Hindawi
Geofluids
Volume 2021, Article ID 6515846, 12 pages
https://doi.org/10.1155/2021/6515846

Research Article

Quantitative Analysis of the Main Controlling Factors of Oil Saturation Variation

Ruijie Huang,[1] Chenji Wei,[1] Jian Yang,[1] Xin Xu,[2] Baozhu Li,[1] Suwei Wu,[1] and Lihui Xiong[1]

[1]*Research Institute of Petroleum Exploration and Development, Beijing 100083, China*
[2]*Bytedance Inc., Hangzhou 310000, China*

Correspondence should be addressed to Chenji Wei; weichenji@petrochina.com.cn

Received 6 October 2021; Revised 28 October 2021; Accepted 11 November 2021; Published 9 December 2021

Academic Editor: Afshin Davarpanah

With the high-speed development of artificial intelligence, machine learning methods have become key technologies for intelligent exploration, development, and production in oil and gas fields. This article presents a workflow analysing the main controlling factors of oil saturation variation utilizing machine learning algorithms based on static and dynamic data from actual reservoirs. The dataset in this study generated from 468 wells includes thickness, permeability, porosity, net-to-gross (NTG) ratio, oil production variation (OPV), water production variation (WPV), water cut variation (WCV), neighbouring liquid production variation (NLPV), neighbouring water injection variation (NWIV), and oil saturation variation (OSV). A data processing workflow has been implemented to replace outliers and to increase model accuracy. A total of 10 machine learning algorithms are tested and compared in the dataset. Random forest (RF) and gradient boosting (GBT) are optimal and selected to conduct quantitative analysis of the main controlling factors. Analysis results show that NWIV is the variable with the highest degree of impact on OSV; impact factor is 0.276. Optimization measures are proposed for the development of this kind of sandstone reservoir based on main controlling factor analysis. This study proposes a reference case for oil saturation quantitative analysis based on machine learning methods that will help reservoir engineers make better decision.

1. Introduction

The variation of oil saturation is a matter of long-standing concern to reservoir engineers. At the middle-to-late development stage, complex geological characteristics and different development scenarios make the oil saturation variation more complex. Clarifying the main controlling factors influencing oil saturation is essential to optimize development plans.

There are three main conventional methods for measuring, analysing, and predicting oil saturation. The first category is the material balance equation method, which estimates average oil saturation of an entire reservoir based on reservoir geological data and development and production data [1]. Deng et al. developed an improved time-differentiated variable multiple material balance model to evaluate residual oil saturation in water-flooded zones by differentiating the water flooding into numerous displacement processes. This method's calculations show that the residual oil saturation is in excellent agreement with the experimental results of the core analysis [2]. Shahamat and Clarkson discussed the application of conventional flowing material balance (FMB) to the analysis of single-phase or multiphase flow in single or multiwell scenarios and proposed a new, comprehensive FMB. The developed FMB can be used to determine the original volumes of hydrocarbons in place in both conventional and unconventional reservoirs [3]. Rahman et al. proposed a new, rigorous material balance equation for gas flow in the presence of a compressible formation and residual fluid saturation [4].

The second category is core analysis and logging analysis. Core analysis is a laboratory method for the direct measurement of oil saturation. Based on the coring tool, core analysis is classified as conventional coring [5, 6], pressure coring [7, 8], and sponge coring [9, 10]. Zhang et al. sampled three-meter-scale core intervals and calculated oil saturation index by X-ray diffraction analysis, TOC analysis, and programmed pyrolysis analysis [11]. Xiao et al. systematically studied the geological and geochemical characteristics of the First Member of the Cretaceous Qingshankou Formation in the Qijia Sag based on core samples and core analysis [12]. Logging is the most widely used method for obtaining reliable oil saturation in oil fields [13], especially pulsed neutron logging [14]. Dong et al. studied the detection principles, modes, advantages, and disadvantages of pulsed neutron logging tools, established a formation model based on the Monte Carlo method, and analysed the sensitivity of detection [14]. Nie et al. introduced a novel oil content model for shale oil reservoirs by analysing the logging and core experimental data and building the relationship between kerogen and the different well logging porosities including nuclear magnetic resonance (NMR) porosity, neutron porosity, and density porosity [15].

The third category is reservoir numerical simulation, which simulates the development process and calculates key parameters (production, injection, saturation, etc.) based on a geological model. Ren and Duncan utilized a commercial reservoir simulator to simulate CO_2-enhanced oil recovery (CO_2-EOR). These simulations explore the effects of strength of aquifer flow, flow direction, and capillary pressure on the nature and distribution of oil saturation in residual oil zones (ROZs) [16]. Ren and Duncan explored the impact of various elements: oil saturation, well patterns, reservoir heterogeneity, and permeability anisotropy based on simulations [17].

In recent years, the world has entered the era of "Industry 4.0", and so has the oil and gas development industry. Machine learning is also widely known for its application in extracting of complex patterns from massive data. Machine learning is widely applied to the upstream oil industry, such as for production prediction and optimization [18–20], geological modelling, managing uncertainty [21, 22], and for characterization of connectivity and heterogeneity [23, 24]. Wang et al. developed a novel equal probability gene expression programming (EP-GEP) method to analyse the production decline of carbonate reservoirs. Validation and comparison showed that this method outperformed the traditional Arps model in perdition accuracy [18]. Liu et al. proposed a well production performance prediction method based on an artificial neural network. This method can help engineers analyse the massive data from unconventional reservoirs to understand the underlying patterns and relationships, especially on enhanced oil recovery (EOR) pilot projects [19]. Niu et al. established a multiparameter comprehensive intelligent prediction method of karst curtain grouting volume (KCGV) based on support vector machine (SVM). The application results show that this method achieves excellent prediction performance on the KCGV and can provide practical and beneficial help for the field karst curtain grouting project [20]. Kang et al. established a classification model for determining the proper geological scenario among plausible geological uncertainties by utilizing machine learning methods including support vector machine (SVM), artificial neural network (ANN), and convolutional neural network (CNN). The results show that this method generates more reliable reservoir models and successfully reduces the uncertainty in the geological scenario [21]. Du et al. utilized deep transfer learning (DTL) to extract features from a training image (TI) of porous media to replace the process of scanning a TI for different patterns as in multiple-point statistical methods. The experimental results show that the proposed method is of high efficiency while preserving similar features with the target image, shortening reconstruction time [22]. Song et al. developed a novel prediction model to forecast vertical heterogeneity of the reservoir based on various deep neural network algorithms. The machine learning models have the ability to learn and capture hidden relationships between dynamic production data and reservoir heterogeneity. The application results show that the proposed models achieve excellent performance in predicting heterogeneity [23]. Liu developed a machine learning method to evaluate the connectivity between injectors and producers based on back propagation (BP) algorithms and convolutional neural networks (CNNs) algorithms in interlayer reservoirs. The model training dataset consists of dynamic production data under different permeability, interlayer dip angle, and injection pressure. The results show that CNN has better prediction performance than BP [24]. Huang et al. developed long short-term memory (LSTM) neural network model to forecast well performance [25].

