Research Article

Prediction of Fluid Viscosity in Multiphase Reservoir Oil System by Machine Learning

Lihua Shao [1], **Ru Ji** [2], **Shuyi Du** [2], **and Hongqing Song** [2]

[1]*School of Mathematics and Physics, University of Science and Technology Beijing, 100083 Beijing, China*
[2]*School of Civil and Resource Engineering, University of Science and Technology Beijing, 100083 Beijing, China*

Correspondence should be addressed to Hongqing Song; songhongqing@ustb.edu.cn

It is important to realize rapid and accurate prediction of fluid viscosity in a multiphase reservoir oil system for improving oil production in petroleum engineering. This study proposed three viscosity prediction models based on machine learning approaches. The prediction accuracy comparison results show that the random forest (RF) model performs accurately in predicting the viscosity of each phase of the reservoir, with the lowest error percentage and highest R^2 values. And the RF model is tremendously fast in a computing time of 0.53 s. In addition, sensitivity analysis indicates that for a multiphase reservoir system, the viscosity of each phase of the reservoir is determined by different factors. Among them, the viscosity of oil is vital for oil production, which is mainly affected by the molar ratio of gas to oil (MR-GO).

1. Introduction

The fluid viscosity of the oil-gas reservoir [1, 2] is the key factor to determine the final development effect and economic benefit of the oil-gas reservoir. Therefore, it has become an important basis performance for formulating oil-gas field development plans [3], studying oil and gas reservoir performance, implementing plan adjustment, and evaluating stimulation [4].

The combination of PVT device and high-pressure falling ball viscometer [5, 6] can realize the laboratory analysis of reservoir samples, to obtain the viscosity value in reservoir environment (high pressure and temperature). PVT device [7, 8] can create specific temperature and pressure to simulate reservoir environmental conditions; therefore, it has been widely used in the oil industry in recent years [9]. However, the acquisition of such data, including sampling and subsequent analysis, will cost considerable cost and time, which is not desirable [10].

In addition, to simplify and quickly obtain the reservoir fluid viscosity and to analyze the influencing factors of viscosity, a large number of simulation studies on viscosity have appeared in the petroleum industry, and many commonly used viscosity models have been proposed, including LBS viscosity model [11], CS viscosity model, LLS viscosity model, Pt viscosity model, and PR viscosity model [12, 13]. The application of these models can realize the viscosity acquisition of reservoir fluid with specific composition and realize viscosity prediction. But the viscosity of oil-gas reservoir fluid, especially in oil-gas-water multiphase reservoir system, is affected by many factors, including reservoir environmental conditions, oil and gas composition, and water and gas injection [14]. At present, it is impossible to find a general viscosity model to describe the viscosity characteristics of fluids in multiphase reservoir systems.

Therefore, the objective of this study is to establish a reliable and accurate machine learning model for predicting the viscosity of each phase in a multiphase mixed oil-gas-water system. Research shows that deep neural networks (DNN), random forests (RF), and support vector regression (SVR) are very good at capturing and learning the nonlinear feature relationships between data, and they can accurately predict parameters in a data-driven manner without physical models. Compared with some classic machine learning algorithms, these machine learning algorithms can often

maintain high prediction accuracy even under small sample conditions, which map more feature spaces. And trained models have higher portability and can quickly adapt to different application scenarios. This paper, therefore, would choose these three machine learning methods to predict and analyze the viscosity of each fluid in a multiphase mixed oil-gas-water system.

To achieve this research purpose, Tarim reservoir oils were taken as an example. Tarim Oilfield is located in Kuqa County, Xinjiang, China. It is urgent to tackle the key problems of enhanced oil recovery. The early gas injection research recognized that gas injection is the practical technical direction of enhanced oil recovery of reservoir in Tarim. This study collects a large number of experimental and simulation data to provide a large amount of data for machine learning. In addition, the sensitivity analysis of influencing factors plays a vital role in the development guidance of the reservoir oils, so the collected data was sorted and analyzed, according to the reservoir environmental factors (temperature T and pressure P) and the reservoir composition (the molar ratio of gas to oil MR-GO and molar ratio of water to oil MR-WO [15, 16]). Therefore, the developed viscosity prediction model covers wider ranges of input data, which is important to the production of oil reservoirs.

The structure of this paper is as follows. In the following section, the background, governing equations, and development methodology of the three presented models, including RF, DNN, and SVR, are introduced and described in detail. In addition, this section will also give the calculation method of the statistical indicators for evaluating the three models. Next, in Section 3, the accuracy and calculation time comparison of these three developed models will be evaluated by the statistical indicators, and the reliability analysis of the calculation process of the RF model will be given. Moreover, the sensitivity analysis of the influencing factors will be carried out, and the influence weight of each influencing factor on the output viscosity in the multiphase system will also be given. Finally, Section 4 will present the key findings of this paper.

2. Methodology

2.1. Prediction Models. Three prediction models were used to predict the viscosity of a multiphase reservoir oil system from input data of crude oil systems, such as MR-GO, MR-WO, reservoir environment pressure (P, MPa), and temperature (T, °C). Table 1 indicates the statistical characteristics of the input data.

