Contents lists available at ScienceDirect

Petroleum Science

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

Original Paper

Machine learning methods for predicting CO_2 solubility in hydrocarbons

Yi Yang [a, b, c], Binshan Ju [a, b, c, *], Guangzhong Lü [d], Yingsong Huang [d]

[a] School of Energy Resources, China University of Geosciences (Beijing), Beijing, 100083, China
[b] Key Laboratory of Marine Reservoir Evolution and Hydrocarbon Enrichment Mechanism, Ministry of Education, Beijing, 100083, China
[c] Key Laboratory of Geological Evaluation and Development Engineering of Unconventional Natural Gas Energy, Beijing, 100083, China
[d] Shengli Oilfield Company, SINOPEC, Dongying, 257015, Shandong, China

ARTICLE INFO

Article history:
Received 29 October 2023
Received in revised form
30 April 2024
Accepted 30 April 2024
Available online 3 May 2024

Edited by Yan-Hua Sun

Keywords:
CO_2 solubility
Machine learning
Support vector regression
Extreme gradient boosting
Random forest
Multi-layer perceptron

ABSTRACT

The application of carbon dioxide (CO_2) in enhanced oil recovery (EOR) has increased significantly, in which CO_2 solubility in oil is a key parameter in predicting CO_2 flooding performance. Hydrocarbons are the major constituents of oil, thus the focus of this work lies in investigating the solubility of CO_2 in hydrocarbons. However, current experimental measurements are time-consuming, and equations of state can be computationally complex. To address these challenges, we developed an artificial intelligence-based model to predict the solubility of CO_2 in hydrocarbons under varying conditions of temperature, pressure, molecular weight, and density. Using experimental data from previous studies, we trained and predicted the solubility using four machine learning models: support vector regression (SVR), extreme gradient boosting (XGBoost), random forest (RF), and multilayer perceptron (MLP). Among four models, the XGBoost model has the best predictive performance, with an R^2 of 0.9838. Additionally, sensitivity analysis and evaluation of the relative impacts of each input parameter indicate that the prediction of CO_2 solubility in hydrocarbons is most sensitive to pressure. Furthermore, our trained model was compared with existing models, demonstrating higher accuracy and applicability of our model. The developed machine learning-based model provides a more efficient and accurate approach for predicting CO_2 solubility in hydrocarbons, which may contribute to the advancement of CO_2-related applications in the petroleum industry.

1. Introduction

Global warming is mainly related to the emission of greenhouse gases, especially carbon dioxide (CO_2) (Solomon et al., 2009). Geological carbon storage has been assessed as a potential technique to mitigate global warming and climate change problems (Aftab et al., 2022; Aslannezhad et al., 2023; Hassanpouryouzband et al., 2021). Oil and gas reservoirs are favorable places for geological storage due to the known reservoir information and the existing infrastructure for CO_2 injection (Kovscek and Cakici, 2005). Due to the increasing applications of CO_2 in enhanced oil recovery and geological sequestration (Ezekiel et al., 2020; Uliasz-Misiak et al., 2021; You et al., 2021), the study of CO_2 solubility in oil has become an important research area.

Currently, there are various methods available for studying the solubility of CO_2 in hydrocarbon systems. However, some methods are time-consuming, such as the pressure drop method (Li et al., 2009), while others can disrupt the system equilibrium during the sampling process, like the equilibrium liquid sampling analysis method (Leu and Robinson, 1987). Additionally, certain methods require extensive data recording and complex calculations using equations of state, such as the gas PVT (pressure–volume–temperature) measurement method (Shah et al., 1991). The challenges in measuring CO_2 solubility in hydrocarbon lie in determining the equilibrium point and analyzing the CO_2 mass. Researchers typically rely on the changes in temperature, pressure, or bubble point (pressure at which the gas phase disappears) before and after the addition of CO_2 to determine the equilibrium point (Wang et al., 2016).

* Corresponding author. School of Energy Resources, China University of Geosciences (Beijing), Beijing, 100083, China.
E-mail address: jubs2936@163.com (B. Ju).

https://doi.org/10.1016/j.petsci.2024.04.018

Welker (1963) conducted experimental measurements to investigate the solubility of CO_2 in several types of oil. They explored the variations in CO_2 solubility with respect to temperature, pressure, and crude oil composition. The results indicated that CO_2 exhibits high solubility in oil, with the solubility decreasing with increasing temperature and increasing with increasing pressure. Additionally, due to the expansion of CO_2 upon dissolution in oil, the experiments also measured the expansion coefficient of the solution, observing its variation with solubility. It was also found that gases are more easily dissolved in light oil. Simon and Graue (1965) proposed correlations for predicting the physical properties of CO_2–oil mixtures, such as the solubility of CO_2 in crude oil systems. The average deviation in solubility estimation was reported to be 2%. Furthermore, the properties of oil also influence the solubility of CO_2, such as in different hydrocarbon systems, where the solubility of CO_2 decreases with an increase in the carbon number of hydrocarbon compounds (Wang et al., 2018). Indeed, hydrocarbons are the primary components of oil. Therefore, when studying the solubility of CO_2 in oil, the focus is primarily on investigating the solubility of CO_2 in hydrocarbons.

The solubility model of CO_2 in hydrocarbons is essentially based on the fundamental principles of thermodynamics (Michelsen, 1990; Wei and Sadus, 2010). It establishes the relationship between observable macroscopic quantities such as pressure, temperature, concentration, and CO_2 solubility when the homogeneous substance system is in thermodynamic equilibrium. By utilizing equations of state, the fugacity of the gas and liquid phases is calculated, and a predictive model for CO_2 solubility is established. Mehrotra and Svrcek (1985) introduced correlations to predict the solubility and various physical properties of pure CO_2 and other gases in bitumen, considering pressure and temperature as variables. Chung et al. (1988) provided predictive tools for determining diverse physical properties, including CO_2 solubility in heavy oils. These correlations utilize only temperature, pressure, and oil specific gravity values for predicting CO_2 solubility and estimating other physical properties. Xue et al. (2005) comprehensively considered various factors affecting gas solubility and derived theoretical equations for the molar solubility of gas in crude oil and the gas–oil ratio. Emera and Sarma (2007) developed correlations using the genetic algorithm (GA) technique to predict CO_2 solubility and other physical properties of CO_2–oil mixtures for both dead and live oils. Table S1 in Supplementary Information summarizes several widely used empirical correlations for calculation of CO_2 solubility in oil.

Machine learning (ML) technology is a data modeling tool that is increasingly used in the oil and gas industry for its ability to discover complex relationships between inputs and outputs in the absence of theoretical models. The solubility of carbon dioxide in hydrocarbons has been regarded as a complex process and cannot be predicted precisely. In this case, it is better to use computer methods to represent this complex relationship (Rostami et al., 2017, 2018).

