Petroleum Science 21 (2024) 2554–2570

Contents lists available at ScienceDirect

Petroleum Science

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

Original Paper

Oilfield analogy and productivity prediction based on machine learning: Field cases in PL oilfield, China

Wen-Peng Bai , Shi-Qing Cheng *, Xin-Yang Guo , Yang Wang , Qiao Guo , Chao-Dong Tan

State Key Laboratory of Petroleum Resources and Engineering, China University of Petroleum (Beijing), Beijing, 102249, China

ARTICLE INFO

Article history:
Received 26 September 2023
Received in revised form
19 December 2023
Accepted 29 February 2024
Available online 1 March 2024

Edited by Meng-Jiao Zhou

Keywords:
Data mining technique
Analogy parameters
Oilfield analogy
Productivity prediction
Software platform

ABSTRACT

In the early time of oilfield development, insufficient production data and unclear understanding of oil production presented a challenge to reservoir engineers in devising effective development plans. To address this challenge, this study proposes a method using data mining technology to search for similar oil fields and predict well productivity. A query system of 135 analogy parameters is established based on geological and reservoir engineering research, and the weight values of these parameters are calculated using a data algorithm to establish an analogy system. The fuzzy matter-element algorithm is then used to calculate the similarity between oil fields, with fields having similarity greater than 70% identified as similar oil fields. Using similar oil fields as sample data, 8 important factors affecting well productivity are identified using the Pearson coefficient and mean decrease impurity (MDI) method. To establish productivity prediction models, linear regression (LR), random forest regression (RF), support vector regression (SVR), backpropagation (BP), extreme gradient boosting (XGBoost), and light gradient boosting machine (LightGBM) algorithms are used. Their performance is evaluated using the coefficient of determination (R^2), explained variance score (EV), mean squared error (MSE), and mean absolute error (MAE) metrics. The LightGBM model is selected to predict the productivity of 30 wells in the PL field with an average error of only 6.31%, which significantly improves the accuracy of the productivity prediction and meets the application requirements in the field. Finally, a software platform integrating data query, oil field analogy, productivity prediction, and knowledge base is established to identify patterns in massive reservoir development data and provide valuable technical references for new reservoir development.

1. Introduction

Oilfield development is a complex and challenging process that requires extensive knowledge of the geological attributes, as well as efficient production strategies. Analogizing similar oilfields has become a popular method for identifying oilfields that share comparable characteristics with the target oilfields. By examining the development experience and mode garnered by the analogy oilfields, through long-term development adjustments and optimizations, it is possible to create a roadmap for productivity prediction and effective full life cycle development of the target oilfields. Analogous reservoirs and the target reservoirs have nearly similar geological characteristics and development methods, and these analogous reservoirs have gone through long-term development adjustment and optimization, and have formed valuable development experience and scientific development mode, which can point out the way to the target reservoirs for efficient development. When facing new reservoirs or reservoirs that have been developed for a period, the analogical method can provide a reference for them in the case of missing dynamic and static data.

Machine learning and data mining techniques have introduced novel tools for analyzing data in the petroleum industry (Ghahramani, 2015; Gurina et al., 2020; Pirizadeh et al., 2021; Werneck et al., 2022). These tools have proven to be efficient in identifying patterns and trends in large datasets, which can aid in the prediction of well performance and the optimization of production strategies. Data mining is the process of extracting implicit and unknown effective information from a large volume of actual data containing noise, using artificial intelligence, machine learning, statistics, and other technologies (Liao et al., 2012; Yuan

* Corresponding author.
E-mail address: chengsq973@163.com (S.-Q. Cheng).

https://doi.org/10.1016/j.petsci.2024.02.018

et al., 2017; Wang and Ayala, 2020; Cai et al., 2020). In the field of oil exploration and development, data mining and artificial intelligence have found application in various fields, including production index prediction and agent model simulation (Bravo et al., 2013; Feng et al., 2020; Wang and Seright, 2021; He et al., 2023). For instance, Lolon et al. (2016) deeply explored the relationship between oil well parameters and production and established predictive statistical models to evaluate various fracture treatment and completion designs. Zhou et al. (2014) combined principal component analysis, clustering method, and regression analysis to investigate the impact of hydraulic fractures, vertical depth, proppant, and fracturing fluid volume on gas production in Marcellus shale. Guo et al. (2018a) integrated the support vector regression surrogate model into the distributed Gauss-Newton method and demonstrated that machine learning algorithms, such as SVR, can be successfully integrated into gradient-based optimization methods to improve overall efficiency. The process of generating accurate prediction models or tools is known as prediction modeling. These models typically include one predictive variable (output) and one or more known independent predictive variables (input) (Awoleke and Lane, 2011; Ma et al., 2015; Lecun et al., 2015; Guo et al., 2019). Similarly, machine learning is frequently used for log analysis, lithofacies, depositional environments, and seismic data inversion (Iraji et al., 2023a; Soltanmohammadi et al., 2024). The study by Iraji et al. (2023b) combines log data, borehole imaging, conventional and micro CT plug data analysis, and thin-section descriptions aimed at characterizing the reservoir in the formation. They used deep learning to predict porosity, permeability, and rock type.

A variety of productivity prediction models have been developed using data mining techniques (Wang et al., 2021a; Eskandarian et al., 2017; Wood, 2020; Handhal et al., 2022). These models provide critical insights into the oil industry's production patterns. They allow for the identification of commonalities and differences between analogous oilfields (Guo et al., 2018b; Bahonar et al., 2022; Wei et al., 2022). Montgomery and O'Sullivan (2017) compared the application of five regression models in predicting the productivity of tight oil wells and confirmed that the linear regression model is more accurate than the generalized linear regression model, support vector regression model, random forest regression model and gradient boosting regression model. Aïfa (2014) developed a productivity prediction model based on artificial neural networks and determined the sensitivity of each influencing factor that affects productivity. Akbilgic et al. (2015) used neural networks to predict the gas-oil ratio of oilfield reservoirs and identified the determinants of SOR (steam-to-oil ratio) to be reservoir depth, gamma curve, and permeability. Wang et al. (2018) compared the performance of several machine learning algorithms, including random forest, adaboost, SVM, and ANN, and developed a productivity prediction model that was optimized by comparing prediction accuracy and error loss on test set data. In reservoir development, statistical models are commonly used to identify the geological and engineering parameters that have the greatest impact on production. While many models can accurately predict productivity, they require a significant amount of data to establish the model, limiting their application in fields with limited data. Moreover, the combination of oilfield analogy and productivity prediction has not been fully explored, and no complete process has been developed. Wang et al. (2021b) presented the application of big data technique in the oil fields of the western South China Sea to build a knowledge base of theoretical/empirical formulas to evaluate well productivity. Guo et al. (2022) used reservoir engineering methods (relationship between core permeability and porosity in similar areas) and the categorical boosting (CatBoost) model to predict reservoir permeability respectively, and finally selected the CatBoost model to predict DLG block reservoir permeability through comparison. However, most of the current reservoir dynamic parameters prediction methods do not consider the importance of field analogies (Wang et al., 2023; Bai et al., 2024), and the existing analogies are difficult to screen many similar reservoirs and lack interpretability quickly and accurately. The parameters considered in the analogical methods are not comprehensive, only for the same type of reservoir, and they do not combine the analogical results with the prediction of dynamic indicators.

This study proposes a method for predicting well productivity based on statistical analysis and data mining techniques, with reservoir geological characteristics and development dynamic data as the core, and similar reservoir analogies as the research samples. Firstly, a comprehensive consideration of eight categories encompassing 135 analogy parameters is taken. Big data algorithms are utilized to calculate the weight values, forming an analogy system. Secondly, methods such as the fuzzy matter-element method and comprehensive evaluation are employed to calculate the similarity, determining similar oilfields. The Pearson correlation coefficient method and MDI are applied to identify the factors influencing productivity. LR, RF, SVR, BP, XGBoost, and LightGBM algorithms are adopted to establish productivity prediction models. R^2, EV, MSE, and MAE are used to optimize productivity prediction models. Finally, a software platform integrating data inquiry, oilfield analogy, and productivity prediction is developed.

2. Workflow

This paper presents an analogy system and oil well productivity prediction model based on big data algorithms, enabling the rapid screening of similar oilfields and accurate prediction of oil well productivity in target oilfields. The workflow is divided into four distinct modules: data query, oilfield analogy, productivity prediction, and knowledge base (Fig. 1).

Data query: This module enables the quick selection of basic parameter tables of oilfields, blocks, and single wells that meet the self-defined data range through multi-attribute filtering, based on the constraint conditions of single or multi-parameter indexes. The statistical analysis of the basic parameter table also generates charts and variation laws of related parameters. The data obtained from this module is subsequently used for oilfield similarity calculation and oilfield single-well productivity prediction.

Oilfield analogy: Initially, the parameters of the target oilfield are imported according to the defined data format. Then, the relevant parameters are selected for analogy between the target oilfield and the similar oilfield. The weight values of the selected parameters are calculated using the big data algorithms, and an analogy system for the target oilfield is established. The similarity between the target oilfield and the analogy oilfield is calculated using the similarity algorithm (fuzzy matter-element method, comprehensive evaluation method, cosine similarity, and Euclidean distance) based on the established analogy system. Finally, the results of the data query are compared with the target oilfield to determine the oilfield similar to the target oilfield from the sample library.

Productivity prediction: In this section, we first define the factors influencing oil well productivity based on the parameter data of similar oilfields. The Pearson correlation coefficient and MDI method are utilized to calculate the correlation coefficients between the selected factors and oil well productivity. The factors with strong correlations are then chosen as input parameters, while

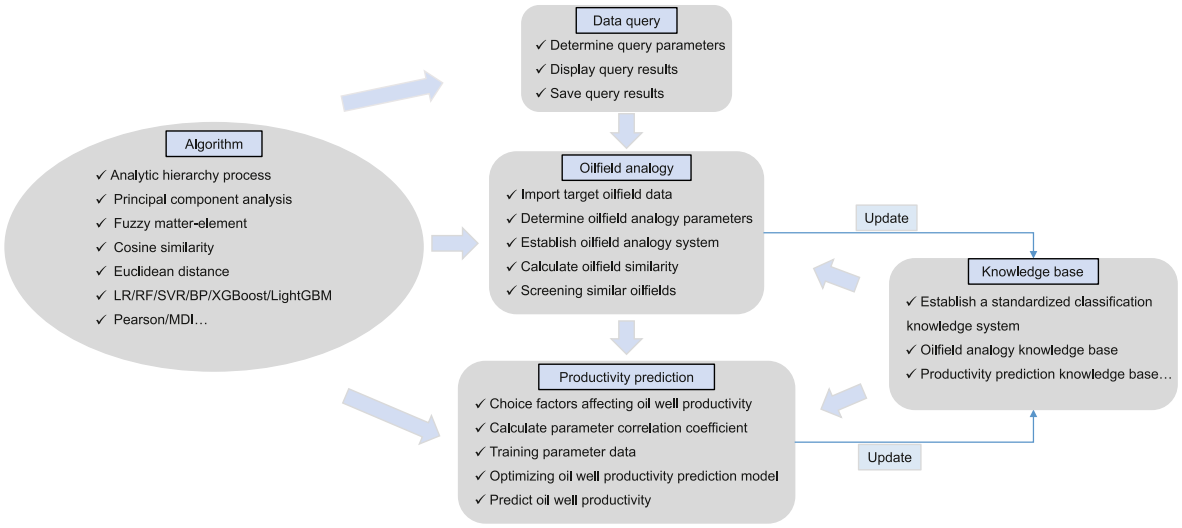


Fig. 1. Workflow of oilfield analogy and productivity prediction in the oilfields.

oil well productivity is designated as the target parameter. Next, 80% of the selected input parameter data is used as the training sample, while the remaining 20% is designated as the test sample. Six machine learning algorithms are employed to build prediction models for oil well productivity. Finally, the model with the best performance is selected based on the evaluation metrics.