2.1.1. Random Forest (RF). Random forest [17], as an ensemble learning algorithm based on classification and regression trees, has been widely used in many fields. Ho [18] first proposed the random forest algorithm in 1995 and improved the algorithm by Breiman [19] in 2001. RF is a machine learning method based on statistical learning theory. First, multiple samples are extracted from the original data set through the bootstrap resampling method. Then, a decision tree model is established for

TABLE 1: Statistical characteristic of the input data.

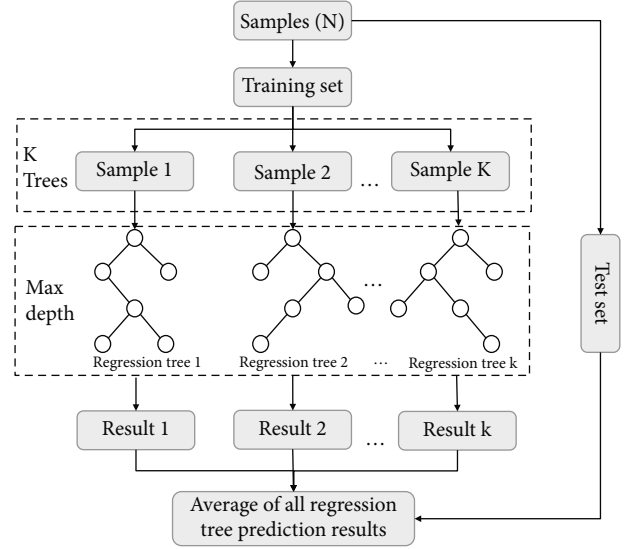| Parameter	Minimum	Maximum	Mean
P	5	95	46.25
T	30	210	120
MR-GO	0.5	6	1.5
MR-WO	0	2.3	1.292

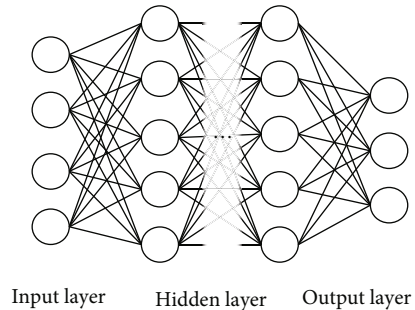FIGURE 1: The training process of random forest.

FIGURE 2: The training process of deep neural network.

TABLE 2: Statistical evaluation results calculated by three models.