Machine learning techniques have been applied to reservoir oil and gas properties estimation (Fazavi et al., 2014), retention and solubility predictions (Ali Ahmadi and Ahmadi, 2016; Kamari et al., 2014; Safari et al., 2014). Mehraein and Riahi (2017) combined the least squares support vector machines (LSSVM) technique with genetic algorithm (GA) multi-layer regression to predict the solubility of CO_2 in ionic liquids (ILs). Yamaguchi et al. (2023) conducted multiscale numerical simulations of CO_2 hydrate storage using two types of neural networks. However, to the best of the authors' knowledge, there are currently relatively few published articles utilizing multiple artificial intelligence methods to simulate the solubility of CO_2 in hydrocarbons. Moreover, the applicable range is also relatively narrow. It is urged to establish a broader model of the solubility of CO_2 in hydrocarbons, in order to provide more references for the study of CO_2 solubility. Temperature (T), pressure (P), molecular weight (M), and density (ρ) are fundamental physical parameters that influence solubility, thus CO_2 solubility in hydrocarbons is affected by these parameters. Furthermore, extensive experimental data from previous studies indicate a correlation between these four parameters and the solubility of CO_2 in hydrocarbons. These parameters are relatively easy to obtain through experimental measurements, and there is already a large amount of data available for modeling purposes. Therefore, we have chosen these four parameters as inputs for the model in this study. The four machine learning models used in this study were support vector regression (SVR), extreme gradient boosting (XGBoost), random forest (RF), and multiple-layers perceptron (MLP). To ensure the accuracy of the results, the hyperparameters for each model were precisely defined, and the results were interpreted and discussed to identify the best predictive model for the dataset.

2. Methodology

2.1. Data set

In this study, a large amount of data on the solubility of CO_2 in hydrocarbons was collected, comprising 2212 datasets (Table S2 in Supplementary Information) from 37 literature sources. After collecting the data, the data is first checked for missing values, corrected for missing values, outliers, and duplicate values. Then, standardization is carried out to transform the data into a range with similar scales, ensuring that different features have equal weighting in the model. The pressures ranged from 0.093 to 40.85 MPa, temperatures ranged from 252.67 to 594.2 K, hydrocarbon molecular weights ranged from 44 to 619.19 g/mol, and hydrocarbon densities ranged from 500 to 919.2 kg/m^3. The data and experimental conditions for each literature source are presented in Table S3 in Supplementary Information. As shown in Fig. 1, the relationship coefficient evaluation between the input parameters and the output parameter (Sol).

2.2. Machine learning algorithms

In this study, four well-established modeling approaches were employed: support vector regression (SVR), extreme gradient

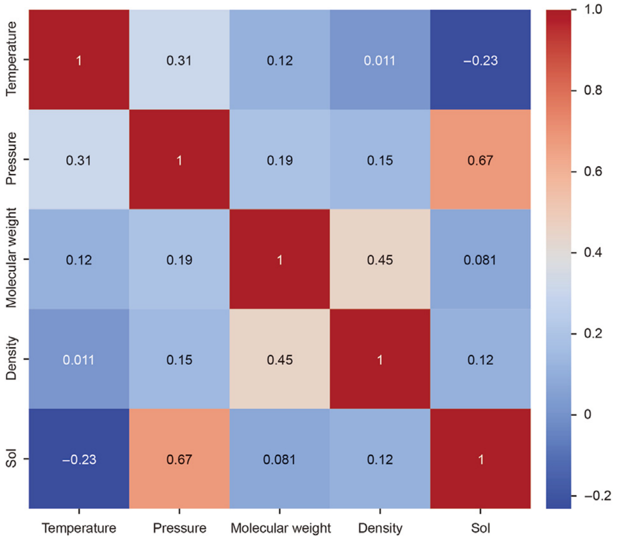

Fig. 1. Heat map implying the correlation between input and output variables.

boosting (XGBoost), random forest (RF), and multiple-layers perceptron (MLP). These approaches are well-known and widely used in machine learning. Four inputs were considered in the model including pressure, temperature, molecular weight, and density, and the CO_2 solubility was considered as the sole predicted output.

2.2.1. Support vector regression (SVR)

Support vector machine (SVM) is a type of supervised learning algorithm that has demonstrated strong predictive performance and stability in both classification and regression tasks. SVM is used for binary classification, and their basic model is a linear classifier that maximizes the distance between the decision boundary and the closest data points in the feature space. The objective of SVM is to find the hyperplane that can best separate the two classes. In the SVM algorithm the margin between the data points and the hyper plane is tried to be maximized (Drucker et al., 1996). A commonly used function to maximize the margin is the radial basis function (RBF) expressed as a gaussian function:

$$K\left(\boldsymbol{x_i}, \boldsymbol{x_j}\right) = \exp\left(- \gamma \| \boldsymbol{x_i} - \boldsymbol{x_j} \|^2\right) \tag{1}$$

where $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ are the feature vectors of two samples; and γ is a parameter controlling the similarity between samples in a higher-dimensional space.

The penalty factor C is a regularization parameter, controlling the trade-off between maximizing the margin and minimizing the classification error:

$$Objective\ function = \frac{1}{2} \| \boldsymbol{w} \|^2 + C \sum\nolimits_{i=1}^{n} \xi_i \tag{2}$$

where $\boldsymbol{w}$ is the normal vector to the hyperplane; ξ_i are slack variables; and C is a tuning parameter balancing the trade-off between margin width and misclassification penalty. Adjusting C allows for flexibility in handling the complexity of the model and the severity of misclassification penalties (Iskandarov et al., 2021).

2.2.2. Extreme gradient boost (XGBoost)

XGBoost is an algorithm based on gradient boosting trees, where the core idea is to "use multiple weak learners to construct a strong learner by gradually optimizing the loss function" (Liang et al., 2021; McCallum et al., 2021; Zhang et al., 2020). Specifically, each weak learner is a decision tree model, and XGBoost employs a customized loss function that simultaneously considers the magnitude of errors and the complexity during the construction of each tree. Additionally, XGBoost utilizes regularization techniques, namely $L1$ and $L2$ regularization, to prevent overfitting.

During each iteration, XGBoost calculates the gradient and Hessian matrix for each sample, which are used to build the decision tree. Then, based on the gradient and Hessian matrix of the loss function, the algorithm computes the split gain for each node to determine which feature and threshold will minimize the loss function (Sutton, 2005). Finally, a new decision tree is generated using a greedy algorithm to select the split points. After multiple iterations, XGBoost combines multiple decision trees to form a strong learner.

2.2.3. Random forest (RF)

Random forest is an algorithm that combines a number of decision trees (DT). It can provide fast and accurate results despite being simple and having only a few parameters to be tuned. It operates a number of DT to do the same tasks, while their outputs are aggregated into the final prediction.

The most important tuning parameters of the algorithm are the number of trees and the minimum samples leaf (Tin Kam, 1998). In this study, five tuning hyperparameter was used, including "n_estimators", "max_depth", "max_features", "min_samples_split", and "min_samples_leaf".

2.2.4. Multilayer perceptron (MLP)

The MLP neural network is a type of feedforward network composed of an input layer, hidden layers, and an output layer (Agirre-Basurko et al., 2006). Each neuron is connected to all neurons in the previous layer, and each connection has a weight. The training of the MLP network utilizes the backpropagation algorithm to adjust the weights and biases in order to minimize prediction errors. The MLP neural network makes predictions by passing input data through the hidden layers and then forwarding the output of the hidden layers to the output layer. There can be multiple hidden layers, and each hidden layer may have a different number of neurons. Neurons apply an activation function to transform input signals into output signals (Ture et al., 2005; Yin et al., 2022).

2.3. Workflow of developing ML models

Assessing the performance of machine learning (ML) models is a crucial stage in the development of robust models. Firstly, preprocess the data, including correcting missing values, handling outliers, and removing duplicates, then normalize the data to make it comparable. In supervised learning, a portion of the available data is utilized for training the ML algorithm, while the remaining data serves as a testing set to evaluate the accuracy of the predictive model. To mitigate potential bias stemming from a specific random split, it is common practice to explore multiple training or testing splits. The flowchart of the ML models employed in this study encompasses several steps, as illustrated in Fig. 2.