Previous studies have looked at the distribution of remaining oil saturation and the influencing factors [26–30]. These studies mainly focused on static characteristics of the reservoir, such as microscopic pore structure and heterogeneity. However, they failed to combine static geological data with dynamic production data, to conduct quantitative analysis, and to calculate main controlling factors.

In this study, machine learning algorithms are employed in the quantitative analysis of main controlling factors of oil saturation variation during the water flooding process of a Middle East reservoir. This study is divided into the following sections. In Section 2, we briefly describe the geological characteristics, development history, well layout, and dynamic production performance of the research reservoir. In Section 3, we propose the workflow of this study. We introduce detailed data gathering and processing, which is significant and the basis of the model training. We screen a total of 10 machine learning algorithms suitable for the analysis and present their basic principles. In addition, model evaluation method is introduced in this section. In Section 4, we obtain the optimal algorithms by comparison and apply them to calculate main controlling factors affecting oil saturation variation. Finally, the discussion and conclusion appear in Section 5.

2. Research Area

Research area (RM reservoir) is a long-axis anticline reservoir located in the Middle East with a NW-SE trend, and its main lithology is sandstone. The sedimentary environment is shallow sea continental slope deposition, open sea, front reef, and lagoon environment. The RM reservoir developed good reservoir continuity, and the physical properties of the north part are better than those of the south part. Core analysis results show a good porosity permeability relationship, with an average porosity range of 10-15% and an average permeability range of 40-300 md. The formation thickness of RM reservoir ranges from 10 to 60 m, and effective section

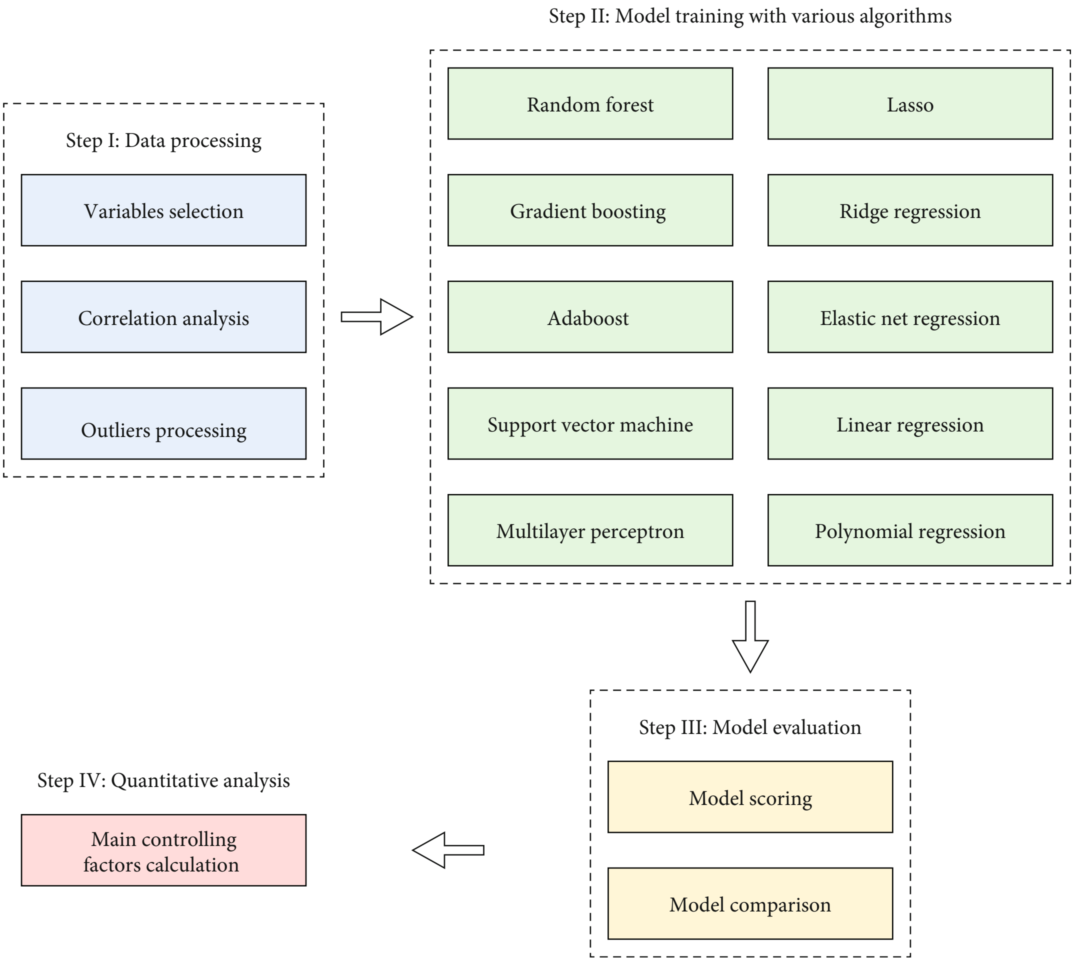

Figure 1: Workflow of the proposed method.

thickness ranges from 6 to 40 m of single well. The target reservoir has been developed by primary depletion for almost 40 years; in 2009, its daily oil production was approximately 200,000 barrels relying on 127 open producers. Water flooding was implemented in this reservoir after 2010 when formation pressure decreased rapidly, resulting in the closure of a large number of producers and declining production. Current reservoir production is approximately 150,000 barrels per day. Long periods of production and development have accumulated a large amount of valuable surveillance and test data, providing a solid foundation for the application of machine learning algorithms. The current development of this reservoir is facing difficulties in evaluating the effect of water flooding and optimizing development measures.

3. Methodology

3.1. Workflow. The specific workflow of this research includes 4 steps as shown in Figure 1. The first step is data processing including variable selection, correlation analysis, and outlier processing. Training models by utilizing different algorithms is the second step. The next step is model evaluation including scoring and comparing different algorithm models. The final step is quantitatively calculating the main controlling factors of oil saturation variation.

3.2. Data Collection and Processing. The most crucial component in machine learning analysis is the database, which is the fundamental of modelling. With the help of data collection and processing, a large number of useful data can be gathered and one can gain meaningful insights from the relationship between different valuables. In this process, inappropriate variables are screened out, outliers are removed, missing values are inserted, and an integrated dataset is established for model training.

The factors that affect oil saturation involve geology, oil reservoirs, engineering, and others, as shown in Figure 2. The dataset established and utilized in this study contains original data from 468 wells representing the actual situation of target reservoir development.

