Petroleum Research 11 (2026) 838–847

Contents lists available at ScienceDirect

Petroleum Research

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

Full Length Article

Application of random forest model optimized with genetic algorithm for determination of minimum miscibility pressure in CO_2 flooding

Siyi Zheng, Mengping Liu, Yuan Zhang *, Jinghong Hu

School of Energy Resources, China University of Geoscience (Beijing), Beijing 100083, China

ARTICLE INFO

Article history:
Received 1 September 2025
Received in revised form
24 January 2026
Accepted 26 January 2026
Available online 30 January 2026

Keywords:
Random forest
Genetic algorithm
Minimum miscibility pressure
Sensitivity analysis
CO_2-EOR

ABSTRACT

CO_2 flooding is an efficient technique for enhancing oil recovery and carbon sequestration, where accurate estimation of the minimum miscibility pressure (MMP) is crucial. Conventional experiment and numerical simulations are time-cost and often inaccurate. Therefore, it is essential to develop an efficient and accurate model for MMP prediction. In this work, we introduced a genetic algorithm-optimized random forest (GA-RF) model using 260 samples with eight input features, including reservoir temperature (T_R), molecular weight of C_{5+} pseudo-components (MWC_{5+}), ratio of the mole percent of volatile to intermediate fractions (vol./int.), and mole fractions of injected gas (CO_2, H_2S, CH_4, C_2–C_4, and N_2). Four algorithms-random forest (RF), multilayer perceptron (MLP), support vector regression (SVR), and gradient boosting decision tree (GBDT) are compared. RF performs best and is further optimized by GA. The resulting GA-RF model outperforms traditional methods, with prediction accuracy of 0.99 and a root mean square error of 0.74. The model was applied to CO_2-hydrocarbons system from different low-permeability reservoirs in China. Results show that the GA-RF model demonstrated an average relative error of 3.43%. These findings highlight that the GA-RF model has great potentials for predicting MMP during CO_2 flooding and improving CO_2 miscible flooding efficiency.

© 2026 Chinese Petroleum Society. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

With increasing global energy demand, CO_2 flooding has been widely adopted as an enhanced oil recovery (EOR) technique due to its high displacement efficiency and cost-effectiveness (Du and Nojabaei, 2019; Louk et al., 2017; Song et al., 2025). Chen et al. (2022) systematically investigated the interaction between displacement efficiency and miscibility degree in CO_2 enhanced oil recovery processes under different pore scales by analyzing relative permeability curves under three distinct miscibility states. Additionally, it contributes to carbon capture and storage, playing a role in climate change mitigation (Tian et al., 2020). The fluid state critically affects crude oil development, with MMP being a key parameter in CO_2 flooding (Liu et al., 2025b). The success of CO_2 flooding techniques is largely dependent on accurate estimates of the MMP, a critical parameter in designing CO_2 miscible flooding determines whether CO_2 can mix with crude oil. Previous research has indicated that reservoir temperature, pressure, and the composition of the injected gas are key parameters influencing MMP (Silva and Orr, 1987). Accurately predicting the MMP is essential. It helps control production costs, improves displacement performance, and is crucial for both theoretical research and field applications. A comprehensive understanding of the parameters that influence MMP, together with refining the CO_2-based enhanced oil recovery process, is essential for improving oil and gas recovery efficiency and promoting environmental sustainability (Oyinloye et al., 2021).

MMP during CO_2 flooding is primarily studied through theoretical models and experimental methods (Dindoruk et al., 2021). Experimental methods such as slim-tube testing, rising bubble apparatus, vapor density analysis, and vanishing interfacial tension measurements are widely used (Johns et al., 2002). Liu et al. (2025a) conducted high-temperature and high-pressure NMR displacement tests in Block L of Jiangsu Oilfield's ultra-low

* Corresponding author. 29#, Xue-Yuan Road, Haidian District, Beijing 100083, China.
E-mail address: yuanzhangpku@hotmail.com (Y. Zhang).
Peer review under the responsibility of Editorial Office of Petroleum Research.

https://doi.org/10.1016/j.ptlrs.2026.01.015
2096-2495/© 2026 Chinese Petroleum Society. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Nomenclature

MMP	minimum miscibility pressure
T_R	reservoir temperature
MWC_{5+}	molecular weight of C_{5+} pseudo-components
vol./int	the ratio of the mole percent of volatile to the intermediate fractions
EOR	enhanced oil recovery
AI	artificial intelligence
ML	machine learning
DL	deep learning
RF	random forest
GA	genetic algorithm
MLP	multilayer perceptron
SVR	support vector regression
GBDT	gradient boosting decision tree
LSSVM	least squares support vector machine
RBFNN	the radial basis function neural network
FNN	the feedforward neural network
ANN	the artificial neural network
SVM	the support vector machine
ANFIS	the adaptive neuro-fuzzy inference systems
PSO	the particle swarm optimization
AdaBoost-CART	Adaptive Boosting with Classification and Regression Trees
GRNN	generalized regression neural network
R^2	coefficient of determination
APRE	the average percentage relative error, %
AAPRE	the average absolute percentage relative error, %
RMSE	the root mean square error
SD	the standard deviation
$H_{(x)}$	the random forest model
$H_{i(x)}$	the *i*th decision tree
X	the sample feature vector
Y	the sample class label
I	the indicator function
N	the total number of data points
y_{iexp}	the experimental or actual values
y_{iped}	the predicted values

permeability reservoir, revealing the mechanism of CO_2-enhanced mass transfer in such reservoirs. Though these experiments can obtain accurate MMP predictions, they are time-consuming, expensive, and sometimes lack of high-precision apparatus and repeatability (Wu and Batycky, 1990; Christiansen and Haines, 1987; Rao, 1997; Ayirala et al., 2003; Franklin and Kristian, 2007). Theoretical approaches encompass empirical relationships, equation of state methods, tie line analysis calculation methods, numerical simulation methods, and artificial intelligence (AI) prediction methods (Ghorbani et al., 2014). Most empirical relationships are only applicable under strict conditions and often produce large errors when applied to diverse reservoir conditions. Equation of state involve computationally intensive thermodynamic calculations (Ashrafizadeh and Ghasrodashti, 2011; Chen et al., 2020). Similarly, numerical simulations demand substantial experimental data in the initial stages to fit PVT properties and the computations involved are resource-intensive.

