors, identify the pairs having high correlation, and then remove one of them. Among the 11 influencing factors, oil saturation exhibits minimal variation (ranging from 50% to 60%, approximately 55%). Moreover, field data indicates that oil saturation, typically calculated from porosity, is highly correlated with porosity; thus it is removed from the influencing factors.

We use the PCC method to estimate the correlation coefficients among the pairs of other 10 factors, and the results are visualized in a heat map (Fig. 6). The values in the heat map represent the correlation coefficient, which ranges from 0 to 1 due to the positive correlation between different factors and production rate. The color gradient signifies the correlation strength, with red and blue indicating higher and lower correlations, respectively. A correlation threshold of 0.5 is set. If the correlation coefficient between the two factors exceeds 0.5, they are considered highly correlated. Fig. 6 reveals four highly correlated factor pairs: (1) porosity and permeability (correlation coefficient: 0.52), (2) lateral length and LLOW (correlation coefficient: 0.91), (3) LLOW and the number of fracture clusters (correlation coefficient: 0.61), and (4) lateral length and the number of fracture clusters (correlation coefficient: 0.58). In actual oil fields, formation permeability is typically estimated from porosity log data using theoretical models or experimental correlations (Glover et al., 2006; Yang and Aplin, 2010). Consequently, we have chosen to focus on porosity rather than permeability to determine the primary controlling factors. In

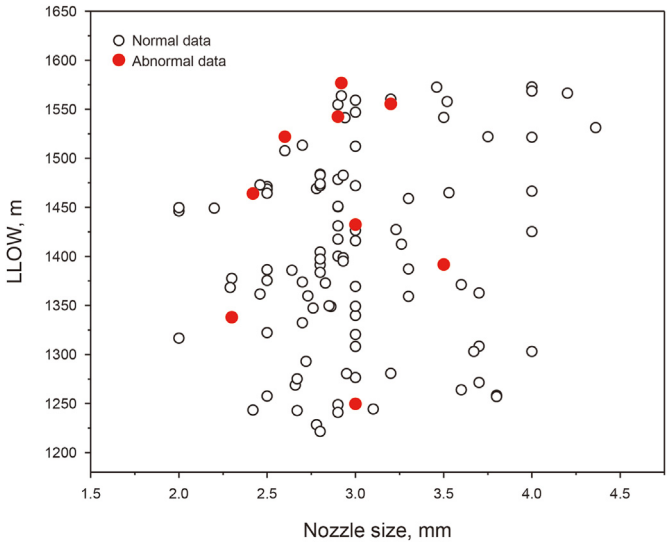

Fig. 5. 2D schematic diagram of isolated forest (IF) detection results.

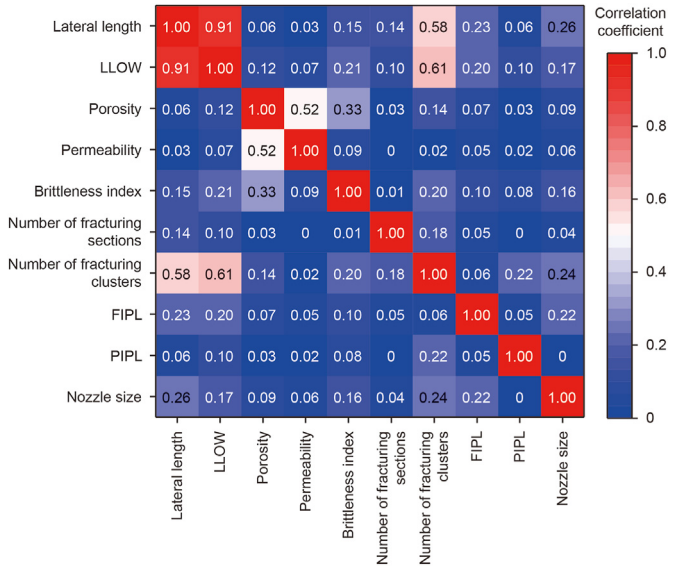

Fig. 6. Results of Pearson correlation analysis.

addition, lateral length and the number of fracture clusters are removed from further analysis. Then we will focus on the contributions of the remaining 7 factors on the productivity: porosity, brittleness index, LLOW, FIPL, PIPL, number of fracturing sections, and nozzle size.

We employ a comprehensive feature selection model to calculate the comprehensive importance coefficient of each influencing factor. Additionally, we estimate the cumulative contribution of different factors to the target production to determine the primary controlling factors. A higher comprehensive importance coefficient indicates a stronger correlation between the influencing factor and the target production. Conversely, smaller coefficients imply lower contributions. Table 3 presents the correlation coefficients calculated by different methods, as well as the comprehensive importance coefficients and cumulative contributions. According to Table 3 and Fig. 7, for this reservoir, the correlations between these factors and production decrease in the following order: nozzle size > PIPL > number of fracturing sections > FIPL > brittleness index > porosity > lateral length. The importance coefficients between these 7 factors and the production are greater than 0.2, indicating their high correlations with productivity. Therefore, all 7 factors including porosity, brittleness index, LLOW, FIPL, PIPL, number of fracturing sections, and nozzle size are utilized as input variables to train the ML model.

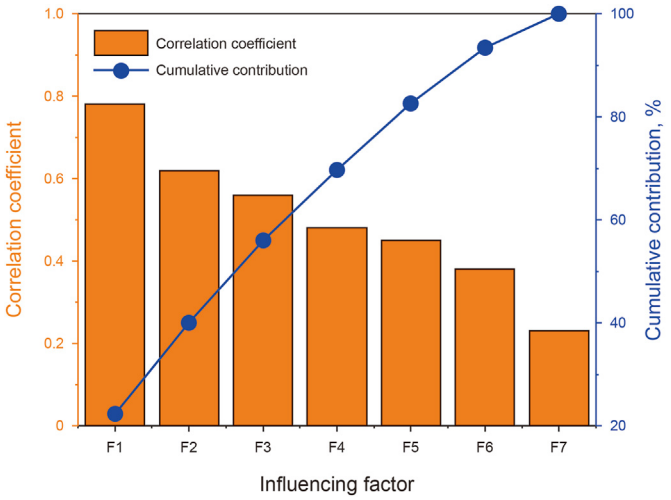

Fig. 7. Correlation and cumulative contribution of distinct influencing factors. F1 = nozzle size, F2 = PIPL, F3 = number of fracturing sections, F4 = FIPL, F5 = brittleness index, F6 = porosity, F7 = LLOW.

4.1.3. Sample expansion

The quality and volume of the dataset significantly impact the performance of ML models. Data preprocessing plays a crucial role in ensuring data quality. Based on the SMOTE algorithm, we determine the appropriate data samples to balance the model's learning ability and training efficiency. Following the data preprocessing and influencing factor analysis, we construct a basic sample set using 7 influencing factors from 106 fractured horizontal wells as model inputs and well productivity as model output. Then we utilize the SMOTE algorithm to expand the samples based on the basic sample set. First, the clustering algorithm is utilized to divide the samples into two types, and the number of samples in these two groups is compared to determine whether the sample set is balanced. If the number ratio of the two types is less than 4:1, the sample set is balanced, and the model's performance can be improved by increasing the number of samples. However, if the ratio exceeds 4:1, it indicates an imbalance. In such cases, it is necessary to expand the minority samples to ensure a balance between the two types and then expand all samples to increase the sample set volume. The clustering results of the basic sample set reveal a balanced data set with a ratio of 71:35 between the two types. Then the SMOTE algorithm is used to expand the data from 106 fractured wells. We expand the data volume to a multiple of the original volume; in other words, the number of new samples is several times that in the basic set. Table 4 shows the different data volume expansion schemes. By comparing the training efficiency and prediction accuracy of the models under distinct schemes, we determine the most reasonable sample size. Then we integrate the original and new samples to construct an expanded sample set to train the prediction model of fractured horizontal wells productivity.