Static geological data include formation thickness, permeability, porosity, and net-to-gross (NTG) ratio. The detailed data come from seismic interpretation, logging interpretation reports, core analysis, and studies. Dynamic production data includes oil production variation (OPV), water production variation (WPV), water cut variation (WCV), neighbouring liquid production variation (NLPV), and neighbouring water injection variation (NWIV), which comes from daily and monthly monitoring. In order to better analyse the impact factors of oil saturation, we also considered the impact of injection and production of neighbouring wells. We define

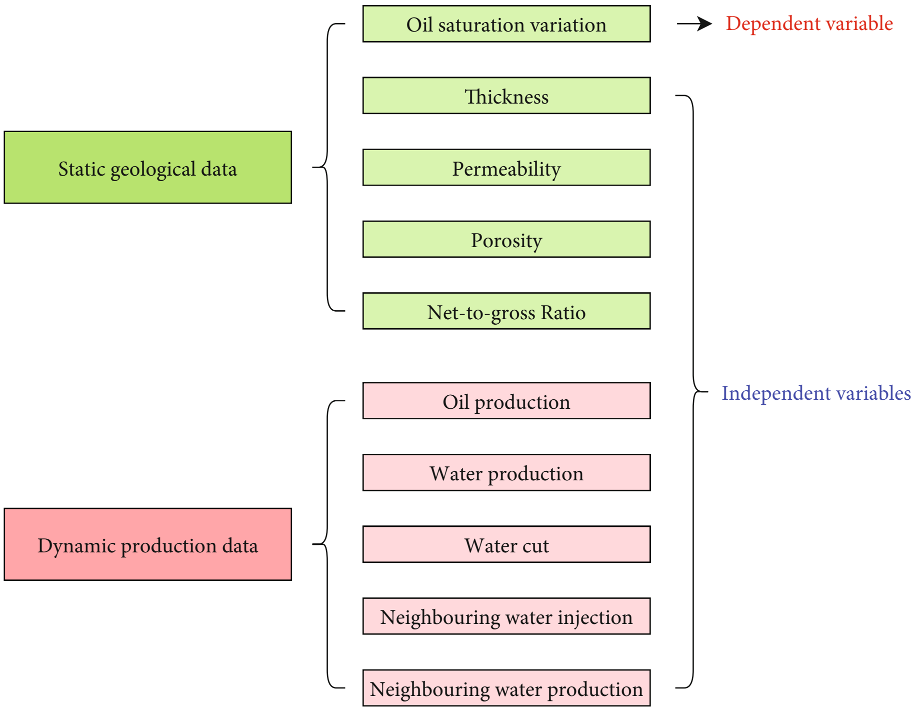

Figure 2: Diagram of neighbouring wells and neighbouring relationship.

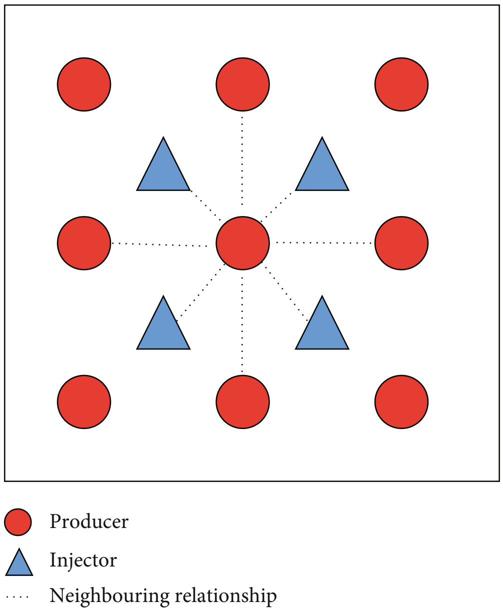

Figure 3: Variables screened for model training.

the neighbouring relationship as two wells are directly adjacent to each other without any well in between. The diagrams of neighbouring wells and neighbouring relationships are shown in Figure 3. Oil saturation variation (OSV) as the dependent variable is also the key parameter in this study. All saturation data comes from production logging test (PLT) and reservoir saturation test (RST) reports. We use the variation of dynamic data instead of rate, because variation can reflect the production performance over a period of time. The variation of dynamic data in this study is equal to the difference between the two saturation tests.

In this study, the Pearson coefficient was employed to analyse correlation between variables; the calculation method is shown in Equation (1). The correlation coefficient varies from

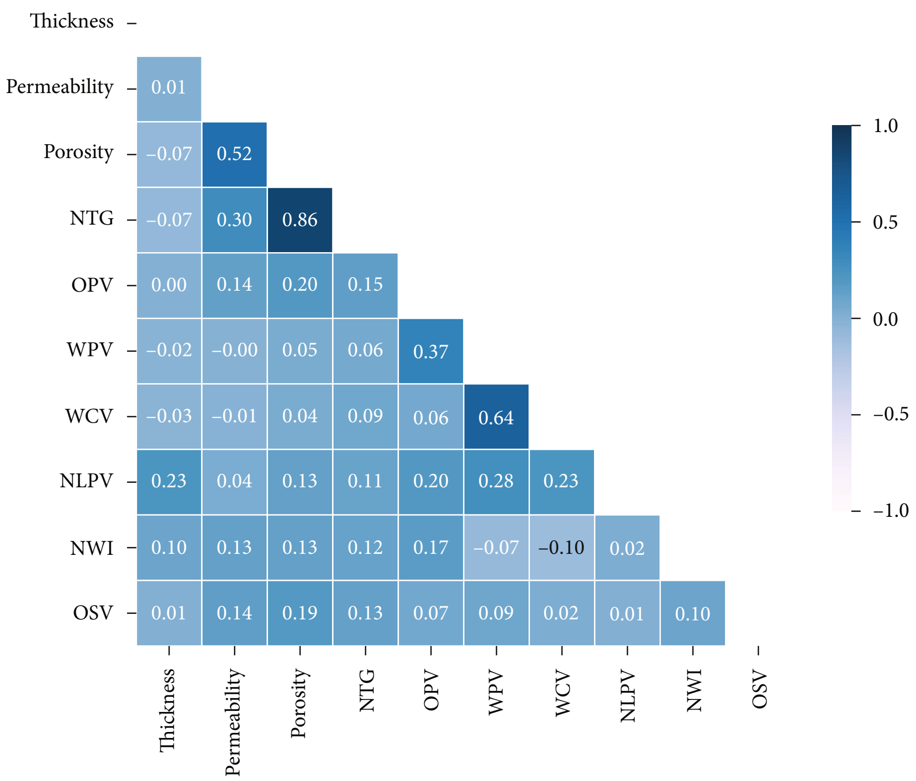

Figure 4: Variables' correlation heat map.