Machine learning (ML) methods usually have numerous hyperparameters (their importances are shown in Table S4 in Supplementary Information) for training models, but only a few need to be carefully selected to optimize model performance. Improper selection of hyperparameters can lead to the issues of underfitting or overfitting. In this study, the grid search technique (GridSearchCV) was employed to identify the best hyperparameters for each machine learning method.

Grid search involves systematically adjusting parameters within a specified range to find the optimal hyperparameters. These hyperparameters are then used to train the model, which is subsequently evaluated on a validation set to determine the parameter set that yields the highest accuracy. This process involves comparing different combinations of arguments through cross-validation.

Finally, the accuracy of each model was assessed using several statistical metrics: r-squared (R^2), mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE). We also compare the selected model with some previous models, and take the average absolute relative deviation (AARD) value as the metrics. These metrics provide quantitative measures of model performance and are expressed as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(out_i^{\text{real}} - out_i^{\text{predicted}} \right)^2}{\sum_{i=1}^{N} \left(out_{\text{ave}}^{\text{real}} - out_i^{\text{predicted}} \right)^2} \tag{3}$$

$$\text{MSE} = \frac{1}{n} \sum\nolimits_{i}^{n} \left(out_i^{\text{real}} - out_i^{\text{predicted}} \right)^2 \tag{4}$$

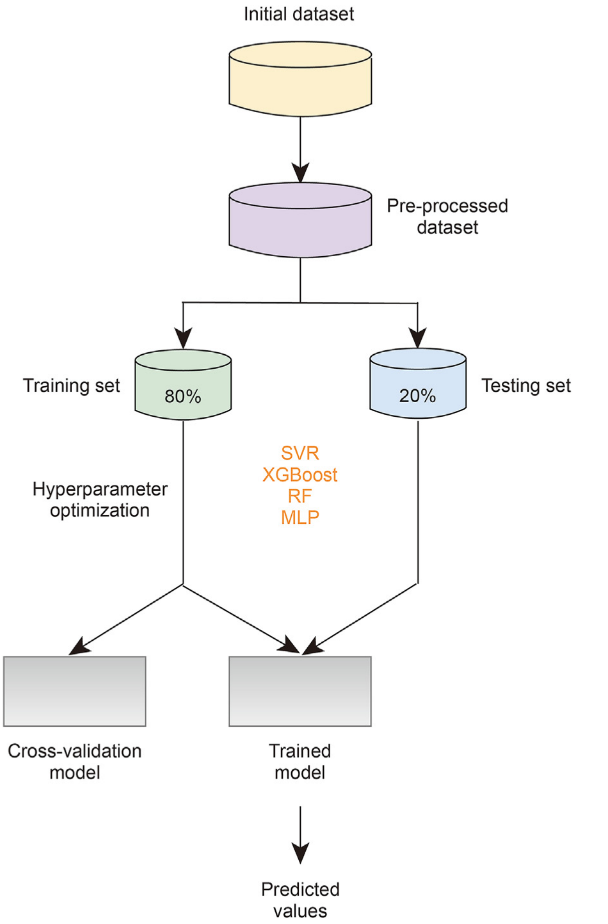

Fig. 2. The flow chart of machine learning (ML) models for this study.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_i^n \left(out_i^{\mathrm{real}} - out_i^{\mathrm{predicted}}\right)^2} \tag{5}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_i^n \left|out_i^{\mathrm{real}} - out_i^{\mathrm{predicted}}\right| \tag{6}$$

$$\mathrm{AARD} = \frac{1}{n}\sum_i^n \left|\frac{out_i^{\mathrm{real}} - out_i^{\mathrm{predicted}}}{out_i^{\mathrm{real}}}\right| \times 100\% \tag{7}$$

where out_i^{real}, $out_i^{\mathrm{predicted}}$, $out_{\mathrm{ave}}^{\mathrm{real}}$, and n are the actual value, predicted value, average actual value and the number of samples in the model, respectively.

Table 1
The control parameter for tuning the ML models.

Model	Parameter	Specific search range	Optimal value
SVR	γ	0.1–50	1
	C	0.1–500	0.5
	epsilon	0.0001–0.1	0.01
XGBoost	n_estimator	200–1000	800
	max_depth	2–10	4
	reg_lambda	0.1–10	3
	reg_alpha	0.1–10	0.1
	learning_rate	0.01–1	0.1
RF	n_estimator	20–500	300
	max_depth	2–14	10
	max_features	2–10	2
	min_samples_split	2–10	2
	min_samples_leaf	2–8	2
MLP	hidden_layer_sizes	50–500	(150, 150)
	alpha	0.00001–0.01	0.0001
	max_iter	500–2000	200
	learning_rate_init	0.001–0.1	0.001

3. Result and discussion

3.1. Hyperparameter selection for ML models

In this study, the grid search technique was employed to identify the optimal hyperparameters for each of the four machine learning models. The tuning parameters used to develop these models are listed in Table 1.

3.2. Model analysis

In this section, we assessed the performance of the predictive machine learning (ML) models described in the previous section on our data. The comparison between the predicted and experimental values for both the training and testing datasets using all four models is shown in Figs. 3–6. Upon comparing the training set and the testing set under the same model, higher prediction accuracy with a good balance is observed, indicating the absence of overfitting and a better generalization ability of the model. Among the four models, the XGBoost and MLP models exhibit superior prediction performance and display closer proximity to the expected results compared to the other two models.

The performance of the predictive ML models was assessed using various statistical metrics, including the R^2 score, MSE, RMSE, and MAE, as shown in Table 2. Notably, the XGBoost and MLP models demonstrated exceptional performance, with highest R^2 values, indicating a high correlation between predicted and experimental values. The evaluation of prediction errors revealed low MSE, RMSE, and MAE, suggesting accurate predictions with minimal deviations. Akaike information criterion (AIC) and Bayesian information criterion (BIC) can be used to explore the likelihood of models that minimize information loss. Therefore, AIC and BIC calculations were also performed for the four models, as shown in Fig. 7. The XGBoost model shows the lowest AIC/BIC values among all models, indicating its best performance on the test data.

The superior performance of the XGBoost and MLP models may be attributed to their robust nonlinear modeling capabilities and excellent generalization performance. XGBoost, as a gradient boosting decision tree model, effectively captures complex relationships in the data and performs well on large-scale datasets. Similarly, the MLP model, as a multilayer perceptron neural network, possesses powerful nonlinear modeling and adaptive learning capabilities, enabling effective learning and adaptation to complex data patterns.

In contrast, the SVR and RF exhibit slightly inferior performance. SVR, a support vector machine regression model, while capable of handling nonlinear problems in some cases, highly depends on the choice of kernel function and hyperparameter tuning, which may limit its flexibility. On the other hand, RF, an ensemble learning model, demonstrates good generalization performance and overfitting resistance but may be less effective in handling high-dimensional sparse data and highly correlated features.