4.2. Results

Upon completion of the sample set preparation, we construct a production prediction model for fractured horizontal wells using four ML algorithms. We optimize the model hyperparameters using PSO, and the prediction model with the most superior performance is obtained after the comparison. We also evaluate different sample expansion schemes based on the best ML model to determine the suitable sample volumes and expansion schemes, further enhancing the model performance.

4.2.1. Comparison of different ML algorithms

To ensure a fair comparison, the trained and validation performance of four ML models are examined using the same dataset. We choose Case 4 as the sample set to avoid the effects of too many or too few samples on the model efficiency and accuracy. The model inputs include porosity, brittleness index, LLOW, FIPL, PIPL, number of fracturing sections, and nozzle size, with the production as the output. In Case 4, 530 samples are randomly divided into training and validation sets in a 7:3 ratio. To minimize the loss function, we utilize PSO to optimize the hyperparameters of the four models and train these models using the training set. Table 5 summarizes the specific hyperparameters for each ML model with their ranges and initial values for optimization. Subsequently, the trained model is combined with a validation set to predict the productivity of horizontal wells.

Table 3
Results of comprehensive feature selection model.

Influencing factor	LCC	PCC	MIC	LR	Lasso	Ridge	RFR	RFE_lr	Comprehensive importance coefficient	Cumulative contribution, %
Nozzle size	1.00	0	0.63	0.61	1.00	1.00	1.00	1.0	0.78	22
PIPL	0.21	0.51	0.97	0.60	0.82	0.43	0.43	1.0	0.62	40
Number of fracturing sections	0.29	0.43	0	1.00	0.87	0.43	0.45	1.0	0.56	56
FIPL	0.32	0.41	1.00	0.41	0.32	0.38	0	1.0	0.48	70
Brittleness index	0	0.91	0.45	0.53	0.34	0.01	0.37	1.0	0.45	83
Porosity	0.09	0.67	0.93	0.04	0.03	0.19	0.58	0.5	0.38	93
LLOW	0	1.00	0.50	0	0	0	0.33	0	0.23	100

Table 4
Different expansion schemes of the sample set. The initial number of the samples is 106 ($n = 106$).

Scheme	Number of samples expanded	Number of total samples
Case 1: Increase the amount of original data by one time	106 (1n)	212 (2n)
Case 2: Increase the amount of original data by two times	212 (2n)	318 (3n)
Case 3: Increase the amount of original data by three times	318 (3n)	424 (4n)
Case 4: Increase the amount of original data by four times	424 (4n)	530 (5n)
Case 5: Increase the amount of original data by five times	530 (5n)	636 (6n)
Case 6: Increase the amount of original data by six times	636 (6n)	742 (7n)

Table 5
Specific hyperparameters for each ML model and their ranges during the optimization.

Model	Hyperparameter	Range	Initial value
SVM	Kernel	RBF	RBF
	C	(0.1, 100)	1
	Gamma	(0.0001, 10)	1
RF	N_estimators	(0, 300]	10
	Bootstrap	[True, False]	True
	Max_depth	(0, 10], None	None
	Max_features	['auto', 'sqrt']	auto
	Random_state	None	None
	Min_samples_leaf	1	1
	Min_samples_split	2	2
XGBoost	N_estimators	(0, 300]	100
	Subsample	0.05, 1, 20	1
	Learning_rate	[0, 1]	0.3
	Gamma	[0, +∞]	0
	Max_depth	(0, 10]	6
	Colsample_bytree	(0, 1]	1
	Min_child_weight	1	1
	Alpha	0	0
	Lambda	1	1
Lasso	Alpha	(0, 1]	1
	Max_iter	(0, 10000]	1000

Fig. 8 illustrates the prediction performance of 4 distinct models. The horizontal axis displays the actual production rate, while the vertical axis represents the production rate predicted using the trained ML models. The black dashed lines correspond to the 45° angles from the horizontal axis, indicating where the prediction results are identical to the actual values. The black solid line represents the absolute error, and the data in its area has an error of less than 10 m^3. In Fig. 8(a), most points are distributed around the 45° line, and only two validation samples fall outside the black solid lines, suggesting that the prediction results of most samples are within an acceptable error, demonstrating good prediction performance of the SVM model. However, in Fig. 8(b) and (d), although most points are within the acceptable error range, they deviate significantly from the 45° line, indicating less accurate predictions. In contrast, in Fig. 8(c), a large number of data points exceed the acceptable error range, indicating poor prediction performance of XGBoost. Fig. 9 presents a bar chart comparing the two evaluation indicators of the four models. The R^2 values for the four models decrease in the order of SVM, XGBoost, RF, and Lasso, with SVM achieving an R^2 greater than 0.9. The order of the RMSE is Lasso > RF > XGBoost > SVM. Both evaluation indicators demonstrate that the model combining PSO and SVM performs the best. Moreover, for SVM, the relative errors of predicted results are more concentrated near the x-axis, indicating a smaller error compared to other models (Fig. 10). Therefore, the SMOTE + PSO + SVM model is recommended to predict the production of fractured horizontal wells in practical applications.

4.2.2. Influences of sample volume

We examine the performance of 6 expansion schemes (Table 4) to determine the appropriate number of expanded samples. The integration of expanded and original samples is used for model training based on the SVM + PSO model. Fig. 11 illustrates the training and validation results of the 6 schemes. Table 6 reveals that the ML model performed well on the training set, maintaining an R^2 over 0.9. However, the difference between training and validation accuracy decreases with increasing the number of samples. For Cases 1, 2, and 3, which have smaller data samples, the difference between training and prediction accuracy is greater, indicating overfitting issues; however, for Cases 4, 5, and 6, the gap between training and validation accuracy is smaller. The comparison manifests that a small amount of data readily leads to the overfitting issues of ML models.

In terms of prediction accuracy, a higher number of samples leads to better model performance (Table 6). However, expanding the total number of samples to five times the original data (Case 4) has met the requirement for practical applications. We observe that the R^2 value of the validation set increases from 0.58 to 0.90 when the total number of the samples is increased from 2 to 5 times. Further expanding the sample set has only a tiny impact on the model accuracy. The R^2 value of the validation set remains almost constant when the total number of samples exceeds six times the original data.

A larger dataset results in lower training efficiency (Table 6). For example, the training time for Case 4 is less than half that of Case 6. However, the model training time in this study does not exceed 1 min, which is within an acceptable range. This is because the model employed here is relatively simple with fewer hyperparameters, and its training efficiency is less affected by the data volume. Conversely, utilizing a complex deep learning model would result in more significant differences in training efficiency based on the number of samples.

Based on the above analysis, it is evident that when the total number of samples is expanded to five times the original data (Case 4), the model shows superior performance in both validation accuracy and training efficiency. Therefore, we use this case to examine the model performance in field applications. Cases 5 and 6 can also be used for the analysis. The only differences from Case 4 are slight improvements in prediction accuracy and longer training time.