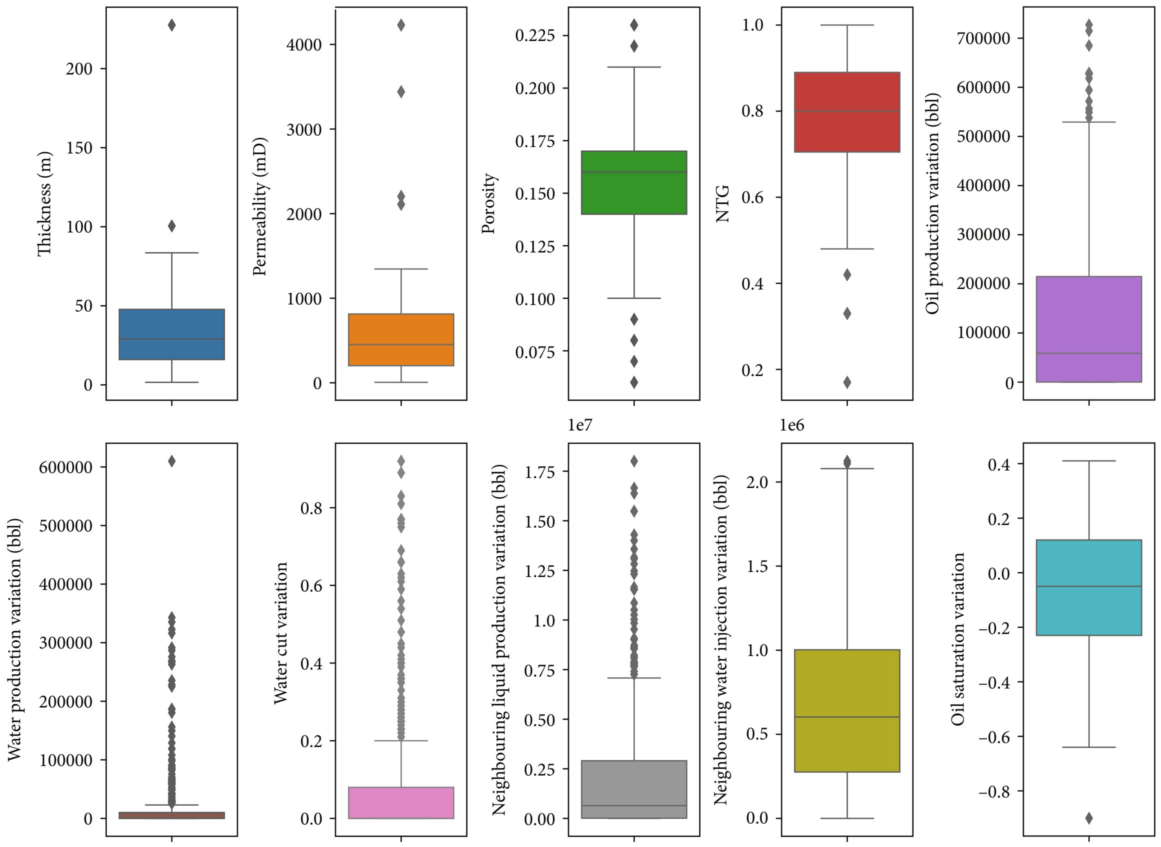

Figure 5: Outlier analysis using boxplot.

-1 (negative correlation) to 1 (positive correlation). When the correlation coefficient is 0, it means that there is no correlation between the two variables.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}, \tag{1}$$

where $\rho_{X,Y}$ is the correlation coefficient of a pair of random variables (X, Y), cov is the covariance, σ_X is the standard deviation of X, and σ_Y is the standard deviation of Y.

The heat map of correlation analysis results is shown in Figure 4. In this reservoir, porosity has a moderate correlation with permeability and a strong correlation with NTG, which indicates that the higher the NTG, the higher the porosity, and the better reservoir properties. Water cut variation is in strong correlation with water production variation, which is in line with the physical law of reservoir development. In addition, the correlation between oil saturation variation and other variables is very weak, so it is impossible to establish a simple linear model to analyse the implicit relationship between them.

The existence of outliers seriously affects the performance of machine learning models, so proper handling of outliers is essential in enhancing model accuracy. Boxplot is a graphical method for depicting the distribution of data through quartiles and averages. Boxplots show the five-number summary of a dataset: minimum, maximum, first quartile (Q_1), third quartile (Q_3), and median. The boxplot method defines an index, interquartile range (IQR) to demarcate outliers; the calculation is shown in Equations (2) and (3). The evaluation index of the model accuracy is the coefficient of determination (R^2), and the calculation formula is shown in Equation (4).

$$\mathrm{IQR} = Q_3 - Q_1, \tag{2}$$

$$\text{Outlier} = \begin{cases} < Q_1 - 1.5 * \mathrm{IQR}, \\ \text{or} \\ > Q_3 + 1.5 * \mathrm{IQR}, \end{cases} \tag{3}$$

$$R^2 = \frac{\sum_{i=1}^{n} (y\wedge_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \tag{4}$$

where y_i and $\hat{y}_i$ are the actual data and prediction. $\bar{y}$ is the average of data.

The overview and outlier distribution of this dataset is shown in Figure 5. The three variables of water production variation (WPV), water cut variation (WCV), and neighbouring liquid production variation (NLPV) have a relatively high proportion of outliers. There are four methods to deal with outliers: Ignore, Mean, Median, and Delete. "Ignore" means to ignore outliers without any processing. "Mean" and "Median" mean to replace outliers with either mean values or median values. "Delete" means removing outlier data points. The four processing methods are tested on the three algorithms of RF, GBT, and ADBT, and the results are shown in Table 1. Using the median values to replace outliers obtains the highest training and testing R^2.

Table 1: R^2 of RF model under different outlier processing methods.

Algorithms	Scoring	Ignore	Mean	Median	Delete
RF	Train R^2	0.96	0.95	0.97	0.96
	Test R^2	0.57	0.51	0.72	0.68
GBT	Train R^2	0.93	0.92	0.93	0.92
	Test R^2	0.60	0.46	0.71	0.65
ADBT	Train R^2	0.78	0.77	0.79	0.72
	Test R^2	0.67	0.46	0.71	0.69

3.3. Machine Learning Algorithms. The Pearson coefficient can only simply analyse the linear effect of a single variable on the oil saturation variation. In the process of reservoir development, the oil saturation variation is affected by multivariable nonlinearity. In order to clarify the contribution of each variable on oil saturation variation, machine learning methods are employed to produce quantitative analysis.

In this study, a total of 10 machine learning algorithms are selected and used to developed regression models based on dataset. These regression models can be used to analyse hidden relationships between the dependent variable "oil saturation variation" and the independent variables and to quantify feature variable importance (impact factor). These 10 algorithms include ensemble algorithms (random forest, AdaBoost, and gradient boost), linear regression algorithms (linear regression, polynomial regression, Lasso, ridge regression, and elastic net regression), support vector machine, and multilayer perceptron.