These findings highlight the effectiveness of ML models in predicting CO_2 solubility in hydrocarbons, providing reliable insights and capturing the underlying patterns in the data. The high accuracy and low error metrics emphasize their capability to

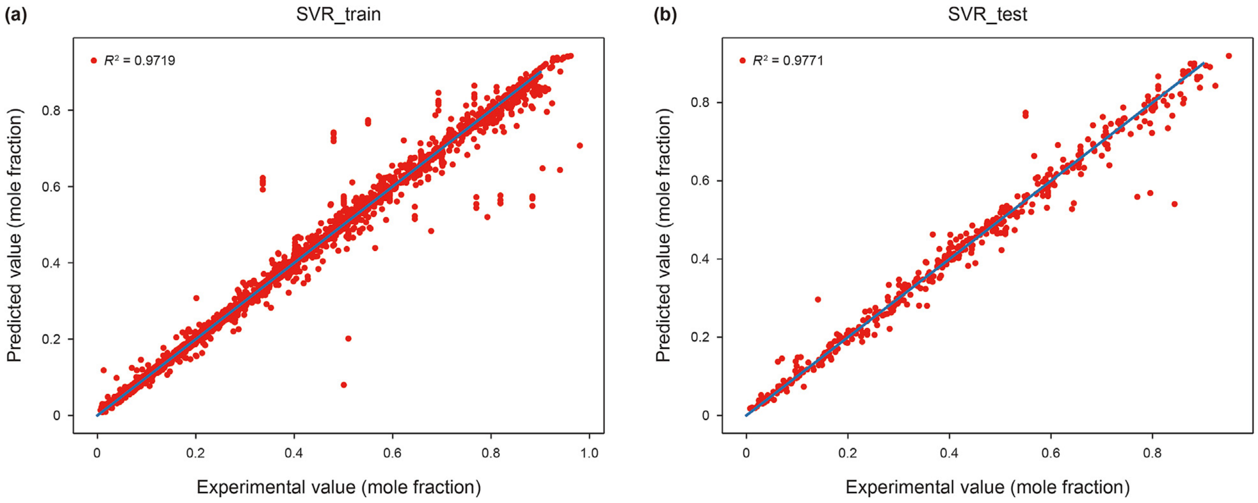

Fig. 3. Comparing prediction with experimental value (SVR model).

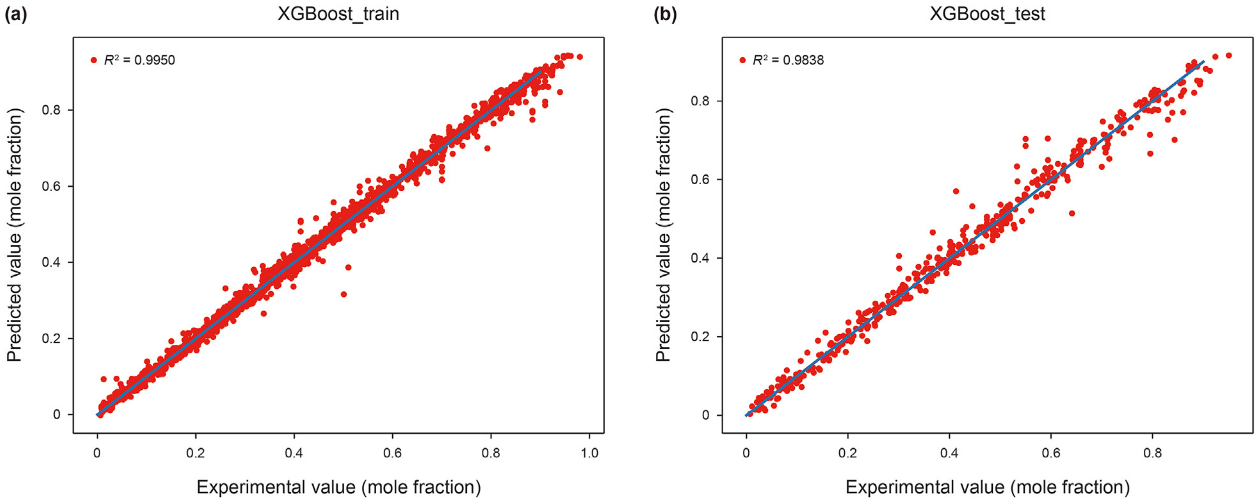

Fig. 4. Comparing prediction with true value (XGBoost model).

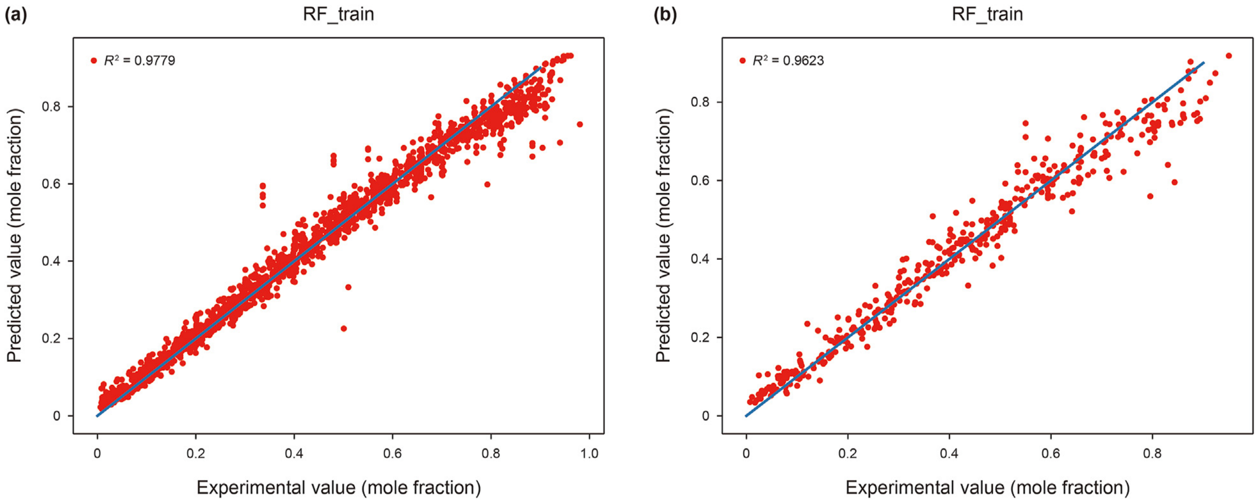

Fig. 5. Comparing prediction with true value (RF model).

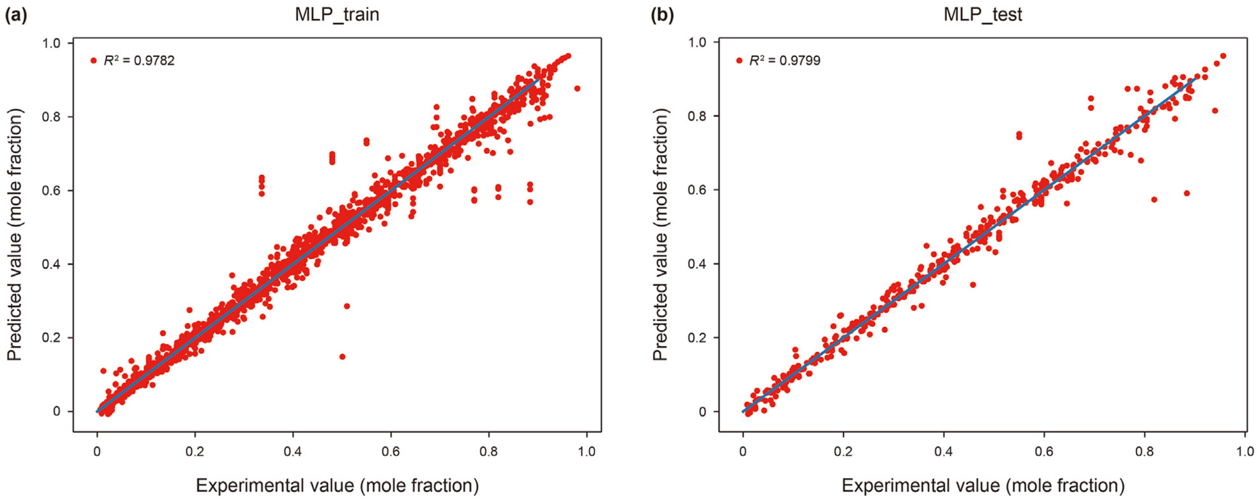

Fig. 6. Comparing prediction with true value (MLP model).