4.2.3. Practical application

To test the performance of the model generated through sample synthesis, the unexpanded sample set (original 106 wells) and the expanded sample set are divided into training, validation, and test sets in a ratio of 6:2:2. Both the training and validation sets have data synthesized using the SMOTE technique, whereas the test set comprises only actual field data. The training, validation, and test processes are conducted using the PSO + SVM model (Figs. 12 and 13). Table 7 compares the predicted and actual productivity of the

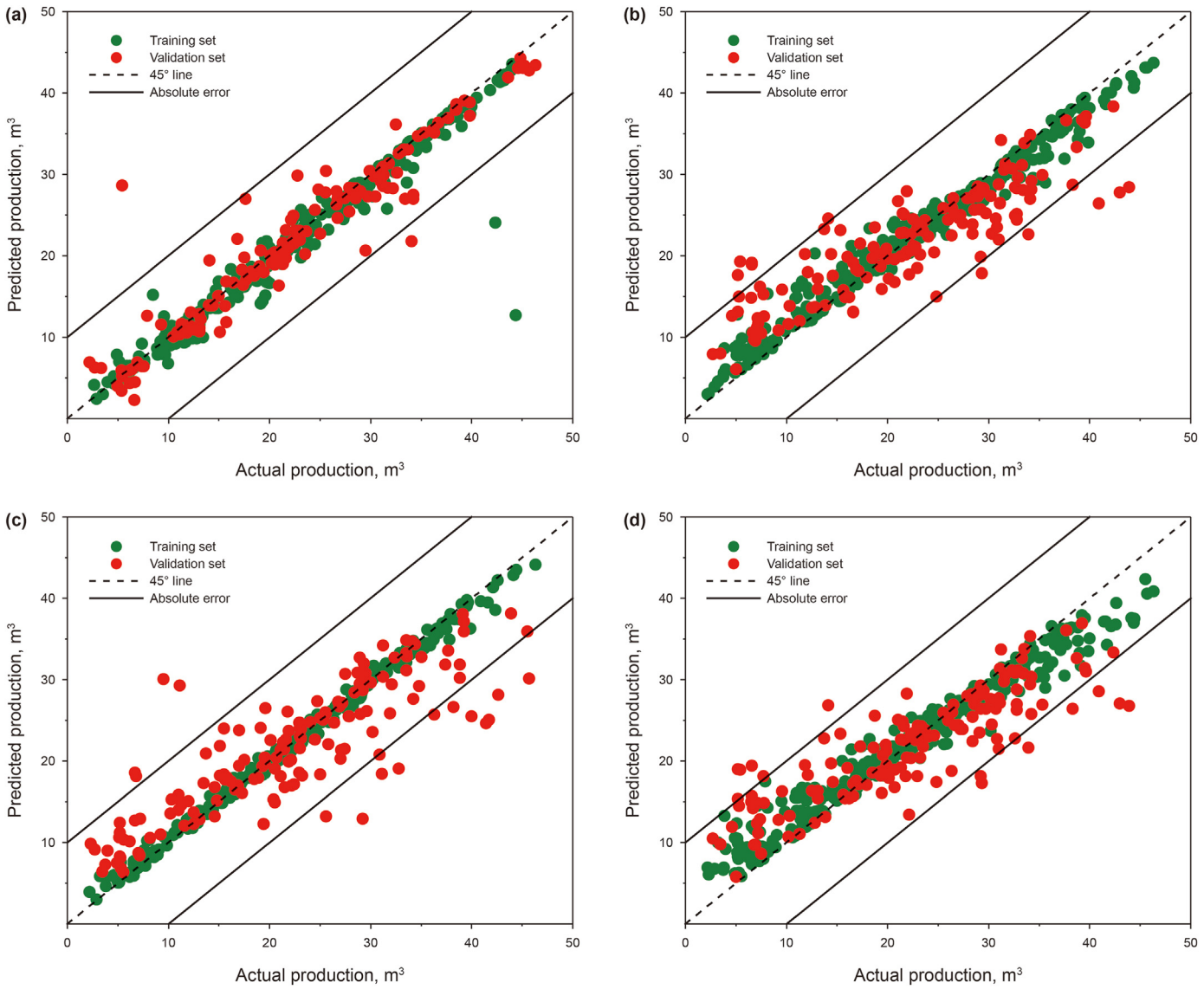

Fig. 8. Training and validation performances of the four models: (**a**) SVM, (**b**) RF, (**c**) XGBoost, and (**d**) Lasso. The number ratio between the training and validation samples is 7:3.

wells in the test set. It is worth noting that the samples in the test set are not used in model training and validation. The comparison between the sample sets before and after the expansion is shown in Table 8. Table 8 and Fig. 13 illustrate that the unexpanded sample set exhibits overfitting issues due to its limited number of samples, leading to significantly lower validation and test accuracy. After sample expansion, the model's predictive performance has notably improved. Nearly all data points are clustered around the 45° line, as opposed to the significant deviation observed before the expansion (Fig. 13).

Fig. 14 illustrates a stacked relative error distribution histogram for the test set. The expanded dataset outperforms the original dataset in terms of relative error reduction. Specifically, before sample expansion, the model's predicted relative error exceeded 5%. After expansion, the predictive performance improves tremendously, with a relative error of less than 5% for 43% of fractured horizontal wells. In addition, the number of samples having relative errors exceeding 20% before sample expansion is four times greater than that observed after expansion. In summary, the sample set enhanced by the SMOTE algorithm resolves the inaccurate production prediction of fractured horizontal wells using small samples.

4.3. Discussion

Our proposed model can be applied to the fracturing-production integration processes. This work aims to develop an accurate and efficient model to forecast the productivity of fractured horizontal wells, even with a small sample dataset. The proposed model allows us to forecast the production rate of fractured wells using geological information, fracturing design, drilling parameters, and well schedule. This model can be used prior to the fracturing process to optimize the fracturing design and well schedules and ultimately maximize the production rate. Because all the data are adopted from an actual oilfield, this model more reliably characterizes the fracturing-production integration processes. Recently, Yu et al. (2023) conducted a comprehensive review of hydraulic fracturing, fluid transport mechanisms, and enhanced recovery technology in shale reservoirs, which provides a basis for understanding the integration and production in shale reservoirs.

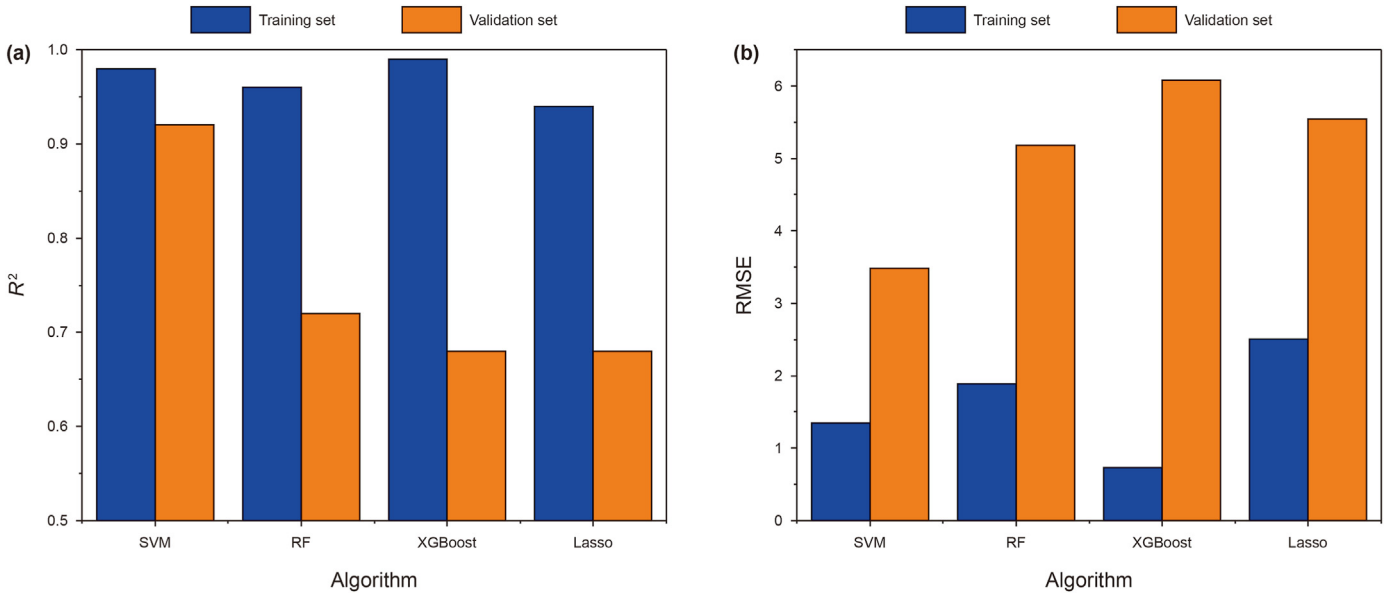

Fig. 9. Bar charts showing the two evaluation indicators of the four models: (**a**) R^2 and (**b**) RMSE.

