ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Support Vector Regression Based on the Particle Swarm Optimization Algorithm for Tight Oil Recovery Prediction

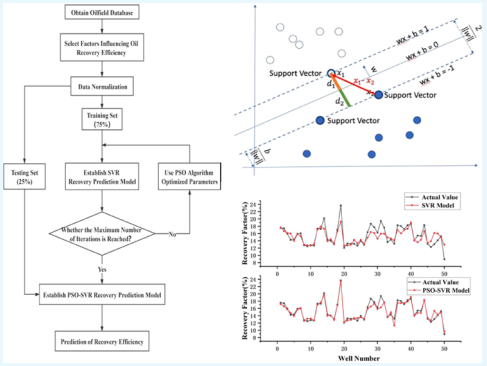
Shihui Huang, Leng Tian,* Jinshui Zhang, Xiaolong Chai, Hengli Wang, and Hongling Zhang

Cite This: *ACS Omega* 2021, 6, 32142−32150

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Tight oil fields are affected by factors such as geology, technology, and development, so it is difficult to directly obtain an accurate recovery rate. The accurate prediction of the recovery rate is very important for measuring reservoir development effects and dynamic analysis. Traditional tight oil recovery predictions are obtained by conventional formula calculations and curve fitting, which are less applicable and very different from actual conditions. Machine learning can make accurate predictions based on a large amount of data, so it is used to predict the recovery rate of tight oil reservoirs. The recovery rate of 200 wells in M tight oil reservoirs ranges widely between 8.8 and 27.6%, with more than 14 factors affecting the recovery rate, and the overall declining rule is not clear. Therefore, this research combines the production data of horizontal wells with random forest, support vector regression (SVR), and other methods, establishing recovery prediction models to gain more accurate recovery predictions. First, the Pearson correlation coefficient and the random forest (RF) machine learning method are used to measure and calculate the degree of nonlinear influence of factors on oil well recovery. Second, SVR and optimization of support vector regression by particle swarm (PSO-SVR) recovery prediction models are developed and tested, with 75% of the data being used to train SVR and PSO-SVR recovery prediction models and 25% to verify the model. Third, the accuracy of the results of these two SVR oil recovery prediction models is compared, suggesting that when the data are scarce, the optimized model is more accurate than the unoptimized one by 10.85%. Thus, this model can assure a relatively more accurate prediction of oil recovery. Machine learning recovery prediction, being more accurate and applicable, enables the data of factors such as construction and production systems to be optimized in the future, enhancing the oil recovery rate.



INTRODUCTION

Since unconventional oil and gas resources are highly valuable, the exploration and development for them have been increasing worldwide. During the development of tight oil, unlike the recovery prediction of conventional reservoirs, the recovery rate of tight oil reservoirs is affected by multiple factors such as geology, fluids, and fracture construction modification, and the analysis of its production dynamics and recovery prediction has become tough in the current stage due to the inconspicuous trap boundary and poor physical properties of the reservoirs.[1,2] M tight oil reservoirs in North China are fracture pore reservoirs with uneven natural fracture development, serious reservoir heterogeneity, and poor physical properties, which lead to complex reservoir seepage characteristics and petrol−gas flow behavior, low natural production from single wells, and low recovery rate, so it is urgent to develop a highly timely and accurate recovery prediction method for M tight oil reservoirs.[3,4]

At present, the recovery rate of tight oil can be divided into microexperimental mechanistic analysis and macroequilibrium analysis after reservoir fracturing and water drive, including the core analysis method, related empirical formula method, water drive characteristic curve method, and numerical reservoir simulation method. Previous researchers applied them to solve the recovery prediction problem.[5−11] The core analysis method is to analyze the taken cores for tests such as simulated water injection to determine the original oil saturation of the reservoir and the residual oil saturation after the test so that recovery prediction can be made in microscopic cores.[5] Based on this method, Hadia et al.[12] analyzed the relative permeability as a function of water saturation using core drive experiments and predicted the recovery rate by establishing a numerical simulation model for the dimensionless Buckley−Leverett equation.[12] Water drive characteristic curves are also widely used in predicting oilfield recovery.[7] Liu et al.[13] formed a new water drive characteristic

Received: September 7, 2021
Accepted: November 3, 2021
Published: November 16, 2021

curve for recovery prediction that can reflect the relationship between oil−water relative permeability and water saturation more accurately for the actual situation of high water-cut oil reservoirs.[13] The production decline method, including the Arps method, Blassingame method, etc, is the main method to predict the recovery rate through numerical reservoir simulations, which can predict future parameters of oil wells by judging the declining type to figure out the law, performing the overall balance analysis, and predicting the recovery rate.[8−11] There are also many solutions based on the production decline method combined with the water drive curve. Chen et al.[14] used the A-type water drive curve combined with Wong's model in numerical simulations for a reasonable recovery rate.[14] Cheng et al.[15] combined water content curves with the exponential decline method, proposing a synchronization iterative oilfield recovery prediction method based on statistical regression experiments and field data through Buckley−Leverett theory, and these methods have improved the accuracy of oilfield recovery prediction.[15] However, the factors influencing tight oil recovery are complicated and diverse, and the above methods cannot take multiple main control factors into consideration and establish reservoir models for the special geological formations and dual-porosity, dual-permeability channels of tight oil reservoirs, so the conventional recovery prediction methods are not applicable to tight oil reservoirs. In contrast, the artificial intelligence method can create a unique prediction model for various tight oilfield characteristics and multiple influencing factors and form a recovery prediction method adapted to the horizontal wells' development in tight oil reservoirs.