Table 2
Final results of four models.

Model	Test R^2	MSE	RMSE	MAE
SVR	0.9771	0.0013	0.0365	0.0205
XGBoost	0.9838	0.0009	0.0307	0.0205
RF	0.9623	0.0022	0.0468	0.0333
MLP	0.9799	0.0012	0.0351	0.0201

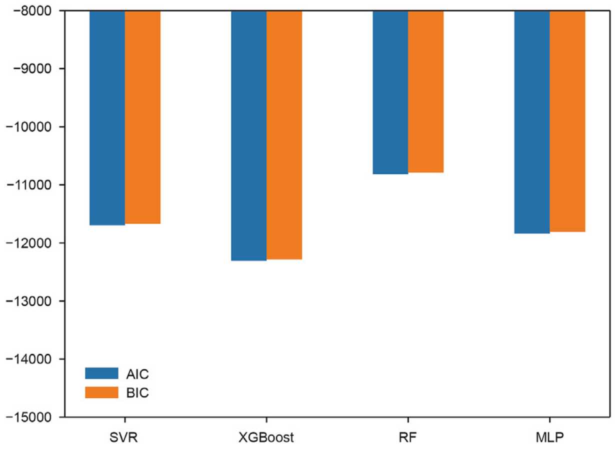

Fig. 7. AIC/BIC comparative performance of SVR, XGBoost, RF and MLP on testing data.

accurately model and predict CO_2 solubility behavior, enhancing our understanding and potential applications in related fields.

3.3. Predictability of models

This study uses a dataset with a wide range of pressure levels, and it would be valuable to investigate whether there are differences in the predicted outcomes at low and high-pressure levels. Therefore, we further divided the dataset into three pressure ranges: 0–15, 15–25, and 25–41 MPa. Under the XGBoost model, using RMSE as a comparison metric, the results obtained are shown in Fig. 8. It can be observed that the accuracy of the model shows a slight increase with increasing pressure. The average RMSE of the model is lowest in the pressure range of 25–41 MPa, at only 0.0333.

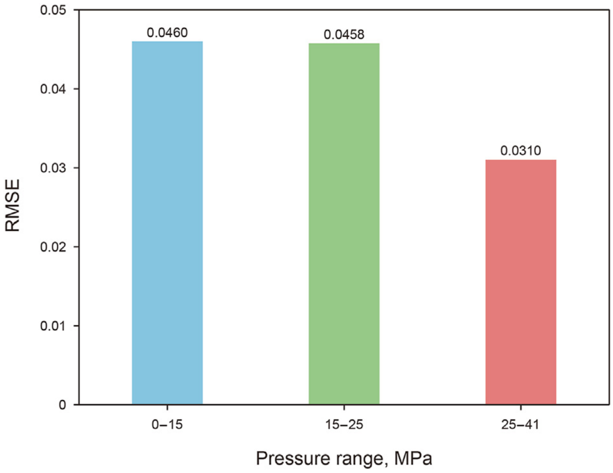

Fig. 8. Performance comparison for the XGBoost model between different ranges of pressure.

3.4. Sensitivity analysis

The Shapley plot is one of the most valuable tools for defining or explaining the influence of each attribute parameter on the output of a machine learning model. The *y*-axis of the plot shows the relevance of each feature; features at the top have the greatest impact on the output, while those at the bottom have less influence. Each feature is represented by a horizontal line in the plot. The length of the bar indicates the extent of the feature's impact on the model output. Positive Shapley values (in red) indicate that the feature enhances the output, while negative Shapley values (in blue) indicate that the feature diminishes the output. Important feature columns are listed on the *y*-axis of the plot. This can reveal which attributes have the greatest impact on the model's predictions. In this section, the XGBoost model is selected to analyze the influence of parameters on the model predictions. Based on this, from Fig. 9, it can be observed that the most important parameter directly affecting the model output for all datasets is pressure.

Fig. 10 provides detailed information on the absolute average Shapley (Shapley explanation plot) values of each input parameter of the XGBoost model. It clearly demonstrates the average absolute

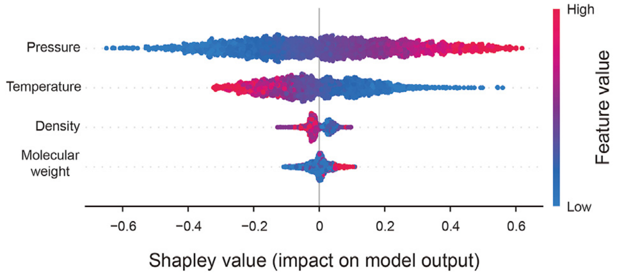

Fig. 9. Shapley plot shows the summary of the input features on output of the XGBoost model.

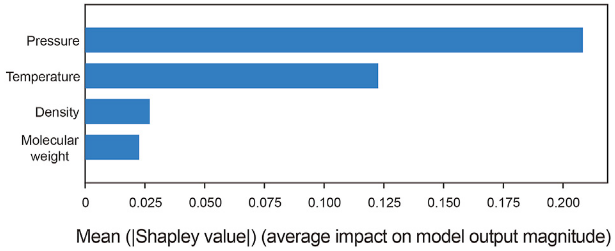

Fig. 10. Absolute average of the Shapley values for each input parameter.

significant impacts of each variable, with pressure being the most critical factor, exerting the greatest influence on CO_2 solubility in hydrocarbons. Following that, temperature also plays significant roles in influencing the solubility. Meanwhile, density and molecular weight have relatively minor effects. Therefore, when studying the solubility of CO_2 in hydrocarbons, particular attention should be paid to changes in pressure.

Based on SHAP and XGBoost models, we drawn a series of SHAP dependence plots to show the relationship between each individual compound and the CO_2 solubility. Since all data points are based on the average of the predicted CO_2 solubility (i.e., under the same criteria), the dependence of the Shapley value on the influencing factors is consistent with the dependence of the CO_2 solubility on the influencing factors. As shown in Fig. 11, with increasing pressure, the Shapley values (i.e., the contribution of a certain feature to predicting CO_2 solubility) exhibit a clear upward trend, while higher temperatures correspond to lower Shapley values. Moreover, it can be observed that the trend of Shapley values with temperature and pressure changes is more pronounced than that of molecular weight and density. This discrepancy arises from variations in the volume and range of analyzed data, the distribution of data within this range, and the specific impact of variables on the minimum miscibility pressure. For molecular weight and density, their influence on solubility is relatively small, making it difficult to accurately quantify their impact on solubility, leading to potential instability. These findings are consistent with experimental results documented in the literature.