Knowledge base: Utilizing the established oilfield analogy index system and model, this study aims to uncover and comprehend the various characteristics of oil and gas fields while integrating the productivity prediction models for different types of oilfields. To achieve this, we construct an extensible oilfield analogy knowledge base and productivity prediction knowledge base. With this knowledge base, we can swiftly match, query, and study the most similar oilfields to the target oilfields, enabling us to predict the dynamic parameters of the target oilfields with accuracy and precision.

3. Setting analogy system

When assessing the similarity between oilfields, both static and dynamic characteristics need to be considered. The selection of analogy parameters is critical to oilfield analogy. In this regard, we have extracted 135 oilfield dynamic and static parameters from 8 categories (profile, reservoir, well, fluid, energy, reserves, development effect, and production performance) based on the research results of reservoir description, oilfield development scheme design, oilfield development adjustment scheme design, reservoir dynamic analysis, and reservoir development effect evaluation. These parameters are provided in a comprehensive parameter query system (Appendix A) that facilitates oilfield analogy and productivity prediction.

In practice, collecting all the analogy parameter data for target oilfields can be challenging, and in some cases, only a limited amount of basic parameter data is available. To improve the operational and practicality of oilfield analogies, we extract a set of analogy parameters applicable to various reservoirs from 135 parameter query systems based on Chinese industry standards SY/T6169 and SY/T6219, combined with expert knowledge and experience (Table 1). Certainly, in the actual process of analogical comparison, adjustments may be made to certain parameters in Table 1 from the 135 parameters by considering the specific data collected from various target oil fields. Then, using the expert scoring method, analytic hierarchy process, and principal component analysis method, we calculate the weights of the analogy parameters and establish an analogy system suitable for all types of reservoirs.

The actual analog process mainly considers static parameters, and for different target oil fields, the system's built-in analog parameters can be appropriately added or deleted from the 135 parameter templates. Take the actual offshore oil field PL as an example, which has a depth of 1000–1500 m, an average porosity of 25.42%, and an average permeability of 1039×10^{-3} μm^2. However, the actual data for dynamic and static parameters are limited, and Appendix B lists the specific parameter data for the target PL oilfield. Based on relevant reservoir knowledge and parameter data obtained from statistical analysis of the PL oilfield, we aim to ensure an ample number of analogous oilfield samples, and appropriate additions or deletions of the parameters involved in the above analog system are made. A similarity system containing the four basic parameters of permeability, porosity, reservoir effective thickness, and oil viscosity is established. Table 2 shows the weight values of analog parameters calculated by different algorithms. Based on this analogy system, big data algorithms are used to screen for similar oilfields to the PL oilfield.

The expert scoring method relies on the expertise of professionals with years of experience in reservoir work to assign scores to the 135 parameters in the system, on a scale of 0–1. The parameter scores are then normalized to derive the weight of each parameter in the analogy system.

Analytic hierarchy process (AHP) is used to determine the appropriate scale for each parameter in the analogy system by comparing them pairwise. The importance of a parameter is directly proportional to the scale value assigned to it, with larger values indicating greater importance. To ensure the rationality of parameter selection, the consistency index (CI) and the consistency ratio (CR) are calculated. CI value of 0 indicates complete consistency, with higher values indicating greater inconsistency. When the CR value is less than or equal to 0.1, the parameter matrix is considered reasonable. In our study, the CI and CR values calculated for the parameter matrix shown are both 0, indicating that the selected parameters are reasonable.

Based on orthogonal transformation, principal component

Table 1
Built-in analogy parameters for all types of reservoirs.

Analogy aspects	Specific parameters
Structural characteristics	Trap type; structural type.
Reserves	Original oil in place; technically recoverable reserves; oil-bearing area; single-well controllable reserves; reserve abundance; original gas saturation.
Drive types	Water volumetric multiple; drive energy; water type.
Fluid properties	Natural gas viscosity; gas-oil ratio; natural gas formation volume factor; formation water type; formation water salinity; formation water pH; crude oil viscosity; asphaltene content in crude oil; wax content in crude oil; crude oil density.
Temperature-pressure system	Original formation temperature; temperature gradient; formation pressure coefficient; formation fracture pressure; pressure gradient; original formation pressure.
Reservoir characteristics	Sedimentary facies; coefficient of variation; permeability contrast; sand ratio; dart coefficient; median pressure; displacement pressure; pore throat ratio; pore type; throat radius; porosity; permeability; irreducible water saturation; reservoir effective thickness; number of oil layers.
Stratigraphic characteristics	Lithology; water depth; offshore/onshore; burial depth; stratigraphic age; stratigraphic thickness.

Table 2
Three methods of oilfield analogy parameter weight allocation results.

Parameter	Expert scoring	Analytic hierarchy process	Principal component analysis
Reservoir effective thickness	0.114	0.165	1.02
Porosity	0.114	0.041	0.12
Oil viscosity	0.114	0.165	0.60
Permeability	0.114	0.041	0.75

analysis (PCA) is a method that can reduce multiple related variables into a few independent variables. By analyzing the compressed variables, the weight of the original features can be calculated. At this time, the calculated weight is the label-free weight, and its significance indicates the proportion of the information expressed by the feature data in all data.

From the weight values calculated by the three methods (Table 2), it can be concluded that the weight value of parameters calculated by the expert scoring method is consistent with the conventional experience, while the results of the analytic hierarchy process and principal component analysis are different from the conventional knowledge. In the PL oilfield analogy system, the weight value of expert scoring is mainly considered, and the three calculation weight results are saved for calculating the similarity between the target oilfield and the analogy oilfield.

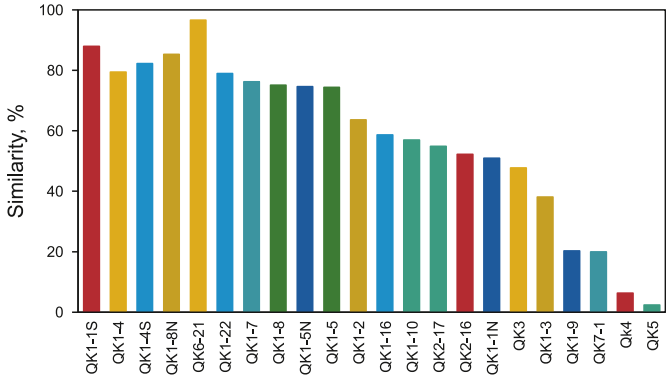


Fig. 2. The sorting results of oilfield similarity using the fuzzy matter-element method.

4. Screening of similar oilfields

This study employs six big data algorithms to calculate the similarity between the target oilfield and analogous oilfields, using the established analogy system. The algorithms used for similarity calculation include the fuzzy matter-element method, comprehensive evaluation method (which includes expert scoring method, analytic hierarchy process, and principal component analysis), cosine similarity, and Euclidean distance. These algorithms are described in detail in Appendix C, including their basic principles, advantages, and disadvantages. The sorted list of similarity scores between the target and analogous oilfields enables the identification of the most similar oilfields, which can provide valuable technical support for the efficient development of the target oilfield throughout its life cycle.

Firstly, we introduce the principle of fuzzy matter-element method. The fuzzy matter-element method is an improvement on Euclidean distance, introducing the concept of weight, which is no longer the absolute spatial distance between two points as represented by Euclidean distance. The basic principle is to calculate the squared difference between each parameter's weighted value multiplied by the normalized value of the corresponding parameter of the oilfield, sum them up to obtain a comprehensive value, take the square root to get a decision coefficient, normalize the decision coefficient to get similarity, and the similarity sorting result is shown in Fig. 2.

From the results, it can be visually observed that the QK6-21 oilfield is the most similar to the target PL oilfield, with a similarity degree of 96.68%. The QK1-1S, QK1-8N, and QK1-4 oilfields have a similarity degree higher than 80% with the target oilfield.

Next is the comprehensive evaluation method, which is based on a certain characteristic of the evaluation index. The whole is composed of multiple related evaluation indexes, integrates multiple indexes into a comprehensive evaluation through a mathematical model, and evaluates the objective entity according to certain standards. This method can judge the advantages and disadvantages of the evaluation object according to the system attributes, and the detailed study is shown in Fig. 3. Fuzzy affiliation is introduced to replace the data normalization in the fuzzy set method, in which the fuzzy set's mathematical tools are mainly applied to the fuzzy affiliation and the degree of affiliation function.

The comprehensive decision value is obtained by accumulating the product of fuzzy affiliation and weight. Due to the uncertainty of the subordinate relationship between factor u and the fuzzy sets on U, to effectively describe this relationship, a value in the interval

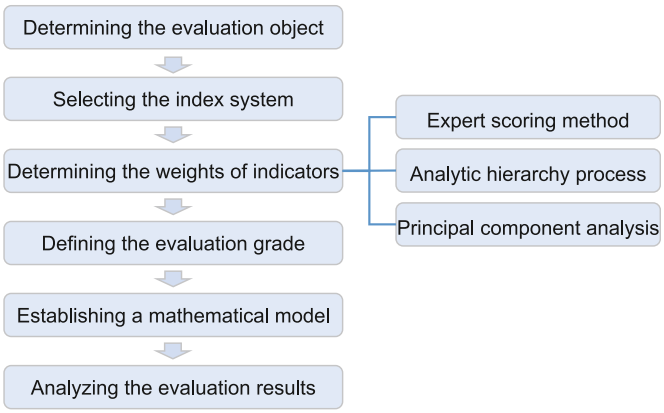


Fig. 3. Detailed process of comprehensive evaluation method.

[0, 1] is selected to replace 0 and 1, thus describing the "truth" that an element belongs to a fuzzy set. The calculation of similarity using this method is shown in Fig. 4.

The oilfield with the highest similarity to the target PL oilfield, calculated using the comprehensive evaluation method (expert scoring), is QK6-21. This result is consistent with the fuzzy matter-element method but differs from the results of the principal component analysis and the analytic hierarchy process. Overall, the computational similarity of the comprehensive evaluation method is high, all around 50%.

Finally, cosine similarity and Euclidean distance are introduced. Cosine similarity, an algorithm commonly used to quantify the degree of similarity between two entities or two indicators, specifically refers to the cosine of the angle between any two vectors in a vector space. The closer the calculated cosine is to the value 1, the more similar the two vectors are in space. The Euclidean distance is the metric we use most frequently to measure the size of a distance, and it quantifies the size of the distance between two points in space when the dimensionality of the space is high. Again, we can think of this method simply as the absolute distance between two points or vectors. The smaller the actual distance calculated, the more similar they are. The results of the similarity between the two methods are shown in Fig. 5.

Based on the similarity calculation principle and sorting results, it can be concluded that the cosine similarity and Euclidean distance methods mainly calculate the similarity between numerical values. Both methods show a high similarity between the actual PL oilfield and the analogy oilfield, with a similarity level above 90%. This similarity, however, does not hold any reference significance.

By comparing and analyzing different similarity algorithm models, Appendix C summarizes the applicability, advantages, and disadvantages of each similarity algorithm. Among them, weights are introduced in the fuzzy matter-element method to eliminate the influence of magnitude on the results, making it more suitable for reservoir analogy. The setting of weights fully utilizes the experience of oil engineers in oilfield development, and the scoring template can be optimized later by using multiple experts' scoring. The comprehensive evaluation method in addition to the consideration of the establishment of the affiliation function, based on the calculation of the weights of the indicators (expert scoring method, analytic hierarchy process, principal component analysis) should not be ignored. If the affiliation function and the weights are set reasonably, theoretically, the effect is better. Otherwise, the effect will be less than optimal.