With the rapid development of AI, intelligent models are increasingly used to predict MMP. AI techniques are generally categorized into machine learning (ML) and deep learning (DL) (Song et al., 2024). Compared to conventional experimental or numerical methods, AI-based approaches offer improved computational efficiency and predictive accuracy (Akbari et al., 2025). Alatefi et al. (2025a, 2025b; 2025c) conducted a series of studies using interpretable AI models, including developing efficient analytical tools for designing and optimizing deep eutectic solvents (DES), predicting the solubility of CO_2–N_2 mixtures in brine, evaluating their impact on impurity-containing CO_2 storage in saline aquifers, and achieving precise predictions of water activity in ternary ionic liquid systems. In the field of oil and gas development research, Shokrollahi et al. developed a least squares support vector machine (LSSVM) model for MMP prediction in both pure and impure CO_2-crude oil systems, achieving an R^2 of 0.90 (Shokrollahi et al., 2013). However, the model lacked input sensitivity analysis and interpretability. Tatar et al. and Hemmati-Sarapardeh et al. proposed radial basis function neural network (RBFNN) and feedforward neural network (FNN) models, respectively, identifying reservoir temperature as the most influential factor on MMP (Hemmati-Sarapardeh et al., 2016; Tatar et al., 2013). Zhang et al. compared the performance of random forest (RF), support vector regression (SVR), and artificial neural network (ANN) models using 170 experimental datasets (Zhang et al., 2020). The RF model demonstrated superior accuracy in both training and testing phases. Alatefi et al. (2024) developed data-driven models to predict CO_2 minimum miscibility pressure (MMP), which could reduce laboratory costs and deliver rapid and accurate predictions.

To enhance model performance further, researchers have integrated intelligent optimization algorithms (Mohammadian et al., 2022; Riazi and Türker, 2018). The Genetic Algorithm (GA) is particularly prevalent due to its effectiveness in solving complex nonlinear problems. Emera and Sarma (2005) were among the first to apply GA for MMP prediction (Emera and Sarma, 2005). Bian et al. developed an SVM model based on 150 experimental samples and improved its accuracy for both pure and impure CO_2-crude oil systems through GA optimization, although the mixed CO_2 system's prediction accuracy was limited to 0.92 (Bian et al., 2016). Shakeel et al. proposed a hybrid forecasting model combining ANN and adaptive neuro-fuzzy inference systems (ANFIS), incorporating ten key variables influencing MMP (Shakeel et al., 2023). However, the study did not disclose the dataset size. Sayyad et al. used particle swarm optimization (PSO) to optimize ANN weights and biases, significantly reducing the mean absolute error (Sayyad et al., 2014).

Stacking ensemble models have been gained significant traction in MMP prediction due to their enhanced generalization capability achieved by integrating multiple base learners. Ghiasi et al. developed an AdaBoost-CART model using 94 datasets to predict MMP during CO_2 flooding, demonstrating superior accuracy compared to individual base models (Ghiasi et al., 2021). Hassan et al. proposed an ensemble framework combining ANN, RBFNN, GRNN, and fuzzy logic systems, trained on 105 experimental samples to estimate MMP in CO_2-crude oil systems (Hassan et al., 2019). However, the study noted that complex models like ANN are prone to overfitting when trained on small datasets, which undermines model stability and predictive reliability (Altman, 1992).

Current studies reveal that the experiments suffer from prolonged cycles and high costs, especially for tight or shale formations. Numerical methods exhibit sensitivity to fluid properties and prohibitively high computational costs. Additionally, they are constrained by the applicability of formulas, and the models

sometimes have poor performance under complex reservoir conditions. Although recent advancements in machine learning have introduced new perspectives, model performance remains heavily dependent on hyperparameter optimization. Conventional grid search or manual parameter tuning are inefficient and fails to guarantee global optimal solutions. Furthermore, most rely on small sample datasets with simplistic data segmentation, limited evaluation metrics, and insufficient validation. All these lead to inconsistent results and potential overfitting. Therefore, in this work, a comprehensive MMP prediction database is constructed using both experimental and field data. Four machine learning models-RF, SVR, MLP, and GBDT-are systematically evaluated for predictive accuracy and generalization. The best-performing model is further optimized using a -GA, and its performance is rigorously validated through multiple techniques. A sensitivity analysis is then conducted to examine the influence of eight key parameters, including reservoir temperature, crude oil composition, and injected gas composition, on MMP. Finally, the proposed models are compared with existing AI-based approaches. The results show notable improvements in both computational efficiency and prediction accuracy. This study innovatively uses genetic algorithms as an efficient global optimization tool and combines it with the random forest model to achieve automatic and precise optimization of key hyperparameters, overcoming the limitations of traditional manual parameter tuning. Meanwhile, by collecting a self-built dataset covering a wide range of reservoir conditions for training and testing, the GA-RF model was applied to the CO_2-hydrocarbon system in China's low-permeability reservoirs, further verifying its robustness. This study provides a robust theoretical foundation and valuable practical insights into CO_2 flooding and enhanced oil recovery.

2. Method and process

2.1. Data selection and process

In this section, we build a dataset from 15 literatures (Shokrollahi et al., 2013; Tatar et al., 2013; Hemmati-Sarapardeh et al., 2016; Shakeel et al., 2023; Sayyad et al., 2014; Ghiasi et al., 2021; Dicharry et al., 1973; Harmon and Grigg, 1988; Thakur et al., 1984; Metcalfe, 1982; Dong et al., 2001; Rathmell et al., 1971; Kamari et al., 2015; Karkevandi-Talkhooncheh et al., 2017; Olukoga and Feng, 2022). The characteristic parameters of the dataset include the T_R, ranging from 22.67 °C to 127.00 °C. the MWC_{5+}, ranging from 136.47 g/mol to 465.59 g/mol. The vol./int., ranging from 0 to 21.1. The mole fraction of the injected gas, which include y_{CO_2}、 y_{H_2S}、 y_{CH_4}、 $y_{C_2-C_4}$、 y_{N_2}. Additionally, the dataset includes the MMP for both pure CO_2 and impure CO_2 systems (6.5–41.15 MPa). After eliminating redundant records, 260 MMP-related data points were obtained, including 156 instances corresponding to pure CO_2 and 104 associated with impure CO_2. The parameter ranges in the dataset are summarized in Table 1.

To mitigate potential bias from fixed data ordering, a randomized auxiliary column was appended to the dataset by scaling the target variable with uniform random values. This column was included solely for structural and shuffling purposes and was not used in model training or prediction. Additionally, to address scale disparities among input features that are common in crude oil composition datasets, all variables were normalized to a common range using Min-Max normalization (Kassambara, 2017). This preprocessing step improves model generalization by ensuring balanced feature contribution and reducing magnitude-induced bias during learning.

Table 1
Statistical description of the database.