With the increasing popularity of artificial intelligence in various fields around the world, such as applications in natural language processing technology, intelligent computing chips, unmanned system driving, and other core technologies, the technological expansion is also becoming more mature.[16−19] While the oil and gas industry acts as the lifeline of national energy development, machine learning also gradually expands deeply into the exploitation of oil and gas resources.[20−24] For example, Ma et al.[20] attempted to correlate stochastic reservoir parameters with observable features in production time series data using artificial intelligence techniques.[20] These techniques can be integrated into the modeling process as an aid to predicting recoverable reserves. Hassan et al.[21] developed models for predicting recovery variability based on reservoir permeability and geomechanic properties, natural fracture properties, and design parameters (e.g., acid injection rate, acid concentration, treatment volume, and acid type).[21] In the area of oil and gas development, the prediction methods of machine learning collect the massive geological information formed by oil and gas exploration and support the establishment of fine reservoir description models, which can improve the accuracy of tight oil reservoir recovery prediction. Cheraghi et al.[25] developed ML models such as shallow and deep neural network models and naive Bayes for predicting candidate reservoirs for enhancing the oil recovery method.[25] Pirizadeh et al.[26] applied bagging, boosting, and stacking algorithms and proposed an efficient ensemble model called B2S to predict an enhanced oil recovery data set, establishing a balance between variance and bias and achieving an average test accuracy of 96.94% on the data set.[26] Support vector regression (SVR) is a high-dimensional nonlinear mapping machine learning method that can be adapted to a small number of sets of trained data, which have low generalization error and low computational complexity.[27] SVR has also been applied widely in the prediction of recovery rates. El-Amin et al.[27] took advantage of machine learning techniques, such as k-nearest neighbor algorithm (k-NN), artificial neural networks (ANNs), support vector machines (SVMs), and random forest (RF) for predicting recovery data and estimating dimensionless oil recovery time in real time.[27] However, SVR prediction accuracy is extremely sensitive to not only the kernel function but also penalty factor parameters and relaxation factor parameters of the stability of the model.[28] So, iterative optimization of key parameters in the kernel function of SVR is also a key issue for SVR to be able to fit a prediction model that best matches the actual tight oil reservoir values.[29] Among the many iterative optimization algorithms, the particle swarm optimization (PSO) algorithm is a class of probabilistic global optimization algorithms, where all of the optimal solution particles have self-organizing, evolutionary, and memory functions in different characteristic environments.[30] The PSO algorithm as an optimization algorithm mixed with other machine learning methods has been used similarly in petroleum engineering. Ahmadi et al.[31] used a hybrid genetic algorithm and particle swarm optimization artificial neural network (PSO-ANN) to predict bottom hole pressure (BPH) in vertical wells, and the error level/rate was reduced to 10% compared with the measured pressure data, which can provide the necessary pressure data for recovery prediction.[31] Among the above results, Yasaman Cheraghi et al. developed a wide range of ML models—shallow and deep artificial neural networks (ANN), simple Bayesian (NB), decision tree (DT), random forest (RF), and principal component analysis (PCA) for the enhanced oil recovery (EOR) method of predicting reservoirs—and also discussed in detail the reliability of using ML techniques after weighing potential disadvantages. Ahmadi et al. extended the innovative approach to upgrade the algorithm structure by optimizing the weights and the biases of ANN models, with hybrid genetic and particle swarm algorithms to improve the prediction performance and accuracy.

Among the above commonly used algorithms, decision tree boasts the features of simple computation and easily understandable interpretation, and therefore, it is more suitable for handling samples with missing attribute values and is also able to handle irrelevant features; however, it easily causes overfitting during training and ignores the correlation between attributes. The artificial neural network has strong nonlinear mapping, self-learning, and self-adaptive abilities; however, it still has the local miniaturization problem and slow convergence speed during iteration. In contrast, SVR solves the problem by converting the multidimensional number problem, which actually contains multiple influencing factors and recovery rate prediction, into a high-dimensional feature space through nonlinear transformation. At the same time, the introduction of the kernel function achieves the purpose of "dimensionality enhancement". The final solution of SVR is theoretically optimal in terms of nonlinear fitting to a high-dimensional feature space, solving the local extreme problem that cannot be avoided in the neural network method. However, its biggest drawback is that it is difficult to find the exact kernel function when the training sample dimension is high, while the particle swarm algorithm starts from a random solution and finds the global optimal solution by fast convergence iterations, so the particle swarm algorithm can be

https://doi.org/10.1021/acsomega.1c04923
ACS Omega 2021, 6, 32142−32150

used to determine the kernel function of SVR to achieve optimization of its nonlinear prediction.

The small amount of data features in the sample wells of M tight oil reservoirs in North China is consistent with the advantages of the data trained by the SVR algorithm. In the field of machine learning, SVR can accurately analyze the nonlinear variation law for prediction in a small sample size. To ensure the authenticity of the training results of the SVR prediction model, we collect the amount of recovery rate-influencing data features as much as possible and use 14 recovery impact factors from 200 single wells in M tight oil reservoirs as the original data set in the end, including reservoir physical factors, fluid characteristic factors, and fracturing engineering factors. The characteristic overlap between each influencing factor was analyzed for data preprocessing and the main control factors were determined. The PSO algorithm is used to optimize the key parameters in the kernel function, and the optimization of support vector regression by the particle swarm optimization (PSO-SVR) is used to train the sample set to establish the recovery rate prediction model of tight oil reservoirs, and the validation set is used to evaluate the accuracy and reasonableness of the recovery rate prediction model of M tight oil reservoirs.

DATA PREPROCESSING

Raw Data Description. For this research, a data set of M tight oil reservoirs was used to estimate tight oil recovery, and the engineering data of 204 producing wells were collected to construct a database. There are three major factors that influence the recovery rate of tight oil wells, including reservoir physical properties, engineering modification factors, and production development factors. The physical properties of the reservoirs have an important impact on the recovery rate and production of tight oil reservoirs. Reservoir porosity, permeability, and thickness are generally important factors that influence recovery rates. For tight reservoirs, natural fractures and fracturing are required to create high-permeability flow channels to increase the conductivity capacity to improve the recovery rate. In terms of fracturing engineering, the key to opening natural fractures to form a more effective fracture network system is larger fluid volumes, and large sand volumes play an important role in the effectiveness of tight oil volume fracturing reconstruction. A larger amount of injected fluid will lead to a larger reconstruction volume in the reservoir. A larger volume of sand will cause more proppant to be added to the fracture; therefore, it will have stronger fracture flow conductivity. The increase in the length of the horizontal section wellbore and the number of clusters will result in an increase in the drainage radius, which will similarly increase the recovery rate of horizontal fractured wells in tight oil reservoirs. In terms of production and development, the higher the oil saturation, the greater the level of production. In addition to the length of the well patterns, the spacing also directly correlates with the reservoir recovery rate. Based on three major factors such as physical properties, engineering modification, and production development, all of them affect the recovery rate of tight reservoirs. Above all, data of 14 influencing factors from 204 wells within the M tight oil reservoirs were selected as the original data set.

Correlation Judgment. The Pearson correlation coefficient was used to determine the linear correlation of each factor affecting the recovery of tight oil reservoirs. The Pearson correlation coefficient measures the linear correlation between two variables X and Y.

If the two data sets X: $\{X_1,X_2,\cdots,X_n\}$ and Y: $\{Y_1,Y_2,\cdots,Y_n\}$ are the overall data, then the overall mean is

$$E(X) = \frac{\sum_{i=1}^{n} X_i}{n},\ E(Y) = \frac{\sum_{i=1}^{n} Y_i}{n} \tag{1}$$

The overall covariance is

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

Thus, the overall Pearson correlation coefficient obtained is

$$R = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n} \frac{(X_i - E(X))}{\sigma_X} \frac{(Y_i - E(Y))}{\sigma_Y}}{n} \tag{3}$$

$$\sigma_X = \sqrt{\frac{\sum_{i=1}^{n} (X_i - E(X))^2}{n}},\ \sigma_Y = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - E(Y))^2}{n}} \tag{4}$$

where R is the magnitude of the Pearson correlation coefficient, and σ_X and σ_Y are the standard deviations of X and Y, respectively.