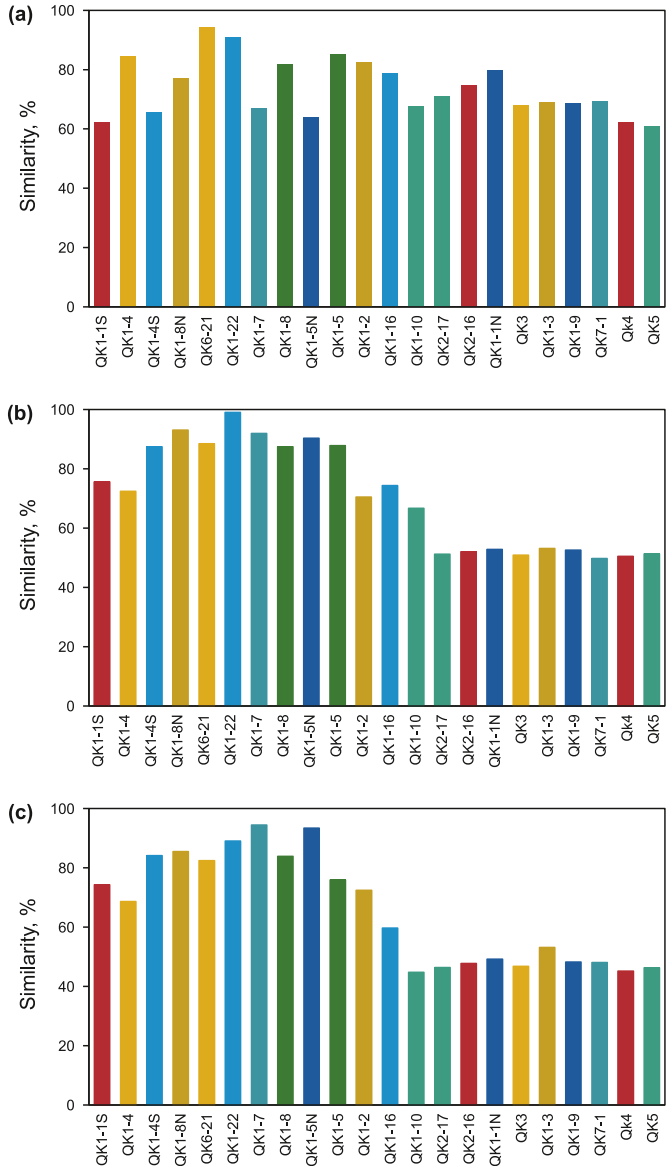


Fig. 4. The sorting results of oilfield similarity using the comprehensive evaluation method: (**a**) The results of expert scoring method. (**b**) The results of principal component analysis method. (**c**) The results of analytic hierarchy process method.

The fuzzy matter-element method and comprehensive scoring method (expert scoring method) use the weight values calculated by the expert scoring method, and the oilfield with the highest similarity in both methods is QK6-21. However, in the latter method, the similarity of 22 analogical oilfields is generally high during the similarity calculation, and the results of the principal component analysis and the analytic hierarchy process are similar, probably because the algorithm's subordination function configuration is unreasonable. By looking at the actual situation of the 22 oilfields in the comparison, the size of the similarity between oilfields calculated by the fuzzy object-element method is most consistent with the actual situation, so the calculation results of the fuzzy object-element method are mainly considered.

To ensure the validity of the results, a similarity threshold of 0.7 is set, meaning that only similarity scores greater than 70% are considered as similar oilfields. The similarity between the target

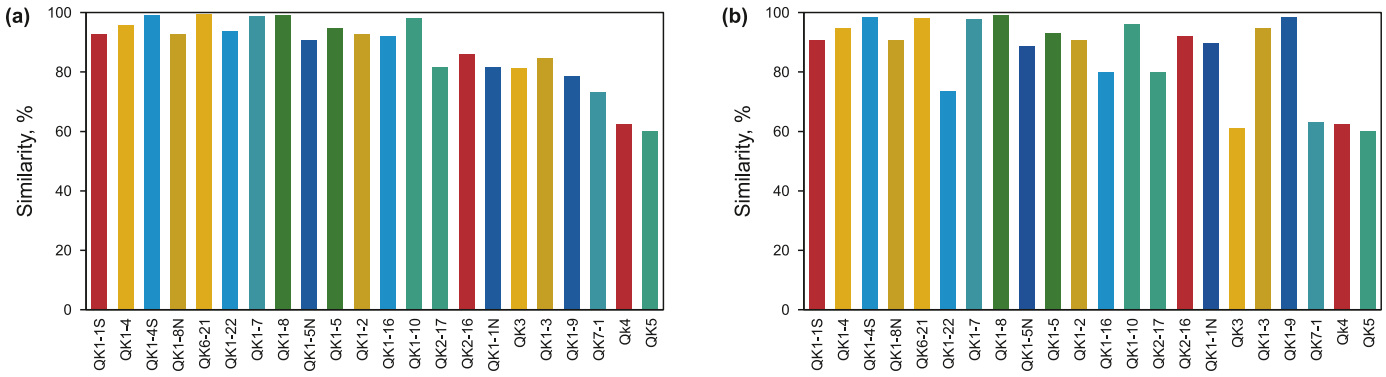


Fig. 5. The sorting results of oilfield similarity using cosine similarity (**a**) and Euclidean distance (**b**).

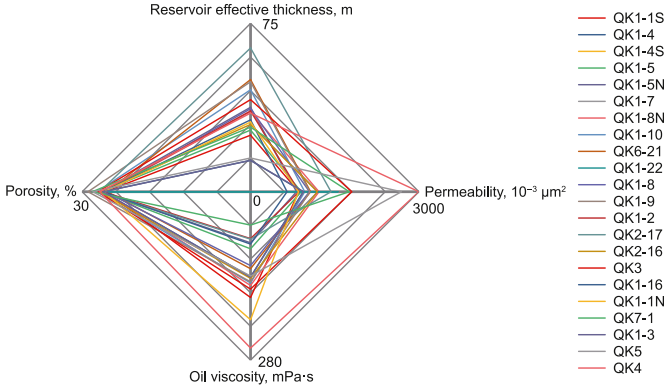


Fig. 6. Radar diagram of four parameters in the target PL oilfield analogy system.

oilfield and analogy oilfields is represented by a radar chart of four parameters from the PL oilfield analogy system. Fig. 6 shows the radar chart of the four parameters and provides an intuitive comparison of the similarity between the target oilfield and analogy oilfield parameters.

By studying the development measures implemented in different stages of QK6-21, valuable development experience can be gained for the PL oilfield. Such measures can include production techniques, drilling methods, and reservoir management strategies that have been successfully applied in QK6-21. This knowledge can be used to optimize the development of the target oilfield in its entire life cycle. To ensure that there are sufficient similar oilfield sample data available for establishing the productivity model, based on the similarity calculation results in Fig. 2, oilfields with similarity scores greater than 70% are selected as similar oilfields. These include a total of 10 oilfields with 394 oil wells, which are

Table 3
Comparison of actual and predicted productivity in the oilfield.

Oil well	2018		2019		2020		2021
	Actual productivity, t/d	Predicted productivity, t/d	Actual productivity, t/d	Predicted productivity, t/d	Actual productivity, t/d	Predicted productivity, t/d	Predicted productivity, t/d
Oil well 1	114.33	112.41	123.24	110.27	110.23	106.70	110.14
Oil well 2	200.16	203.74	202.36	205.46	201.41	201.30	201.56
Oil well 3	110.00	108.69	111.02	109.29	100.24	106.40	100.12
Oil well 4	121.80	119.87	119.25	121.45	112.25	121.43	100.89
Oil well 5	110.01	107.24	108.56	110.23	100.25	98.70	100.14
Oil well 6	135.60	134.69	142.21	123.63	132.56	129.46	129.85
Oil well 7	123.38	125.40	100.21	97.20	98.45	99.63	97.54
Oil well 8	154.82	168.30	145.89	135.90	125.63	119.36	132.13
Oil well 9	124.45	114.56	120.34	123.40	110.28	105.21	108.56
Oil well 10	121.80	117.29	115.26	109.76	100.56	97.42	99.52
Oil well 11	92.30	94.27	90.12	94.21	85.23	82.13	80.57
Oil well 12	210.00	215.29	208.25	205.23	200.24	197.23	196.54
Oil well 13	94.09	96.28	94.12	90.24	92.16	105.12	89.52
Oil well 14	55.81	47.20	52.21	24.76	42.23	51.42	38.98
Oil well 15	144.44	135.90	142.56	134.27	140.52	138.70	135.56
Oil well 16	120.32	162.45	118.56	110.46	98.56	102.40	95.47
Oil well 17	201.78	197.50	200.23	197.23	200.56	194.20	197.21
Oil well 18	123.58	116.87	120.14	116.42	121.14	116.23	116.28
Oil well 19	132.22	135.21	130.23	132.71	126.24	114.78	120.58
Oil well 20	140.09	134.70	138.56	129.80	137.25	126.54	132.41
Oil well 21	117.72	100.20	110.23	105.24	109.56	106.42	105.26
Oil well 22	126.32	106.80	123.26	134.10	121.26	103.27	120.17
Oil well 23	127.96	156.30	119.56	115.42	118.78	112.70	118.12
Oil well 24	219.62	236.20	210.23	209.58	209.25	202.30	200.23
Oil well 25	100.13	105.60	95.64	89.45	94.89	102.30	90.56
Oil well 26	114.33	121.36	110.23	97.23	109.23	104.21	104.21
Oil well 27	193.80	195.74	192.13	204.10	191.56	196.45	190.56
Oil well 28	106.48	100.20	100.45	102.70	97.23	101.20	93.24
Oil well 29	56.93	67.40	45.26	38.25	42.13	39.78	40.12
Oil well 30	134.91	129.63	120.41	113.27	119.56	121.03	110.34

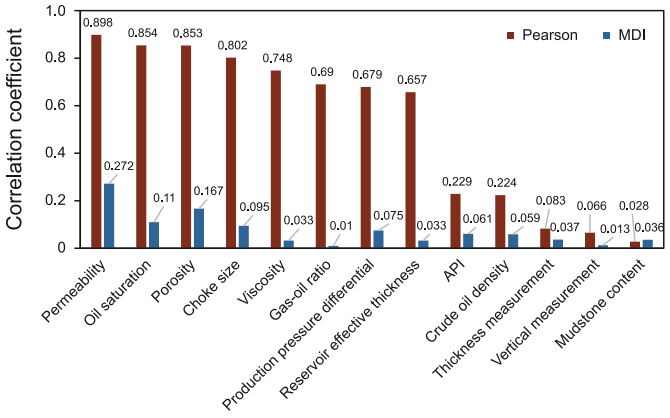


Fig. 7. Correlation analysis chart of Pearson correlation coefficient method and MDI method.

used as sample data for building the productivity prediction model.

5. Analysis of influencing factors

To analyze the correlations between the actual dynamic and static data of the selected similar oilfields and the oil well productivity, we use the Pearson correlation coefficient method. The mean decrease impurity (MDI) method is then applied for verification of the selected influencing factors. To reduce the short-term impact of many factors on production performance, this study uses the average productivity of 1 year as the main evaluation metric for determining the influencing factors.