Models		MSE	MAE	R^2
RF	v_g	$2.8979e-06$	0.0005	0.9824
	v_o	0.0008	0.0093	0.9623
	v_w	$4.3221e-06$	0.0008	0.9999
DNN	v_g	$9.2978e-05$	0.0063	0.4364
	v_o	0.0022	0.0301	0.8942
	v_w	0.0099	0.0463	0.7759
SVR	v_g	0.0002	0.0112	-0.1083
	v_o	0.0108	0.0855	0.4798
	v_w	0.0118	0.0767	0.7340
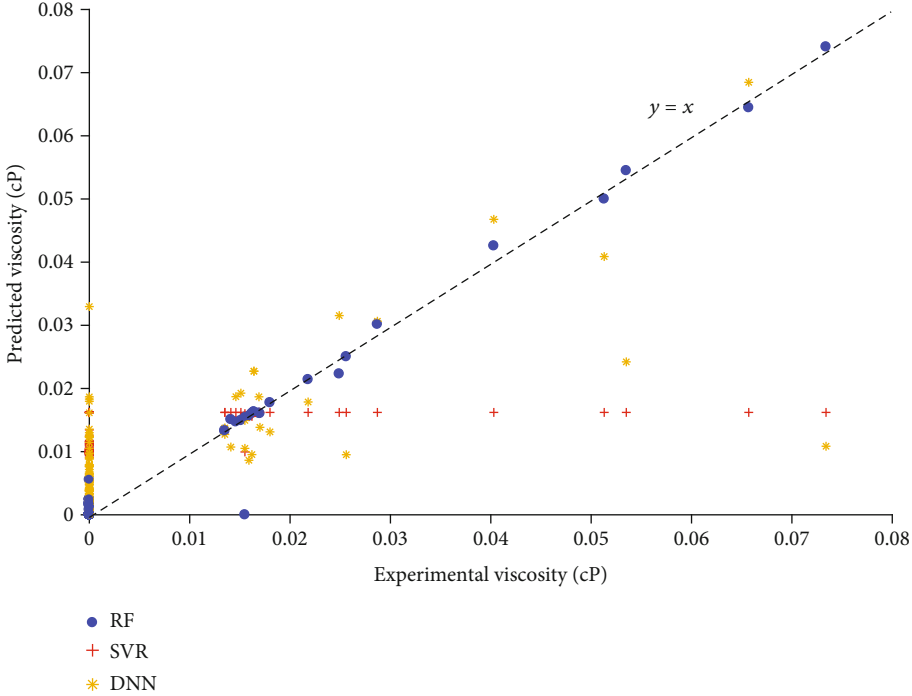
(a) Gas viscosity (v_g)
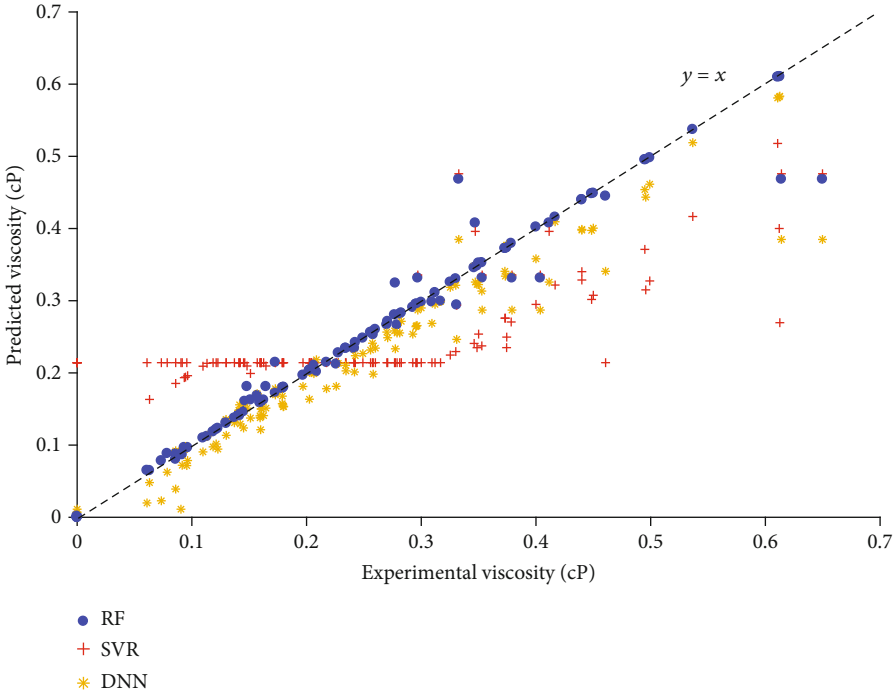
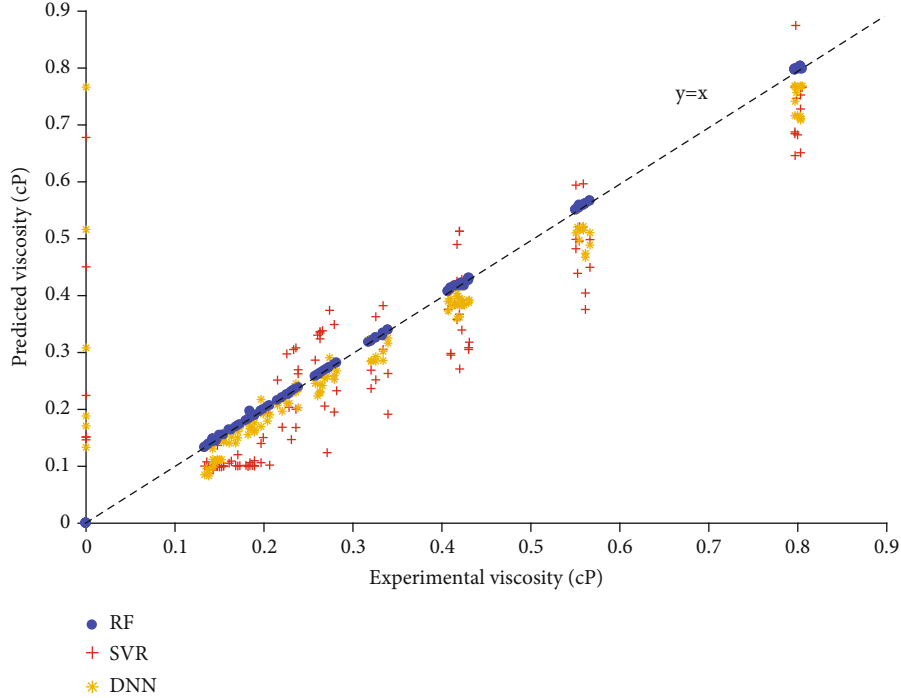(b) Oil viscosity (v_o)

FIGURE 3: Continued.

(c) Water viscosity (v_w)

FIGURE 3: The cross-plot pictures between model predictions and the corresponding experimental values.

each bootstrap sample. Finally, the predictions of multiple decision trees are combined and averaged to obtain the final prediction result. The ultimate regression decision formula is as follows:

$$M_{rf}(x) = \frac{1}{K}\sum_{k=1}^{K} t_i(x), \tag{1}$$

where M_{rf} represents the calculation result of the random forest model, K is the number of regression trees required, and t_i represents a single regression tree model. In the calculation process of the random forest model, there are two extremely critical hyperparameters, which are the number of regression trees (n-estimators) and the number of random variables at the nodes (max depth). Too little number of regression trees will affect the accuracy of the calculation. Similarly, too many numbers will increase the complexity of the calculation. The training process of random forest is shown in Figure 1.