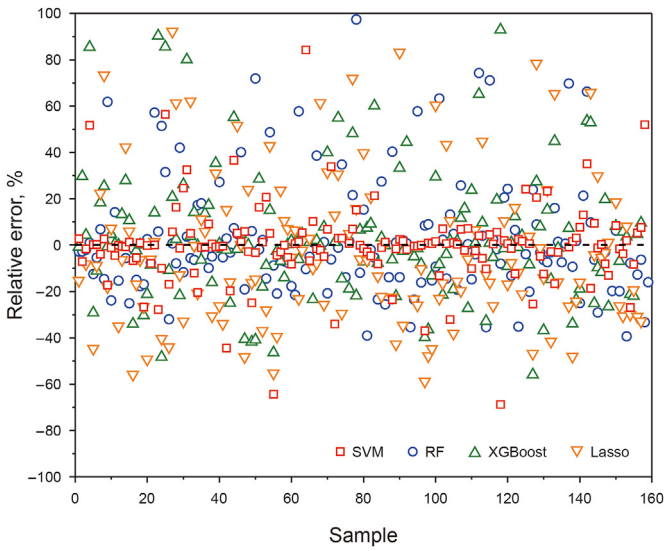

Fig. 10. Relative errors of the prediction results of the four models.

Owing to the widespread nanopores, which leads to the breakdown of classical fluid flow theory such as Darcy's law, understanding the nanoscale transport mechanisms is crucial for efficiently exploiting unconventional reservoirs (Wang et al., 2022b). However, the storage and transport behavior in the nanopores of unconventional resources like shales is tremendously complex, involving multiple rock surfaces, high pressure/temperature conditions, and multiphase interactions. Common techniques for studying nanoscale transport behavior include molecular simulations, nanofluidic chips, etc. Interested readers may refer to the reviews by Yu et al. (2020) and Yang et al. (2023). However, the costly and time-consuming nature of traditional methods has limited their widespread use. Machine learning has offered a way to quickly predict the existence state and transport characteristics at the micro/nanoscale. For example, Huang et al. (2022) combined machine learning and kinetic theory to develop a framework that enhances the computational efficiency of CH_4 adsorption behavior in shale nanopores. Zhang et al. (2023) proposed an intelligent interpolation function to improve the finite element computational efficiency for voxel-based irregular structures. The machine learning models can be further integrated into a multiscale framework to replace the traditional laborious techniques and improve prediction efficiency. As all the data in this work are adopted from an actual oilfield, the multiscale transport principles, including the fluid flow mechanisms in the confined nanopores, have been implicitly taken into account in the dataset. However, if one wants to use simulation data to predict production performance, the transport mechanisms at micro/nanoscale should be accounted for via multiscale modeling.

5. Conclusions

A comprehensive framework for predicting the productivity of fractured horizontal wells is proposed based on few-shot learning. The highlight of this study lies in the expansion of a small unbalanced sample dataset and the prediction through our proposed SMOTE-SVM-PSO model. We examine the performance by varying sample sizes and the ML models used for training (SVM, RF, XGBoost, and Lasso). We also conduct feature fusion and dimensionality reduction to enhance the training performance. We reach the following conclusions.

(1) The PSO + SVM model, which combines the particle swarm optimization algorithm and support vector machine, demonstrates the most favorable performance for small samples among the evaluated algorithms. Compared with RF and XGBoost, SVM has fewer hyperparameters, and the PSO excels in rapidly and effectively finding the optimal hyperparameters. The Lasso algorithm is more suitable for dealing with simple linear problems and the prediction accuracy of nonlinear problems is lower than that of SVM.

(2) Using the SMOTE algorithm to augment the number of samples enhances the robustness of ML models. When the total number of samples is expanded to five times that of the original dataset, the predicted R^2 value of the trained model can reach 0.9. As the number of samples continues to grow, the model performance further improves.

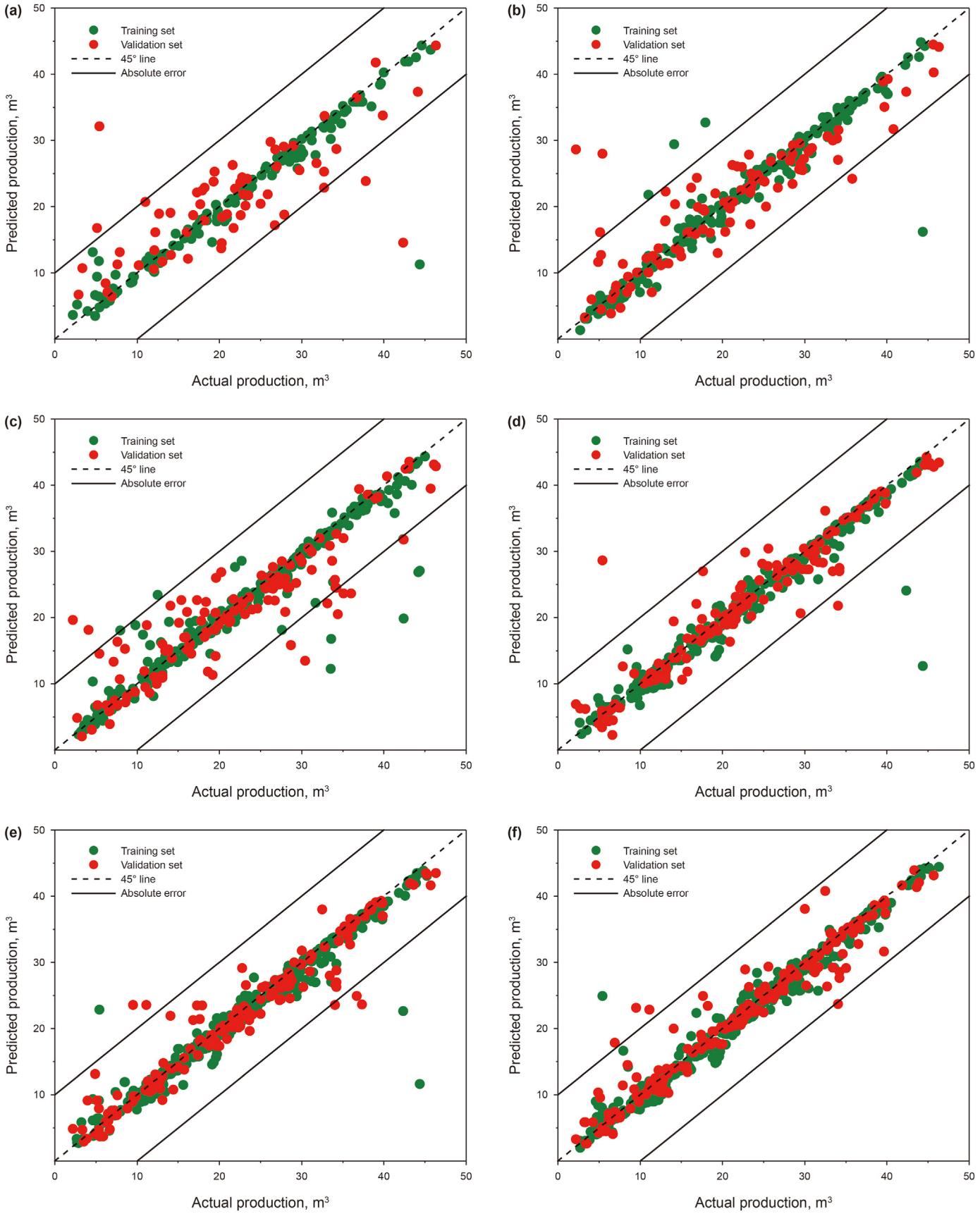

Fig. 11. Training and validation performances of the six schemes: (**a**) Case 1, (**b**) Case 2, (**c**) Case 3, (**d**) Case 4, (**e**) Case 5, and (**f**) Case 6. The number ratio between the training and validation samples is 7:3.