3.5. Comparison of CO_2 solubility in hydrocarbon with previous models

To assess the accuracy of our model, we compared the selected XGBoost model with other established models (Emera and Sarma, 2007; Xue et al., 2005), using experimental data from two items of literature (Gasem et al., 1989; Mutelet et al., 2005) as references. Fig. 12 shows a comparison of the predicted values of each model under different pressures when the temperature is 323.2 K, the molecular weight is 100.2 g/mol and the density is 683 kg/m^3. The XGBoost model predicts CO_2 solubility closest to the experimental value, with the smallest AARD value, while the other two models perform poorly. Fig. 13 displays the predicted values of different models under different pressures when the temperature is 344.3 K, the molecular weight is 198.39 g/mol and the density is 765.3 kg/m^3. XGBoost model continues to show the best prediction accuracy, followed by the Emera and Sarma model (Emera and Sarma, 2007), while Xue model (Xue et al., 2005) deviates significantly. Compared to the other two models, XGBoost model offers a broader range of applications, simpler computations, and more accurate results. In summary, our model is a robust and accurate predictive model for estimating the solubility of CO_2 in hydrocarbons.

Emera and Sarma model and Xue model are widely used models for predicting the solubility of CO_2 in hydrocarbons. However, both models have limitations in terms of their applicability and complex calculations, leading to significant solubility errors. The limitations of empirical models primarily lie in their assumptions about the data and the requirements for feature engineering. Empirical models are often based on simplistic mathematical formulas or empirical rules, making simplified assumptions about the data distribution and feature relationships, which may hinder their ability to capture complex patterns and nonlinear relationships in the data.

In contrast, our XGBoost model offers a broader range of applications and simpler computations. It is a robust and accurate predictive model for estimating the solubility of CO_2 in hydrocarbons.

4. Conclusions

This study performed comprehensive artificial intelligence-based models to predict CO_2 solubility in hydrocarbons using previous literature data. Besides, the hyperparameters are adjusted to improve the expected results. Furthermore, the model was compared with other established models, revealing higher accuracy and applicability. Conclusions drawn from the results are as follows.

(1) The R^2 values of the four models are all more than 0.9, with XGBoost and MLP performing exceptionally well, which prove good predictive abilities.
(2) The importance of each parameter is evaluated. The result shows that pressure has the most important influence, while molecular weight shows the least important effect.
(3) Compared with other established models, the model obtained in this study has higher accuracy and applicability.

Our model provides an accurate method to predict the solubility of CO₂ in hydrocarbons, which is crucial for optimizing and planning the CO₂-EOR and storage processes. By accurately predicting the solubility of CO₂, we can effectively optimize CO₂ geological storage schemes in CO₂-EOR operations. Future research efforts can focus on further improving the accuracy and applicability of the model. For example, exploring the integration of different types of models to obtain more accurate and reliable predictions, and optimizing the structure and parameters of the model to enhance its performance. Additionally, emphasis can be placed on validating and applying the model in real-world CO₂-EOR and geological storage projects to further assess its reliability and practicality, and

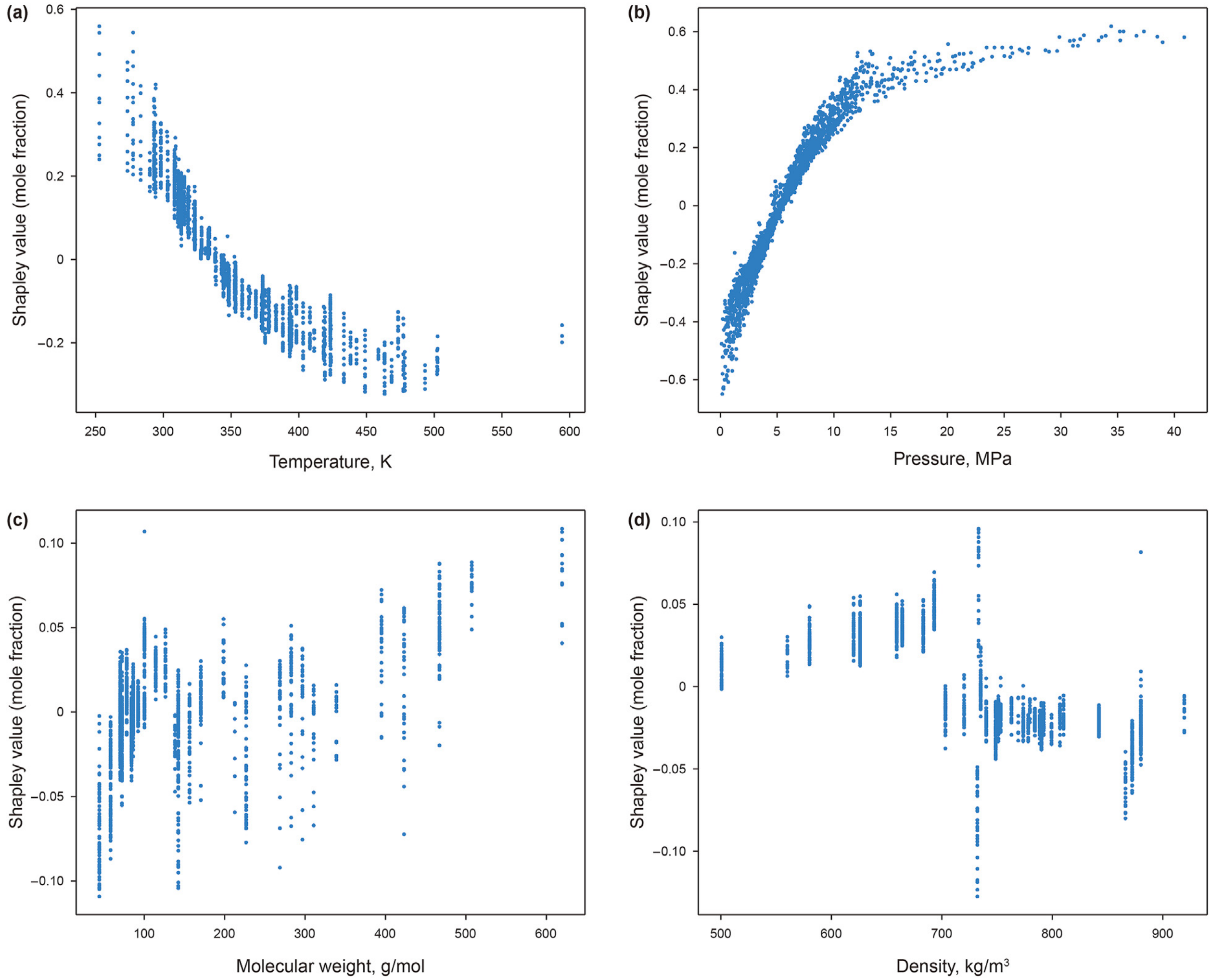

Fig. 11. SHAP dependence plot for each input parameter for the XGBoost model.