	CO_2 Gas Stream Composition					T_R (°C)	Oil Composition		MMP(MPa)
	CO_2 (%)	H_2S (%)	C_1 (%)	C_2–C_4 (%)	N_2 (%)		MWC_{5+}(g/mol)	vol./int.	
MAX	100	0	0	0	0	22.67	136.47	0	6.5
MIN	40	50	20	25	8.8	127.00	465.59	21.10	41.45
MEAN	70	25	1	12.5	4.4	149.67	301.03	10.55	23.98
STD	13.54	10.73	5.22	4.17	0.73	25.44	44.80	2.38	6.75

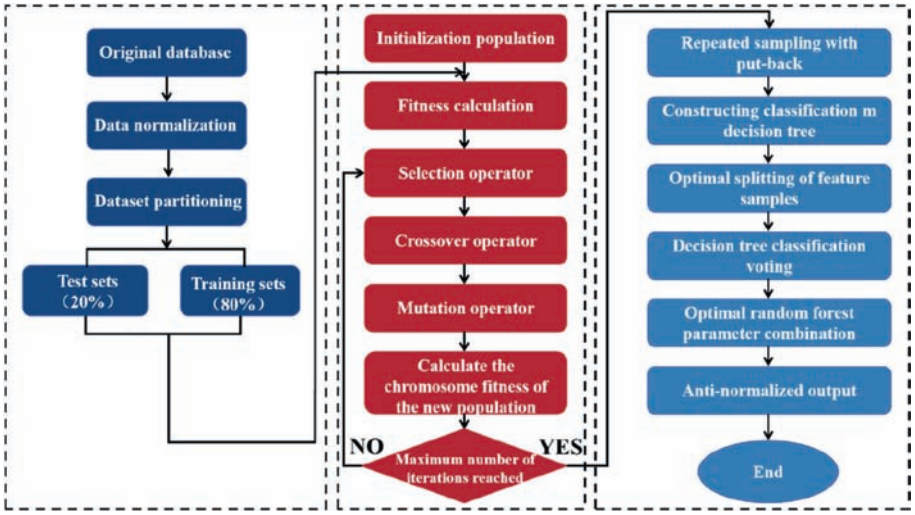


Fig. 1. Flow chart of GA-RF prediction model.

Table 2
Average results of 10-cross-validation.

Model	R^2	AAPRE%	RMSE	SD
RF	0.96	5.13	5.13	5.13
SVR	0.88	8.79	1.98	1.98
MLP	0.89	8.89	2.18	2.19
GBDT	0.84	10.01	2.67	2.31

2.2. Random forest (RF)

RF is an ensemble learning algorithm proposed by Breiman (Breiman, 2001). Its basic building block is a decision tree, and two types of randomness are introduced during model construction: bootstrap sampling of the dataset and random selection of input features. These random processes result in the generation of multiple diverse decision trees. The predicted class is obtained through a majority voting mechanism. The final output aggregates the predictions from all trees in the forest, thereby improving predictive accuracy and mitigating the risk of overfitting. RF is capable of performing both classification and regression tasks, with the primary difference lying in the method used to aggregate the outputs of individual trees. In addition, RF have the important capability of evaluating feature importance (Rodriguez-Galiano et al., 2012; Strobl et al., 2007).

For classification problems, each tree in Random Forest acts as an independent classifier, and the final prediction is determined by majority voting among the individual tree outputs.

$$H(x) = \operatorname{argmax} \sum_{i=1}^{N} I(h_i(x) = y) \quad (1)$$

For regression tasks, the prediction is the average of outputs from all decision trees.

$$H(x) = \frac{1}{N} \sum_{i=1}^{N} h_i(x) \quad (2)$$

Where $H_{(x)}$ is the random forest model, $H_{i(x)}$ represents the ith decision tree, X refers to the sample feature vector, Y denotes the sample class label, and I is the indicator function.

2.3. GA-RF model process

The GA is inspired by Darwinian evolution, simulating natural selection and survival of the fittest (Holland, 1975). It functions as a parallel and adaptive iterative algorithm, mimicking the genetic evolution of organisms. Individuals, or chromosomes, are encoded as binary strings. Starting from an initial population, the algorithm iteratively applies selection, crossover, and mutation operations. Selection retains promising solutions, crossover generates new combinations, and mutation introduces diversity. Through successive generations, the population evolves toward better fitness and adaptability. GA is inherently random and adaptive, with its core mechanism being the continuous refinement of solutions through simulated evolution. As a global optimization technique, the GA model effectively addresses the limitations of manual parameter tuning in RF, enhancing both prediction accuracy and generalization capability. The flowchart of the GA-RF optimization algorithm is shown in Fig. 1, and the specific steps are as follows:

(1) A total of 260 datasets were normalized to reduce parameter scale effects.
(2) Data were split into training (80%) and test (20%) sets, with 10-fold cross-validation applied to the training set.
(3) Three RF hyperparameters, max_d, n_trees, and min_leaf, were optimized using GA.

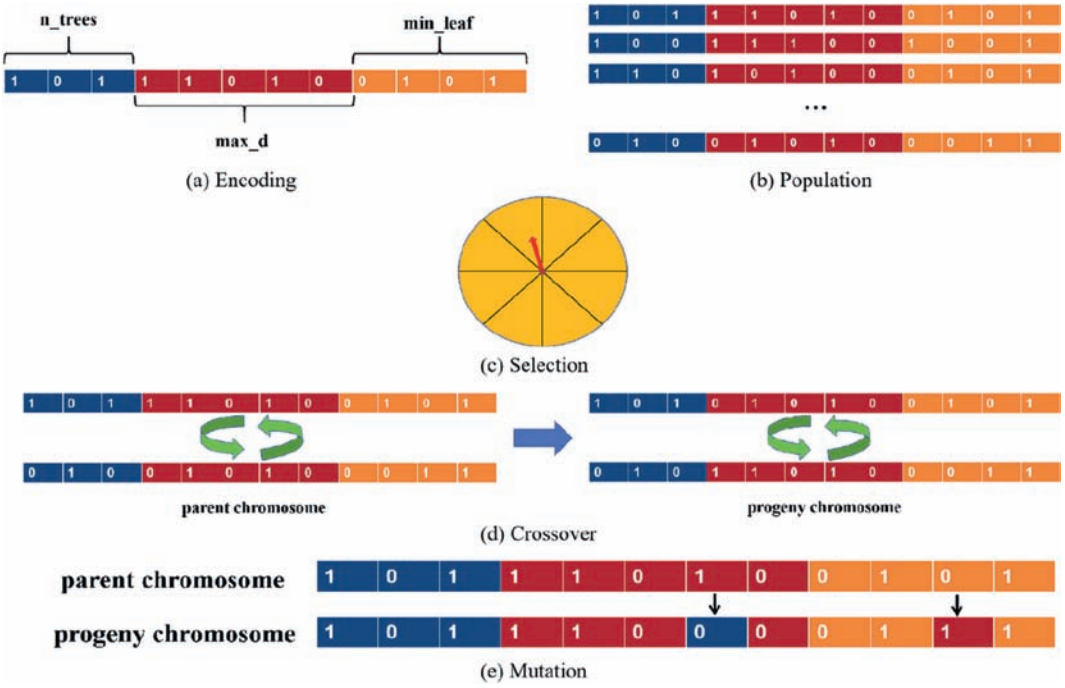


Fig. 2. Schematic diagram of genetic coding processing of random forest parameters.