2.1.2. Deep Neural Network (DNN). A deep neural network [20, 21] is a machine learning method that combines a multilayer perceptron structure and a backpropagation algorithm. It is mainly composed of three parts, including the input layer, hidden layer, and output layer (Figure 2). The key structure of a deep neural network is called a neuron, which can characterize the nonlinear mapping

relationship between input and output. The output equation of each neuron is as follows:

$$O(x) = g\left(\sum_i w_i x_i + b\right), \tag{2}$$

where O is the output value, g is the activation function, w is the weight of the input parameter, and b is the threshold. The training process of the neural network model is divided into two steps: forward propagation and backward propagation. First, forward propagation is used to calculate the predicted value of the model. The error gradient between the predicted value and the true value is obtained by the loss function. According to the error gradient, the weights and thresholds in the neural network are adjusted through the backpropagation algorithm. Repeating this process can make the network continuously learn the hidden features between the data.

2.1.3. Support Vector Regression (SVR). Support vector machine [22–24] is a machine learning method proposed by Cortes and Vapnik [25] in 1995 to learn the mapping relationship between parameters. The core idea of support vector regression is to find the nonlinear mapping relationship between input space and output space. Relying on the nonlinear mapping, data is mapped to a high-dimensional characteristic space. The estimating
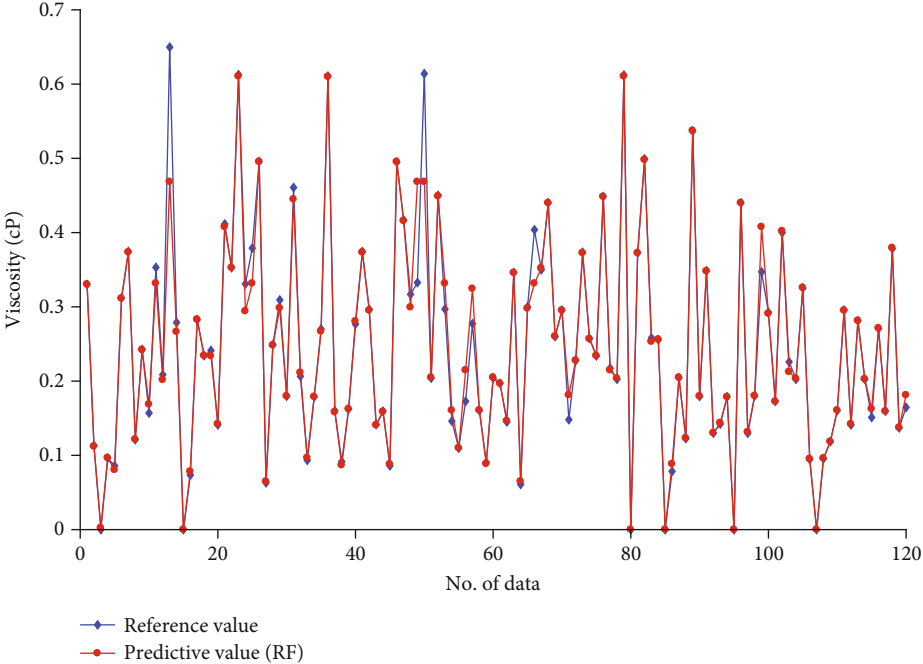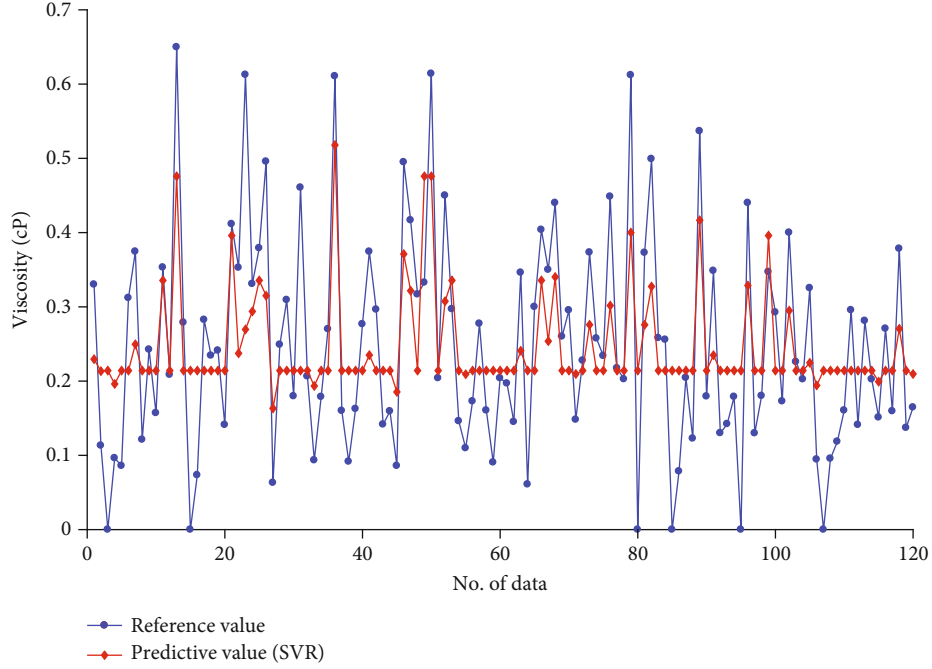
(a) RF model
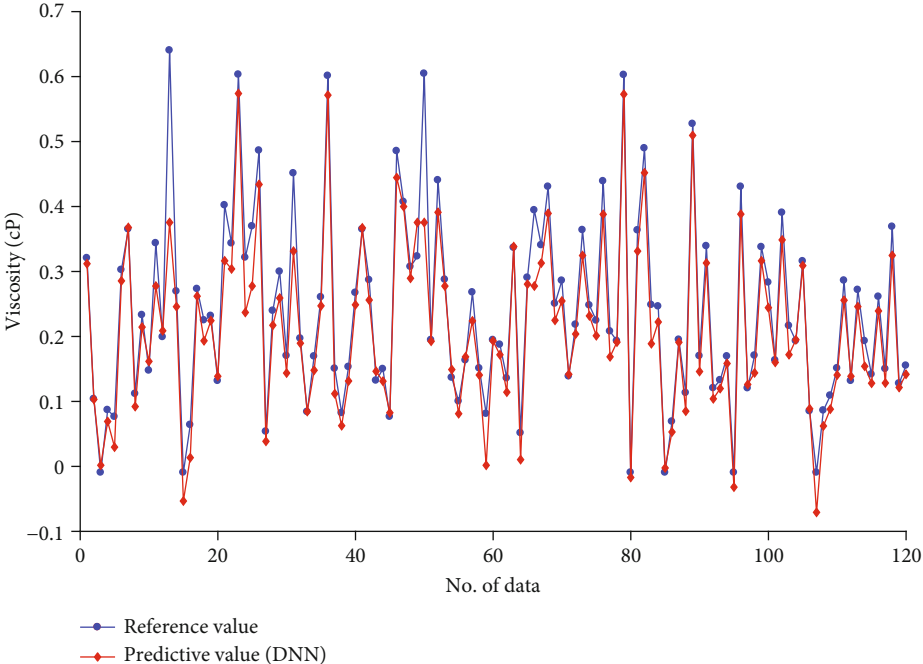
(b) DNN model

Figure 4: Continued.

(c) SVR model

FIGURE 4: The comparison diagram between model predictions and the corresponding experimental values of oil viscosity (v_o).

function of linear regression in characteristic space is as follows:

$$f(x) = \langle w, x \rangle + b, \tag{3}$$

where f is the linear function and w and b are the identified weight vector and the bias term, respectively.

In the high-dimensional feature space, the optimization problem for SVR with ε-insensitive loss function is as follows:

$$\min \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{N} (\xi_i + \zeta_i)$$

$$\text{s.t.} \quad (w^T x_i + b) - y_i \le \varepsilon + \xi_i \tag{4}$$

$$y_i - (w^T x_i + b) \le \varepsilon + \zeta_i$$

$$\xi_i, \zeta_i \ge 0, \quad i = 1, 2, \cdots, N,$$

where $\|w\|^2$ in the objective function is the confidence range reflecting the generalization ability, ξ_i and ζ_i are the slack variables that represent the upper and lower limits of allowable error, $\sum_{i=1}^{N}(\xi_i + \zeta_i)$ denotes the experimental risk reflecting the learning capacity of function, $\varepsilon > 0$ is an insensitive loss coefficient, and parameter $C(C \ge 0)$ is a penalty factor. In SVR, the dual problem of Equation (4) is often derived by using the Lagrange multiplier method, based on which a linear regression function can finally be constructed.

2.2. Statistical Evaluation of Three Models. To evaluate the accuracy of machine learning models, three statistical indi-

TABLE 3: The prediction calculation time of different models.

Methods	Prediction time
RF	0.53 s
DNN	0.82 s
SVR	0.76 s
TDM	52.8 s

TABLE 4: Statistical error analysis results calculated by RF model.

Parameter	Value
Training set	
R^2	0.9921
MAE	0.0021
MSE	$9.942E-5$
Testing set	
R^2	0.9816
MAE	0.0035
MSE	0.0003
Total	
R^2	0.9811
MAE	0.0023
MSE	0.0001

cators were used, including mean square error (MSE), mean absolute error (MAE), and coefficient of determination (R^2) [26–28]. This experiment also utilizes these indicators to
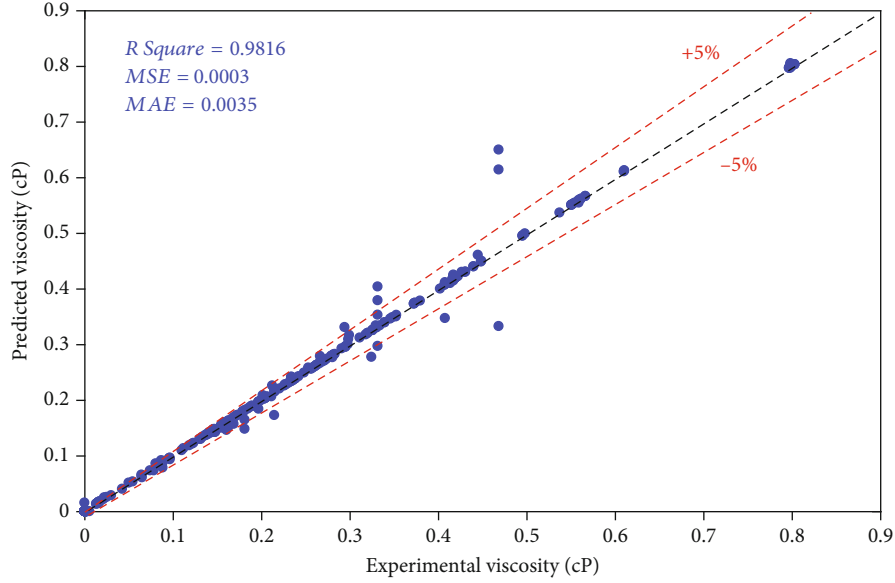
FIGURE 5: The cross-plot figure of the proposed RF model.

calculate the error and test the robustness of the machine learning model.

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} \left(X_i^{\text{data}} - X_i^{\text{model}} \right)^2, \tag{5}$$

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} \left| X_i^{\text{real}} - X_i^{\text{mod el}} \right|, \tag{6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \left(X_i^{\text{data}} - X_i^{\text{model}} \right)^2}{\sum_{i=1}^{N} \left(X_i^{\text{data}} - \text{average} \left(X_i^{\text{model}} \right) \right)^2}, \tag{7}$$

where N represents the total number of samples, X_i^{data} is the referenced parameter that is the actual expected value, and X_i^{model} represents the predicted value of the machine learning methods.

3. Results and Discussion

3.1. Comparison of Three Proposed Models. To evaluate the precision of each constructed network, MSE, MAE, and R^2 are calculated, based on different output data of gas viscosity (v_g), oil viscosity (v_o), and water viscosity (v_w), respectively. The calculated evaluation results are presented in Table 2.

According to the results of Table 2, the proper prediction model for viscosity modeling is RF. Take the oil viscosity modeling as an example; among these three prediction models, the presented RF model has the lowest MSE of 0.008, lowest MAE of 0.0093, and highest R^2 value of 0.9623. Compared to the other two developed models, the SVR model has the worst prediction results.

To reveal and visualize the performance of each prediction model, a cross-plot picture between model predictions and the corresponding experimental values is drawn in Figure 3, for gas viscosity (v_g), oil viscosity (v_o), and water
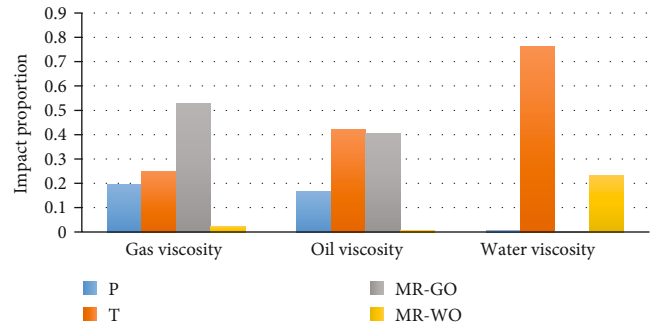

FIGURE 6: Importance analysis of influencing factors on the viscosity of multiphase gas-oil-water systems.

viscosity (v_w), respectively. In this curve, there are three color points: blue points for estimated data of RF, red points for estimated data of SVR, and yellow points for estimated data of DNN, and $Y = X$ line for experimental values. In a cross-plot picture, a higher precision is attained when the data is closer to the $Y = X$ line. From Figures 3(a)–3(c), in all prediction results for gas viscosity (v_g), oil viscosity (v_o), and water viscosity (v_w), only in the RF model, there is an adequate closeness of the majority of the data points to the line $Y = X$, showing a very good agreement between model predictions and the corresponding experimental values. For the RF model, the coefficient of determination (R^2) is high to 0.9824, 0.9623, and 0.9999 for gas viscosity (v_g), oil viscosity (v_o), and water viscosity (v_w), respectively, which is much higher than the other models.

Moreover, in Figure 4, model predictions and the corresponding experimental values of oil viscosity (v_o) by three proposed models are also compared and depicted. From these results, the predicted results using the RF model are very close to the corresponding experimental results. It can be seen that the RF model achieves a good accuracy of

predicting the oil viscosity (v_o) of the multiphase oil-gas-water systems.

Finally, taking 1200 sets of data as an example, the prediction calculation time of machine learning models was compared with the tradition numerical method (TDM), as shown in Table 3. In this paper, the experiment is based on TensorFlow and Sk-learn learning library with Python language. The hardware resources include Intel i7-7700hq@2.8G processor, 16 G memory, and Nvidia GTX 1060 (6 G) graphics card. Table 3 indicates that machine learning models only need 0.53 s, 0.82 s, and 0.76 s for the prediction calculation of RF, DNN, and SVR, respectively. Prediction time by machine learning models is tremendously fast, while for the tradition numerical method, the prediction time is high up to 52.8 s. Machine learning models, therefore, have advantages in terms of computing time.

3.2. Results Analysis by RF Model.

Statistical evaluation parameters of the best network RF model for training, testing, and total data sets are presented in Table 4. Table 4 reports that for training set, R^2 of the RF model is high to 0.9921, while MAE and MSE are very low with the values of 0.0021 and $9.942E - 5$. For testing set, the R^2, MAE, and MSE values are 0.9816, 0.0035, and 0.0003. While for a total set, the corresponding statistical evaluation results are 0.9811, 0.0023, and 0.0001, respectively, indicating good evaluation parameters for the RF model.

Cross-plot figure of the proposed RF model is also shown in Figure 5. It could be seen that there is a dense point distribution around the $Y = X$ line for all of the data, and the error is basically maintained within 5%, indicating the sufficient accuracy and reliability of the developed RF model.

3.3. Importance Analysis.

Importance analysis of influencing factors was carried out to figure out a sensitivity analysis on viscosity in a multiphase gas-oil-water system. For each viscosity, such as gas, oil, and water, each of the independent influencing factors, such as P, T, MR-GO, and MR-WO, was evaluated in this part. The results of the importance analysis are indicated in Figure 6. For each output result, the sum of the influence proportion of the influencing factors (P, T, MR-GO, and MR-WO) is 1. The higher the impact proportion value, the stronger the relationship between the input parameter and the output function.

As it is expected, for gas viscosity, the MR-GO is the most significant factor influencing the output result with an impact proportion of 0.53. The second is environmental factors, including P and T, and the proportions are 0.20 and 0.25, respectively. Here, the water content has a litter effect. For oil viscosity, the output value is mainly affected by ambient temperature T and gas content MR-GO, and the specific impact proportion is 0.42 and 0.40, respectively. Secondly, pressure P also plays an important role with the affecting proportion of 0.17. Finally, the viscosity of water in the multiphase system is mainly determined by the ambient temperature T, and the influence proportion is 0.76, followed by its content MR-WO in the system, and the ratio is 0.23.

4. Conclusions

In this study, three machine learning models, namely, random forest (RF), deep neural network (DNN), and support vector regression (SVR), were proposed for calculating and predicting the phase viscosity of multiphase reservoir oils systems.

To make a judgment of the accuracy of each developed model, various statistical evaluation indicators, including mean square error (MSE), mean absolute error (MAE), and coefficient of determination (R^2), were applied. The results show that the RF model has higher accuracy compared with the other two models. Specifically, for the RF model, the R^2, MAE, and MSE values are 0.9811, 0.0023, and 0.0001, respectively, indicating good evaluation performance. Moreover, machine learning models have advantages in computing time for the RF model, which only needs 0.53 s for 1200 sets of data prediction.

Moreover, an importance analysis of influencing factors was carried out on viscosity in this multiphase gas-oil-water system. For gas viscosity, the MR-GO is the most significant influencing factor with an impact proportion of 0.53, followed by P and T. Next, oil viscosity is mainly affected by T and MR-WO, and P also has an effect. Finally, for water, the viscosity is mainly determined by T, followed by MR-WO.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

All authors confirm that there is no financial/personal interest or belief that could affect our objectivity, and no conflicts exist.

Acknowledgments

References

[1] A. Hemmati-Sarapardeh, A. Shokrollahi, A. Tatar, F. Gharagheizi, A. H. Mohammadi, and A. Naseri, "Reservoir oil viscosity determination using a rigorous approach," *Fuel*, vol. 116, no. 1, pp. 39–48, 2014.

[2] H. Chen, "Oil-gas-water system study in fine reservoir description researches—taking Yulou oil-bearing sets in the West Depression in Liaohe Basin in China as an example," *Arabian Journal of Geosciences*, vol. 14, no. 4, pp. 1–16, 2021.

[3] L. Chen, Z. Liu, and N. Ma, "Optimize production allocation for the oil-gas field basing on a novel grey model," *Journal of Natural Gas Geoscience*, vol. 4, no. 2, pp. 121–128, 2019.

[4] S. K. Matthai, "Reservoir simulation: mathematical techniques in oil recovery," *Geofluids*, vol. 8, no. 4, 345 pages, 2008.

[5] B. Calvignac, E. Rodier, J. J. Letourneau, P. Vitoux, C. Aymonier, and J. Fages, "Development of an improved

falling ball viscometer for high-pressure measurements with supercritical CO_2," *The Journal of Supercritical Fluids*, vol. 55, no. 1, pp. 96–106, 2010.

[6] S. Swamura, "An improvement of high-pressure rolling-ball viscometer and evaluation of the equipment constant," *Journal of the Society of Materials Science Japan*, vol. 63, no. 10, pp. 729–734, 2014.

[7] I. A. Munz, B. W. D. Yardley, and S. A. Gleeson, "Petroleum infiltration of high-grade basement, South Norway: pressure-temperature-time-composition (P–T–t–X) constraints," *Geofluids*, vol. 2, no. 1, 53 pages, 2002.

[8] C.-Y. Sun G.-J. Chen et al., "Measurement of interfacial tension for the CO_2 injected crude oil + reservoir water system," *Journal of Chemical & Engineering Data*, vol. 50, no. 3, pp. 936–938, 2005.

[9] R. A. Almehaideb, I. Ashour, and K. A. El-Fattah, "Improved K-value correlation for UAE crude oil components at high pressures using PVT laboratory data," *Fuel*, vol. 82, no. 9, pp. 1057–1065, 2003.

[10] M. G. Rezk and J. Foroozesh, "Phase behavior and fluid interactions of a CO_2-light oil system at high pressures and temperatures," *Heliyon*, vol. 5, no. 7, p. e02057, 2019.

[11] V. A. Tarasov, M. A. Komkov, V. A. Romanenkov, R. V. Boyarskaya, and D. S. Boyarskii, "A rheological model of the influence of thermal exposure on the viscosity of a an LBS-4 varnish during impregnation of a multilayer woven blank of a heat-protective coating," *Polymer Science Series D*, vol. 11, no. 2, pp. 179–182, 2018.

[12] S. Khosharay, M. Pierantozzi, and G. Di Nicola, "Modeling investigation on the viscosity of pure refrigerants and their liquid mixtures by using the Patel–Teja viscosity equation of state," *International Journal of Refrigeration*, vol. 85, pp. 255–267, 2018.

[13] G. Feng, C. Shuang, and X. Guo, "Comparison and analysis of the viscosity forecasting models of reservoir fluids," *Petroleum Geology and Recovery Efficiency*, vol. 5, 2007.

[14] H. Wang, J. G. Wang, F. Gao, and X. Wang, "A two-phase flowback model for multiscale diffusion and flow in fractured shale gas reservoirs," *Geofluids*, vol. 2018, Article ID 5910437, 15 pages, 2018.

[15] J. G. Heuer, C. L. Jacocks, and Inventors; Conoco Inc., "Control of gas-oil ratio in producing wells," *U.S. Patent 3,368,624*, vol. 13, 1968.

[16] K. S. Bajaalah, "New correlations for gas oil ratio and oil formation volume factor for gas condensate reservoirs," *International Journal of Innovation and Applied Studies*, vol. 13, no. 4, 2015.

[17] M. Daviran, A. Maghsoudi, R. Ghezelbash, and B. Pradhan, "A new strategy for spatial predictive mapping of mineral prospectivity: automated hyperparameter tuning of random forest approach," *Computers & Geosciences*, vol. 148, 2021.

[18] T. K. Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Montreal, QC, Canada, 1995.

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

[20] M. Hauser, S. Gunn, S. Saab, and A. Ray, "State-space representations of deep neural networks," *Neural Computation*, vol. 31, no. 3, pp. 538–554, 2019.

[21] "Deep neural networks-based classification optimization by reducing the feature dimensionality with the variants of gravitational search algorithm," *International Journal of Modern Physics C*, vol. 32, no. 10, p. 2150137, 2021.

[22] A. C. Serfidan, F. Uzman, and M. Tuerkay, "Optimal estimation of physical properties of the products of an atmospheric distillation column using support vector regression," *Computers & Chemical Engineering*, vol. 134, article 106711, 2020.

[23] L. Wang, Y. Ma, X. Chang, C. Gao, Q. Qu, and X. Chen, "Projection wavelet weighted twin support vector regression for OFDM system channel estimation," *Artificial Intelligence Review*, vol. 54, no. 1, pp. 469–489, 2021.

[24] M. Asadolahi, M. G. Akbari, G. Hesamian, and M. Arefi, "A robust support vector regression with exact predictors and fuzzy responses," *International Journal of Approximate Reasoning*, vol. 132, pp. 206–225, 2021.

[25] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[26] R. M. Khalifa, S. Yacout, and S. Bassetto, "Developing machine-learning regression model with logical analysis of data (LAD)," *Computers & Industrial Engineering*, vol. 151, article 106947, 2021.

[27] B. Kumi-Boateng and M. S. Peprah, "Modelling local geometric geoid using soft computing and classical techniques: a case study of the University of Mines and Technology (UMaT) local geodetic reference network," *International Journal of Earth Sciences*, vol. 2, no. 3, pp. 166–177, 2020.

[28] H. Gong, R. Pishgar, and J. H. Tay, "Artificial neural network modelling for organic and total nitrogen removal of aerobic granulation under steady-state condition," *Environmental Technology*, vol. 40, no. 24, pp. 3124–3139, 2019.