Random forest (RF) is an ensemble learning algorithm used for classification and regression and is composed of decision trees [31–33]. In the classification or regression process, the output of RF is based on the output value of its internal decision trees. RF outperforms decision tree due to the ability to avoid overfitting. RF has the following advantages:

(i) It is unsurpassed in accuracy among current algorithms

(ii) It can process large datasets and thousands of input variables efficiently without dimensionality reduction

(iii) It can handle datasets with missing data and maintains accuracy when a large proportion of the data is missing

(iv) It can calculate estimates of variable importance in the classification

Adaptive boosting or AdaBoost (ADBT) is an iterative algorithm that trains different weak classifiers for the same training dataset and then combines these weak classifiers to form a stronger final classifier [34, 35]. ADBT classifier has high accuracy and is not prone to overfitting.

Gradient boosting (GBT) is a classification and regression method based on boosting technique [36, 37]. The core idea of GBT is to train the newly added weak classifier based on the negative gradient information of the current model loss function and then integrate the trained weak classifier into the existing model in an accumulated form. In each iteration, GBT first calculates the negative gradient of the current model on all samples, and then trains a new weak classifier to fit the negative gradient with this value, and calculates the weight of the weak classifier to update the model. Compared with AdaBoost which only uses exponential loss function, GBT can use any differentiable loss function, so GBT can be applied from binary classification to regression and multiclass classification.

Least absolute shrinkage and selection operator (Lasso) is a regression analysis method that performs feature selection and regularization at the same time to enhance accuracy and interpretability [38–40]. Lasso uses L1 regularization, which will make some learned feature weights 0, so as to achieve the purpose of sparseness and feature selection.

Ridge regression is a biased estimation regression method dedicated to collinearity data analysis [41, 42]. It is essentially an improved least squares estimation method. Ridge regression obtains more realistic regression coefficients by abandoning the lack of bias of the least square method and at the cost of losing part of the information and reducing the accuracy. The fitting of ridge regression to abnormal data is stronger than the least square method.

Elastic net regression (ENR) is a hybrid of ridge regression and Lasso. ENR is a linear regression model trained using L1 and L2 regularization as a priori regularization term [43, 44].

Support vector machine (SVM) is a supervised learning algorithm used for analysing data, classification, and regression. SVM establishes a hyperplane in a high-dimensional space to make a good separation which has the largest distance to the nearest data point [45–47].

Multilayer perceptron (MLP) is a multilayer fully connected feedforward neural network. The basic structure of MLP consists of three layers: the first input layer, the middle hidden layer, and the last output layer. Each node is a neuron that uses a nonlinear activation function. This enables MLP to process complex linear inseparable data [48, 49].

Linear regression (LR) is a regression analysis that uses the least square function to model the relationship between one or more independent variables and dependent variables [50].

Polynomial regression (PR) is considered a special form of multiple linear regression. PR modelled a nonlinear relationship between the independent variable x and the dependent variable y as an nth degree polynomial in x [51].

4. Results

4.1. Accuracy Comparison of Different Algorithms. The 10 algorithms mentioned above have been trained and tested on the same processed dataset. The best one can be screened out through the horizontal comparison of multiple algorithms. Train and test performances of 10 algorithms are shown in Figure 6, the detailed values are shown in Table 2, and the crossplot of the actual value and the model prediction value are shown in Figure 7.

Random forest (RF) and gradient boosting (GBT) are significantly better than the other algorithms. These two accurately capture the features of the dataset and fit it well. AdaBoost (ADBT) performs well in test processing, but obtains low fitting in model training. Support vector machine (SVM), multilayer perceptron (MLP), linear regression (LR), and least absolute shrinkage and selection operator (Lasso) rank in the middle. The fitting effect of polynomial regression (PR) is very poor, and the crossplot presents a divergent shape, as shown in Figure 7. Elastic net regression (ENR) and ridge regression are completely inapplicable to this dataset.

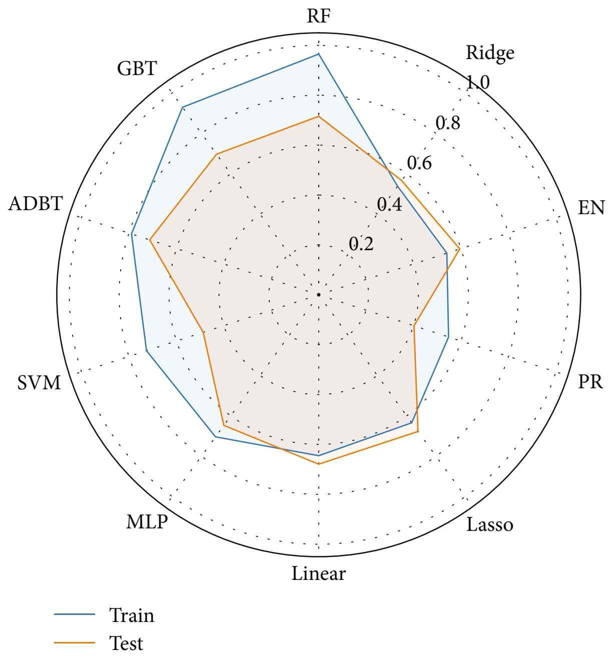

Figure 6: $R^2$of all algorithms.

Table 2: R^2 of all algorithms' train and test.

	RF	GBT	ADBT	SVM	MLP	Linear	Lasso	PR	EN	Ridge
Train R^2	0.97	0.93	0.79	0.73	0.70	0.65	0.63	0.55	0.54	0.54
Test R^2	0.72	0.70	0.71	0.49	0.65	0.68	0.68	0.40	0.59	0.57

In comparison, more advanced ensemble learning algorithms defeated the traditional regression algorithms. RF and GBT are utilized to quantitatively analyse the main controlling factors of oil saturation variation.

4.2. Main Controlling Factors. RF and GBT are used to calculate the impact factor of each variable on OSV, as shown in Figure 8. "Average" is the mean of the results of RF and BGT. Neighbouring water injection effect is the variable with the highest degree of impact on OSV, which is consistent with the development dynamics of the target reservoir after 2010, as shown in Figure 9. It shows the positive effect of water flooding in this reservoir.