The correlation coefficient R has a value between + 1 and −1, where 1 represents a perfectly positive correlation, 0 represents an uncorrelated correlation, and −1 represents a perfectly negative correlation. Table 1 displays the degree of Pearson correlation coefficient judgments. The Pearson correlation coefficient eliminates the effect of the magnitude of the two variables, and it is typically used to determine the degree of linear correlation between the two variables.

Table 1. Pearson Correlation Coefficient

R	0.75−1.0	0.5−0.75	0.25−0.5	0.0−0.25
degree of correlation	extremely strong correlation	strong correlation	weak correlation	no correlation

Based on the Pearson correlation coefficients, the correlation coefficients between the factors affecting the recovery of each well are presented in Figure 1. Furthermore, the findings indicate that in tight oil reservoirs, high linear correlations between porosity and permeability exist, and the overlap between the two factors is high, with correlation coefficients greater than 0.75, meaning the factors are redundant, so one of them may be eliminated. It has also been established that the horizontal section length of the well is highly correlated with the amount of fracturing fluid and the amount of fracturing sand needed, so a factor with a similar influence could also be eliminated in the same manner. The correlation coefficients between each influencing factor and the recovery rate show that there is a stronger correlation between horizontal section length, fracturing fluid addition, sand addition, and the oil-bearing saturation and thickness of the formation in relation to the tight oil recovery rate.

Screening of Main Factors. The Pearson correlation coefficient can reflect the degree of linear correlation between the influencing factors, but it cannot precisely determine the degree of contribution of each influencing factor to the recovery rate. To further analyze the degree of their weight, the machine learning method of random forest (RF) can be used to determine the degree of nonlinear influence on the recovery

https://doi.org/10.1021/acsomega.1c04923
ACS Omega 2021, 6, 32142−32150

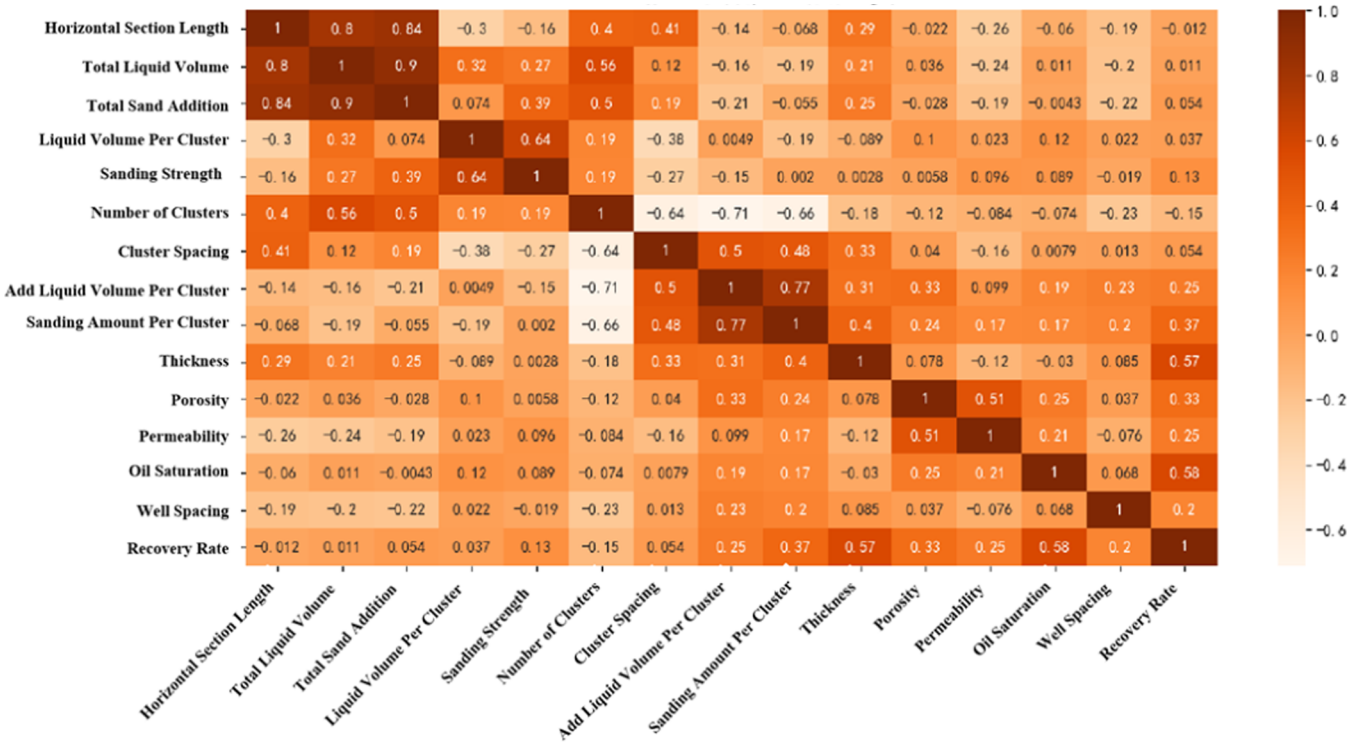


Figure 1. Thermodynamic diagram of the Pearson correlation coefficient matrix.

rate to find the main controlling factors affecting the recovery rate of oil fields. Random forest is an algorithm that integrates multiple trees into one decision tree through the idea of integrated learning, and its basic unit is a decision tree that obtains the final result by voting. The Gini index can be used as an evaluation index to measure when ranking the importance of features.

The variable importance measures are denoted by VIM and the Gini index is denoted by GI. Suppose there are m features $x_1, x_2, x_3,..., x_m$, then the Gini index is calculated by the following formula.

$$\mathrm{GI}_m = \sum_{k=1}^{|K|} \sum_{k' \neq k} p_{mk} p_{mk'} = 1 - \sum_{k=1}^{|K|} p_{mk}^2 \tag{5}$$

where K shows that there are k categories and p_{mk} is the proportion of category k in node m.

The importance measures of feature x_j at node m and at the ith tree are

$$\mathrm{VIM}_{jm}^{(\mathrm{Gini})} = \mathrm{GI}_m - \mathrm{GI}_l - \mathrm{GI}_r \tag{6}$$

$$\mathrm{VIM}_{ij}^{(\mathrm{Gini})} = \sum_{m \in M} \mathrm{VIM}_{jm}^{(\mathrm{Gini})} \tag{7}$$

where GI_l and GI_r are the Gini indices of the two new nodes after branching.