Pearson correlation coefficient is a measure of similarity that quantifies the linear correlation between two random variables *X* and *Y*. It ranges from 0 to 1, and a higher absolute value indicates a stronger correlation. In this study, the Pearson correlation coefficient is utilized to determine whether there exists a strong

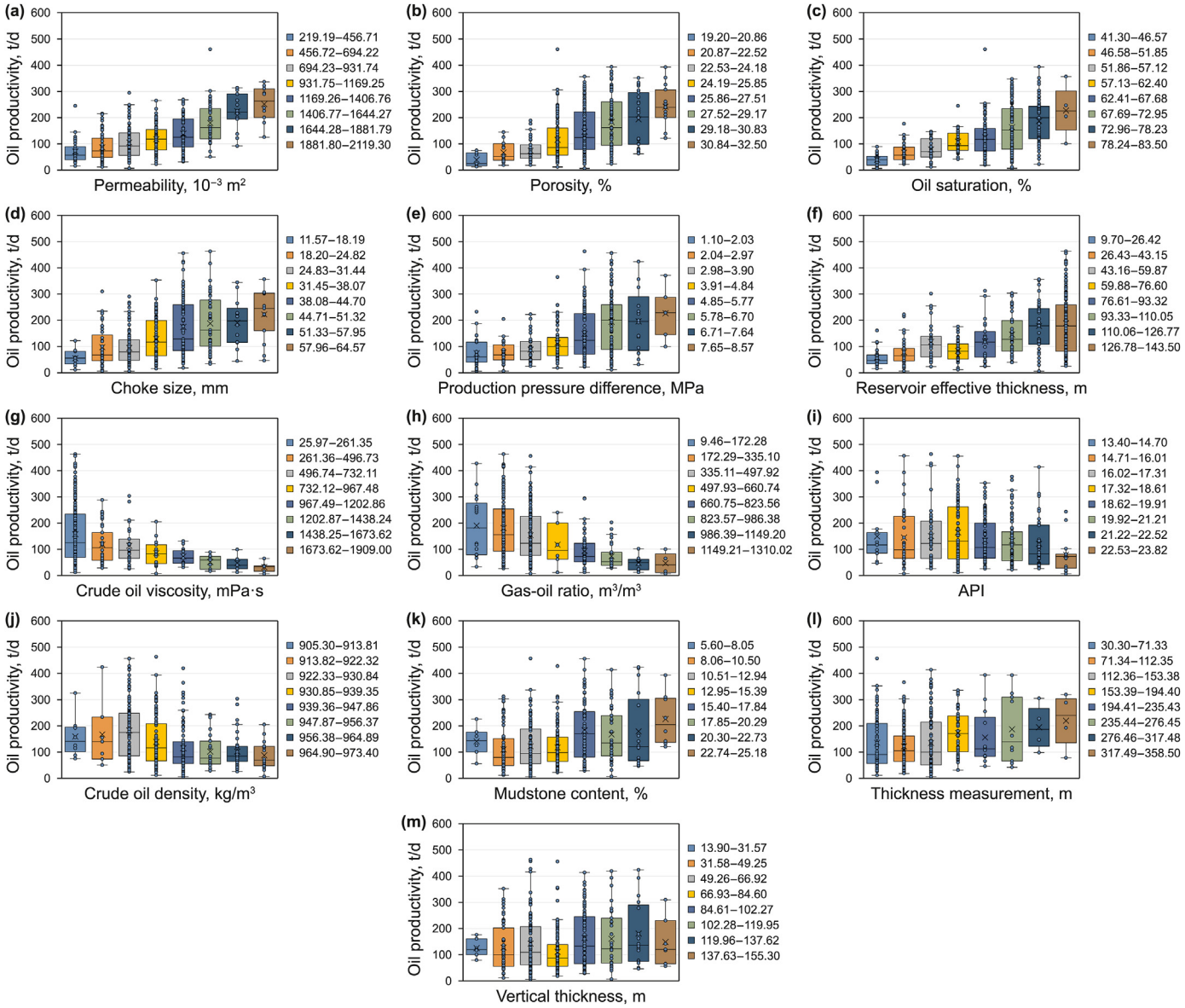


Fig. 8. Statistical diagram of productivity changing with 13 parameters: (**a**) permeability; (**b**) porosity; (**c**) oil saturation; (**d**) choke size; (**e**) production pressure difference; (**f**) reservoir effective thickness; (**g**) crude oil viscosity; (**h**) gas-oil ratio; (**i**) API; (**j**) crude oil density; (**k**) mudstone content; (**l**) thickness measurement; (**m**) vertical thickness.

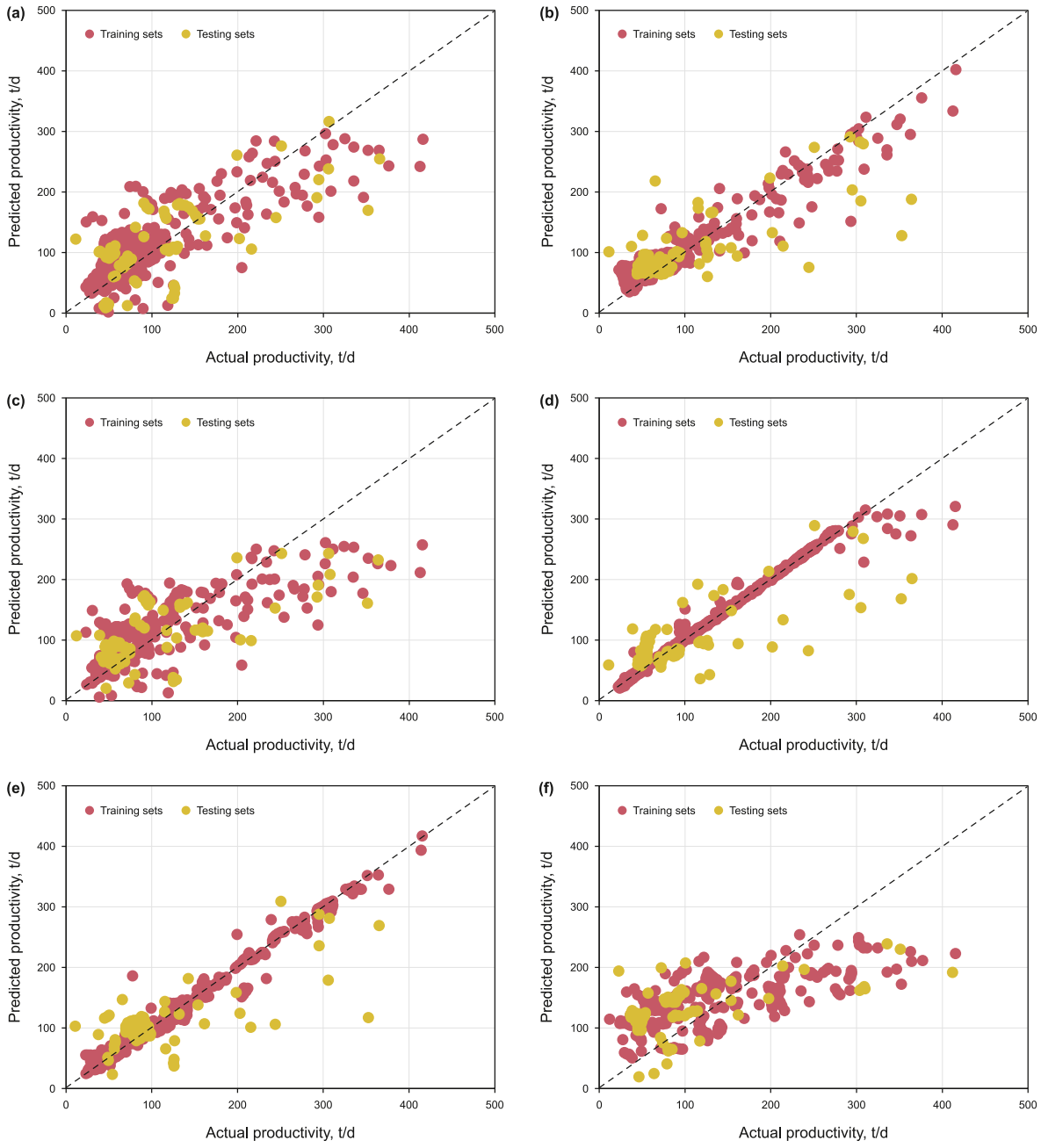


Fig. 9. Productivity prediction results of machine learning models: (**a**) LR; (**b**) RF; (**c**) SVR; (**d**) XGBoost; (**e**) LightGBM; (**f**) BP.

correlation between the statistical oilfield parameters and the productivity of oil wells. Table 3 presents the criteria for determining the strength of correlation in different ranges of correlation coefficient. MDI method is a random forest feature selection method, which belongs to the label training method. In this paper, oil well productivity prediction is a regression problem, usually using variance or least squares fitting to measure error. When training the decision tree, we can calculate the change in each feature's impure value. For the decision tree forest, we can calculate how much the impure of each feature is reduced, which can be used as a measure of the importance of features, and the calculation formula is shown in Appendix D.

To identify the factors that have a great impact on oil well productivity, 13 parameters are considered in the analysis of actual dynamic and static data from oil wells in similar oilfields identified through the similarity ranking method. These parameters include porosity, permeability, oil saturation, gas-oil ratio, crude oil density, oil API degree, choke size, thickness measurement, vertical thickness, reservoir effective thickness, mudstone content, crude oil viscosity, and production pressure difference. The Pearson correlation coefficient method is employed to assess the strength of the linear correlation between these parameters and oil well productivity. The results show that porosity, permeability, oil saturation, choke size, and crude oil viscosity are strongly correlated with productivity, while gas-oil ratio, production pressure difference, and reservoir effective thickness have moderate correlations. Crude

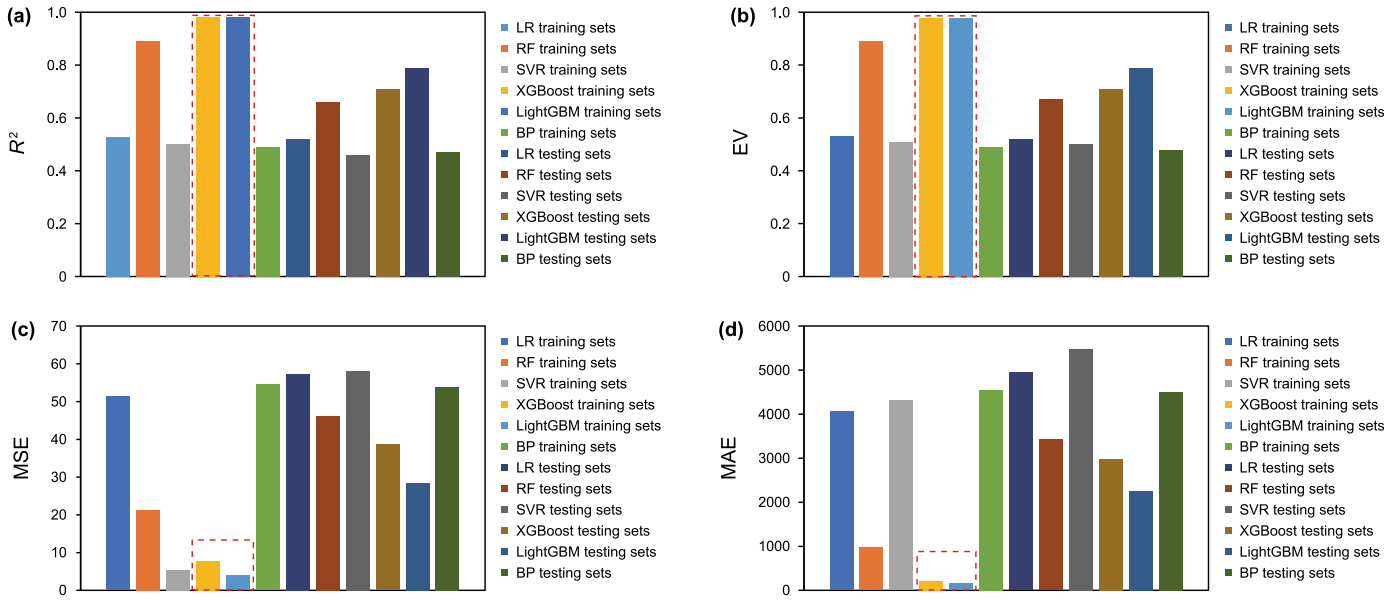


Fig. 10. Column diagrams of four evaluation metrics: (**a**) R^2; (**b**) EV; (**c**) MSE; (**d**) MAE.