Permeability and oil production rank second and third, respectively, in importance, which is also in line with the

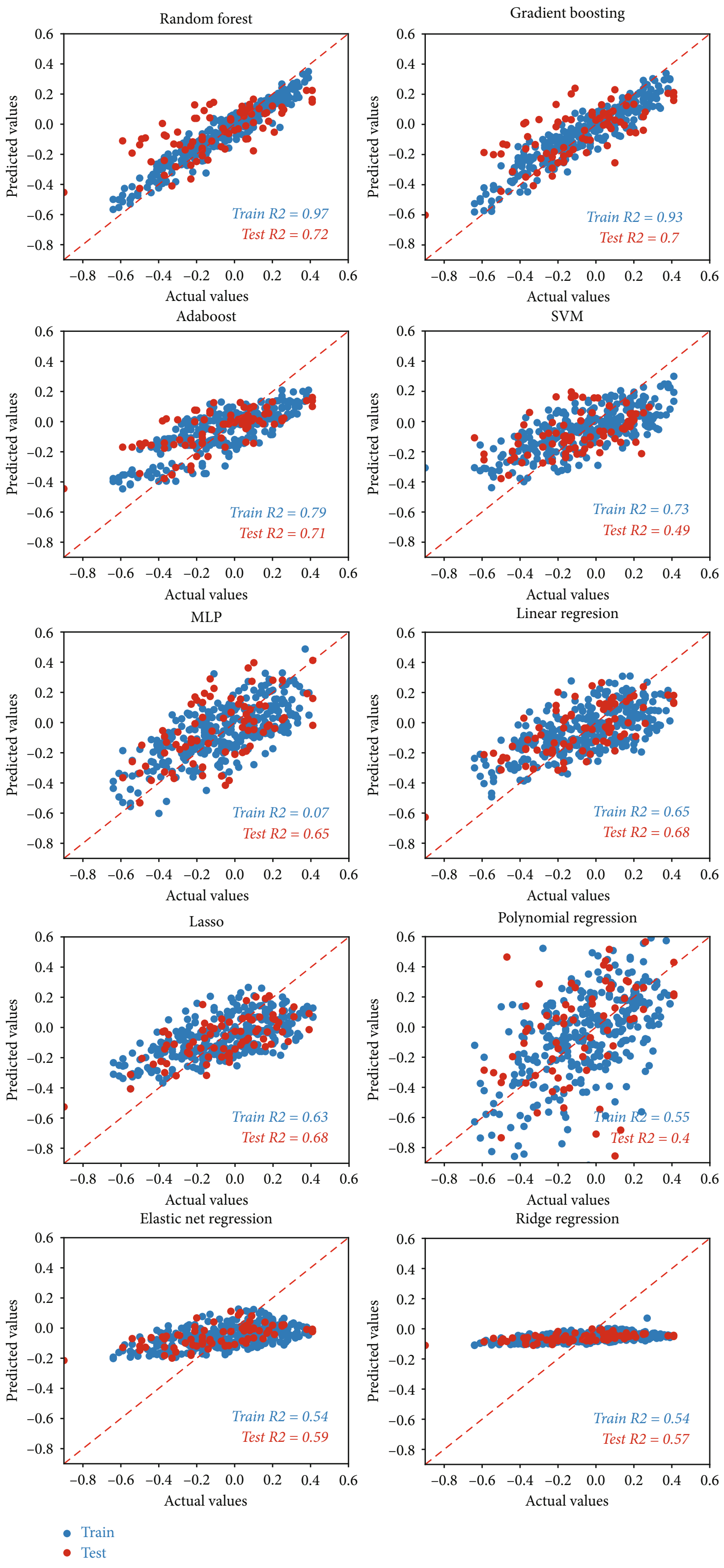

Figure 7: Crossplot between actual values and predicted values.

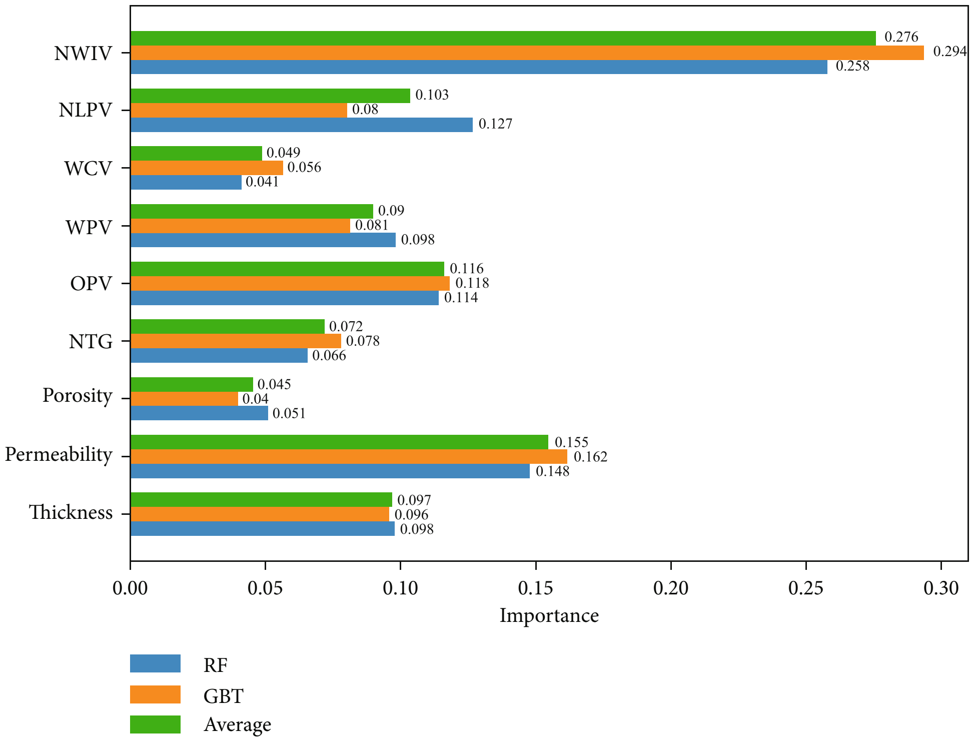

Figure 8: Importance of impact factor.

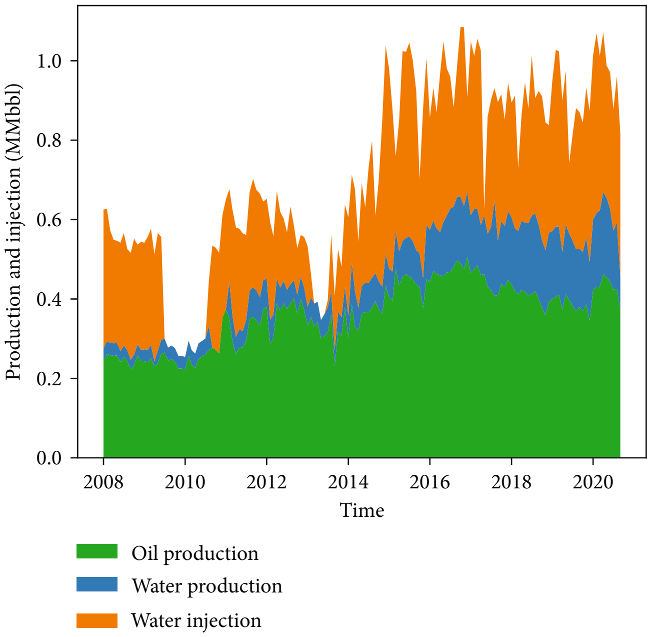

Figure 9: Reservoir development performance after water flooding implementation.

law of reservoir development. The neighbouring liquid production variation ranks fourth, indicating that complete injection-production well patterns are essential to improve oil recovery and tap the potential of remaining oil.