Table 6
Evaluation indicators under different expansion schemes.

Scheme	Total number of samples	Training set		Validation set		Training efficiency, s
		R^2	RMSE	R^2	RMSE	
Case 1	212 ($2n$)	0.92	3.10	0.58	7.04	6.3
Case 2	318 ($3n$)	0.96	2.12	0.76	6.51	9.5
Case 3	424 ($4n$)	0.90	3.30	0.82	3.95	12.3
Case 4	530 ($5n$)	0.94	2.36	0.90	3.48	19.9
Case 5	636 ($6n$)	0.96	2.03	0.92	3.12	29.9
Case 6	742 ($7n$)	0.97	1.64	0.92	3.08	40.5

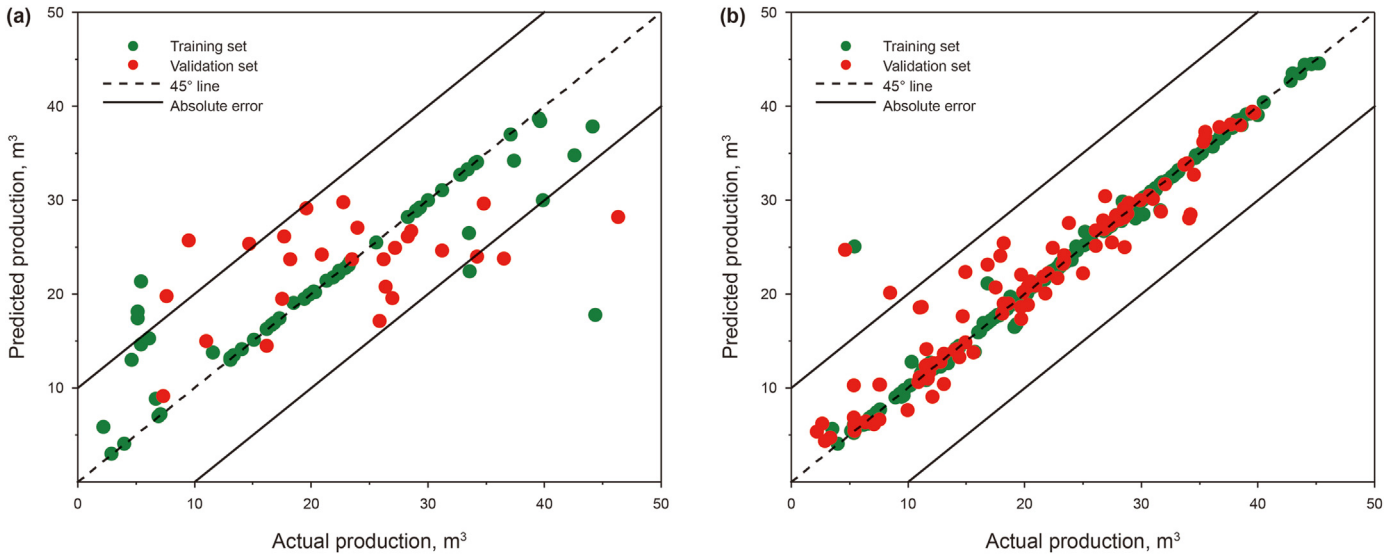

Fig. 12. Training and validation performances before (**a**) and after (**b**) sample expansion.

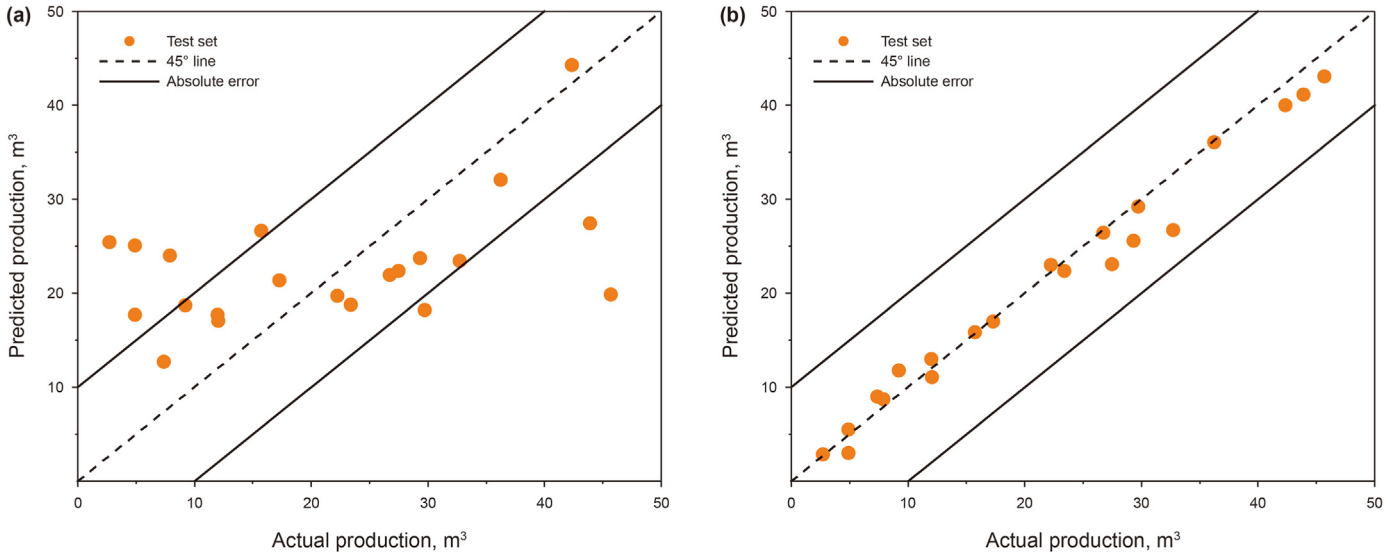

Fig. 13. Prediction effect of the test set before (**a**) and after (**b**) sample expansion. The dashed 45° lines indicates the predictions are identical to the true values.

(3) A small amount of samples readily leads to overfitting of ML models. After sample expansion, the training, validation, and test accuracy of the model significantly improved, with a test set R^2 of up to 0.9. The accuracy is more than twice that before sample expansion. This method effectively addresses the common issue in reservoir engineering studies of small sample sizes, which results in inaccurate prediction performance.

The proposed model serves as an effective tool for accurately predicting the production of fractured horizontal wells without sufficient samples, thereby providing a theoretical basis for

Table 7
Comparison between the predicted and actual productivity of the wells in the test set. Both the predictions before and after the sample expansion are given.

Sample number	True value	Prediction before sample expansion	Prediction after sample expansion
1	23.40	18.81	22.33
2	17.30	21.36	17.00
3	7.40	12.75	9.01
4	27.50	22.35	23.05
5	26.70	21.90	26.41
6	15.70	26.64	15.89
7	2.70	25.45	2.84
8	9.20	18.72	11.79
9	4.90	17.71	3.00
10	4.90	25.10	5.50
11	43.90	27.44	41.16
12	7.90	24.02	8.69
13	29.30	23.75	25.54
14	22.22	19.68	23.01
15	12.06	17.08	11.10
16	12.00	17.70	13.00
17	42.33	44.26	40.00
18	45.71	19.85	43.10
19	29.73	18.22	29.19
20	32.73	23.40	26.71
21	36.24	32.08	36.07

Table 8
Comparison of the prediction performance before and after sample expansion.

Data set	R^2 before sample expansion	R^2 after sample expansion
Training set	0.77	0.97
Validation set	0.35	0.92
Test set	0.13	0.90

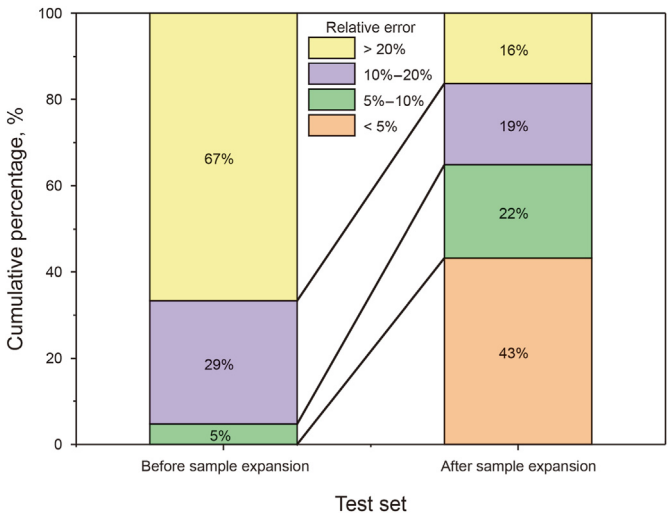

Fig. 14. Stacked histograms showing the relative error distributions of the results for the test set.

reservoir production optimization and new well fracturing decisions. Moreover, the model can be used to address small sample problems in other fields as well.

CRediT authorship contribution statement

Sen Wang: Writing – review & editing, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Wen Ge:** Writing – original draft, Methodology, Formal analysis, Data curation. **Yu-Long Zhang:** Writing – review & editing, Visualization, Methodology, Formal analysis. **Qi-Hong Feng:** Supervision, Resources, Funding acquisition. **Yong Qin:** Methodology, Formal analysis, Data curation. **Ling-Feng Yue:** Visualization, Methodology. **Renatus Mahuyu:** Writing – review & editing, Formal analysis. **Jing Zhang:** Formal analysis, Data curation.

Declaration of competing interest

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

Acknowledgments

This work is supported by the National Natural Science Foundation of China (52274055), the Shandong Provincial Natural Science Foundation (ZR2022YQ50), and the Taishan Scholar Program of Shandong Province (tsqn202408088).

References

Afagwu, C., Alafnan, S., Weijermars, R., et al., 2023. Multiscale and multiphysics production forecasts of shale gas reservoirs: new simulation scheme based on Gaussian pressure transients. Fuel 336, 127142. https://doi.org/10.1016/j.fuel.2022.127142.

Akbarabadi, M., Alizadeh, A.H., Piri, M., et al., 2023. Experimental evaluation of enhanced oil recovery in unconventional reservoirs using cyclic hydrocarbon gas injection. Fuel 331, 125676. https://doi.org/10.1016/j.fuel.2022.125676.

Alom, M.S., Tamim, M., Rahman, M.M., 2017. Decline curve analysis using rate normalized pseudo-cumulative function in a boundary dominated gas reservoir. J. Petrol. Sci. Eng. 150, 30–42. https://doi.org/10.1016/j.petrol.2016.11.006.

Arps, J.J., 1945. Analysis of decline curves. Transactions of the AIME 160 (1), 228–247. https://doi.org/10.2118/945228-G.

Azom, P.N., Javadpour, F., 2012. Dual-continuum modeling of shale and tight gas reservoirs. In: SPE Annual Technical Conference and Exhibition. https://doi.org/10.2118/159584-MS.

Baihly, J.D., Malpani, R., Altman, R., et al., 2015. Shale gas production decline trend comparison over time and basins—revisited. In: SPE/AAPG/SEG Unconventional Resources Technology Conference. https://doi.org/10.15530/URTEC-2015-2172464.

Bassey, M., Akpabio, M.G., Agwu, O.E., 2024. Enhancing natural gas production prediction using machine learning techniques: a study with random forest and artificial neural network models. In: SPE Nigeria Annual International Conference and Exhibition. https://doi.org/10.2118/221577-MS.

Bergstra, J., Bardenet, R., Bengio, Y., et al., 2011. Algorithms for hyperparameter optimization. Adv. Neural Inf. Process. Syst. 24, 2546–2554. https://dl.acm.org/doi/10.5555/2986459.2986743.

Bhattacharya, S., Mishra, S., 2018. Applications of machine learning for facies and fracture prediction using Bayesian Network Theory and Random Forest: case studies from the Appalachian basin, USA. J. Petrol. Sci. Eng. 170, 1005–1017.

https://doi.org/10.1016/j.petrol.2018.06.075.
Bhattacharya, S., Ghahfarokhi, P.K., Carr, T.R., et al., 2019. Application of predictive data analytics to model daily hydrocarbon production using petrophysical, geomechanical, fiber-optic, completions, and surface data: a case study from the Marcellus Shale, North America. J. Petrol. Sci. Eng. 176, 702–715. https://doi.org/10.1016/j.petrol.2019.01.013.
Blasingame, T.A., Johnston, J.L., Lee, W.J., 1989. Type-curve analysis using the pressure integral method. In: SPE California Regional Meeting. https://doi.org/10.2118/18799-MS.
Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140. https://doi.org/10.1007/BF00058655.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.
Cao, C., Jia, P., Cheng, L., et al., 2022. A review on application of data-driven models in hydrocarbon production forecast. J. Petrol. Sci. Eng. 212, 110296. https://doi.org/10.1016/j.petrol.2022.110296.
Chahar, J., Verma, J., Vyas, D., et al., 2022. Data-driven approach for hydrocarbon production forecasting using machine learning techniques. J. Petrol. Sci. Eng. 217, 110757. https://doi.org/10.1016/j.petrol.2022.110757.
Chawla, N.V., Bowyer, K.W., Hall, L.O., et al., 2002. SMOTE: synthetic minority oversampling technique. J. Artif. Intell. Res. 16, 321–357. https://doi.org/10.1613/jair.953.
Chen, T., Guestrin, C., 2016. Xgboost: a scalable tree boosting system. In: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. https://doi.org/10.1145/2939672.2939785.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297. https://doi.org/10.1007/BF00994018.
Cui, Y., Cai, M., Stanley, H.E., 2017. Comparative analysis and classification of cassette exons and constitutive exons. BioMed Res. Int., 7323508 https://doi.org/10.1155/2017/7323508.
Deng, J., Zhu, W., Ma, Q., 2014. A new seepage model for shale gas reservoir and productivity analysis of fractured well. Fuel 124, 232–240. https://doi.org/10.1016/j.fuel.2014.02.001.
Díez-Pastor, J.F., Rodríguez, J.J., García-Osorio, C.I., et al., 2015. Diversity techniques improve the performance of the best imbalance learning ensembles. Inf. Sci. 325, 98–117. https://doi.org/10.1016/j.ins.2015.07.025.
Fan, D., Sun, H., Yao, J., et al., 2021. Well production forecasting based on ARIMA-LSTM model considering manual operations. Energy 220, 119708. https://doi.org/10.1016/j.energy.2020.119708.
Fetkovich, M.J., 1980. Decline curve analysis using type curves. J. Petrol. Technol. 32 (6), 1065–1077. https://doi.org/10.2118/4629-PA.
Fraim, M.L., Wattenbarger, R.A., 1987. Gas reservoir decline-curve analysis using type curves with real gas pseudopressure and normalized time. SPE Form. Eval. 2 (4), 671–682. https://doi.org/10.2118/14238-PA.
Genuer, R., Poggi, J.M., Tuleau-Malot, C., et al., 2017. Random forests for big data. Big Data Research 9, 28–46. https://doi.org/10.1016/j.bdr.2017.07.003.
Glover, P.W., Zadjali, I.I., Frew, K.A., 2006. Permeability prediction from MICP and NMR data using an electrokinetic approach. Geophysics 71 (4), F49–F60. https://doi.org/10.1190/1.2216930.
Hakimi, M.H., Hamed, T.E., Lotfy, N.M., et al., 2023. Hydraulic fracturing as unconventional production potential for the organic-rich carbonate reservoir rocks in the Abu El Gharadig Field, north western Desert (Egypt): evidence from combined organic geochemical, petrophysical and bulk kinetics modeling results. Fuel 334, 126606. https://doi.org/10.1016/j.fuel.2022.126606.
Hawkins, D.M., 1980. Identification of Outliers. Chapman and Hall. https://doi.org/10.1007/978-94-015-3994-4.
He, Y.W., He, Z.Y., Tang, Y., et al., 2023. Shale gas production evaluation framework based on data-driven models. Petrol. Sci. 20 (3), 1659–1675. https://doi.org/10.1016/j.petsci.2022.12.003.
Huang, M., Xu, H., Yu, H., et al., 2022. Fast prediction of methane adsorption in shale nanopores using kinetic theory and machine learning algorithm. Chem. Eng. J. 446, 137221. https://doi.org/10.1016/j.cej.2022.137221.
Hui, G., Chen, Z., Wang, Y., et al., 2023. An integrated machine learning-based approach to identifying controlling factors of unconventional shale productivity. Energy 266, 126512. https://doi.org/10.1016/j.energy.2022.126512.
Ilk, D., Rushing, J.A., Perego, A.D., et al., 2008. Exponential vs. hyperbolic decline in tight gas sands: understanding the origin and implications for reserve estimates using Arps' decline curves. In: SPE Annual Technical Conference and Exhibition. https://doi.org/10.2118/116731-MS.
Jia, C., Huang, Z., Sepehrnoori, K., et al., 2021. Modification of two-scale continuum model and numerical studies for carbonate matrix acidizing. J. Petrol. Sci. Eng. 197, 107972. https://doi.org/10.1016/j.petrol.2020.107972.
Jiang, W., Wang, X., Zhang, S., 2023. Integrating multi-modal data into AFSA-LSTM model for real-time oil production prediction. Energy 127935. https://doi.org/10.1016/j.energy.2023.127935.
Jović, A., Brkić, K., Bogunović, N., 2015. A review of feature selection methods with applications. In: 38th International Convention on Information and Communication Technology, Electronics and Microelectronics. MIPRO, pp. 1200–1205. https://doi.org/10.1109/MIPRO.2015.7160458.
Kamrava, S., Tahmasebi, P., Sahimi, M., 2019. Enhancing images of shale formations by a hybrid stochastic and deep learning algorithm. Neural Network. 118, 310–320. https://doi.org/10.1016/j.neunet.2019.07.009.
Kennedy, J., Eberhart, R., 1995. Particle swarm optimization. Proceedings of ICNN'95-international conference on neural networks 4, 1942–1948. https://doi.org/10.1109/ICNN.1995.488968.
Klie, H., Florez, H., 2020. Data-driven prediction of unconventional shale-reservoir dynamics. SPE J. 25 (5), 2564–2581. https://doi.org/10.2118/193904-PA.
Li, W., Wang, L., Dong, Z., et al., 2022. Reservoir production prediction with optimized artificial neural network and time series approaches. J. Petrol. Sci. Eng. 215, 110586. https://doi.org/10.1016/j.petrol.2022.110586.
Li, W., Zhang, T., Liu, X., et al., 2024. Machine learning-based fracturing parameter optimization for horizontal wells in Panke field shale oil. Sci. Rep. 14 (1), 6046. https://doi.org/10.1038/s41598-024-56660-8.
Liang, T., Chang, Y., Guo, X., et al., 2013. Influence factors of single well's productivity in the Bakken tight oil reservoir, Williston Basin. Petrol. Explor. Dev. 40 (3), 383–388. https://doi.org/10.1016/S1876-3804(13)60047-6.
Lin, W.C., Tsai, C.F., Hu, Y.H., et al., 2017. Clustering-based undersampling in class-imbalanced data. Inf. Sci. 409, 17–26. https://doi.org/10.1016/j.ins.2017.05.008.
Liu, F.T., Ting, K.M., Zhou, Z.H., 2008. Isolation forest. In: Eighth IEEE International Conference on Data Mining 413-422. https://doi.org/10.1109/ICDM.2008.17.
Liu, Y., Zeng, J., Qiao, J., et al., 2023. An advanced prediction model of shale oil production profile based on source-reservoir assemblages and artificial neural networks. Appl. Energy 333, 120604. https://doi.org/10.1016/j.apenergy.2022.120604.
Liu, Y.Y., Ma, X.H., Zhang, X.W., et al., 2021. A deep-learning-based prediction method of the estimated ultimate recovery (EUR) of shale gas wells. Petrol. Sci. 18 (5), 1450–1464. https://doi.org/10.1016/j.petsci.2021.08.007.
Lu, C., Jiang, H., Yang, J., et al., 2022. Shale oil production prediction and fracturing optimization based on machine learning. J. Petrol. Sci. Eng. 217, 110900. https://doi.org/10.1016/j.petrol.2022.110900.
Luo, G., Tian, Y., Bychina, M., et al., 2019. Production-strategy insights using machine learning: application for bakken shale. SPE Reservoir Eval. Eng. 22 (3), 800–816. https://doi.org/10.2118/195681-PA.
Manfroni, M., Bukkens, S.G.F., Giampietro, M., 2022. Securing fuel demand with unconventional oils: a metabolic perspective. Energy 261, 125256. https://doi.org/10.1016/j.energy.2022.125256.
Manjunath, G.L., Liu, Z., Jha, B., 2023. Multi-stage hydraulic fracture monitoring at the lab scale. Eng. Fract. Mech. 289, 109448. https://doi.org/10.1016/j.engfracmech.2023.109448.
Micheal, M., Xu, W.L., Xu, H.Y., et al., 2021. Multi-scale modelling of gas transport and production evaluation in shale reservoir considering crisscrossing fractures. J. Nat. Gas Sci. Eng. 95, 104156. https://doi.org/10.1016/j.jngse.2021.104156.
Moinfar, A., Narr, W., Hui, M.H., et al., 2011. Comparison of discrete-fracture and dual-permeability models for multiphase flow in naturally fractured reservoirs. In: SPE Reservoir Simulation Symposium. https://doi.org/10.2118/142295-MS.
Nguyen-Le, V., Shin, H., 2022. Artificial neural network prediction models for Montney shale gas production profile based on reservoir and fracture network parameters. Energy 244, 123150. https://doi.org/10.1016/j.energy.2022.123150.
Ning, Y., Kazemi, H., Tahmasebi, P., 2022. A comparative machine learning study for time series oil production forecasting: ARIMA, LSTM, and Prophet. Comput. Geosci. 164, 105126. https://doi.org/10.1016/j.cageo.2022.105126.
Niu, W., Lu, J., Sun, Y., et al., 2022. Development of visual prediction model for shale gas wells production based on screening main controlling factors. Energy 250, 123812. https://doi.org/10.1016/j.energy.2022.123812.
Pan, S., Yang, B., Wang, S., et al., 2023. Oil well production prediction based on CNN-LSTM model with self-attention mechanism. Energy 128701. https://doi.org/10.1016/j.energy.2023.128701.
Pan, Y., Bi, R., Zhou, P., et al., 2019. An effective physics-based deep learning model for enhancing production surveillance and analysis in unconventional reservoirs. In: SPE/AAPG/SEG Unconventional Resources Technology Conference. https://doi.org/10.15530/urtec-2019-145.
Pan, Y., Deng, L., Zhou, P., et al., 2021. Laplacian Echo-State Networks for production analysis and forecasting in unconventional reservoirs. J. Petrol. Sci. Eng. 207, 109068. https://doi.org/10.1016/j.petrol.2021.109068.
Pedregosa, F., Varoquaux, G., Gramfort, A., et al., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830. https://dl.acm.org/doi/10.5555/1953048.2078195.
Raza, M.Y., Lin, B., 2023. Development trend of Pakistan's natural gas consumption: a sectorial decomposition analysis. Energy 278, 127872. https://doi.org/10.1016/j.energy.2023.127872.
Shi, Y., Eberhart, R., 1998. A modified particle swarm optimizer. IEEE World Congress on Computational Intelligence (Cat. No.98TH8360) 69–73. https://doi.org/10.1109/ICEC.1998.699146.
Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Stat. Comput. 14, 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88.
Song, X., Liu, Y., Xue, L., et al., 2020. Time-series well performance prediction based on Long Short-Term Memory (LSTM) neural network model. J. Petrol. Sci. Eng. 186, 106682. https://doi.org/10.1016/j.petrol.2019.106682.
Tangirala, S., Sheng, J.J., 2019. Investigation of oil production and flowback in hydraulically-fractured water-wet formations using the Lab-on-a-Chip method. Fuel 254, 115543. https://doi.org/10.1016/j.fuel.2019.05.126.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc. B 58 (1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
Tontiwachwuthikul, P., Chan, C., Zeng, F., et al., 2020. Recent progress and new developments of applications of artificial Intelligence (AI), knowledge-based systems (KBS), machine learning (ML) in the petroleum industry. Petroleum 6 (4), 319–320. https://doi.org/10.1016/j.petlm.2020.08.001.
Valkó, P.P., Lee, W.J., 2010. A better way to forecast production from unconventional gas wells. In: SPE Annual Technical Conference and Exhibition. https://doi.org/10.2118/134231-MS.