Based on the fact that there are n decision trees in RF, we obtain

$$\mathrm{VIM}_{j}^{(\mathrm{Gini})} = \sum_{i=1}^{n} \mathrm{VIM}_{ij}^{(\mathrm{Gini})} \tag{8}$$

Finally, the normalization of all of the obtained importance scores gives the magnitude of the influence of each factor on the recovery rate of the gas well.

$$\mathrm{VIM}_j = \frac{\mathrm{VIM}_j}{\sum_{i=1}^{m} \mathrm{VIM}_i} \tag{9}$$

In terms of sample data processing, RF can handle sample data with high dimensions and multiple features, which is consistent with the actual situation that the recovery rate is affected by various factors. Because of the randomness of random forest feature selection and of selecting training samples in the training set, random forest is not easy to fall into overfitting. It can handle discrete and continuous data sets that do not need to be normalized, so it has better noise immunity. After training, the RF algorithm can give specific data indicating the influence of each feature as the output result. As an integrated learning algorithm, it can build and combine multiple classifiers to finish the learning task, which have better generalization performance than a single learner and therefore will obtain more accurate results.

In Figure 2, the random forest method was applied to calculate the importance ranking of the factors influencing the recovery rate of wells in the M tight oil reservoirs. Among each factor influencing the recovery rate of tight oil reservoirs, a threshold value $\alpha = 0.075$ was determined, and there were five factors with importance indices above it. The top five characteristic factors were oil saturation, permeability, thickness, total sand addition, and horizontal section length. Combined with the Pearson correlation coefficient, the correlation coefficient between porosity and permeability is more than 0.75, and the redundancy degree is higher. The permeability was selected as the model input factor between the two, and the sand addition volume was selected as the input factor between the sand addition volume and the liquid addition volume. At the same time, it is verified from the side that the oil saturation and thickness factors have a greater influence on the recovery factor. It can be introduced that in the actual development of tight oil reservoirs, the new well

https://doi.org/10.1021/acsomega.1c04923
ACS Omega 2021, 6, 32142−32150

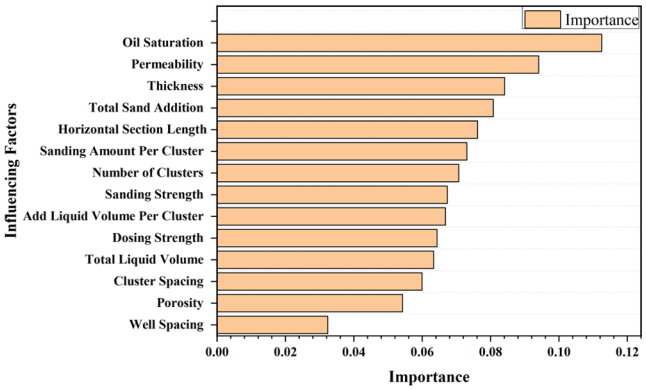


Figure 2. Ranking the importance of factors affecting the rate of recovery.

location should be selected in the area with large effective thickness and high permeability. The oil saturation of the well also has a greater influence on the recovery rate. According to the different oil saturation of each well, it is necessary to adjust the corresponding well spacing and horizontal section length, so a reasonable distribution of well spacing and horizontal section length is essential to maintain and improve the recovery rate of the well. At the same time, the recovery rate of tight oil reservoir wells can be improved by increasing the amount of sand addition and improving the strength of sand addition during fracturing construction.

Based on the importance indices of the factors influencing the recovery rate of random forest calculated from the data of this oil field, five main control factors, namely, oil content saturation, permeability, thickness, total sand addition, and horizontal section length, were selected as parameters of the support vector machine (SVM), for establishing a model to predict the recovery rate.

RESULTS AND DISCUSSION

The recovery rate and 14 factors affecting the recovery rate of 204 production wells in the tight oil reservoirs were counted as the sample database. According to the previous feature factor selection method, five main control factors, oil content saturation, permeability, thickness, total sand addition, and horizontal section length, were selected as the input parameters for establishing the recovery rate prediction model. In total, 153 wells were considered as the training set and 51 wells as the test set. The SVR and PSO-SVR were used to build the tight oil recovery rate prediction model. The SVR model with fixed-value vector machine parameters C = 52.4 and g = 0.01 was used for the prediction.

A training parameter batch size of 120 is determined by data size adaptation because the total sample data volume is 204, the training and validation sets are 75 and 25%, respectively, and the total training set volume is 153. According to eq 10, the number of iterations can be derived as 127.5, and to ensure a certain buffer space, the final number of iterations is chosen to be 150.

$$n = N \times \frac{100}{\text{batch size}} \tag{10}$$

where n is the number of iterations, N is the total number of training sets; and batch size is the size of the crawled data sets in each training.

The SVR optimized by the PSO algorithm was also set to have a population size of 20, fivefold cross-validation was used, and the PSO algorithm was used to iterate 150 times. The support vector machine parameter results are C = 25.3613 and g = 4.5786.

The sum of squares due to error (SSE), mean absolute error (MAE), mean square error (MSE), root-mean-square error (RMSE), mean absolute percentage error (MAPE), and coefficient of determination (R) were chosen as parameters to evaluate the two prediction models. The correlation coefficient R is usually used to indicate the degree of model fitness. The comparison of the prediction result error using the SVR and PSO-SVR models is listed in Table 2; it can be seen that the PSO-SVR builds a recovery prediction model with a much smaller error than the SVR prediction result.

Table 2. Comparison of the Prediction Errors of the SVR Model and the PSO-SVR Model

model type	SSE	MAE	MSE	RMSE	MAPE (%)	R
SVR prediction model	111.747	1.072	2.235	1.495	7.009	0.855
PSO-SVR prediction model	27.436	0.498	0.549	0.741	3.289	0.964

Therefore, the SVR model has a larger error in the well recovery prediction results, and the accuracy of the model is just at 85.47%, while the accuracy of the PSO-SVR model is 96.32%. Figure 3 shows the prediction errors of the PSO-SVR

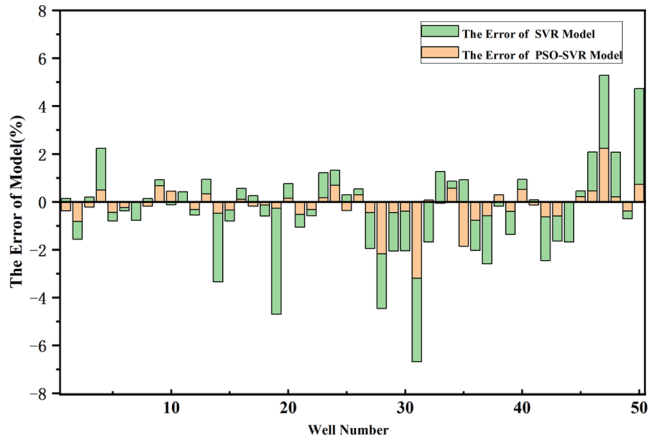


Figure 3. Plot of the prediction errors for the SVR and PSO-SVR models.