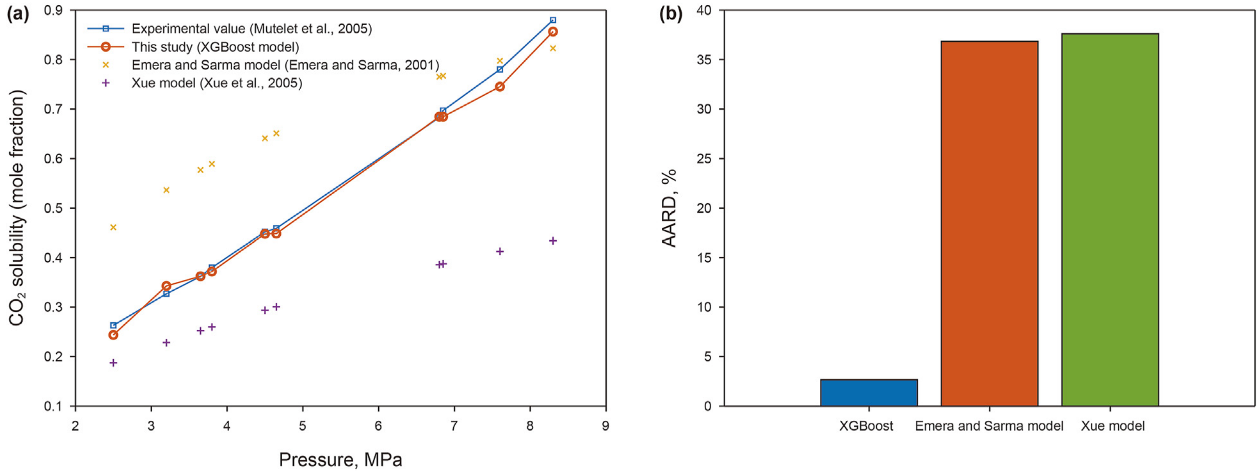

Fig. 12. Comparison of CO_2 solubility prediction at temperature of 323.2 K, molecular weight of 100.2 g/mol, and density of 683 kg/mol.

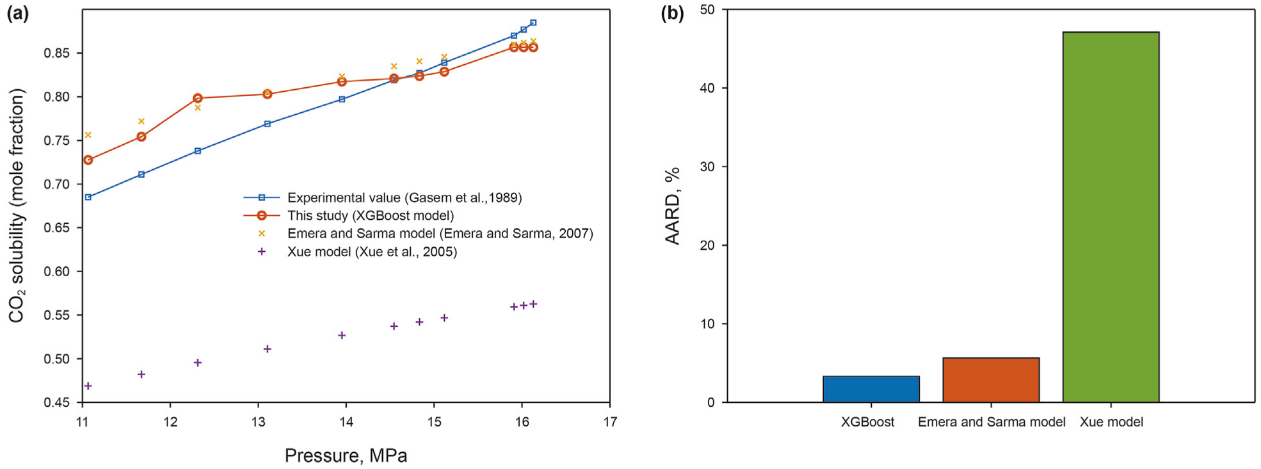

Fig. 13. Comparison of CO_2 solubility prediction at temperature of 344.3 K, molecular weight of 198.39 g/mol, and density of 765.3 kg/mol.

to propose improvements and optimization suggestions for practical applications.

Data availability

Data will be made available on request.

CRediT authorship contribution statement

Yi Yang: Writing – original draft, Methodology, Investigation, Formal analysis. **Binshan Ju:** Supervision, Methodology. **Guangzhong Lü:** Conceptualization. **Yingsong Huang:** Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The research was supported by the Fundamental Research Funds for the National Major Science and Technology Projects of China (No. 2017ZX05009-005). The authors would like to thank the editors and anonymous referees for their valuable comments and suggestions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.petsci.2024.04.018.

References

Aftab, A., Hassanpouryouzband, A., Xie, Q., et al., 2022. Toward a fundamental understanding of geological hydrogen storage. Ind. Eng. Chem. Res. 61, 3233–3253. https://doi.org/10.1021/acs.iecr.1c04380.

Agirre-Basurko, E., Ibarra-Berastegi, G., Madariaga, I., 2006. Regression and multilayer perceptron-based models to forecast hourly O_3 and NO_2 levels in the Bilbao area. Environ. Model. Software 21, 430–446. https://doi.org/10.1016/j.envsoft.2004.07.008.

Ali Ahmadi, M., Ahmadi, A., 2016. Applying a sophisticated approach to predict CO_2 solubility in brines: application to CO_2 sequestration. Int. J. Low Carbon Technol. 11, 325–332. https://doi.org/10.1093/ijlct/ctu034.

Aslannezhad, M., Ali, M., Kalantariasl, A., et al., 2023. A review of hydrogen/rock/brine interaction: implications for hydrogen geo-storage. Prog. Energy Combust. Sci. 95, 101066. https://doi.org/10.1016/j.pecs.2022.101066.

Chung, F.T., Jones, R.A., Nguyen, H.T., 1988. Measurements and correlations of the physical properties of CO_2-heavy crude oil mixtures. SPE Reservoir Eng. 3, 822–828. https://doi.org/10.2118/15080-PA.

Drucker, H., Burges, C., Kaufman, L., et al., 1996. Linear support vector regression machines. Adv. Neural Inf. Process. Syst. 9. NIPS, Denver, CO, USA, December 2-5, 1996.

Emera, M.K., Sarma, H.K., 2007. Prediction of CO_2 solubility in oil and the effects on the oil physical properties. Energy Sources, Part A Recovery, Util. Environ. Eff. 29, 1233–1242. https://doi.org/10.1080/00908310500434481.

Ezekiel, J., Ebigbo, A., Adams, B.M., et al., 2020. Combining natural gas recovery and CO_2-based geothermal energy extraction for electric power generation. Appl. Energy 269, 115012. https://doi.org/10.1016/j.apenergy.2020.115012.

Fazavi, M., Hosseini, S.M., Arabloo, M., et al., 2014. Applying a smart technique for accurate determination of flowing oil-water pressure gradient in horizontal pipelines. J. Dispersion Sci. Technol. 35, 882–888. https://doi.org/10.1080/01932691.2013.805653.

Gasem, K.A.M., Dickson, K.B., Dulcamara, P.B., et al., 1989. Equilibrium phase compositions, phase densities, and interfacial tensions for carbon dioxide + hydrocarbon systems. 5. Carbon dioxide + *n*-tetradecane. J. Chem. Eng. Data 34, 191–195. https://doi.org/10.1021/je00056a013.

Hassanpouryouzband, A., Joonaki, E., Edlmann, K., et al., 2021. Offshore geological storage of hydrogen: is this our best option to achieve net-zero. ACS Energy Lett. 6, 2181–2186. https://doi.org/10.1021/acsenergylett.1c00845.

Iskandarov, J., Fanourgakis, G., Alameri, W., et al., 2021. Machine learning application to CO_2 foam rheology. In: Abu Dhabi International Petroleum Exhibition & Conference. https://doi.org/10.2118/208016-MS.

Kamari, A., Gharagheizi, F., Bahadori, A., et al., 2014. Determination of the equilibrated calcium carbonate (calcite) scaling in aqueous phase using a reliable approach. J. Taiwan Inst. Chem. Eng. 45, 1307–1313. https://doi.org/10.1016/j.jtice.2014.03.009.