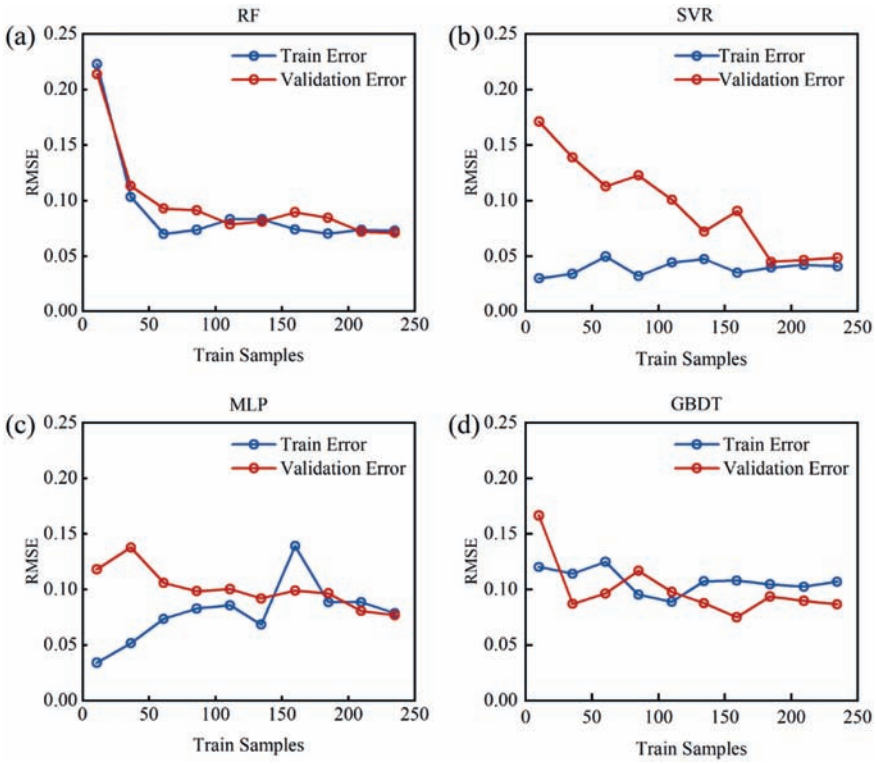


Fig. 3. Learning curves of training RMSE.

(4) Hyperparameters were encoded as binary chromosomes; an initial population was generated randomly, as shown in Fig. 2(a)–(b).
(5) Selection was done via roulette wheel, and crossover with a probability of 0.8 and mutation with a probability of 0.001 were applied to ensure diversity, as shown in Fig. 2(c)–(e).
(6) Fitness was evaluated to evolve the population; the algorithm stopped when convergence or iteration limits were reached.

2.4. Model evaluation

The evaluation metrics included the average percentage relative error (APRE), average absolute percentage relative error (AAPRE), root mean square error (RMSE), standard deviation (SD), and determination coefficient (R^2). Among these, R^2 and APRE are key indicators for evaluating the overall performance of the model. The R^2 value approaching to 1 signifies higher predictive accuracy, whereas an RMSE value of 0 indicates a perfect model fit. These metrics are employed to quantify the model's predictive capabilities, with their specific calculations detailed in Equations (3)–(7):

$$APRE = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{y_{i\,\mathrm{exp}} - y_{iped}}{y_{i\,\mathrm{exp}}}\right] \tag{3}$$

$$AAPRE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_{i\,\mathrm{exp}} - y_{iped}}{y_{i\,\mathrm{exp}}}\right| \tag{4}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_{i\,\mathrm{exp}} - y_{iped}\right)^2} \tag{5}$$

$$SD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{y_{i\,\mathrm{exp}} - y_{iped}}{y_{i\,\mathrm{exp}}}\right)^2} \tag{6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_{i\,\mathrm{exp}} - y_{iped}\right)^2}{\sum_{i=1}^{N}\left(y_{iped} - y_{i\,\mathrm{exp}}\right)^2} \tag{7}$$

Where N represents the total number of data points, y_{iexp} denotes the experimental or actual values, y_{iped} refers to the predicted values.

3. Result and discussion

3.1. Basic model selection

Table 2 presents the 10-fold cross-validation results for four machine learning models: RF, SVR, MLP, and GBDT (Vapnik, 1995; Friedman, 2001). This validation method reduces the randomness caused by data partitioning and provides a more reliable evaluation of model performance. Among the models, RF achieved the best results, with an AAPRE of 5.13%, RMSE and SD of 1.35, and an R^2 of 0.96, indicating strong generalization and robustness. In contrast, MLP, SVR, and GBDT showed relatively lower performance. Based on overall metrics, the ranking is: RF > MLP > SVR > GBDT.

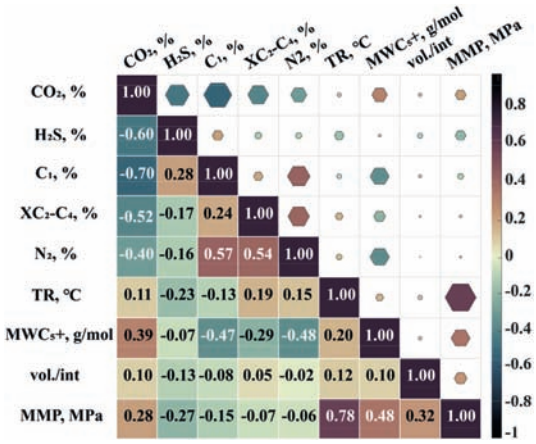


Fig. 4. Relevancy factor for input variables while estimating MMP.

Fig. 3 displays the RMSE learning curves for each model on the training set, used to assess learning ability and overfitting risk. The RF model shows a rapid RMSE drop and convergence in both training and validation sets as sample size increases, indicating stable MMP prediction once key features are captured. This reflects RF's robustness against high-dimensional and noisy data, leading to consistently strong performance across random seeds. SVR maintains a stable validation error without clear overfitting, but slightly underperforms RF, likely due to sensitivity to kernel selection and its limitations in capturing complex nonlinearities. MLP displays clear overfitting at small sample sizes, with large validation error fluctuations, possibly due to dependence on initial weights and architecture. As the data increases, performance improves. GBDT performs well on training data but suffers from unstable validation results. Overall, RF shows superior and stable convergence and is thus chosen for further optimization.

3.2. Correlation analysis

To analyze the influence of input parameters on MMP, Spearman's rank correlation analysis was performed, as shown in Fig. 4. The mole fractions of H_2S, C_1, N_2, and C_2–C_4 in the gas phase show negative correlations with MMP (Kendall, 1938), indicating that increases in these components reduce MMP. Higher CO_2 content enhances interactions between CO_2 and crude oil, promoting miscibility and further lowering MMP. In contrast, the remaining four parameters show positive correlations, suggesting that their increase leads to higher MMP. Among all variables, temperature exhibits the strongest impact. Elevated reservoir temperatures reduce CO_2 solubility in crude oil, hindering miscibility and thus raising MMP. Similarly, higher molecular weight of the C_{5+} fraction impairs mixing, increasing MMP as molecular weight rises. Additionally, N_2, a non-polar gas, decreases CO_2 solubility under certain pressures, enlarging the two-phase region and complicating multiple contact miscibility. These results align with established phase behavior principles and validate the model's interpretability.