and SVR models, and it can be seen that the error fluctuation of the PSO-SVR model is smaller than that of the SVR model. Figure 4 shows the comparison between the predictions of the PSO-SVR and SVR models and actual values, and it can be clearly seen that the PSO-SVR model has more accurate results for the prediction of the recovery rate. Figure 5 shows the adaptation curve for 150 iterations of the PSO iterative algorithm, indicating that with the increase of the number of iterations, the adaptation decreases and gradually approaches the best adaptation. The parameters are continuously optimized, and the accuracy of the test set evaluation model is higher.

https://doi.org/10.1021/acsomega.1c04923
ACS Omega 2021, 6, 32142−32150

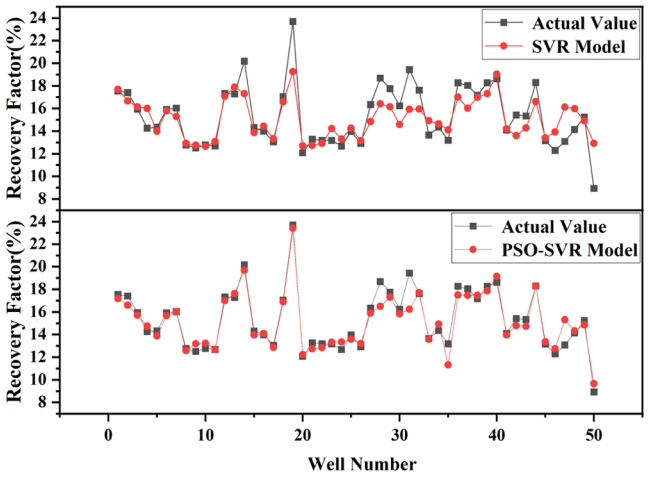


Figure 4. Comparison between SVR and PSO-SVR predictions.

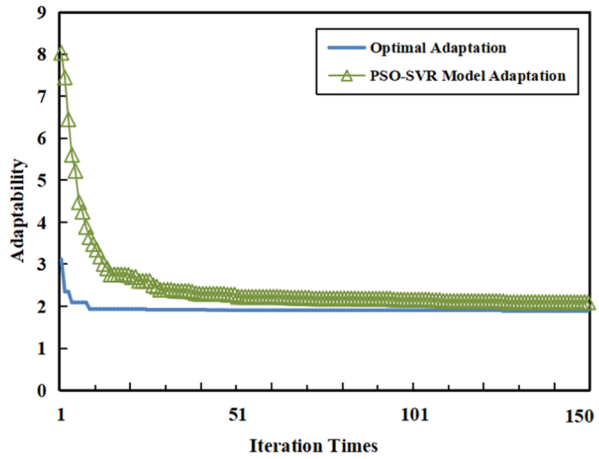


Figure 5. PSO-SVR recovery model optimization algorithm adaptation curve.

CONCLUSIONS

In this study, the Pearson correlation coefficient and the random forest method were used to select the main factors, and the SVR prediction model was optimized using the PSO algorithm based on MATLAB software. In addition, the SVR recovery prediction model and the PSO-SVR prediction model were developed and compared. The following are the main conclusions:

(1) In the Pearson correlation analysis, it was found that permeability, porosity, horizontal section length, fracturing fluid addition volume, and sand addition volume strongly correlate.
(2) Based on the random forest calculation results, the five main factors with the greatest influence weights were determined as oil saturation, permeability, thickness, total sand addition, and horizontal section length.
(3) The PSO-SVR recovery prediction model has high accuracy with an average percentage error of only 3.29%. Compared with the support vector regression machine without optimization, its accuracy improved from 85.47 to 96.32%.

These data mining-based analysis and machine learning prediction methods have flexible operations and can achieve high prediction accuracy, providing a new idea for tight oil recovery prediction and improving the recovery prediction efficiency of oil and gas reservoirs.

THEORY AND METHODS

Support Vector Regression (SVR). Support vector regression (SVR) is a machine learning method developed on the basis of statistical learning theory. Its basic idea is to transform the input space to a high-dimensional space by a nonlinear transformation defined by the inner product function and then find a linear relationship between the input data and the output data in this high-dimensional space. Selecting a suitable kernel function can easily be able to transform data from the input space to the corresponding nonlinear high-dimensional space.

The SVR algorithm is used to achieve a nonlinear fit between the input data and the output data, and its fitting function $f(x,w)$ can be expressed as

$$f(x, w) = w\cdot\phi(x) + b = (w, \phi) + b \tag{11}$$

where w is the vector of weights, $\phi(x)$ is the nonlinear mapping that generates vector which has the same dimension with input vector x, b is the deviation, and $w\cdot\phi(x)$ is the inner product of w and ϕ.

The SVR optimization problem is to find a function that is centered on $f(x,w)= w\cdot\phi(x) + b = (w\cdot\phi) + b$; by taking into account the allowed fitting error, a band of width 2ε of the interval band is constructed. By introducing relaxation factors ξ_i, and ξ_i^*, the original optimization problem with the relaxation factor can be expressed as

$$\min_{w,b,\xi_i,\xi_i^*} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$
$$s.t.\begin{cases} y_i - w\cdot\phi(x_i) - b \le \varepsilon + \xi_i \\ w\cdot\phi(x_i) + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \tag{12}$$

where ξ_i and ξ_i^* are the relaxation factors and C is the penalty factor, which is used as the weight between the error and the optimization objective.

The Lagrangian function constrained optimization is used to solve the original problem. We consider the Lagrangian minimal−extreme problem and introduce a kernel function to solve the quadratic programming problem, and the following results are obtained.

$$f(x, \alpha_i, \alpha_i^*) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)k(x, x_i) + b \tag{13}$$

where function f is completely determined by α_i and α_i^*. By the nature of the SVR regression function, only a few α_i and α_i^* are not zero, and deviation b can also be calculated from the standard support vector. The introduced kernel function $k(x,x_i)$ is a function that must satisfy the Mercer condition. There are many forms of kernel functions, and the radial basis function (RBF) that can well solve complex nonlinear problems is used in the study for predicting the recovery of tight oil reservoirs.

$$k(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right) \tag{14}$$

https://doi.org/10.1021/acsomega.1c04923
ACS Omega 2021, 6, 32142−32150

Thus, the final estimation function of the hyperplane support vector regression is obtained as follows

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*)\exp\left(-\frac{\| x - x_i \|^2}{2\sigma^2}\right) + b \quad (15)$$

Particle Swarm Optimization (PSO) Algorithm. Particle swarm optimization (PSO), an evolutionary computation technique, was first proposed by Eberhart and Kennedy in 1995, and its basic concept is derived from the study of the foraging behavior of bird flocks. The basic idea of PSO is to find the optimal solution by collaboration and information sharing among individuals in a population, and the optimal solution searched by each particle individually is the current global optimal solution. Finally, the optimal solution that satisfies the termination condition is obtained.