Kovscek, A.R., Cakici, M.D., 2005. Geologic storage of carbon dioxide and enhanced oil recovery. II. Cooptimization of storage and recovery. Energy Convers. Manag. 46, 1941–1956. https://doi.org/10.1016/j.enconman.2004.09.009.

Leu, A.D., Robinson, D.B., 1987. Equilibrium phase properties of the n-butane-carbon dioxide and isobutane-carbon dioxide binary systems. J. Chem. Eng. 32, 444–447. https://doi.org/10.1021/je00050a017.

Li, D.-D., Hou, J.R., Zhao, F.L., et al., 2009. Study of molecular diffusion coefficients and solubility of carbon dioxide in a Jinlin crude oil. Oilfield Chem. 26, 405–408. https://doi.org/10.19346/j.cnki.1000-4092.2009.04.016 (in Chinese).

Liang, H., Jiang, K., Yan, T.-A., et al., 2021. XGBoost: an optimal machine learning model with just structural features to discover MOF adsorbents of Xe/Kr. ACS Omega 6, 9066–9076. https://doi.org/10.1021/acsomega.1c00100.

McCallum, C., Gabardo, C.M., O'Brien, C.P., et al., 2021. Reducing the crossover of carbonate and liquid products during carbon dioxide electroreduction. Cell Rep.Phys. Sci. 2, 100522. https://doi.org/10.1016/j.xcrp.2021.100522.

Mehraein, I., Riahi, S., 2017. The QSPR models to predict the solubility of CO_2 in ionic liquids based on least-squares support vector machines and genetic algorithm-multi linear regression. J. Mol. Liq. 225, 521–530. https://doi.org/10.1016/j.molliq.2016.10.133.

Mehrotra, A.K., Svrcek, W.Y., 1985. Viscosity, density and gas solubility data for oil sand bitumens. Part I: athabasca bitumen saturated with CO and C_2H_6. AOSTRA J. Res. 1, 263–268.

Michelsen, M.L., 1990. A modified Huron-Vidal mixing rule for cubic equations of state. Fluid Phase Equil. 60, 213–219. https://doi.org/10.1016/0378-3812(90)85053-D.

Mutelet, F., Vitu, S., Privat, R., et al., 2005. Solubility of CO_2 in branched alkanes in order to extend the PPR78 model (predictive 1978, Peng–Robinson EOS with temperature-dependent kij calculated through a group contribution method) to such systems. Fluid Phase Equil. 238, 157–168. https://doi.org/10.1016/

j.fluid.2005.10.001.

Rostami, A., Ebadi, H., Arabloo, M., et al., 2017. Toward genetic programming (GP) approach for estimation of hydrocarbon/water interfacial tension. J. Mol. Liq. 230, 175–189. https://doi.org/10.1016/j.molliq.2016.11.099.

Rostami, A., Anbaz, M.A., Erfani Gahrooei, H.R., et al., 2018. Accurate estimation of CO_2 adsorption on activated carbon with multi-layer feed-forward neural network (MLFNN) algorithm. Egypt.J.Petrol. 27, 65–73. https://doi.org/10.1016/j.ejpe.2017.01.003.

Safari, H., Shokrollahi, A., Jamialahmadi, M., et al., 2014. Prediction of the aqueous solubility of $BaSO_4$ using pitzer ion interaction model and LSSVM algorithm. Fluid Phase Equil. 374, 48–62. https://doi.org/10.1016/j.fluid.2014.04.010.

Shah, N.N., Zollweg, J.A., Streett, W.B., 1991. Vapor-liquid equilibrium in the system carbon dioxide + cyclopentane from 275 to 493 K at pressures to 12.2 MPa. J. Chem. Eng. Data 36, 188–192. https://doi.org/10.1021/je00002a014.

Simon, R., Graue, D., 1965. Generalized correlations for predicting solubility, swelling and viscosity behavior of CO_2-crude oil systems. J. Petrol. Technol. 17, 102–106. https://doi.org/10.2118/917-PA.

Solomon, S., Plattner, G.-K., Knutti, R., et al., 2009. Irreversible climate change due to carbon dioxide emissions. Proc. Natl. Acad. Sci. USA 106, 1704–1709. https://doi.org/10.1073/pnas.081272110.

Sutton, C.D., 2005. Classification and regression trees, bagging, and boosting. Handb. Stat. 24, 303–329. https://doi.org/10.1016/S0169-7161(04)24011-1.

Tin Kam, H., 1998. The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20, 832–844. https://doi.org/10.1109/34.709601.

Ture, M., Kurt, I., Turhan Kurum, A., et al., 2005. Comparing classification techniques for predicting essential hypertension. Expert Syst. Appl. 29, 583–588. https://doi.org/10.1016/j.eswa.2005.04.014.

Uliasz-Misiak, B., Lewandowska-Śmierzchalska, J., Matuła, R., 2021. Criteria for selecting sites for integrated CO_2 storage and geothermal energy recovery. J. Clean. Prod. 285, 124822. https://doi.org/10.1016/j.jclepro.2020.124822.

Wang, J., Li, G., Zhou, Y., et al., 2018. Research progress of dissolved physical properties of CO_2 during geological storage in oil and gas fields. Oilfield Chem. 35, 550–561. https://doi.org/10.19346/j.cnki.1000-4092.2018.03.032 (in Chinese).

Wang, W., Gao, Q., Gui, X., et al., 2016. Determination and model prediction of solubilities of CO_2 in heavy oil under high pressure. CIESC J. 67, 442–447. https://doi.org/10.11949/j.issn.0438-1157.20151209 (in Chinese).

Wei, Y.S., Sadus, R.J., 2010. Equations of state for the calculation of fluid-phase equilibria. AIChE J. 46. https://doi.org/10.1002/aic.690460119.

Welker, J., 1963. Physical properties of carbonated oils. J. Petrol. Technol. 15, 873–876. https://doi.org/10.2118/567-PA.

Xue, H., Lu, S., Fu, X., 2005. Forecasting model of solubility of CH_4, CO_2 and N_2 in crude oil. Oil Gas Geol. 26, 444–449. https://doi.org/10.1016/j.molcatb.2005.02.001.

Yamaguchi, A.J., Sato, T., Tobase, T., et al., 2023. Multiscale numerical simulation of CO_2 hydrate storage using machine learning. Fuel 334, 126678. https://doi.org/10.1016/j.fuel.2022.126678.

Yin, G., Jameel Ibrahim Alazzawi, F., Bokov, D., et al., 2022. Multiple machine learning models for prediction of CO_2 solubility in potassium and sodium based amino acid salt solutions. Arab. J. Chem. 15, 103608. https://doi.org/10.1016/j.arabjc.2021.103608.

You, J., Ampomah, W., Morgan, A., et al., 2021. A comprehensive techno-eco-assessment of CO_2 enhanced oil recovery projects using a machine-learning assisted workflow. Int. J. Greenh. Gas Control 111, 103480. https://doi.org/10.1016/j.ijggc.2021.103480.

Zhang, J., Feng, Q., Zhang, X., et al., 2020. A supervised learning approach for accurate modeling of CO_2–brine interfacial tension with application in identifying the optimum sequestration depth in saline aquifers. Energy & Fuels. 34, 7353–7362. https://doi.org/10.1021/acs.energyfuels.0c00846.