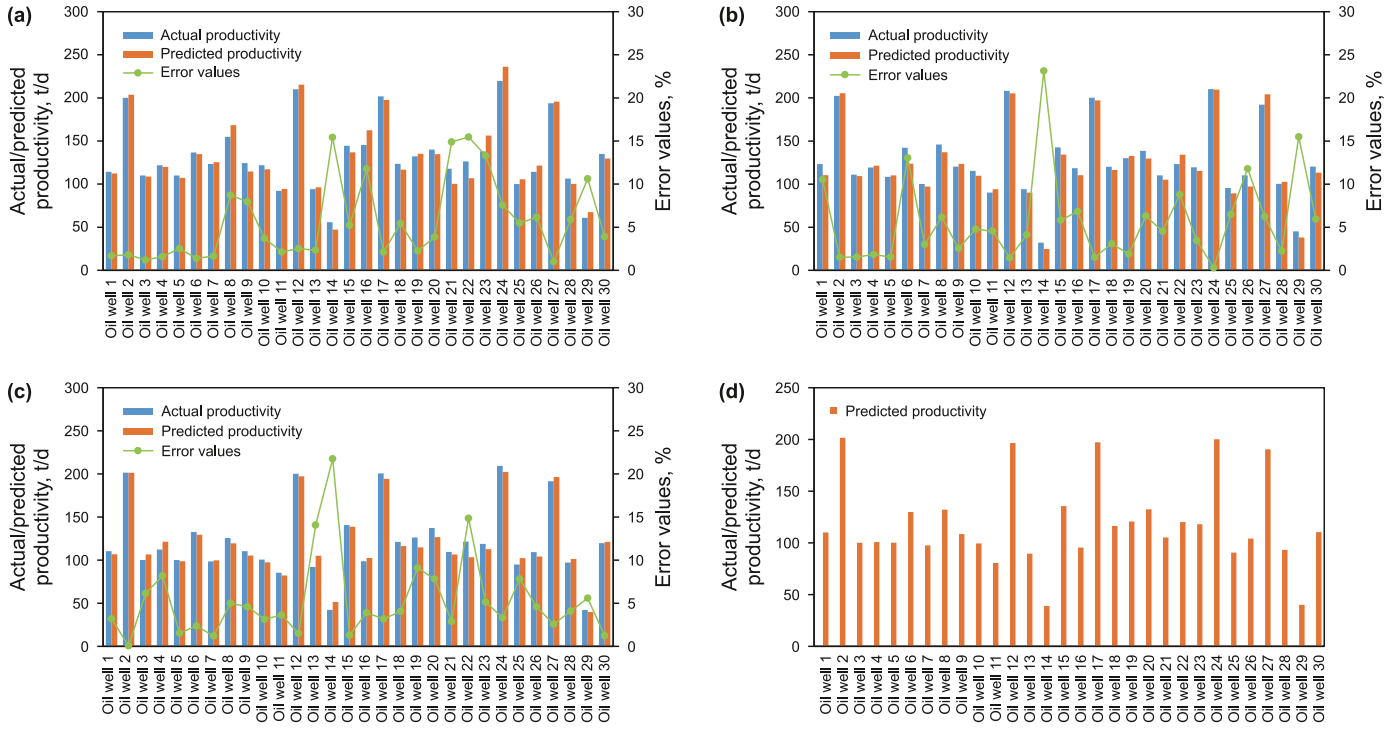
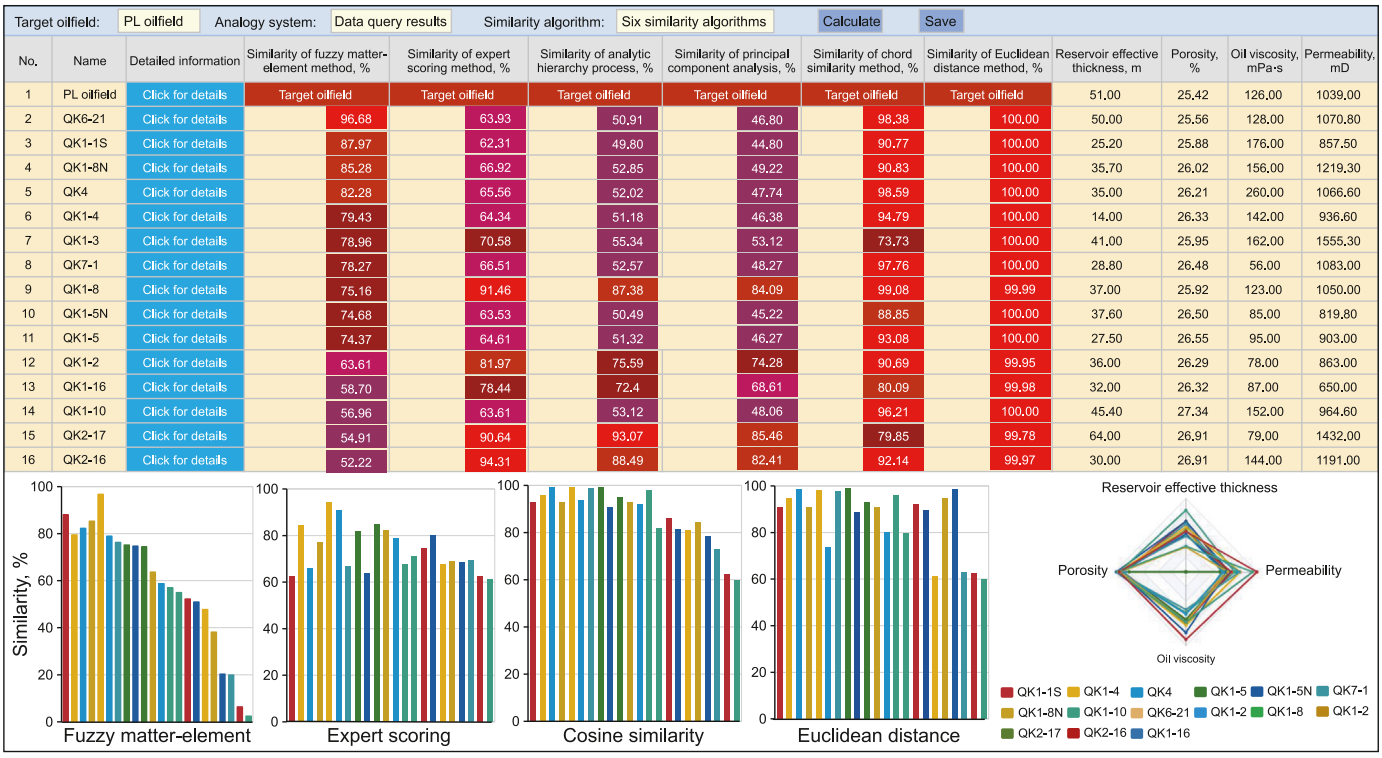


Fig. 11. Comparison of productivity prediction results in PL oilfield: (**a**) Actual and predicted productivity in 2018. (**b**) Actual and predicted productivity in 2019. (**c**) Actual and predicted productivity in 2020. (**d**) Predicted productivity in 2021.

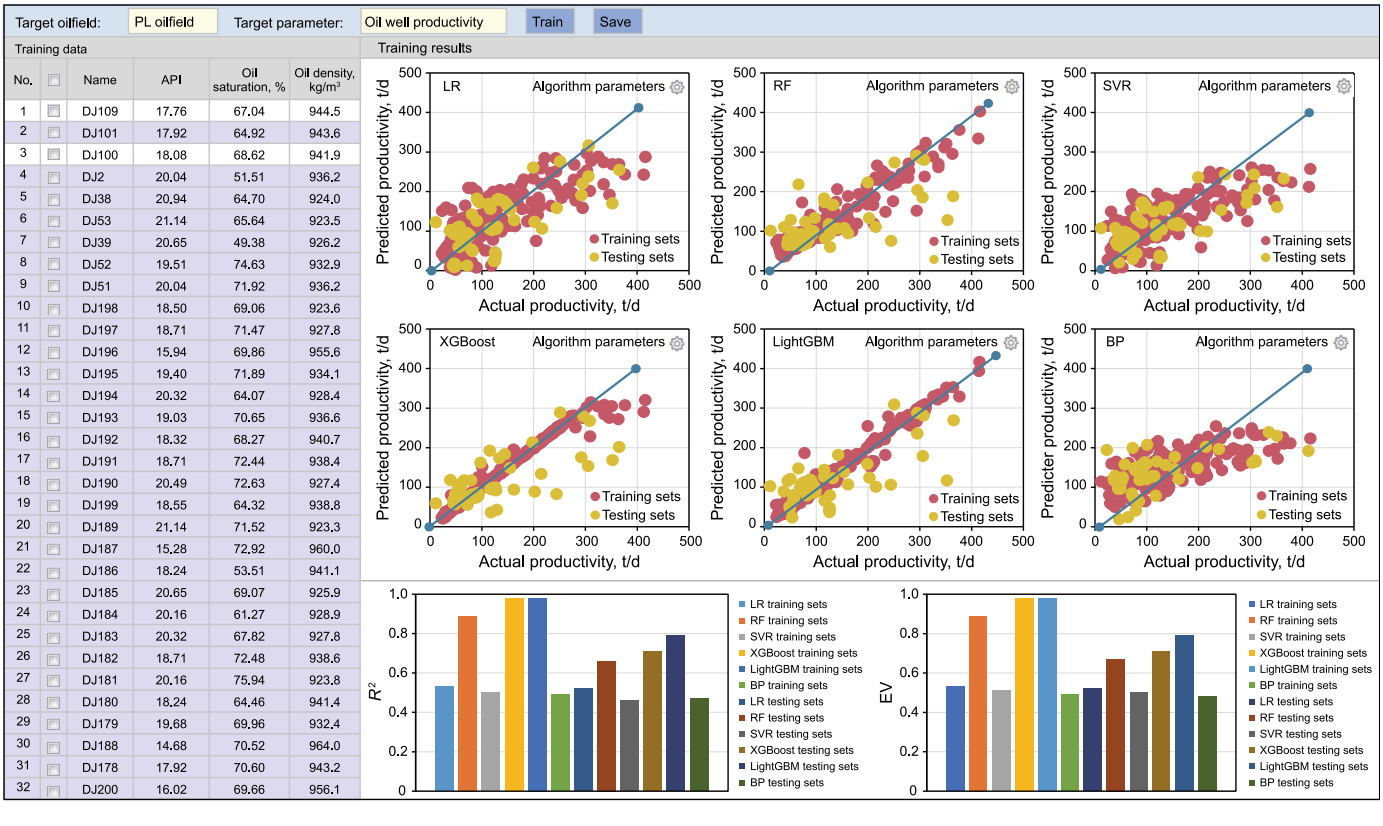
oil density and API degree have weak correlations, and mudstone content, thickness measurement, and vertical thickness are extremely weakly correlated with productivity. Additionally, the MDI method is used to measure the importance of these parameters in predicting oil well productivity. The MDI method is also applied to verify the results, which yield low correlation coefficients but are consistent with the Pearson method. Based on these findings, eight parameters (porosity, permeability, oil saturation, choke size, crude oil viscosity, gas-oil ratio, production pressure difference, and reservoir effective thickness) are selected as the influencing factors for single well productivity in the target PL oilfield. The results of the analysis are presented in Fig. 7.