The particles of the simulated birds have only two properties: velocity V represents the speed of movement and position X represents the direction of movement. Each particle individually searches for the optimal solution P_{best} in the search space and shares the optimal solution with the other particles in the whole swarm, and the optimal solution G_{best} found is the current global optimal solution of the whole swarm. The particles update their velocities and new positions to find the two optimal solutions according to the following equations

$$\begin{aligned} V_{id} &= \omega V_{id} + C_1\ \text{random}(0, 1)(P_{id} - X_{id}) + C_2 \\ &\quad \text{random}(0, 1)(P_{gd} - X_{id}) \\ X_{id} &= X_{id} + V_{id} \end{aligned} \quad (16)$$

where V_{id} is the velocity of the particle, ω is the inertia weight, X_{id} is the current position of the particle, P_{id} is the dth dimension of the individual extremum of the ith variable, P_{gd} is the dth dimension of the global optimal solution, C_1 is the individual learning factor of each particle, and C_2 is the social learning factor of each particle. Suganthan's experiments show that better solutions can be obtained when C_1 and C_2 are constants, and usually set $C_1 = C_2 = 2$ random (0,1) denotes a random number in the interval [0,1]. Figure 6 shows the flow chart of the PSO algorithm.

Model Development. Since SVR constructs a kernel function instead of a linear equation to achieve linear regression, constructing a suitable kernel function is of the greatest importance for prediction of the recovery rate in the SVR model. The RBF is a localized kernel function that can map a sample into a higher dimensional space, which is able to fit data sets with multiple features better. It had a relatively good performance in the data samples of this study. In the kernel function, the parameters are chosen to be optimized by the PSO algorithm to make it fit better.

The PSO algorithm is simpler than the genetic algorithm because it does not have the "crossover" and "mutation" operations in the genetic algorithm. It finds the global optimum by achieving the current searched optimal value. This algorithm is easier to implement and has simpler operations than the genetic algorithm. Compared with other algorithms, this algorithm is able to remember the velocity and position of all particles, while other algorithms change as the population changes, so it has higher accuracy and faster convergence, which can gain superiority of prediction with the global optimal solution found by the SVR model and improve the accuracy.

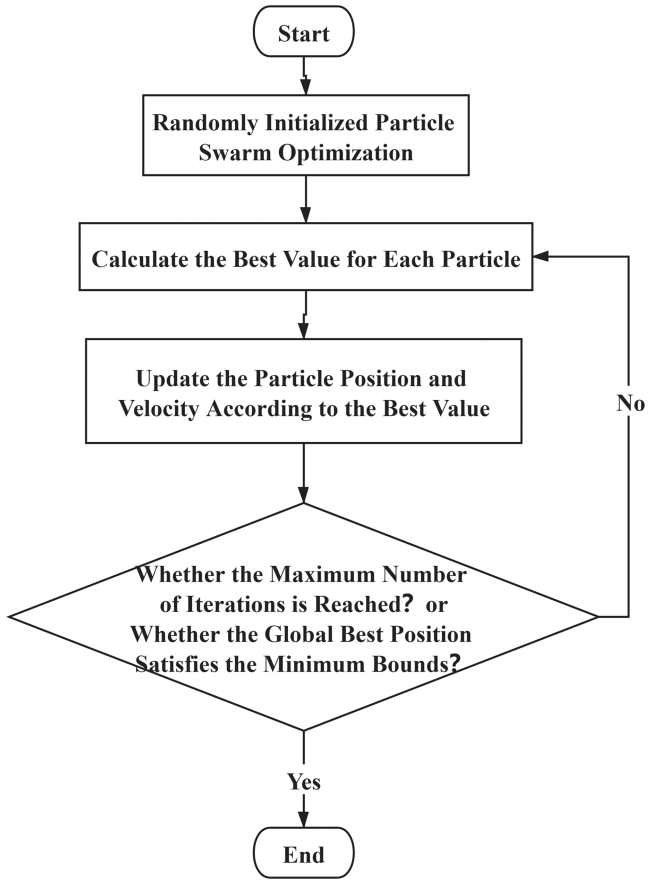


Figure 6. Flow chart of the PSO algorithm.

The optimization of the support vector regression by particle swarm (PSO-SVR) recovery rate prediction model building process is shown in Figure 7. First, the collected data were normalized, and then 75% of them were used as the training set and 25% as the test set, and the support vector machine recovery rate prediction model was built using the training set. The radial basis function (RBF) was selected as the kernel function and optimized by the PSO algorithm, and the two key parameters of penalty parameter C and kernel function parameter g in the support vector machine were calculated by optimization iterations. A certain number of iterations were set, and the PSO algorithm kept finding the optimal penalty parameter C and the kernel function parameter g through the particle swarm within the allowed number of iterations and would keep approximating the two optimal parameters that best fit the training set because of the convergence of the algorithm. Later, the test set was applied to the trained model to calculate the recovery rate results of the PSO-SVR prediction model.

Model Evaluation. The accuracy of the model in the test set is evaluated by the calculation of the model and true values. The sum of squares due to error (SSE), mean absolute error (MAE), mean square error (MSE), root-mean-square error (RMSE), mean absolute percentage error (MPAE), and coefficient of determination (R) were used to evaluate the SVR and PSO-SVR recovery rate prediction models.

SSE can be calculated as follows

$$\text{SSE} = \sum_{i=1}^{n} (R_{\text{FActual } i} - R_{\text{FPredict } i})^2 \quad (17)$$

https://doi.org/10.1021/acsomega.1c04923
ACS Omega 2021, 6, 32142−32150

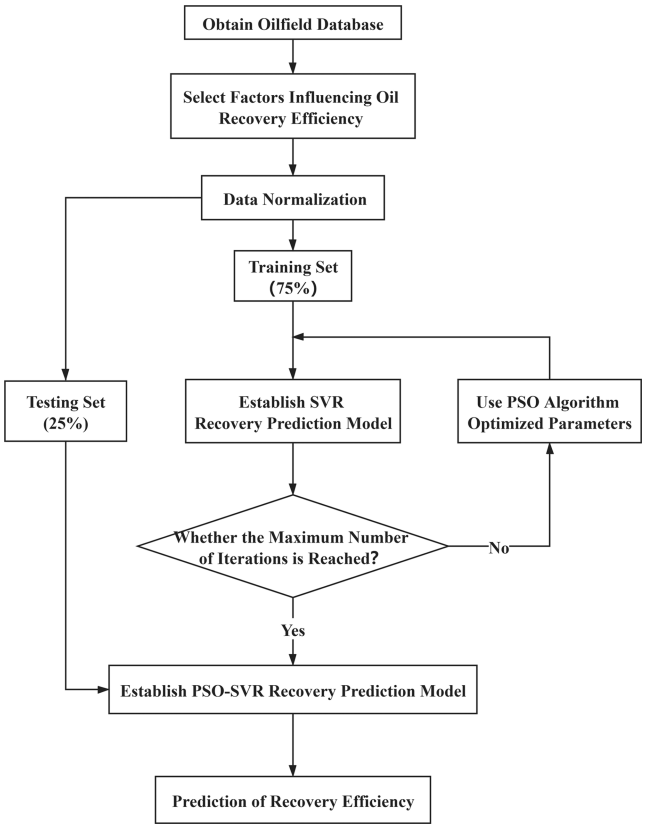


Figure 7. Flow chart of the PSO-SVR recovery prediction model.

where $R_{\text{F Actual}\,i}$ and $R_{\text{F Predict}\,i}$ are the actual and predicted recovery rates, respectively, and n is the number of data points.