Vyas, A., Datta-Gupta, A., Mishra, S., 2017. Modeling early time rate decline in unconventional reservoirs using machine learning techniques. In: Abu Dhabi International Petroleum Exhibition & Conference. https://doi.org/10.2118/188231-MS.

Wang, L., Yao, Y., Wang, K., et al., 2022a. Hybrid application of unsupervised and supervised learning in forecasting absolute open flow potential for shale gas reservoirs. Energy 243, 122747. https://doi.org/10.1016/j.energy.2021.122747.

Wang, S., Chen, S., 2019. Insights to fracture stimulation design in unconventional reservoirs based on machine learning modeling. J. Petrol. Sci. Eng. 174, 682–695. https://doi.org/10.1016/j.petrol.2018.11.076.

Wang, S., Chen, Z., Chen, S., 2019. Applicability of deep neural networks on production forecasting in Bakken shale reservoirs. J. Petrol. Sci. Eng. 179, 112–125. https://doi.org/10.1016/j.petrol.2019.04.016.

Wang, S., Qin, C., Feng, Q., et al., 2021. A framework for predicting the production performance of unconventional resources using deep learning. Appl. Energy 295, 117016. https://doi.org/10.1016/j.apenergy.2021.117016.

Wang, S., Liang, Y., Feng, Q., et al., 2022b. Sticky layers affect oil transport through the nanopores of realistic shale kerogen. Fuel 310, 122480. https://doi.org/10.1016/j.fuel.2021.122480.

Wang, S., Zhang, Z., Wen, Z., et al., 2023. Inferring the interwell connectivity of multilayer waterflooded reservoirs accounting for incomplete injection/production profiles. Geoenergy Science and Engineering 227, 211897. https://doi.org/10.1016/j.geoen.2023.211897.

Wang, S., Xiang, J., Wang, X., et al., 2024. A deep learning based surrogate model for reservoir dynamic performance prediction. Geoenergy Science and Engineering 233, 212516. https://doi.org/10.1016/j.geoen.2023.212516.

Werneck, R.O., Prates, R., Moura, R., et al., 2022. Data-driven deep-learning forecasting for oil production and pressure. J. Petrol. Sci. Eng. 210, 109937. https://doi.org/10.1016/j.petrol.2021.109937.

Wu, H., Zhang, N., Lou, Y., et al., 2024. Optimization of fracturing technology for unconventional dense oil reservoirs based on rock brittleness index. Sci. Rep. 14 (1), 15214. https://doi.org/10.1038/s41598-024-66114-w.

Wu, Y., An, S., Tahmasebi, P., et al., 2023. An end-to-end approach to predict physical properties of heterogeneous porous media: coupling deep learning and physics-based features. Fuel 352, 128753. https://doi.org/10.1016/j.fuel.2023.128753.

Xu, S., Feng, Q., Wang, S., et al., 2018. Optimization of multistage fractured horizontal well in tight oil based on embedded discrete fracture model. Comput. Chem. Eng. 117, 291–308. https://doi.org/10.1016/j.compchemeng.2018.06.015.

Xue, L., Liu, Y., Xiong, Y., et al., 2021. A data-driven shale gas production forecasting method based on the multi-objective random forest regression. J. Petrol. Sci. Eng. 196, 107801. https://doi.org/10.1016/j.petrol.2020.107801.

Xue, L., Wang, J., Han, J., Yang, M., Mwasmwasa, M.S., Nanguka, F., 2023. Gas well performance prediction using deep learning jointly driven by decline curve analysis model and production data. Advances in Geo-Energy Research 8 (3), 159–169. https://doi.org/10.46690/ager.2023.06.03.

Yan, B., Harp, D.R., Chen, B., et al., 2022. A physics-constrained deep learning model for simulating multiphase flow in 3D heterogeneous porous media. Fuel 313, 122693. https://doi.org/10.1016/j.fuel.2021.122693.

Yang, Y., Aplin, A.C., 2010. A permeability–porosity relationship for mudstones. Mar. Petrol. Geol. 27 (8), 1692–1697. https://doi.org/10.1016/j.marpetgeo.2009.07.001.

Yang, Y., Wang, S., Feng, Q., et al., 2023. Imbibition mechanisms of fracturing fluid in shale oil formation: a review from the multiscale perspective. Energy & Fuels 37 (14), 9822–9840. https://doi.org/10.1021/acs.energyfuels.3c00502.

Yehia, T., Khattab, H., Tantawy, M., et al., 2022. Removing the outlier from the production data for the decline curve analysis of shale gas reservoirs: a comparative study using machine learning. ACS Omega 7 (36), 32046–32061. https://doi.org/10.1021/acsomega.2c03238.

Yu, H., Xu, H., Fan, J., et al., 2020. Transport of shale gas in microporous/nanoporous media: molecular to pore-scale simulations. Energy & Fuels 35 (2), 911–943. https://doi.org/10.1021/acs.energyfuels.0c03276.

Yu, H., Xu, W., Li, B., et al., 2023. Hydraulic fracturing and enhanced recovery in shale reservoirs: theoretical analysis to engineering applications. Energy & Fuels 37 (14), 9956–9997. https://doi.org/10.1021/acs.energyfuels.3c01029.

Yu, W., Xu, Y., Weijermars, R., et al., 2018. A numerical model for simulating pressure response of well interference and well performance in tight oil reservoirs with complex-fracture geometries using the fast embedded-discrete-fracture-model method. SPE Reservoir Eval. Eng. 21 (2), 489–502. https://doi.org/10.2118/184825-PA.

Zha, W., Liu, Y., Wan, Y., et al., 2022. Forecasting monthly gas field production based on the CNN-LSTM model. Energy 260, 124889. https://doi.org/10.1016/j.energy.2022.124889.

Zhang, H., Yu, H., Wang, Q., et al., 2023. How to achieve the fast computation for voxel-based irregular structures by few finite elements? Extreme Mechanics Letters 65, 102103. https://doi.org/10.1016/j.eml.2023.102103.

Zhou, G., Guo, Z., Sun, S., et al., 2023. A CNN-BiGRU-AM neural network for AI applications in shale oil production prediction. Appl. Energy 344, 121249. https://doi.org/10.1016/j.apenergy.2023.121249.

Zhu, L., Li, M.S., Wu, Q.H., et al., 2015. Short-term natural gas demand prediction based on support vector regression with false neighbours filtered. Energy 80, 428–436. https://doi.org/10.1016/j.energy.2014.11.083.

Zou, C., Yang, Z., Hou, L., et al., 2015. Geological characteristics and “sweet area” evaluation for tight oil. Petrol. Sci. 12, 606–617. https://doi.org/10.1007/s12182-015-0058-1.