To analyze the influence of the selected parameters on the productivity of offshore oil wells, 394 wells in similar oilfields are studied using their actual dynamic and static data. Thirteen box diagrams (Fig. 8) are created to illustrate the productivity changes



No.	Name	Detailed information	Similarity of fuzzy matter-element method, %	Similarity of expert scoring method, %	Similarity of analytic hierarchy process, %	Similarity of principal component analysis, %	Similarity of chord similarity method, %	Similarity of Euclidean distance method, %	Reservoir effective thickness, m	Porosity, %	Oil viscosity, mPa·s	Permeability, mD
1	PL oilfield	Click for details	Target oilfield	Target oilfield	Target oilfield	Target oilfield	Target oilfield	Target oilfield	51.00	25.42	126.00	1039.00
2	QK6-21	Click for details	96.68	63.93	50.91	46.80	98.38	100.00	50.00	25.56	128.00	1070.80
3	QK1-1S	Click for details	87.97	62.31	49.80	44.80	90.77	100.00	25.20	25.88	176.00	857.50
4	QK1-8N	Click for details	85.28	66.92	52.85	49.22	90.83	100.00	35.70	26.02	156.00	1219.30
5	QK4	Click for details	82.28	65.56	52.02	47.74	98.59	100.00	35.00	26.21	260.00	1066.60
6	QK1-4	Click for details	79.43	64.34	51.18	46.38	94.79	100.00	14.00	26.33	142.00	936.60
7	QK1-3	Click for details	78.96	70.58	55.34	53.12	73.73	100.00	41.00	25.95	162.00	1555.30
8	QK7-1	Click for details	78.27	66.51	52.57	48.27	97.76	100.00	28.80	26.48	56.00	1083.00
9	QK1-8	Click for details	75.16	91.46	87.38	84.09	99.08	99.99	37.00	25.92	123.00	1050.00
10	QK1-5N	Click for details	74.68	63.53	50.49	45.22	88.85	100.00	37.60	26.50	85.00	819.80
11	QK1-5	Click for details	74.37	64.61	51.32	46.27	93.08	100.00	27.50	26.55	95.00	903.00
12	QK1-2	Click for details	63.61	81.97	75.59	74.28	90.69	99.95	36.00	26.29	78.00	863.00
13	QK1-16	Click for details	58.70	78.44	72.4	68.61	80.09	99.98	32.00	26.32	87.00	650.00
14	QK1-10	Click for details	56.96	63.61	53.12	48.06	96.21	100.00	45.40	27.34	152.00	964.60
15	QK2-17	Click for details	54.91	90.64	93.07	85.46	79.85	99.78	64.00	26.91	79.00	1432.00
16	QK2-16	Click for details	52.22	94.31	88.49	82.41	92.14	99.97	30.00	26.91	144.00	1191.00

Fig. 12. Software interface of oilfield analogy module.



No.	Name	API	Oil saturation, %	Oil density, kg/m³
1	DJ109	17.76	67.04	944.5
2	DJ101	17.92	64.92	943.6
3	DJ100	18.08	68.62	941.9
4	DJ2	20.04	51.51	936.2
5	DJ38	20.94	64.70	924.0
6	DJ53	21.14	65.64	923.5
7	DJ39	20.65	49.38	926.2
8	DJ52	19.51	74.63	932.9
9	DJ51	20.04	71.92	936.2
10	DJ198	18.50	69.06	923.6
11	DJ197	18.71	71.47	927.8
12	DJ196	15.94	69.86	955.6
13	DJ195	19.40	71.89	934.1
14	DJ194	20.32	64.07	928.4
15	DJ193	19.03	70.65	936.6
16	DJ192	18.32	68.27	940.7
17	DJ191	18.71	72.44	938.4
18	DJ190	20.49	72.63	927.4
19	DJ199	18.55	64.32	938.8
20	DJ189	21.14	71.52	923.3
21	DJ187	15.28	72.92	960.0
22	DJ186	18.24	53.51	941.1
23	DJ185	20.65	69.07	925.9
24	DJ184	20.16	61.27	928.9
25	DJ183	20.32	67.82	927.8
26	DJ182	18.71	72.48	938.6
27	DJ181	20.16	75.94	923.8
28	DJ180	18.24	64.46	941.4
29	DJ179	19.68	69.96	932.4
30	DJ188	14.68	70.52	964.0
31	DJ178	17.92	70.60	943.2
32	DJ200	16.02	69.66	956.1

Fig. 13. Software interface of productivity prediction module.

across different parameter ranges.

The results of the box diagrams indicate that productivity increases with higher levels of porosity, permeability, oil saturation, choke size, production pressure difference, and reservoir effective thickness. Permeability, porosity, oil saturation, and reservoir effective thickness are all physical parameters of the reservoir, and productivity will increase with increasing values of these parameters under conventional production operations. The choke size is controlled by human factors, and the larger the nozzle size, the higher the productivity. The higher the production differential pressure, the lower the bottomhole pressure, the faster the fluid in the reservoir flows to the bottom of the well, and the productivity rises sharply. Conversely, productivity shows a negative correlation with crude oil viscosity and gas-oil ratio. The greater the crude oil viscosity, the poorer the fluid mobility, and the lower the productivity of the well in the same situation. The greater the gas-oil ratio, the sharper the gas production and the lower the oil production. Nevertheless, there is no significant correlation between productivity and oil API, crude oil density, mudstone content, thickness measurement, and vertical thickness. The box diagrams' trends are consistent with the Pearson method and MDI method, which verifies the reasonableness of the selection of productivity influencing factors.

Based on the patterns of actual parameter data, as well as the results of correlation ranking, and combined with relevant reservoir experience and knowledge, we consider parameters that have a significant impact on productivity during the actual mining process of the reservoir. Ultimately, we select eight parameters including porosity, permeability, oil saturation, choke size, crude oil viscosity, gas-oil ratio, production pressure difference, and reservoir effective thickness as the factors affecting PL oilfield's single-well productivity, which participate in training to establish the productivity prediction model.

6. Productivity prediction models

The input parameters for the prediction models are the eight influencing factors in the 394 single well samples determined previously, and the output parameter is the oil well productivity. We have selected six machine learning algorithms, namely LR, RF, SVR, XGBoost, LightGBM, and BP, to establish the productivity prediction models. Appendix E provides the principles, advantages, and disadvantages of each algorithm. We use four evaluation metrics to assess the performance of the models and select the best one to save the model parameters. The evaluation metrics are determination coefficient (R^2), explained variance score (EV), mean squared error (MSE), and mean absolute error (MAE). The specific calculation method is provided in Appendix F. We train and test the models using 80% of the collected data as training samples and the remaining 20% as testing samples. Fig. 9 depicts the exponential curve-fitting relationship between the actual productivity and the predicted productivity of the six models. The red point set represents the training set, and the yellow point set represents the testing set. The upper right corner of each box shows the algorithm parameters and their respective values. Specific parameters for the different regression models can be found in Appendix G. It can be observed that XGBoost and LightGBM outperform the other models, indicating a better prediction effect.

Fig. 10 presents a comparison of the evaluation metrics for the six regression models, where the LightGBM algorithm outperformed the other models with the highest R^2 and EV values and the lowest MSE and MAE values. Thus, we chose LightGBM as the optimal model for predicting the productivity of the target PL oilfield. Table 3 displays the predicted productivity results using the LightGBM model.

Fig. 11 illustrates the error analysis of the LightGBM productivity prediction model on 30 oil wells in the target PL oilfield from 2018 to 2020. The plot reveals that the predicted values from the model are largely in agreement with the actual productivity values.

The results demonstrate that the predicted values of the LightGBM productivity prediction model are consistent with the actual values. The average error of oil well prediction in the past three years is 6.31%, and the actual statistics of the oilfield production data are relatively accurate. The results of the 2021 productivity data prediction show that the productivity of the 30 wells in the coming year is not much changed compared with the previous three years, and the original mining scheme can be maintained to continue mining, which can assist in the efficient development of the oilfield and reduce the cost of crude oil exploitation.

7. Software platform

To support efficient oilfield development, a data warehouse is established based on the development data of oilfields. Utilizing big data algorithms, a geological reservoir analogy research platform is designed and developed with four functional modules: data query module, oilfield analogy module, productivity prediction module, and knowledge base. The data query module allows users to set parameter constraint conditions to query the oilfields, blocks, and single wells that meet the conditions from the data warehouse. The results are presented in a list, and the geographic location can be displayed on the GIS geographic information map of the main interface. In the oilfield analogy section, users can import target oilfield parameter data based on the specified data format. They can choose to use the built-in analogy system (Table 1) or customize the analogy parameters from 135 reservoir parameters to establish an analogy system. The similarity is calculated using big data algorithms to find oilfields that are similar to the target oilfield (Fig. 12). In the oil well productivity prediction module, users can select and click on the oil well parameters and productivity for correlation analysis, save the parameters with high correlation, and then train the model. The performance of different models on the test set is evaluated by comparing their prediction results. Subsequently, the optimal model is selected to predict the oil well productivity, as shown in Fig. 13. During the training phase, we leverage a historical dataset to train the model and evaluate its performance to select the best-performing model. During the inference phase, we utilize the pre-trained model to make predictions on new data. As such, retraining for every query is unnecessary; instead, we can rely on the pre-trained model for inference. This approach not only saves time and computational resources but also enhances efficiency. The knowledge base module includes the establishment of a standardized classification knowledge system, which includes the collation of structured, semi-structured data, and unstructured data, and it can be used for knowledge acquisition, query, and push.

The software platform is designed to provide users with a range of functionality to facilitate the identification of potential reservoirs, prediction of production dynamics, and analysis of the performance of similar oilfields. The data query module allows for easy access to relevant data from the data warehouse. The oilfield analogy module utilizes the prediction models to identify reservoirs with similar characteristics, which can then be used to extrapolate information about the performance of the target oilfield. The productivity prediction module enables users to predict the productivity of specific oil wells using a range of machine

learning algorithms. Finally, the knowledge base provides users with a wealth of information on oilfield development, including best practices and case studies. The platform includes similar oilfield screening and production dynamic index prediction models at its core.

8. Conclusions

This paper presents a novel approach for utilizing data technology to improve oilfield analogy and oil well productivity prediction. The following conclusions can be drawn.

(1) To establish the analogy index system, we customize the relevant parameters for different oilfields by considering 135 static and dynamic characteristic parameters. The parameter weights are calculated using a big data algorithm, allowing for flexibility to meet the unique characteristics of each oilfield. This approach enables the establishment of a corresponding analogy index system that is tailored to the specific parameters needs of each oilfield.
(2) Based on the established analogy model, six algorithms are employed to calculate the similarity between the target oilfield and the analogy oilfields. The approach enables quick and accurate quantification of the differences between the target oilfield and the analogy oilfields, leading to the identification of the most similar oilfield. The development strategy of the most similar oilfield provides an excellent technical reference for efficient development of the target oilfield.
(3) The oilfields that exhibit similarity greater than 70% are selected as sample data. Relevant parameters that affect oil well productivity can be quickly screened out using big data algorithms. These parameters are then used to develop six productivity prediction models, which are optimized based on four evaluation metrics to ensure the accuracy of the model's prediction for different data samples.
(4) An oilfield analogy research platform covering data query, setting target oilfields, setting analogy system, screening similar oilfields, influencing factor analysis, and optimization of oil well productivity prediction model is established, which can rapidly and accurately identify the similar oilfields of a target oilfield and predict its oil well productivity. This tool is designed to improve the efficiency and accuracy of oilfield development planning and reduce the cost of crude oil exploitation.

The oil industry is currently home to vast amounts of production data, which can be analyzed using machine learning algorithms to facilitate oilfield analogy and oil well productivity prediction in newly developed oilfields, thereby reducing the cost of oil companies. This paper presents the application of data mining technology in oilfield analogy and oil well productivity prediction, which involves setting up an analogy system, screening similar oilfields, analyzing the influencing factors of oil well productivity, and optimizing the prediction model of oil well productivity. The proposed method is not only suitable for different types of oil fields but also for gas fields. By quickly and accurately finding similar oilfields and predicting oil well productivity, this approach provides robust technical support for efficient development of the target oilfields. Our team is continually enriching and refining the software platform to further enhance and promote the application program.

CRediT authorship contribution statement

Wen-Peng Bai: Conceptualization, Writing – original draft. **Shi-Qing Cheng:** Conceptualization, Funding acquisition. **Xin-Yang Guo:** Data curation, Investigation. **Yang Wang:** Methodology, Resources. **Qiao Guo:** Software, Validation, Visualization, Writing – review & editing. **Chao-Dong Tan:** Methodology, Project administration, Resources.

Declaration of interest statement

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 research is supported by the National Natural Science Fund of China (No. 52104049), and the Science Foundation of China University of Petroleum, Beijing (No. 2462022BJRC004).

Appendix A. Parameter query system

Table A1
Parameter query system.