MAE can be calculated as follows

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |(R_{\text{FActual}\,i} - R_{\text{FPredict}\,i})| \tag{18}$$

where $R_{\text{F Actual}\,i}$ and $R_{\text{F Predict}\,i}$ are the actual and predicted recovery rates, respectively, and n is the number of data points.

MSE can be calculated as follows

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} (R_{\text{FActual}\,i} - R_{\text{FPredict}\,i})^2 \tag{19}$$

where $R_{\text{F Actual}\,i}$ and $R_{\text{F Predict}\,i}$ are the actual and predicted recovery rates, respectively, and n is the number of data points.

RMSE can be calculated as follows

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (R_{\text{FPredict}\,i} - R_{\text{FActual}\,i})^2} \tag{20}$$

where $R_{\text{F Actual}\,i}$ and $R_{\text{F Predict}\,i}$ are the actual and predicted recovery rates, respectively, and n is the number of data points.

MAPE can be calculated as follows

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n} \left|\frac{R_{\text{FPredict}\,i} - R_{\text{FActual}\,i}}{R_{\text{FActual}\,i}}\right| \tag{21}$$

where $R_{\text{F Actual}\,i}$ and $R_{\text{F Predict}\,i}$ are the actual and predicted recovery rates, respectively, and n is the number of data points.

R is used to indicate the degree of fit of the model, which can be calculated as follows

$$R^2 = \left[\left[n\sum_{i=1}^{n}(R_{\text{FActual}\,i}\times R_{\text{FPredicted}\,i})\right] - \left[\sum_{i=1}^{n} R_{\text{FActual}\,i}\times\sum_{i=1}^{n} R_{\text{FPredicted}\,i}\right]\right] / \left[\left\{\left[n\sum_{i=1}^{n}(R_{\text{FActual}\,i})^2 - \left(\sum_{i=1}^{n} R_{\text{FActual}\,i}\right)^2\right] \times \left[n\sum_{i=1}^{n}(R_{\text{FPredicted}\,i})^2 - \left(\sum_{i=1}^{n} R_{\text{FPredicted}\,i}\right)^2\right]\right\}^{1/2}\right] \tag{22}$$

where $R_{\text{F Actual}\,i}$ and $R_{\text{F Predicted}\,i}$ are the actual and predicted recovery rates, respectively, and n is the number of data points.

A comparison was made between SVR and PSO-SVR prediction models using SSE, MAE, MSE, RMSE, MAPE, and R.

In the PSO-SVR model, the optimization parameters are mainly the penalty parameter C and the kernel function g of the SVM. During the training process, the MSE of the C and g of the PSO-SVR model is used as the adaptation function to determine the accuracy of the model.

The fitness function is designed as follows

$$\text{fitness} = \text{argmin}(\text{MSE}_{\text{predict}}) \tag{23}$$

where $\text{MSE}_{\text{predict}}$ is the mean square error of the SVR parameter.

With less MSE error, the predicted data are more likely to converge to the original data.

AUTHOR INFORMATION

Corresponding Author

Leng Tian − *State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum (Beijing), Beijing 102249, China; Department of Petroleum Engineering, China University of Petroleum (Beijing), Beijing 102249, China;* orcid.org/0000-0002-5796-7154; Email: tianleng2008@126.com

Authors

Shihui Huang − *State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum (Beijing), Beijing 102249, China; Department of Petroleum Engineering, China University of Petroleum (Beijing), Beijing 102249, China*

Jinshui Zhang − *State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum (Beijing), Beijing 102249, China; Department of Petroleum Engineering, China University of Petroleum (Beijing), Beijing 102249, China*

Xiaolong Chai − *State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum (Beijing), Beijing 102249, China; Department of Petroleum Engineering, China University of Petroleum (Beijing), Beijing 102249, China*

Hengli Wang − *State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum (Beijing), Beijing 102249, China; Department of Petroleum*

https://doi.org/10.1021/acsomega.1c04923
ACS Omega 2021, 6, 32142−32150

Engineering, China University of Petroleum (Beijing), Beijing 102249, China; orcid.org/0000-0001-5997-8474

Hongling Zhang − *State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum (Beijing), Beijing 102249, China; Department of Petroleum Engineering, China University of Petroleum (Beijing), Beijing 102249, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.1c04923

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was financially supported by the Natural Science Foundation of China (No. 51974329).

■ REFERENCES

(1) Sun, L. D.; Zou, C. N.; Jia, A. L.; Wei, Y. S.; Zhu, R. K.; Wu, S. T.; Guo, Z. Development characteristics and orientation of tight oil and gas in China. *Pet. Explor. Dev.* **2019**, *46*, 1073−1087.

(2) Hu, S. Y.; Zhu, R. K.; Wu, S. T.; Bai, B.; Yang, Z.; Cui, J. W. Exploration and development of continental tight oil in China. *Pet. Explor. Dev.* **2018**, *45*, 790−802.

(3) Li, Z. X.; Qu, X. F.; Liu, W. T.; Lei, Q. H.; Sun, H. L.; He, Y. A. Development modes of Triassic Yanchang Formation Chang 7 Member tight oil in Ordos Basin, NW China. *Pet. Explor. Dev.* **2015**, *42*, 241−246.

(4) Lyu, W. Y.; Zeng, L. B.; Liao, Z. H.; Ji, Y. Y.; Lyu, P.; Dong, S. Q. Fault damage zone characterization in tight-oil sandstones of the Upper Triassic Yanchang Formation in the southwest Ordos Basin, China: Integrating cores, image logs, and conventional logs. *Interpretation* **2017**, *5*, SP27−SP39.

(5) Guo, C. F.; Li, H. B.; Tao, Y.; Lang, L. Y.; Niu, Z. X. Water invasion and remaining gas distribution in carbonate gas reservoirs using Water invasion and remaining gas distribution in carbonate gas reservoirs using core displacement and NMR. *J. Cent. South. Univ.* **2020**, *27*, 531−541.