3.3. Model performance

To assess the predictive performance of the model, 52 sets of previously untrained MMP data were used to test the trained model. The predictive performance of the RF model and the GA-RF model was evaluated by comparing the differences between their predicted values and the experimental values. The results show that the RF model achieves an AAPRE of 5.13 %, an APRE of 1.21 %, an RMSE and a SD of 1.35. After optimization, the GA-RF model exhibits enhanced performance, with an AAPRE of 4.47 %, an APRE of −0.57 %, an RMSE of 0.74, and a SD of 0.06. Compared to the standalone RF model, the optimized GA-RF model demonstrates significantly reduction in error metrics. This indicates that the GA algorithm effectively improves the accuracy of intelligent prediction models.

The intersection results between the predicted MMP values and the experimental values for both the RF model and the GA-RF model are presented in Fig. 5. The findings indicate that R^2 of the RF model is 0.96 (Fig. 5(a)), while that of the GA-RF model increases to 0.99 (Fig. 5(b)), reflecting an improvement of 3.15%. This demonstrates that GA-optimized model provides a significant enhancement in the correlation coefficient. Additionally, the experimental data and predicted values align closely with the diagonal line (x = y), illustrating the high predictive accuracy of the GA-RF model.

3.3.1. Model superiority to correlations

To evaluate the performance of the GA-RF intelligent model in predicting MMP, we compare the prediction with six empirical correlation models (Metcalfe, 1982; Alston et al., 1985; Orr et al. ,1984). Fig. 6 displays the prediction accuracy of these models. As shown, significant discrepancies between the predicted and actual values for both pure and impure CO_2 injections, with many data points significantly deviating from the ideal x = y line. Large deviations also illustrate the limitations of empirical correlations, particularly with complex injected gas and oil components.

3.3.2. Model superiority to ML

To further comprehensively evaluate the performance of proposed model, in this study, we compared with existing machine

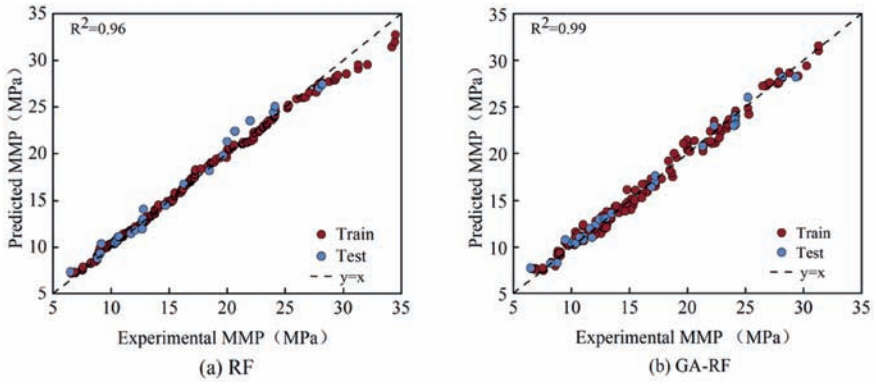


Fig. 5. Comparison between the results of the RF and GA-RF model for all data set.

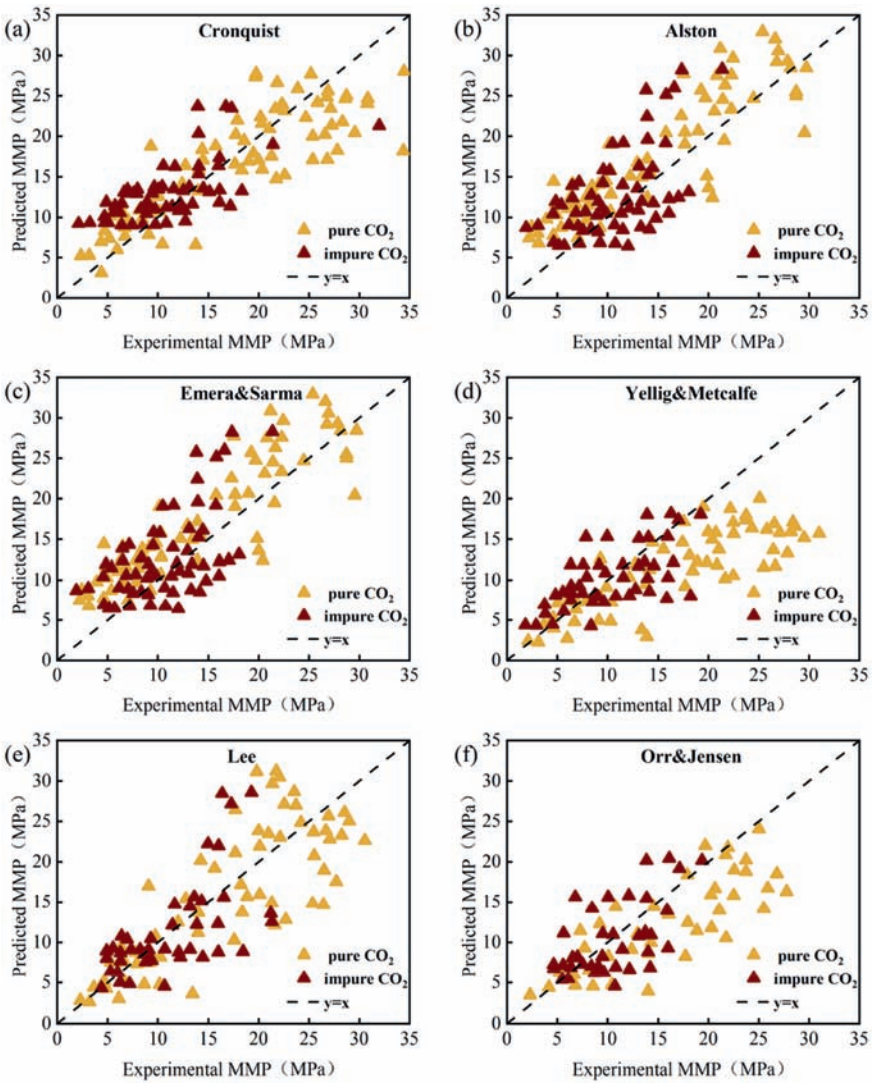


Fig. 6. Intersection diagram of prediction results of different types of models.

learning models. Since most models are not explicitly presented, their practical application is not feasible. Consequently, this study employs the statistical error metrics for the published models for evaluation. The results of 4 statistical error indexes are summarized in Table 3, demonstrating that the developed GA-RF model performs better than the existing models.