Profile	Reservoir	Well	Fluid
Sedimentary facies	Average reservoir thickness	Pump frequency	Formation water mineralization
Strata age	Reservoir effective thickness	Pump inlet pressure	Formation water salinity
Main drive mechanism	Effective porosity	Thickness measurement	Formation water type
Structure type	Pore throat radius	Tested productivity	The potential of hydrogen in water
Offshore/onshore oilfield	Alkali sensitivity	Initial gas productivity	Oil density
Development method	Cementation type	Initial oil productivity	Oil viscosity
Pore type	Pore-throat volume ratio	Vertical measurement	Wax content of formation crude oil
Burial depth	Porosity	Commissioning date	Asphaltene in formation crude oil
Gas layers temperature	Coefficient of variation	Oil well productivity	Crude oil density in stock tank
Drive type	Displacement pressure	Choke size	Surface crude oil viscosity (0 °C)
Trap type	Sand ratio	Total number of wells	Mobility
Water depth	Sandstone thickness	Well distance	Capillary pressure data
Hydrocarbon type	Permeability	Wellhead pressure	Gas-oil ratio
Production time	Permeability contrast	Well pattern type	Natural gas density
Lithology	Irreducible water saturation	Well spacing density	Gas formation volume factor
Reservoir type	Water sensitivity	Well type	Viscosity of natural gas
Abandonment time	Speed sensitivity	Duration of the well opening	Surface crude oil density (50 °C
	Acid sensitivity	Argillaceous content	Surface crude oil viscosity (50 °C

(*continued on next page*)

Table A1 (*continued*)

Profile	Reservoir	Well	Fluid
	Total compressibility	Daily gas production	Oil saturation
	Dart coefficient	Daily water production	Water saturation
	Temperature gradient	Penetration thickness	Original gas saturation
	Salt sensitivity	API degree of oil	Oil volume factor
	Stress sensitivity	Casing pressure	
	Number of oil layers		
	Reservoir perforation thickness		
	Median pressure		
Energy	**Reserves**	**Development effect**	**Production performance**
Formation fracture pressure	Reserves abundance	Recovery	Recovery degree
Formation pressure factor	Dynamic reserves	Oil production rate	Accumulative condensate oil
Water volumetric multiple	Oil-bearing area	Productivity index	Cumulative gas production
Water type	Condensate oil recoverable reserves	Water cut rising rate	Cumulative water production
Pressure gradient	Gas recoverable reserves	Flushing efficiency	Cumulative oil production
Production pressure difference	Oil recoverable reserves	Remaining recoverable reserve	Average daily fluid production
Present formation pressure	Proved reserves	Water flooding utilization degree	Average daily oil production
Present formation temperature	Original condensate oil in place	Control of water drive reserves	Average daily injection volume
Original formation temperature	Original gas in place	Reserve-production ratio	Average injection pressure
Original formation pressure	Original oil in place	Injection-production ratio	Pressure drawdown
Saturation pressure		Yearly decline rate	Average daily water production
Bottomhole pressure		Natural decline rate	Water injection mode
		Composite decline rate	

Appendix B. Basic parameter information of the PL oilfield

Table B1
Basic parameter information of the PL oilfield.

Oilfield parameter	Value	Individual well parameter	Value
Reservoir types	Lithologic structural reservoir	Porosity, %	23.00–29.38
Drive types	Artificial water injection	Permeability, 10^{-3} μm^2	420.4–1792.5
Sedimentary facies	Meandering river	Oil saturation, %	41.30–80.15
Lithology	Sedimentary rock	Gas-oil ratio, m^3/m^3	28.69–127.49
Reservoir effective thickness, m	51	Crude oil density, kg/m^3	905.3–945.1
Porosity, %	25.42	API	17.60–21.31
Permeability, 10^{-3} μm^2	1039	Duration of well production, h	22.99–23.94
Crude oil viscosity, mPa·s	126	Pump frequency, Hz	52–90
Mobility, 10^{-3} μm^2/mPa·s	8.24	Pump intake pressure, MPa	0.10–0.71
Vertical thickness, m	32.0–137.6	Wellhead pressure, MPa	1.22–5.18
Perforation thickness, m	25.3–134.6	Casing pressure, MPa	0.61–5.23
Mudstone content, %	7.11–68.04	Pressure gradient, MPa/m	0.86–1.10
Crude oil viscosity, mPa·s	61.14–231.10	Choke size, mm	10.07–18.14
Water saturation, %	1.97–30	Thickness measurement, m	34.6–343.5
Pressure drawdown, MPa	2.01–8.04		

Appendix C. Comparison of similarity algorithms

Table C1
Comparison of similarity algorithms.

Similarity algorithm	Algorithm principle	Advantages	Disadvantages
Fuzzy matter-element	First, the matter-element model is established. Then the results of the expert scoring method to determine the weight of each factor to obtain a fuzzy matter element matrix. Finally, the fuzzy evaluation matrix and the weight vector of factors are calculated and normalized to obtain the comprehensive results of the fuzzy evaluation.	The mathematical model is simple and easy to master. It has a good effect on multi-factor and multi-level complex problems.	It cannot solve the problem of evaluation information repetition caused by the correlation between evaluation metrics. Since the determination of the weight of each factor is subjective, the subjective judgment of people is widely used in the evaluation process.
Comprehensive evaluation	Expert scoring: Firstly, the factor set and evaluation set of the evaluated object are determined. Then the weights and their membership vectors	In the absence of sufficient statistical data and raw data, quantitative evaluation can be made with simple and intuitive features.	It has the characteristics of strong subjective factors and insufficient explanatory power.
	Analytic hierarchy process: calculated by expert scoring method, analytic hierarchy process, and principal component analysis are used to obtain the evaluation	This algorithm requires a small amount of data and takes a very short time for decision-	It is subjective to use AHP to make decisions. When there are many factors and large-scale

Table C1 (*continued*)

Similarity algorithm	Algorithm principle	Advantages	Disadvantages
	matrix. Finally, the evaluation matrix and the weight vector of the factors are calculated and normalized to get the comprehensive evaluation results.	making. The decision-making process is highly organized and scientific.	evaluation problems, the model is prone to problems.
	Principal component analysis	It can deeply analyze the essence, influencing factors, and internal relations of complex decision problems, and provide a simple decision method for complex decision problems with multi-objective, multi-criteria, and unstructured characteristics.	It is divided into qualitative analysis and quantitative analysis. The method of qualitative evaluation efficiency is less subjective, and the evaluation results are accidental.
Cosine similarity	The chord angle of two vectors in vector space is used to measure the difference between two individuals. This method focuses on the difference between the two vectors in the direction, rather than the distance or length.	It can offset the high-dimensional problem of Euclidean distance.	It ignores the size of vectors, only considers their directions, and doesn't consider the scale differences between different parameters.
Euclidean distance	The most common measure of distance is the absolute distance between points in a multidimensional space.	It is intuitive, easy to implement, and can reflect the absolute difference of individual numerical characteristics.	The calculated distance may be skewed according to the unit of elements. With the increase of data dimension, the use of Euclidean distance is smaller.

Appendix D. Calculation formula of correlation algorithm

Table D1
Calculation formula of correlation algorithm.

Correlation algorithm	Calculation formula	Symbolic meaning of the formula
Pearson correlation coefficient	$r = \frac{\sum_{i=1}^{n_{samples}}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n_{samples}}(X_i-\overline{X})^2}\sqrt{\sum_{i=1}^{n_{samples}}(Y_i-\overline{Y})^2}}$	r is the correlation coefficient, X_i represents the observed value of variable X at point i, Y_i represents the observed value of variable Y at point i, $\overline{X}$ represents the average number of X samples, and $\overline{Y}$ represents the average number of Y samples.
Mean decrease accuracy (MDI)	$Error_{std} = \frac{\left\|y_{pred} - y_{obs}\right\|}{y_{obs}}$	y_{pred} represents the predicted value of the random forest model established by using the selected M parameters, y_{obs} represents the observed value, and $Error_{std}$ represents the prediction error value of the random model.
	$MDI_i = \frac{\lvert Error_i - Error_{std}\rvert}{Error_{std}}$	MDI_i represents the predicted value of the random forest model established by removing one characteristic parameter and using the remaining $M-1$ parameters, and $Error_i$ represents the random forest prediction error with one characteristic parameter removed.

Appendix E. Comparison of six machine algorithms

Table E1
Comparison of six machine algorithms.

Productivity prediction model	Algorithm principle	Advantages	Faults
LR	A method to study the linear relationship between a dependent variable Y and multiple independent variables X.	There is no need to normalize the data, and the result is the original data, so there is no dimension problem.	It has high time complexity. When the number of samples increases, the computing time will enhance.
RF	Multiple decision trees are constructed. When a sample needs to be predicted, the prediction results of each tree in the forest for the sample are counted, and then the result is selected from these prediction results by voting.	Training can be highly parallelized, which has advantages for large sample training speed in the era of big data. When the sample feature dimension is high, the model can still be trained efficiently. The trained model has a small variance and strong generalization ability.	RF model is prone to overfitting on some noisy sample collection.
SVR	An interval band is made on both sides of the linear function, and the loss is not calculated for all samples falling into the interval band. Only outside the interval band, the loss function is included.	The algorithm is suitable for small sample data with low computational complexity. It can solve high dimensional problems and overcome the	The model is sensitive to the selection of parameters and kernel functions.

(*continued on next page*)

Table E1 (*continued*)

Productivity prediction model	Algorithm principle	Advantages	Faults
	Then the model is optimized by minimizing the width and total loss of the interval.	defects of slow convergence and local minima of neural networks.	
BP	The learning process consists of two processes: signal forward propagation and error backpropagation. In forward propagation, the input sample is introduced from the input layer and processed layer by layer by layer to the output layer. If the actual output of the output layer is not consistent with the expected output, the reverse propagation stage of the steering error. The backpropagation of error is to back-transmit the output error layer by layer to the input layer in some form through the hidden layer, and allocate the error to all the units of each layer, to obtain the error signal of each unit. This error signal is used as the basis for correcting the weight of each unit.	The model has an extremely strong adaptive function for many non-structural and inaccurate laws and has the characteristics of information memory, autonomous learning, knowledge reasoning, and optimization calculation. The model has self-learning and adaptive functions and has certain generalization abilities.	The algorithm is complex, and the network structure is difficult to determine. It is easy to fall into local minima, and the learning speed is very slow. The possibility of network training failure is large.
XGBoost	In the training process, multiple classifiers are learned by changing the weights of training samples, and the optimal classifier is finally obtained. After each round of training, the weights of the correctly classified training samples are reduced, and the weights of the wrongly classified training samples are increased. After multiple trainings, some wrongly classified training samples will get more attention, while the weights of the correct training samples are close to 0, and multiple simple classifiers are obtained. By combining these classifiers, a final prediction model is obtained.	The model is more accurate and flexible, and the regularization reduces the variance of the model. It helps prevent overfitting, reduces computation, processes missing values, and supports parallelism.	The data set needs to be traversed in the process of node splitting, so the calculation is large. The spatial complexity of the pre-sorting process is too high. The algorithm consumes large space and time and is not friendly with cache optimization.
LightGBM	Based on the histogram decision tree algorithm, the LightGBM algorithm is further optimized. The algorithm uses the leaf-wise algorithm with depth limitation, uses GOSS (gradient-based one-side sampling, gradient-based unilateral sampling) to sample the samples to calculate the gradient, and uses EFB (exclusive feature bundling, mutually exclusive feature binding) to bind some features together to reduce the dimension of features, so that the consumption of finding the best segmentation point is reduced.	The model has high-efficiency parallel training and has faster training speed, lower memory consumption, better accuracy, and support for distributed features. It can quickly process massive data.	It is based on a deviation algorithm, so it is more sensitive to noise. It may grow deeper decision trees, resulting in overfitting.

Appendix F. Regression model evaluation index

Table F1
Regression model evaluation index.

Evaluation index	Formula	Evaluation criteria
R^2 score	$R^2 = 1 - \frac{\sum_{i=1}^{n_{\text{samples}}}(y_{\text{ture}} - y_{\text{pred}})^2}{\sum_{i=1}^{n_{\text{samples}}}(y_{\text{ture}} - \overline{y}_{\text{ture}})^2}$	The results are between 0 and 1. The larger the value is, the better the effect is.
Explained variance score	$EV = 1 - \frac{\text{Var}\{y_{\text{ture}} - y_{\text{pred}}\}}{\text{Var}\{y_{\text{ture}}\}}$	The result is between 0 and 1. The smaller the value is, the worse the result is.
Mean squared error	$MSE = \frac{1}{n_{\text{samples}}}\sum_{i=1}^{n_{\text{samples}}}(y_{\text{ture}} - y_{\text{pred}})^2$	The smaller the calculation result, the smaller the error.
Mean absolute error	$MAE = \frac{1}{n_{\text{samples}}}\sum_{i=1}^{n_{\text{samples}}}\left\lvert y_{\text{ture}} - y_{\text{pred}}\right\rvert$	The larger the calculation results, the greater the error.

Notes: y_{ture} is the actual output value (observed value) for data y, $\overline{y}_{\text{ture}}$ is the average actual output value for data y, and y_{pred} is the corresponding predicted value.

Appendix G. Different regression algorithm model parameters

Table G1
Linear regression algorithm parameters.

Parameter	Value
Proportion of testing sets	0.2
Proportion of training sets	0.8

Table G2
Random forest regression algorithm parameters.

Parameter	Value
Proportion of testing sets	0.2
Number of weak learning machines	6
Number of processors available	−1
Random state	30
Maximum number of features to consider when dividing	Automation
Minimum number of samples for leaf nodes	5

Table G3
Support vector regression algorithm parameters.

Parameter	Value
Proportion of testing sets	0.2
Kernel function selection	Linear kernel

Table G4
Extreme gradient boosting algorithm parameters.

Parameter	Value
Proportion of testing sets	0.2
Classifier	Gbtree
Learning rate	0.1
Maximum delta steps	0.1
Maximum depth of tree	6
The largest leaf node	3
L2 regularized weights	1
Evaluation indicators	R_{mse}
Random tree seed	0

Table G5
Light gradient boosting machine algorithm parameters.

Parameter	Value
Proportion of testing sets	0.2
Model upgrade method	Gbdt
The largest leaf of the base learner	10
Minimum data required for leaves	30
Learning tasks and learning objectives	Regression model
Maximum depth of base learner tree	3
Learning rate	0.05
Random selection of feature proportion in iteration	0.8
Sample sampling ratio of guided clustering algorithm	105
Random selection of partial sample proportion in each iteration	0.8
Random number seeds of guided aggregation algorithm	11
Evaluation index	Mean square error
L1 regularization	0.1
Whether to output intermediate information	−1

Table G6
Back propagation algorithm parameters.

Parameter	Value
Proportion of testing sets	0.2
Proportion of training sets	0.8
Number of iterations	200

References

Aïfa, T., 2014. Neural network applications to reservoirs: physics-based models and data models. J. Petrol. Sci. Eng. 123, 1–6. https://doi.org/10.1016/j.petrol.2014.10.015.

Akbilgic, O., Zhu, D., Gates, I.D., et al., 2015. Prediction of steam-assisted gravity drainage steam to oil ratio from reservoir characteristics. Energy 93, 1663–1670. https://doi.org/10.1016/j.energy.2015.09.029.

Awoleke, O.O., Lane, R.H., 2011. Analysis of data from the Barnett shale using conventional statistical and virtual intelligence techniques. SPE Reservoir Eval. Eng. 14 (5), 544–556. https://doi.org/10.2118/127919-pa.

Bahonar, E., Chahardowli, M., Ghalenoei, Y., et al., 2022. New correlations to predict oil viscosity using data mining techniques. J. Petrol. Sci. Eng. 208, 109736. https://doi.org/10.1016/j.petrol.2021.109736.

Bravo, C., Saputelli, L., Rivas, F., et al., 2013. State of the art of artificial intelligence and predictive analytics in the E&P industry: a technology survey. SPE J. 19 (4), 547–563. https://doi.org/10.2118/150314-pa.

Bai, W., Cheng, S., Wang, Y., et al., 2024. A transient production prediction method for tight condensate gas wells with multiphase flow. Petrol. Explor. Dev. 51 (1), 1–7. https://doi.org/10.11698/PED.20230382 in Chinese.

Cai, J., Hajibeygi, H., Yao, J., et al., 2020. Advances in porous media science and engineering from InterPore2020 perspective. Adv. Geo-Energy Res. 4 (4), 352–355. https://doi.org/10.46690/ager.2020.04.02.

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

Feng, Q., Xu, S., Xing, X., et al., 2020. Advances and challenges in shale oil development: a critical review. Adv. Geo-Energy Res. 4 (4), 406–418. https://doi.org/10.46690/ager.2020.04.06.

Ghahramani, Z., 2015. Probabilistic machine learning and artificial intelligence. Nature 521 (7553), 452–459. https://doi.org/10.1038/nature14541.

Guo, Q., Cheng, S., Zeng, F., et al., 2022. Reservoir permeability prediction based on analogy and machine learning methods: field cases in DLG Block of Jing'an Oilfield, China. Lithosphere 2022 (Special 12), 5249460. https://doi.org/10.2113/2022/5249460.

Guo, X., Hu, D., Li, Y., et al., 2019. Theoretical progress and key technologies of onshore ultra-deep oil/gas exploration. Engineering 5 (3), 458–470. https://doi.org/10.1016/j.eng.2019.01.012.

Guo, Z., Chen, C., Gao, G., et al., 2018a. Integration of support vector regression distributed Gauss-Newton optimization method and its application to the uncertainty assessment of unconventional assets. SPE Reservoir Eval. Eng. 21 (4), 1007–1026. https://doi.org/10.2118/191373-PA.

Guo, Z., Chen, C., Gao, G., et al., 2018b. Enhancing the performance of the distributed Gauss-Newton optimization method by reducing the effect of numerical noise and truncation error with support-vector regression. SPE J. 23 (6), 2428–2443. https://doi.org/10.2118/187430-PA.

Gurina, E., Klyuchnikov, N., Zaytsev, A., et al., 2020. Application of machine learning to accidents detection at directional drilling. J. Petrol. Sci. Eng. 184, 106519. https://doi.org/10.1016/j.petrol.2019.106519.

Handhal, A.M., Ettensohn, F.R., Al-Abadi, A.M., et al., 2022. Spatial assessment of gross vertical reservoir heterogeneity using geostatistics and gis-based machine-learning classifiers: a case study from the zubair formation, rumaila oil field, southern Iraq. J. Petrol. Sci. Eng. 208, 109482. https://doi.org/10.1016/j.petrol.2021.109482.

He, Y., He, Z., Tang, Y., et al., 2023. Shale gas production evaluation framework based on data-driven models. Petrol. Sci. 20, 1659–1675. https://doi.org/10.1016/j.petsci.2022.12.003.

Iraji, S., Soltanmohammadi, R., Munoz, E.R., et al., 2023a. Core scale investigation of fluid flow in the heterogeneous porous media based on X-ray computed tomography images: upscaling and history matching approaches. Geoenergy Sci. Eng. 225, 211716. https://doi.org/10.1016/j.geoen.2023.211716.

Iraji, S., Soltanmohammadi, R., Matheus, G.F., et al., 2023b. Application of unsupervised learning and deep learning for rock type prediction and petrophysical characterization using multi-scale data. Geoenergy Sci. Eng. 230, 212241. https://doi.org/10.1016/j.geoen.2023.212241.

Lecun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444. https://doi.org/10.1038/nature14539.

Liao, S.H., Chu, P.H., Hsiao, P.Y., 2012. Data mining techniques and application—a decade review from 2000 to 2011. Expert Syst. Appl. 39, 11303–11311. https://doi.org/10.1016/j.eswa.2012.02.063.

Lolon, E., Hamidieh, K., Weijers, L., et al., 2016. Evaluating the relationship between well parameters and production using multivariate statistical models: a middle Bakken and Three Forks case history. In: SPE Hydraulic Fracturing Technology Conference. https://doi.org/10.2118/179171-MS.

Ma, Z., Leung, J.Y., Zanon, S., et al., 2015. Practical implementation of knowledge-based approaches for steam-assisted gravity drainage production analysis. Expert Syst. Appl. 42 (21), 7326–7343. https://doi.org/10.1016/j.eswa.2015.05.047.

Montgomery, J.B., O'Sullivan, F.M., 2017. Spatial variability of tight oil well productivity and the impact of technology. Appl. Energy 195, 344–355. https://doi.org/10.1016/j.apenergy.2017.03.038.

Pirizadeh, M., Alemohammad, N., Manthouri, M., et al., 2021. A new machine learning ensemble model for class imbalance problem of screening enhanced oil recovery methods. J. Petrol. Sci. Eng. 198, 108214. https://doi.org/10.1016/j.petrol.2020.108214.

Soltanmohammadi, R., Iraji, S., de Almeida, T.R., et al., 2024. Investigation of pore geometry influence on fluid flow in heterogeneous porous media: a pore-scale study. Energy Geosci 5 (1), 100222. https://doi.org/10.1016/j.engeos.2023.100222.

Wang, D., Seright, R.S., 2021. Examination of literature on colloidal dispersion gels for oil recovery. Petrol. Sci. 18 (4), 1097–1114. https://doi.org/10.1016/j.petsci.2021.07.009.

Wang, F., Xu, H., Liu, Y., et al., 2023. Mechanism of low chemical agent adsorption by high pressure for hydraulic fracturing-assisted oil displacement technology: a study of molecular dynamics combined with laboratory experiments. Langmuir 39 (46), 16628–16636. https://doi.org/10.1021/acs.langmuir.3c02634.

Wang, X., Yang, S., Zhao, Y., et al., 2018. Improved pore structure prediction based on MICP with a data mining and machine learning system approach in Mesozoic strata of Gaoqing field, Jiyang depression. J. Petrol. Sci. Eng. 171, 362–393. https://doi.org/10.1016/j.petrol.2018.07.057.

Wang, Y., Ayala, L.F., 2020. Explicit determination of reserves for variable bottom-hole pressure conditions in gas well decline analysis. SPE J. 25 (1), 369–390. https://doi.org/10.2118/195691-PA.

Wang, Y., Cheng, S., Wei, C., et al., 2021a. Gas rate decline analysis for boundary-dominated flow with fractal reservoir properties under constant/variable bottom-hole pressure conditions. J. Nat. Gas Sci. Eng. 88, 103823. https://doi.org/10.1016/j.jngse.2021.103823.

Wang, Y., Cheng, S., Zhang, F., et al., 2021b. Big data technique in the reservoir parameters' prediction and productivity evaluation: a field case in western South China Sea. Gondwana Res. 96, 22–36. https://doi.org/10.1016/j.gr.2021.03.015.

Wei, C., Liu, Y., Deng, Y., , et al.Hassanzadeh, H., 2022. Temperature transient analysis of naturally fractured geothermal reservoirs. SPE J. 27 (5), 2723–2745. https://doi.org/10.2118/205862-PA.

Werneck, R.D.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.

Wood, D.A., 2020. Predicting porosity, permeability and water saturation applying an optimized nearest-neighbour, machine-learning and data-mining network of well-log data. J. Petrol. Sci. Eng. 184, 106587. https://doi.org/10.1016/j.petrol.2019.106587.

Yuan, Z., Qin, W., Zhao, J., 2017. Smart manufacturing for the oil refining and petrochemical industry. Engineering 3 (2), 179–182. https://doi.org/10.1016/J.ENG.2017.02.012.

Zhou, Q., Dilmore, R., Kleit, A., et al., 2014. Evaluating gas production performance in Marcellus using data mining technologies. J. Nat. Gas Sci. Eng. 20, 109–120. https://doi.org/10.1016/j.jngse.2014.06.014.