(6) Yuan, Z. X.; Wang, J. Y.; Li, S. X.; Ren, J. H.; Zhou, M. Q. A new approach to estimating recovery factor for extra-low permeability water-flooding sandstone reservoir. *Pet. Explor. Dev.* **2014**, *41*, 377−386.

(7) Baili, W.; Qiao, F.; Haiying, J.; Chuanrui, D. Application of water drive characteristic curve in the development of low permeability reservoirs. *Spec. Oil Gas Reservoirs* **2019**, *26*, 82−87.

(8) Al-Jifri, M.; Al-Attar, H.; Boukadi, F. New proxy models for predicting oil recovery factor in waterflooded heterogeneous reservoirs. *J. Pet. Explor. Prod.* **2021**, *11*, 1443−1459.

(9) Fathaddin, M. T.; Thomas, M. M.; Pasarai, U. In *Predicting Oil Recovery through CO2 Flooding Simulation Using Methods of Continuous and Water Alternating Gas*, 4th Annual Applied Science and Engineering Conference (AASEC), Apr 25 2019.

(10) Teng, L.; Zhang, D. T.; Li, Y. X.; Wang, W. C.; Wang, L.; Hu, Q. H.; Ye, X.; Bian, J.; Teng, W. C. Multiphase mixture model to predict temperature drop in highly choked conditions in CO_2 enhanced oil recovery. *Appl. Therm. Eng.* **2016**, *108*, 670−679.

(11) Miura, K.; Wang, J. An Analytical Model To Predict Cumulative Steam/Oil Ratio (CSOR) in Thermal-Recovery SAGD Process. *J. Can. Pet. Technol.* **2012**, *51*, 268−275.

(12) Hadia, N.; Chaudhari, L.; Aggarwal, A.; et al. Experimental and numerical investigation of one-dimensional waterflood in porous reservoir. *Exp. Therm. Fluid Sci.* **2007**, *32*, 355−361.

(13) Liu, Z. B.; Liu, H. H. An effective method to predict oil recovery in high water cut stage. *J. Hydrodyn.* **2015**, *27*, 988−995.

(14) Chen, Y. R.; Wang, Y. F. Research on quantitative evaluation method of reasonable oil recovery rate in oil and gas field development. *Fresenius Environ. Bull.* **2020**, *29*, 2540−2546.

(15) Cheng, M. M.; Lei, G. L.; Gao, J. B.; Xia, T.; Wang, H. S. Laboratory Experiment, Production Performance Prediction Model, and Field Application of Multi-slug Microbial Enhanced Oil Recovery. *Energy Fuels* **2014**, *28*, 6655−6665.

(16) Roh, T.; Jeong, Y.; Jang, H.; Yoon, B. Technology opportunity discovery by structuring user needs based on natural language processing and machine learning. *PLoS One* **2019**, *14*, No. e0223404.

(17) Hung, J. M.; Li, X. Q.; Wu, J. J.; Chang, M. F. Challenges and Trends in Developing Nonvolatile Memory-Enabled Computing Chips for Intelligent Edge Devices. *IEEE Trans. Power Electron* **2020**, *67*, 1444−1453.

(18) Schiestl, M.; Marcolini, F.; Incurvati, M.; Capponi, F. G.; Starz, R.; Caricchi, F.; Rodriguez, A. S.; Wild, L. Development of a High Power Density Drive System for Unmanned Aerial Vehicles. *IEEE Trans. Power Electron* **2021**, *36*, 3159−3171.

(19) Liu, G. S.; Liu, H. T.; Xu, Z. J. Design of motor drive system for quadrotor unmanned aerial vehicle. In *Basic & Clinical Pharmacology & Toxicology*; WILEY, 2020; Vol. *127*, pp 254.

(20) Ma, Z. W.; Leung, J. Y.; Zanon, S. Integration of artificial intelligence and production data analysis for shale heterogeneity characterization in steam-assisted gravity-drainage reservoirs. *J. Pet. Sci. Eng.* **2018**, *163*, 139−155.

(21) Hassan, A.; Aljawad, M. S.; Mahmoud, M. An Artificial Intelligence-Based Model for Performance Prediction of Acid Fracturing in Naturally Fractured Reservoirs. *ACS Omega* **2021**, *6*, 13654−13670.

(22) Feng, X.; Feng, Q.; Li, S. H.; Hou, X. W.; Zhang, M. Q.; Liu, S. G. Automatic Deep Vector Learning Model Applied for Oil-Well-Testing Feature Mining, Purification and Classification. *IEEE Access* **2020**, *8*, 151634−151649.

(23) Liu, S. S.; Zhao, Y. P.; Wang, Z. M. Artificial Intelligence Method for Shear Wave Travel Time Prediction considering Reservoir Geological Continuity. *Math. Probl. Eng.* **2021**, 2021.

(24) Kozel, T.; Stary, M. Adaptive stochastic management of the storage function for a large open reservoir using an artificial intelligence method. *J. Hydrol. Hydromech.* **2019**, *67*, 314−321.

(25) Cheraghi, Y.; Kord, S.; Mashayekhizadeh, V. Application of machine learning techniques for selecting the most suitable enhanced oil recovery method; challenges and opportunities. *J. Pet. Sci. Eng.* **2021**, *205*, No. 108761.

(26) Pirizadeh, M.; Alemohammad, N.; Manthouri, M.; Pirizadeh, M. A new machine learning ensemble model for class imbalance problem of screening enhanced oil recovery methods. *J. Pet. Sci. Eng.* **2021**, *198*, No. 108214.

(27) El-Amin, M. F.; Subasi, A. Developing a Generalized Scaling-Law for Oil Recovery Using Machine Learning Techniques. *Procedia Comput. Sci.* **2019**, *163*, 237−247.

(28) Ozturk, U.; Cicek, K.; Celik, M. In *An Intelligent Fault Diagnosis System on Ship Machinery Systems Based on Support Vector Machine Principles*, 26th Conference on European Safety and Reliability (ESREL), Glasgow, SCOTLAND, 2017, Sep 25−29; Glasgow, SCOTLAND, 2016; pp 1949−1953.

(29) Fan, J.; Li, X. Prediction of the productivity of steam flooding production wells using Gray Relation Analysis and Support Vector Machine. *J. Comput. Methods Sci. Eng.* **2015**, *15*, 499−506.

(30) Yin, W.-J.; Ming, Z.-F. Electric vehicle charging and discharging scheduling strategy based on local search and competitive learning particle swarm optimization algorithm. *J. Energy Storage* **2021**, *42*, No. 102966.

(31) Ahmadi, M. A.; Chen, Z. X. Machine learning models to predict bottom hole pressure in multi-phase flow in vertical oil production wells. *Can. J. Chem. Eng.* **2019**, *97*, 2928−2940.