3.3.3. Model verification

To assess the generalization capability of the GA-RF model, an independent dataset of 13 samples was used for external validation. These samples were collected from the Jilin, Changqing, Xinjiang, and Tuha oilfields (Liao et al., 2014). Table 4 summarizes the compositions and properties of the crude oils. The volatile content ranges from 11.48% to 44.48%, while the intermediate fraction varies from 2.76% to 17.33%. Notably, the mole ratio of volatile to intermediate components (vol./int.) exceeds 1 in all cases, indicating a predominance of light oil characteristics. The formation temperatures of the samples vary widely, from 45 °C to 108.4 °C. Several reservoirs in this study have higher formation temperatures than typical low-permeability reservoirs, which generally range between 30 °C and 90 °C.

The data shown in Table 4 are then fed into the optimized model, which leverages its parallel ensemble learning capability to deliver highly accurate minimum miscibility pressure predictions. The low time costs significantly enhance decision-making speed for project design. Furthermore, the reliable predictions are directly applied to the design and optimization of CO_2 injection, enabling engineers to scientifically determine the injection pressure and gas slug sizing. Therefore, CO_2 miscible or immiscible flooding can be effectively designed for enhanced oil recovery.

Fig. 7 presents a comparison between the predicted and experimental MMP values for 13 crude oil samples using the GA-RF model. As shown, the predicted values closely match the experimental data, and the overall trend is well captured. This indicates that the model exhibits strong generalization capability and high predictive accuracy. The RMSE of 2.4 MPa further confirms that the model meets the required precision level for practical applications.

3.4. Feature importance

Fig. 8 presents the relative importance of the eight input variables in the GA-RF model for MMP prediction. The results show

Table 3
Comparison of the predictive performances of developed models with machine learning models.

Model Sources		Performance			
Author	ML method	R^2	RMSE	AAPRE (%)	AAPE (%)
Shokrollahi et al., 2013	LSSVM	0.903	1.69	9.6	/
Sinha et al., 2021	SVR	/	/	/	5.00
Hassan et al., 2019	RBF	0.96	/	/	8.67
	ANN	0.95	/	/	7.61
	GRNN	0.87	/	/	14.49
	Fuzzy Logic	0.85	/	/	12.73
Huang et al., 2003	ANN	0.987	/	/	5.91
Zhang et al., 2020	RF	0.91	/	/	5.00
	SVM	0.88	/	/	/
	ANN	0.874	/	/	/
Bian et al., 2016	GA-SVR	0.978	0.79	/	4.75
Kamari et al., 2015	GEP	0.828	2.26	/	10.47
Karkevandi-Talkhooncheh et al., 2017	ANIFS-BP	/	1.98	10.87	−3.85
	ANIFS-PSO	/	1.68	7.53	−1.36
	ANIFS-GA	/	1.61	8.10	−0.85
	ANIFS-ACO	/	1.87	9.16	−2.18
	ANIFS-DE	/	2.03	11.49	−1.19
Ghiasi et al., 2021	AdaBoost-CART	0.99	/	/	0.39
	Hybrid-ANFIS	0.93	/	/	6.90
This work	GA-RF	0.99	0.74	4.47	0.06

Table 4
Oil compositions and MMP_{pure} of the representative low permeability oil reservoirs in China.

Samples	T_R/°C	MWC_{5+}/(g mol^{-1})	Vol./%	Int./%	vol./int.	MMP_{pure}/MPa	GA-RF MMP_{pred}/MPa
1	98.9	248.2	18.71	11.78	1.59	22.30	23.51
2	97.3	265.9	19.01	11.97	1.59	22.10	22.77
3	108.4	277.5	26.82	6.80	3.94	27.90	27.85
4	76	259.0	25.35	8.74	2.90	20.20	19.44
5	83.9	288.8	11.48	2.76	4.16	20.30	20.84
6	112	271.6	14.33	5.68	2.52	22.90	22.77
7	71	227.3	13.71	6.66	2.05	16.70	16.39
8	74.8	237.2	36.33	9.76	3.72	26.10	26.21
9	45	350.3	19.22	2.95	6.51	29.80	29.04
10	53	272.4	42.75	7.12	6.00	19.10	19.82
11	62.2	249.3	30.35	4.33	7.00	21.60	20.88
12	92	224.9	26.67	17.33	1.53	19.80	19.36
13	71.6	233.4	18.79	15.65	1.20	19.50	19.16

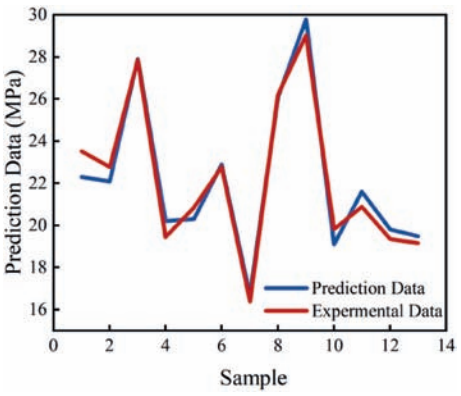


Fig. 7. External validation results.

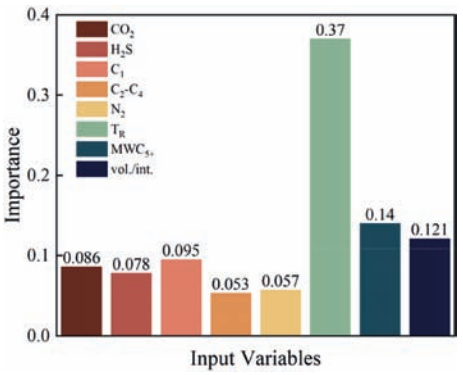


Fig. 8. Sensitivity analysis for the impact of independent variables on MMP prediction.

that T_R has the highest importance, with a relative importance score of 0.37. This value is significantly higher than those of the other features, indicating that T_R plays a dominant role in controlling the miscibility process between CO_2 and crude oil. The MWC_{5+} and vol./int. follow in importance. In contrast, C_2–C_4 hydrocarbons and N_2 exhibit the lowest importance scores in the model. The overall importance ranking of the eight variables is as follows: T_R > MW_{C5+}>vol./int> y_{CH_4} >y_{CO_2} > y_{H_2S} >y_{N_2} >$y_{C_2-C_4}$.

4. Conclusions

In this study, a novel GA-optimized-RF model is proposed for predicting MMP in CO_2-oil systems, different injected components are included. The key findings of this research are as follows:

(1) The model is built using a dataset of 260 data points for training and validation. Under robust cross-validation, the RF

model outperformed MLP, SVR, and GBDT, demonstrating its superior predictive capability and suitability for further optimization.