Based on the analysis results of the main controlling factors of oil saturation, we propose optimization measures for the development of this kind of sandstone reservoir.

(i) Continue to implement water flooding development, establish effective displacement system, and closely monitor the performance of producers. Reservoir saturation test (RST) and production logging test (PLT) need to be conducted to monitor oil saturation variation and prevent water breakthrough

(ii) In areas with high permeability, optimize the injection-production well patterns by infilling wells and others to tap the potential of remaining oil

(iii) In areas with low permeability or poor properties, reservoir reconstruction measures such as hydraulic fracturing and EOR measures such as low-salinity water flooding can be applied

5. Discussion and Conclusion

Machine learning is a data-driven analysis method. It can process massive data and clarify hidden relationships between variables. The two traditional methods fail to quantitatively calculate the impact factor of the influence that affects oil saturation. The main advantage of this research is that all data used by machine learning analysis come from actual sandstone reservoir. Machine learning has been widely used in the field of oil and gas development, but not all algorithms are perfectly applicable. The purpose of this research is to select suitable algorithms through comparing and testing 10 different machine learning algorithms, make full use of real oil field data, and conduct quantitative analysis of the main controlling factors of oil saturation. This research can provide strong support for the further research that characterizes oil saturation distribution for the whole reservoir by developing neural network model.

This article proposes a method for analysing the main controlling factors of oil saturation variation. Actual static geological data and dynamic production data are gathered to establish machine learning analysis models. The specific conclusions are as follows.

(1) Established a data processing workflow, including correlation analysis and outlier processing. The median method was the most successful outlier processing method in this study

(2) In comparison and testing of 10 machine learning algorithms, RF and GBT were the optimal algorithms and obtained the highest accuracy in modelling

(3) The oil saturation analysis model was established using RF and GBT algorithms, and the main controlling factors were quantitatively calculated. NWIV is the most important factor, with an impact factor of 0.276

(4) The ranking of the variables provides the basis of the proposal for optimizing reservoir development. The workflow is also an advanced and complex data analysis method, which provides a foundation for the subsequent establishment of a neural network saturation prediction model

(5) Continue to implement water flooding development, establish effective displacement system, and closely monitor the performance of wells to prevent water breakthrough

Data Availability

The manuscript is a self-contained data article; the entire data used to support the findings of this study are included within the article. If any additional information is required, this is available from the corresponding author upon request to weichenji@petrochina.com.cn.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors thank the support from the National Natural Science Foundation of China (Grant No. 51974357).

References

[1] M. M. Chang, N. L. Maerefat, L. Tomutsa, and M. M. Honarpour, "Evaluation and comparison of residual oil saturation determination techniques," *SPE Formation Evaluation*, vol. 3, no. 1, pp. 251–262, 1988.

[2] Z. Deng, L. Ding, H. Zhang, W. Tan, and W. Yuan, "Assessment of residual oil saturation with time-differentiated variable multiple material balance model," *Energy Geoscience*, vol. 3, no. 1, pp. 1–7, 2022.

[3] M. S. Shahamat and C. R. Clarkson, "Multiwell, multiphase flowing material balance," *SPE Reservoir Evaluation & Engineering*, vol. 21, no. 2, pp. 445–461, 2018.

[4] A. Rahman, S. Ahmed, M. E. Hossain, and F. A. Happy, "Development of scaling criteria for steam flooding EOR process," *Journal of Petroleum Exploration and Production Technologies*, vol. 10, no. 8, pp. 3849–3863, 2020.

[5] J. A. Shirer, E. P. Langston, and R. B. Strong, "Application of field-wide conventional coring in the Jay-Little Escambia Creek unit," *Journal of Petroleum Technology*, vol. 30, no. 12, pp. 1774–1780, 1978.

[6] S. Li, V. Vaziri, M. Kapitaniak, J. M. Millett, and M. Wiercigroch, "Application of resonance enhanced drilling to coring," *Journal of Petroleum Science and Engineering*, vol. 188, article 106866, 2020.

[7] Y. Kubo, Y. Mizuguchi, F. Inagaki, and K. Yamamoto, "A new hybrid pressure-coring system for the drilling vessel *Chikyu*," *Scientific Drilling*, vol. 17, pp. 37–43, 2014.

[8] N. Inada and K. Yamamoto, "Data report: hybrid pressure coring system tool review and summary of recovery result from gas-hydrate related coring in the Nankai project," *Marine and Petroleum Geology*, vol. 66, pp. 323–345, 2015.

[9] A. Park and C. A. Devier, *Improved Oil Saturation Data Using Sponge Core Barrel SPE Production Operations Symposium SPE Production Operations Symposium*, Society of Petroleum Engineers, Oklahoma City, Oklahoma, 1983.

[10] L. Shale, S. Radford, T. Uhlenberg, J. Rylance, A. Kvinnesland, and C. Rengel, *New Sponge Liner Coring System Records Step-Change Improvement in Core Acquisition and Accurate Fluid Recovery SPE/EAGE European Unconventional Resources Conference and Exhibition vol 2014*, European Association of Geoscientists & Engineers, 2014.

[11] T. Zhang, Q. Fu, X. Sun, P. C. Hackley, L. T. Ko, and D. Shao, "Meter-scale lithofacies cycle and controls on variations in oil saturation, Wolfcamp A, Delaware and Midland Basins," *American Association of Petroleum Geologists Bulletin*, vol. 105, no. 9, pp. 1821–1846, 2021.

[12] F. Xiao, J. Yang, S. Li et al., "Geological and geochemical characteristics of the first member of the Cretaceous Qingshankou Formation in the Qijia Sag, Northern Songliao Basin, Northeast China: implication for its shale oil enrichment," *Geofluids*, vol. 2021, Article ID e9989792, 20 pages, 2021.

[13] F. E. R. T. L. WH and C. H. I. L. I. N. G. A. R. I. A. N. GV, "Determination of residual oil saturation from well logs," *Energy Sources*, vol. 10, no. 2, pp. 95–101, 1988.

[14] J. Dong, R. Deng, Z. Quanying, J. Cai, Y. Ding, and M. Li, "Research on recognition of gas saturation in sandstone reservoir based on capture mode," *Applied Radiation and Isotopes*, vol. 178, article 109939, 2021.

[15] X. Nie, J. Lu, J. Chi, P. Wang, and C. Zhang, "Oil content prediction method based on the TOC and porosity of organic-rich shales from wireline logs: a case study of lacustrine intersalt shale plays in Qianjiang Sag, Jianghan Basin, China," *Geofluids*, vol. 2021, Article ID e9989866, 8 pages, 2021.

[16] B. Ren and I. Duncan, "Modeling oil saturation evolution in residual oil zones: implications for CO_2 EOR and sequestration," *Journal of Petroleum Science and Engineering*, vol. 177, p. 528, 2019.

[17] B. Ren and I. J. Duncan, "Maximizing oil production from water alternating gas (CO2) injection into residual oil zones: the impact of oil saturation and heterogeneity," *Energy*, vol. 222, p. 119915, 2021.

[18] L. Wang, M. Shao, G. Kou et al., "Time series analysis of production decline in carbonate reservoirs with machine learning," *Geofluids*, vol. 2021, Article ID e6638135, 8 pages, 2021.

[19] K. Liu, B. Xu, C. Kim, and J. Fu, "Well performance from numerical methods to machine learning approach: applications in multiple fractured shale reservoirs," *Geofluids*, vol. 2021, Article ID e3169456, 13 pages, 2021.

[20] J. Niu, B. Wang, H. Wang et al., "An intelligent prediction method of the karst curtain grouting volume based on support vector machine," *Geofluids*, vol. 2020, Article ID e8892106, 14 pages, 2020.

[21] B. Kang and K. Lee, "Managing uncertainty in geological scenarios using machine learning-based classification model on production data," *Geofluids*, vol. 2020, Article ID e8892556, 16 pages, 2020.

[22] Y. Du, J. Chen, and T. Zhang, "Reconstruction of three-dimensional porous media using deep transfer learning," *Geofluids*, vol. 2020, Article ID e6641642, 22 pages, 2020.

[23] H. Song, S. Du, R. Wang et al., "Potential for vertical heterogeneity prediction in reservoir basing on machine learning methods," *Geofluids*, vol. 2020, Article ID e3713525, 12 pages, 2020.

[24] J. Liu, "Potential for evaluation of interwell connectivity under the effect of intraformational bed in reservoirs utilizing machine learning methods," *Geofluids*, vol. 2020, Article ID e1651549, 10 pages, 2020.

[25] R. Huang, C. Wei, B. Wang et al., "Well performance prediction based on long short-term memory (LSTM) neural network," *Journal of Petroleum Science and Engineering*, vol. 208, p. 109686, 2022.

[26] Z. Fan, K. Li, J. Li, H. Song, L. He, and X. Wu, "A study on remaining oil distribution in a carbonate oil reservoir based on reservoir flow units," *Petroleum Exploration and Development*, vol. 41, no. 5, pp. 634–641, 2014.

[27] S. Zheng, M. Yang, Z. Kang et al., "Controlling factors of remaining oil distribution after water flooding and enhanced oil recovery methods for fracture-cavity carbonate reservoirs in Tahe oilfield," *Petroleum Exploration and Development*, vol. 46, no. 4, pp. 786–795, 2019.

[28] J. Li, Y. Liu, Y. Gao, B. Cheng, F. Meng, and H. Xu, "Effects of microscopic pore structure heterogeneity on the distribution and morphology of remaining oil," *Petroleum Exploration and Development*, vol. 45, no. 6, pp. 1112–1122, 2018.

[29] M. Yue, W. Zhu, H. Han, H. Song, Y. Long, and Y. Lou, "Experimental research on remaining oil distribution and recovery performances after nano-micron polymer particles injection by direct visualization," *Fuel*, vol. 212, pp. 506–514, 2018.

[30] J. Zhang, F. Fang, J. Wang et al., "Prediction of intraformational remaining oil distribution based on reservoir heterogeneity: application to the J-field," *Advances in Civil Engineering*, vol. 2021, Article ID e8870274, 10 pages, 2021.

[31] T. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[32] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[33] R. Genuer, J.-M. Poggi, C. Tuleau-Malot, and N. Villa-Vialaneix, "Random forests for big data," *Big Data Research*, vol. 9, pp. 28–46, 2017.

[34] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[35] W. Gao and Z.-H. Zhou, "On the doubt about margin explanation of boosting," *Artificial Intelligence*, vol. 203, pp. 1–18, 2013.

[36] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[37] T. Chen and C. Guestrin, *XGBoost: A Scalable Tree Boosting System Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD '16: the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, San Francisco California USA, 2016.

[38] F. Santosa and W. W. Symes, "Linear inversion of band-limited reflection seismograms," *SIAM Journal on Scientific Computing*, vol. 7, no. 4, pp. 1307–1330, 1986.

[39] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, 1996.

[40] R. Tibshirani, "The Lasso method for variable selection in the Cox model," *Statistics in Medicine*, vol. 16, no. 4, pp. 385–395, 1997.

[41] A. E. Hoerl and R. W. Kennard, "Ridge regression: applications to nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 69–82, 1970.

[42] G. C. McDonald, "Ridge regression," *WIREs Computational Statistics*, vol. 1, no. 1, pp. 93–100, 2009.

[43] C. De Mol, E. De Vito, and L. Rosasco, "Elastic-net regularization in learning theory," *Journal of Complexity*, vol. 25, no. 2, pp. 201–230, 2009.

[44] Z. Zhang, Z. Lai, Y. Xu, L. Shao, J. Wu, and G.-S. Xie, “Discriminative elastic-net regularized linear regression,” *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1466–1481, 2017.

[45] D. Meyer, F. Leisch, and K. Hornik, “The support vector machine under test,” *Neurocomputing*, vol. 55, no. 1-2, pp. 169–186, 2003.

[46] W. S. Noble, “What is a support vector machine?,” *Nature Biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[47] D. A. Pisner and D. M. Schnyer, *Chapter 6 - Support Vector Machine Learning ed a Mechelli and S Vieira*, Academic Press, 2020.

[48] M. W. Gardner and S. R. Dorling, “Artificial neural networks (the multilayer perceptron)–a review of applications in the atmospheric sciences,” *Atmospheric Environment*, vol. 32, no. 14-15, pp. 2627–2636, 1998.

[49] J. Tang, C. Deng, and G.-B. Huang, “Extreme learning machine for multilayer perceptron,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 4, pp. 809–821, 2016.

[50] C. Yu and W. Yao, “Robust linear regression: a review and comparison,” *Communications in Statistics: Simulation and Computation*, vol. 46, no. 8, pp. 6261–6282, 2017.

[51] E. Ostertagová, “Modelling using polynomial regression,” *Procedia Engineering*, vol. 48, pp. 500–506, 2012.