(2) The GA-RF model achieves a prediction accuracy of 0.99 and an RMSE of 0.74, which is more accurate than the empirical correlations and other machine learning models examined. It also reduces manual parameter tuning and lowers computational time costs.

(3) Sensitivity analysis identified key parameters affecting MMP, with reservoir temperature (TR) and heavy crude oil components (MWC_{5+}) showing significant positive correlations with MMP. The importance of the eight variables affecting MMP can be ranked as follows: $T_R > MW_{C5+}$>vol./int> $y_{CO_2} > y_{H_2S} > y_{CH_4} > y_{C_2-C_4} > y_{N_2}$.

(4) In practical applications, the GA-RF model has a better performance in the low-permeability reservoirs in China, achieving an RMSE of only 2.4 MPa.

(5) The performance of the proposed GA-RF model is dependent upon training data quality, necessitating future development of an expanded, standardized MMP database to improve generalizability. In the future, this intelligent framework will be extended for other gas injection methods and key reservoir parameters, establishing an integrated predictive toolchain.

CRediT authorship contribution statement

Siyi Zheng: Writing – original draft, Validation. **Mengping Liu:** Writing – original draft, Data curation. **Yuan Zhang:** Writing – review & editing, Supervision. **Jinghong Hu:** Methodology, Conceptualization.

Declaration of competing interest

The authors report there are no competing interests to declare.

Acknowledgements

This work was supported by National Natural Science Foundation of China (52474052), Young Elite Scientists Sponsorship Program by BAST (BYESS2023414), and Scientific Research Innovation Capability Support Project for Young Faculty (ZYGXQNJSKYCXNLZCXM-E14).

References

Alatefi, S., Agwu, O.E., Amar, M.N., Djema, H., 2025a. Explainable artificial intelligence models for estimating the heat capacity of deep eutectic solvents. Fuel 394, 135073.

Alatefi, S., Amar, M.N., Agwu, O.E., Alkouh, A., 2025b. Toward explicit learning frameworks for predicting the solubility of CO_2–N_2 gas mixtures in brine: implication for impure CO_2 storage in saline aquifers. J. Contam. Hydrol. 274, 104660.

Alatefi, S., Amar, M.N., Agwu, O.E., Alkouh, A., 2025c. Accurate prediction of water activity in ionic liquid-based aqueous ternary solutions using advanced explainable artificial intelligence frameworks. Chem. Eng. Sci. 318, 122218.

Alatefi, S., Agwu, O.E., Azim, R.A., Dzulkarnain, I., 2024. Development of multiple explicit data-driven models for accurate prediction of CO_2 minimum miscibility pressure. Chem. Eng. Res. Des. 205, 672–694.

Alston, R.B., Kokolis, G.P., James, C.F., 1985. CO_2 minimum miscibility pressure: a correlation for impure CO_2 streams and live oil systems. SPE J. 25 (2), 268–274.

Altman, N.S., 1992. An introduction to kernel and nearest-neighbor nonparametric regression. Am. Statistician 46 (3), 175–185.

Akbari, A., Ranjbar, A., Kazemzadeh, Y., Mohammadinia, F., Borhani, A., 2025. Estimation of minimum miscible pressure in carbon dioxide gas injection using machine learning methods. J. Pet. Explor. Prod. Technol. 15 (2), 25.

Ayirala, S.C., Rao, D.N., Casteel, J., 2003. Comparison of minimum miscibility pressures determined from gas-oil interfacial tension measurements with equation of state calculations. In: Proceedings of the SPE Annual Technical Conference and Exhibition.

Ashrafizadeh, S.N., Ghasrodashti, A.A., 2011. An investigation on the applicability of parachor model for the prediction of MMP using five equations of state. Chem. Eng. Res. Des. 89 (6), 690–696.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Bian, X.Q., Han, B., Du, Z.M., 2016. Integrating support vector regression with genetic algorithm for CO_2-Oil minimum miscibility pressure (MMP) in pure and impure CO_2 streams. Fuel 182, 550–557.

Christiansen, R.L., Haines, H., 1987. Rapid measurement of minimum miscibility pressure with the rising-bubble apparatus. SPE Reserv. Eng. 2 (4), 523–527.

Chen, H., Li, B., Duncan, I., Elkhider, M., 2020. Empirical correlations for prediction of minimum miscible pressure and near miscible pressure interval for oil and CO_2 systems. Fuel 278, 118272.

Chen, H., Liu, X., Zhang, C., Tan, X., Yang, R., Yang, S., Yang, J., 2022. Effects of miscible degree and pore scale on seepage characteristics of unconventional reservoirs fluids due to supercritical CO_2 injection. Energy 239, 122287.

Dicharry, R.M., Perryman, T.L., Ronquille, J.D., 1973. Evaluation and design of a CO_2 miscible flood Project-SACROC unit, kelly-snyder field. J. Petrol. Technol. 25 (11), 1309–1318.

Dong, M., Huang, S., Dyer, S.B., 2001. A comparison of CO_2 minimum miscibility pressure determinations for weyburn crude oil. J. Petrol. Sci. Eng. 31, 13–22.

Du, F., Nojabaei, B., 2019. A review of gas injection in shale reservoirs: enhanced oil/gas recovery approaches and greenhouse gas control. Energies 12 (12), 2355.

Dindoruk, B., Johns, R.T., Orr, F.M., 2021. Measurement and modeling of minimum miscibility pressure: a state-of-the-art review. SPE Reservoir Eval. Eng. 24 (2), 367–389.

Emera, M., Sarma, H., 2005. Use of genetic algorithm to estimate CO_2-Oil minimum miscibility Pressure-A key parameter in design of CO_2 miscible flood. J. Petrol. Sci. Eng. 46 (1–2), 37–52.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29 (5), 1189–1232.

Franklin, M.O.B., Kristian, J., 2007. An analysis of the vanishing interfacial tension technique for determination of minimum miscibility pressure. Fluid Phase Equilib. 255 (2), 99–109.

Ghorbani, M., Momeni, A., Safavi, S., 2014. Modified Vanishing Interfacial Tension (VIT) test for CO_2-Oil minimum miscibility pressure (MMP) measurement. J. Nat. Gas Sci. Eng. 20, 92–98.

Ghiasi, M.M., Mohammadi, A.H., Zendehboudi, S., 2021. Use of Hybrid-ANFIS and ensemble methods to calculate minimum miscibility pressure of CO_2-Reservoir oil system in miscible flooding process. J. Mol. Liq. 331, 115369.

Holland, J.H., 1975. Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor, MI.

Harmon, R.A., Grigg, R.B., 1988. Vapor-density measurement for estimating minimum miscibility pressure. SPE Reserv. Eng. 3 (4), 1215–1220.

Huang, Y., Huang, G., Dong, M., 2003. Development of an artificial neural network model for predicting minimum miscibility pressure in CO_2 flooding. J. Petrol. Sci. Eng. 37 (1–2), 83–95.

Hemmati-Sarapardeh, A., Ghazanfari, M.H., Ayatollahi, S., Masihi, M., 2016. Accurate determination of the CO_2-Crude oil minimum miscibility pressure of pure and impure CO_2 streams: a robust modelling approach. Can. J. Chem. Eng. 94 (2), 253–261.

Hassan, A., Elkatatny, S., Abdulraheem, A., 2019. Intelligent prediction of Minimum Miscibility Pressure (MMP) during CO_2 flooding using artificial intelligence techniques. Sustainability 11 (24), 7020.

Johns, R., Sah, P., Solano, R., 2002. Effect of dispersion on local displacement efficiency for multicomponent enriched-gas floods above the minimum miscibility enrichment. SPE Reservoir Eval. Eng. 5 (1), 4–10.

Kendall, M.G., 1938. A new measure of rank correlation. Biometrika 30 (1–2), 81–93.

Kamari, A., Arabloo, M., Shokrollahi, A., 2015. Rapid method to estimate the Minimum Miscibility Pressure (MMP) in live reservoir oil systems during CO_2 flooding. Fuel 153, 310–319.

Karkevandi-Talkhooncheh, A., Hajirezaie, S., Hemmati-Sarapardeh, A., 2017. Application of adaptive neuro-fuzzy interface system optimized with evolutionary algorithms for modeling CO_2-Crude oil minimum miscibility pressure. Fuel 205, 34–45.

Kassambara, A., 2017. Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning. Sthda.

Liao, C., Liao, X., Chen, J., 2014. Correlations of minimum miscibility pressure for pure and impure CO_2 in low permeability oil reservoirs. J. Energy Inst. 87 (3), 208–214.

Liu, X., Chen, H., Yang, L., Zhu, Y., Liao, H., Zhao, Q., Zhou, X., Zeng, H., 2025a. Oil production characteristics and CO_2 storage mechanisms of CO_2 flooding in ultra-low permeability sandstone oil reservoirs. Petrol. Explor. Dev. 52 (1), 196–207.

Liu, X., Chen, H., Cheng, W., Li, Y., Zhao, Y., Zhu, Y., Zeng, H., 2025b. Occurrence states and transport behavior of crude oil in different permeability oil reservoirs during depletion development. Geoenergy Sci. Eng., 213944

Louk, K., Ripepi, N., Luxbacher, K., Gilliland, E., Tang, X., Keles, C., Schlosser, C., Diminick, E., Keim, S., Amante, J., 2017. Monitoring CO_2 storage and enhanced gas recovery in unconventional shale reservoirs: results from the Morgan County, Tennessee injection test. J. Nat. Gas Sci. Eng. 45, 11–25.

Metcalfe, R.S., 1982. Effects of impurities on minimum miscibility pressures and minimum enrichment levels for CO_2 and rich-gas displacements. SPE J. 22 (2), 219–225.

Mohammadian, E., Liu, B., Riazi, A., 2022. Evaluation of Different Machine Learning

Frameworks to Estimate CO_2 Solubility in Nacl Brines: Implications for CO_2 Injection into Low-Salinity Formations. *Lithosphere*.
Oyinloye, O., Darmaki, N.A., Zarooni, M.A., Boukadi, F., Nantongo, H., 2021. Estimation of minimum miscibility pressure for flue gas injection using soft experimentations. Nat. Resour. 12 (11), 363–380.
Olukoga, T.A., Feng, Y., 2022. Determination of miscible CO_2 flooding analogue projects with machine learning. J. Petrol. Sci. Eng. 209, 109826.
Rathmell, J.J., Stalkup, F.I., Hassinger, R.C., 1971. A laboratory investigation of miscible displacement by CO_2. In: Proceedings of the SPE 1971 Annual Fall Meeting.
Rao, D.N., 1997. A new technique of vanishing interfacial tension for miscibility determination. Fluid Phase Equilib. 139, 311–324.
Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol-Sanchez, J.P., 2012. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J. Photogramm. 67, 93–104.
Riazi, A., Türker, U., 2018. A genetic algorithm-based search space splitting pattern and its application in hydraulic and coastal engineering problems. Neural Comput. Appl. 30 (12), 3603–3612.
Silva, M.K., Orr, F.M., 1987. Effect of oil composition on minimum miscibility pressure part 1: solubility of hydrocarbons in dense CO_2. SPE Reserv. Eng. 2 (4), 468–478.
Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinf. 8 (1), 25.
Sayyad, H., Manshad, A.K., Rostami, H., 2014. Application of hybrid neural particle swarm optimization algorithm for prediction of MMP. Fuel 116, 625–633.
Shokrollahi, A., Arabloo, M., Gharagheizi, F., 2013. Intelligent model for prediction of CO_2-Reservoir oil minimum miscibility pressure. Fuel 112, 375–384.
Sinha, U., Dindoruk, B., Soliman, M., 2021. Prediction of CO_2 minimum miscibility pressure using an augmented machine learning-based model. SPE J. 26 (4), 1666–1678.
Shakeel, M., Khan, M.R., Kalam, S., 2023. Machine learning for prediction of CO_2 minimum miscibility pressure. In: Proceedings of the SPE Annual Technical Conference and Exhibition.
Song, G., Meng, Y., Zhang, C., Zhao, Z., Yang, Q., 2024. Comprehensive review of the determination and reduction of the minimum miscibility pressure during CO_2 flooding. ACS Omega 9 (13), 14747–14765.
Song, Y., Song, Z., Mo, Y., Zhou, Q., Jing, Y., Chen, F., Tian, S., Chen, Z., 2025. Determination of minimum miscibility and near-miscibility pressures for CO_2-oil mixtures in shale reservoirs. Fuel 388, 134531.
Thakur, G.C., Lin, C.J., Patel, Y.R., 1984. CO_2 minitest, little knife field, ND: a case history. In: *Proceedings of* the SPE/DOE Fourth Symposium on Enhanced Oil Recovery.
Tatar, A., Shokrollahi, A., Mesbah, M., 2013. Implementing radial basis function networks for modeling CO_2-Reservoir oil minimum miscibility pressure. J. Nat. Gas Sci. Eng. 15, 82–92.
Tian, Y., Ju, B., Yang, Y., Wang, H., Dong, Y., Liu, N., Ma, S., Yu, J., 2020. Estimation of minimum miscibility pressure during CO_2 flooding in hydrocarbon reservoirs using an optimized neural network. Energy Explor. Exploit. 38 (6), 2485–2506.
Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer, New York.
Wu, R.S., Batycky, J.P., 1990. Evaluation of miscibility from slim tube tests. J. Can. Petrol. Technol. 29 (6), 63–70.
Zhang, J., Zhang, X., Dong, S., 2020. Estimation of crude oil minimum miscibility pressure during CO_2 flooding: a comparative study of random forest, support vector machine, and back propagation neural network. In: Proceedings of the 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference.