The integration of domain knowledge and data-driven modeling evaluation process for predicting minimum miscible pressure of CO$_2$-oil systems in CO$_2$-EOR

Xing Zhang$^a$, Hang Xu$^{a,*}$, Kunjuan Wang$^b$, Bin Shen$^c$, Ruiming Zhao$^d$, Huiying Guo$^e$, Yuankai Zhang$^e$, Xinyuan Gao$^e$.

$^a$College of Petroleum, China University of Petroleum, -Beijing at Karamay, 834000, PR China.

$^b$College of Information Science and Engineering, China University of Petroleum, Beijing, 102249, PR China.

$^c$China University of Petroleum, Beijing, 102249, PR China.

$^d$Sinopec Northwest Oilfield Branch, Urumqi, 830000, PR China.

$^e$China National Petroleum Corp Xinjiang Oil Field Research Institute of Petroleum Exploration and Development, Karamay, 834000, PR China.

*corresponding author, hangxu0630@163.com.
ABSTRACT

To address the problems of underfitting or overfitting of traditional algorithms, the traditional minimum miscible pressure (MMP) intelligent prediction model ignores the influence of fluid full-component, lacks the input feature screening and has single evaluation method. In this study, based on the high-dimensional nonlinear full component data, the intermediate heavy hydrocarbons in crude oil in the traditional sense are considered separately by combining domain knowledge and mathematical statistics. Adaptive Boosting (AdaBoost), Adaptive Boosting with Support Vector Regression (AdaBoost-SVR), Extreme Gradient Boosting (XGBoost), and Deep Belief Network with Whale Optimization Algorithm (WOA-DBN) were used to analyze the results of different models using three evaluation methods: model accuracy, learning curve, and single control variable. The results of statistical analysis show that: WOA-DBN is the best solution for predicting MMP due to its excellent performance, fast convergence and high accuracy in finding the best value, the R-Square (R^2) is 0.9792, the Theil inequality coefficient (TIC) is 0.0266 and the Mean Relative Error (MRE) is 0.0553; the data points visualized by this model are all densely distributed on both sides of the 45° line; learning curve shows that the WOA-DBN model performs well on the dataset. Deep Belief Network (DBN) model performs reasonable learning on the dataset. The prediction model proposed in this study conforms to certain mechanistic rules and has high applicability. The modeling and evaluation process in this study can handle high-dimensional nonlinear data and guide field production and related theoretical research.

Keywords: Minimum miscible pressure (MMP); Machine learning; Domain knowledge; Data-driven; WOA-DBN

1. Introduction

In recent years, the global warming problem has attracted worldwide attention. The emissions of other greenhouse gases such as carbon dioxide (CO₂) are the main causes of global warming (Schmalensee et al., 1998). With carbon neutralization and carbon peaks, that is, the concept of dual carbon is proposed, carbon capture, utilization and storage (CCUS) have received unprecedented
attention. In the field of oilfield development, carbon dioxide enhanced oil recovery (CO₂-EOR) can find a balance point between improving oil recovery and reducing CO₂ emissions. During the CO₂ oil driven process, accompanied by the decrease in the viscosity of crude oil, the ratio of crude oil and water, the inflation of the volume of crude oil, and the reduction of interface tension. Therefore, the CO₂ oil-driven oil has received wide attention.

Generally speaking, there are three ways of CO₂ oil-driven machine: mixed-phase drive, non-mixed-phase drive and near-mixed-phase drive. Theoretical and practical studies have shown that compared with non-hybrid drives, the efficiency of oil recovery can be achieved more significantly compared to non-hybrid drives (Liu et al., 2020). In order to achieve a mixed-phase drive, the pressure must reach a threshold, that is, minimum miscible pressure (MMP). The minimum mixed pressure can be defined as the minimum pressure of dynamic mixing gas in the injection gas at the reservoir temperature (Chen et al., 2014), which is the pressure that makes the interface tension of crude oil and CO₂ a zero or local drive efficiency of nearly 100% (Tian et al., 2020). Therefore, accurately predict that MMP is significant for the reduction of the cost of reducing the cost of oilfields, increasing crude oil harvesting, and screening CO₂ injection of the reservoir (Belhaj et al., 2013).

At present, MMP methods are mainly traditional experimental methods, experience formula methods, status equation methods (Oyinloye et al., 2021), and numerical simulation methods. In the experimental method, fine tube experiments are the most widely used. In addition, it also includes the foaming instrument method, the interface tension disappearance method, and the rock heart drive method. The influence (Yellig and Metcalf, 1980). Experience formula method is easy to calculate, but poor universality and large results; the state equation method is fast, but there is no clear standard for determination of the mixed phase function (Ahmed, 1997); although the numerical simulation method saves effort, there is an impact of the values, low accuracy. Therefore, it is urgent to establish a method of fast, accurate, simple, and generally applicable to determining MMP.

With the rapid development of machine learning (ML), intelligent algorithms have been widely used in the field of oil collection engineering. Because of its robustness, efficiency, and accuracy
To predict the MMP of CO$_2$-crude oil system with intelligent algorithms (Sohn, 2023), a large number of domestic and experts and scholars can try to use intelligent algorithms. Zakaria et al. (Hamdi and Dong, 2019) constructed the adaptive fuzzy reasoning system (ANFIS) to predict the four factors that affect MMP. The results show that the model is more accurate and time-saving than traditional related methods. Abdorreza et al. (Karkevandi-Talkhooncheh et al., 2017) used a large amount of data points to build an adaptive fuzzy reasoning system (ANFIS) based on Back Propagation (BP), Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), Differential Evolution (DE) has high accuracy. Tian et al. (Tian et al., 2020) built a back propagation network based on the Dragonfly Algorithm, and the results showed that DA-BPNN had high accuracy. Huang et al. (Huang et al., 2022) builds a confrontation network based on the conditions based on the Bayesian Optimization Algorithm, and compared with the existing artificial neural network and support vector regression model based on the radial basal function core and polynomial function core, there is a lower error and higher accuracy. Chen et al. (Chen et al., 2014) uses a genetic algorithm to optimize the weight and parameters of the reverse communication network to predict MMP. Osamah et al. (Osamah A. Alomair et al., 2016) provide a cost-effective alternative for estimating MMP based on generalized regressive neural network (GRNN) prediction of MMP and evaluating the accuracy of the model and models in the literature. Considering the difference in the results of the narrow tube experiment method and the foam riser method, Huang et al. (Can et al., 2023) established a one-dimensional convolutional neural network model, and used The Shapley Additive Explanations (SHAP) interpretable method to analyze the influence of various factors on the prediction of MMP, and obtained a unified and consistent machine learning framework.

Studies such as Schuetter (Schuetter et al., 2018) have shown that there are many problems in the field of oil and gas collection. When building a reasonable machine learning model, there are many steps. However, in the current MMP research, the lack of screening of input features is simply dividing the factors that affect the MMP according to past experience; secondly, only a few influencing factors are selected, thereby ignoring the fluid full group of MMP to MMP. Impact; after
the establishment of models, most studies generally use evaluation indicators as the final inspection standard for model verification, and have not adopted other evaluation methods to evaluate the model performance. This may have a lot of discrepancy with the actual situation.

In view of this, this article widely collects experimental or predicted data in domestic and foreign literature, and establishes a set of WOA-DBN models with high applicability, strong robustness and various evaluation methods based on the input feature screening, with a view to providing CO₂ drive high harvesting shall be provided to provide high income yam to provide high harvesting rate

Theoretical foundation and practical guidance. Figure 1 shows the overall workflow of this study.

Overall, this study aims to accomplish the following contributions:

- Collect a more comprehensive MMP database of pure and impure CO₂ and crude oil systems for ML technique study.
- Screening main controlling factors based on feature engineering. Based on four characteristic screening methods and partial correlation analysis methods, 21 influencing factors in the database were evaluated, and the traditional intermediate heavy hydrocarbons(C₅-C₆) in crude oil were considered separately to explore their independent influence on MMP, and 10 groups of characteristic factors were comprehensively screened out.
- Develop a knowledge-guided intelligent framework and integrate the swarm intelligence optimization algorithm with deep learning models for modeling the minimum miscible pressure has been investigated simultaneously in CO₂-EOR projects.
- Build a variety of evaluation methods, determine the most recommended model. Base on three evaluation methods of model accuracy evaluation, learning curve evaluation and single control variable evaluation to offer the most effective intelligent model for modeling MMP.
2. Methodology

2.1. Data collection

Generally speaking, the smallest mixing pressure mainly depends on the temperature, crude oil composition, and injection gas composition. On the basis of investigating a large number of literatures (Adekunle and Hoffman, 2014; Ahmed, 1994; Bui et al., 2010; Coelho et al., 2015; Dicharry et al., 1973; Frimodig et al., 1983; Ghiasi, 2021; Jacobson and H., 1972; Jr and Watkins, 1980; Rathmell et al., 1971; Shokir and Engineering, 2007; Yellig and Metcalfe, 1980), a total of 180 groups of streaming data were obtained. It contains 160 groups of non-pure CO$_2$ data, and 20 groups of pure CO$_2$ data. Most of these data come from the on-site data of various oil fields at home and abroad, including the real data of the laboratory. Specifically, this sample data set mainly covers 21 influencing factors: reservoir temperature ($T_R$), the molecular fraction of the full component of the crude oil ($O_{-N_2}$, $O_{-CO_2}$, $O_{-C_1}$, $O_{-C_2}$, $O_{-C_3}$, $O_{-C_4}$, $O_{-C_5}$, $O_{-C_6}$, $O_{-C_7}$), the molecular fraction of the injection gas ($G_{-N_2}$, $G_{-CO_2}$, $G_{-H_2S}$, $G_{-C_1}$, $G_{-C_2}$, $G_{-C_3}$, $G_{-C_4}$, $G_{-C_5}$, $G_{-C_6}$, $G_{-C_7}$), and the C$_7$+.
molecular weight (MW\textsubscript{C7+}) in crude oil. The dataset contains a wide range of scope, complete consideration, and has certain reference value. The statistical results of each parameter are shown in Table 1.

<table>
<thead>
<tr>
<th>factor</th>
<th>number</th>
<th>Parameter</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Average</th>
<th>Upper quartile</th>
<th>Lower quartile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reservoir temperature</td>
<td>1</td>
<td>T\textsubscript{R}</td>
<td>8.95</td>
<td>115.6</td>
<td>84.06</td>
<td>64.2</td>
<td>115.6</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>O\textsubscript{N2}</td>
<td>0</td>
<td>2.13</td>
<td>0.34</td>
<td>0.01</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>O\textsubscript{CO2}</td>
<td>0</td>
<td>15.57</td>
<td>3.81</td>
<td>0.65</td>
<td>6.66</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>O\textsubscript{C1}</td>
<td>4.07</td>
<td>54.4</td>
<td>18.9</td>
<td>6.35</td>
<td>32.98</td>
</tr>
<tr>
<td>Composition of crude oil</td>
<td>5</td>
<td>O\textsubscript{C2}</td>
<td>1.06</td>
<td>23.16</td>
<td>10.85</td>
<td>6.28</td>
<td>23.16</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>O\textsubscript{C3}</td>
<td>0.58</td>
<td>12.39</td>
<td>6.73</td>
<td>5.58</td>
<td>8.39</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>O\textsubscript{C4}</td>
<td>0.34</td>
<td>7.82</td>
<td>4.56</td>
<td>4.23</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>O\textsubscript{C5}</td>
<td>0.49</td>
<td>5.89</td>
<td>3.48</td>
<td>3.09</td>
<td>4.24</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>O\textsubscript{C6}</td>
<td>0.21</td>
<td>7.35</td>
<td>3.23</td>
<td>1.88</td>
<td>3.96</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>O\textsubscript{C7+}</td>
<td>19.59</td>
<td>80.75</td>
<td>48.11</td>
<td>206</td>
<td>281</td>
</tr>
<tr>
<td>C\textsubscript{7+} molecular weight</td>
<td>11</td>
<td>MW\textsubscript{C7+}</td>
<td>153.9</td>
<td>402.7</td>
<td>225.96</td>
<td>206</td>
<td>281</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>G\textsubscript{N2}</td>
<td>0</td>
<td>80.1</td>
<td>3.59</td>
<td>0</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>G\textsubscript{CO2}</td>
<td>1.05</td>
<td>100</td>
<td>55.5</td>
<td>6.48</td>
<td>90.84</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>G\textsubscript{H2S}</td>
<td>0</td>
<td>50</td>
<td>4.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Composition of injected gas</td>
<td>15</td>
<td>G\textsubscript{C1}</td>
<td>0</td>
<td>83.21</td>
<td>20.82</td>
<td>0</td>
<td>43.17</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>G\textsubscript{C2}</td>
<td>0</td>
<td>29.85</td>
<td>8.46</td>
<td>0</td>
<td>19.49</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>G\textsubscript{C3}</td>
<td>0</td>
<td>28</td>
<td>4.62</td>
<td>0</td>
<td>8.33</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>G\textsubscript{C4}</td>
<td>0</td>
<td>7.08</td>
<td>1.49</td>
<td>0</td>
<td>2.89</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>G\textsubscript{C5}</td>
<td>0</td>
<td>4.76</td>
<td>0.89</td>
<td>0</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>G\textsubscript{C6}</td>
<td>0</td>
<td>0.72</td>
<td>0.11</td>
<td>0</td>
<td>0.136</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>G\textsubscript{C7+}</td>
<td>0</td>
<td>0.06</td>
<td>0.01</td>
<td>0</td>
<td>0.011</td>
</tr>
<tr>
<td>MMP</td>
<td>22</td>
<td>MMP</td>
<td>6.55</td>
<td>41.47</td>
<td>20.7</td>
<td>14.07</td>
<td>25.55</td>
</tr>
</tbody>
</table>

2.2. Feature engineering

The selection of main control factors has a great impact on the accuracy of the model. In order to reasonably screen the main control factors, this study screened and evaluated based on domain knowledge and four correlation ranking methods. When establishing the sample data set, fully consider the influence of all fluid components on MMP is then sorted and comprehensively screened.
using four screening methods: gray correlation analysis, Pearson correlation coefficient, Kendall correlation coefficient and Spearman correlation coefficient.

The basic principle of gray relational analysis is to compare several series with a set reference series, and judge the degree of closeness by the geometric shape between them (Amine and Mohammed, 2023). Gray relational analysis is a simple and reliable common data analysis method (Zhuojie, 2022). The Pearson correlation coefficient is mainly used to study the linear correlation between data (Sedgwick, 2012), and the Kendall correlation coefficient and Spearman correlation coefficient are used to solve the rank of the data matrix to measure the correlation between data (Kendall, 1990). The size of the gray correlation analysis coefficient indicates the degree of correlation; the positive and negative of Pearson correlation coefficient, Kendall correlation coefficient and Spearman correlation coefficient represent positive and negative correlation, and the size represents the degree of correlation.

Partial correlation analysis refers to the process of removing the third variable when two variables are correlated with the third variable at the same time, and only analyzing the degree of correlation between the other two variables. Using this method in this study, the aim is to investigate whether there are characteristic factors that exert the same influence on MMP.

Table 2 shows the results of using four correlation analysis methods based on the data set of this study. It can be seen from Table 2 that the Pearson, Kendall and Spearman correlation coefficients are completely consistent in terms of positive and negative correlations, which shows that the results are accurate and reliable, and can better reflect the correlation between various factors and MMP. Among them, reservoir temperature, the molecular fraction of CO₂, C₁, C₂ and C₃ in crude oil, the molecular fraction of N₂ in injected gas, the molecular fraction of C₁-C₇⁺, and the molecular weight of C₇⁺ are positively correlated with MMP. On the contrary, the molecular fraction of N₂ and C₄-C₇⁺ in crude oil, the molecular fraction of CO₂ in injected gas are negatively correlated with MMP.

<table>
<thead>
<tr>
<th>correlation coefficient</th>
<th>T_R</th>
<th>O⁻N₂</th>
<th>O⁻C₀₂</th>
<th>O⁻C₁</th>
<th>O⁻C₂</th>
<th>O⁻C₃</th>
<th>O⁻C₄</th>
<th>O⁻C₅</th>
<th>O⁻C₆</th>
<th>O⁻C₇⁺</th>
</tr>
</thead>
</table>

Table 2 Correlation coefficient of influencing factors.
<table>
<thead>
<tr>
<th>correlation coefficient</th>
<th>MW_{C7+}</th>
<th>G_{N2}</th>
<th>G_{CO2}</th>
<th>G_{H2S}</th>
<th>G_{C1}</th>
<th>G_{C2}</th>
<th>G_{C3}</th>
<th>G_{C4}</th>
<th>G_{C5}</th>
<th>G_{C6}</th>
<th>G_{C7+}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grey</td>
<td>0.755</td>
<td>0.618</td>
<td>0.485</td>
<td>0.607</td>
<td>0.754</td>
<td>0.681</td>
<td>0.665</td>
<td>0.661</td>
<td>0.637</td>
<td>0.644</td>
<td>0.643</td>
</tr>
<tr>
<td>Pearson</td>
<td>0.490</td>
<td>0.504</td>
<td>-0.602</td>
<td>-0.141</td>
<td>0.460</td>
<td>0.476</td>
<td>0.268</td>
<td>0.316</td>
<td>-0.179</td>
<td>0.232</td>
<td>0.225</td>
</tr>
<tr>
<td>Kendall</td>
<td>0.587</td>
<td>0.587</td>
<td>-0.528</td>
<td>0.006</td>
<td>0.346</td>
<td>0.596</td>
<td>0.474</td>
<td>0.502</td>
<td>0.053</td>
<td>0.313</td>
<td>0.302</td>
</tr>
<tr>
<td>Spearman</td>
<td>0.412</td>
<td>0.484</td>
<td>-0.368</td>
<td>0.007</td>
<td>0.288</td>
<td>0.406</td>
<td>0.322</td>
<td>0.343</td>
<td>0.023</td>
<td>0.231</td>
<td>0.223</td>
</tr>
</tbody>
</table>

Continued Table 2 Correlation coefficient of influencing factors.

On the basis of analyzing the correlation coefficient of the whole components of crude oil and injected gas, 21 factors affecting MMP are further divided. For crude oil composition, most scholars (Chen et al., 2020; Kun et al., 2013; Salim et al., 2016) divide the volatile components C1 and N2 together, because they have a negative impact on MMP due to their volatility. In this study, there is a strong correlation between average values of the four correlation coefficients of CO2 and the correlation coefficients of volatile components, on the basis of previous research (Huang et al., 2022), CO2, C1, N2 are regarded as a group of characteristic factors; according to the laboratory measurement results (Alston et al., 1985), C2-C4 and C6 in crude oil are positively correlated with MMP, because they are more miscible with CO2, so they have a positive impact on MMP. Therefore, the intermediate components C2-C4 and C6 are regarded as a group of characteristic factors. Usually, researchers consider the intermediate heavy hydrocarbons (C5-C6) as a group, but as mentioned above, on the one hand, the correlation coefficient values of C5 and C6 are quite different, on the other hand, the presence of colloidal asphaltene will seriously affecting the size of MMP, studies (Keshmiri et al., 2019) have shown that n-pentane can produce tiny precipitates, leading to higher aggregation and flocculation, thus negatively affecting MMP. This shows that C5 and C6 have different effects on MMP production, and it is inappropriate to simply regard them as a group of characteristic factors, and should be...
considered separately. It can be seen from the values in Table 1 shows that the molecular fraction of
C_{7+} and the molar weight of C_{7+} are different from other hydrocarbon components, therefore, they
have different effects on MMP, so they are listed separately as a set of characteristic factors.

For injected gas, CO_2 is the most important gas, so it is regarded as an independent characteristic
factor; the molecular fraction of C_1 with fewer carbon atoms and the molecular fraction of N_2 will
increase MMP, so it is regarded as a group of characteristic factors; Generally speaking, for the
intermediate group C_2-C_4 and H_2S have a positive impact on MMP, and as the number of carbon
atoms increases, the miscibility of CO_2 and crude oil is stronger; in the dataset of this study, the
molecular fraction of C_5-C_{7+} is relatively small, which can be approximately regarded as a component,
so it can be divided into a group of characteristic factors.

Based on the above analysis, a total of 10 groups of characteristic factors were screened and
divided. Respectively, the reservoir temperature (T_R), the molecular fraction of all components of
crude oil (O_{N_2-C_1-CO_2}, O_{C_2-C_3-C_4-C_6}, O_{C_5}, O_{C_{7+}}), the molecular fraction of all components of
.injected gas (G_{CO_2}, G_{C_1-N_2}, G_{H_2S-C_2-C_3-C_4}, G_{C_5-C_6-C_{7+}}) and the molecular weight of C_{7+} in crude
oil (MW_{C_{7+}}).

In order to further determine the degree of influence of the above 10 groups of characteristic
factors on MMP, four correlation coefficient methods were used to sort the characteristic factors, and
the comprehensive ranking of the correlation degrees of different characteristic factors shown in
Figure 2.

The following conclusions can be drawn from Figure. 2: reservoir temperature (T_R) has the
greatest impact on MMP, followed by the molecular fraction of N_2 and C_1 in injected gas (G_{C_1-N_2}),
and the molecular fraction of C_5 in crude oil has the least influence, which also confirms the necessity
of dividing C_5 and C_6 in crude oil separately.
The results of partial correlation analysis of the above 10 groups of characteristic factors are shown in Figure 3. It can be found that the influence of each characteristic factor on MMP is unique, and there is no intersection and overlap point, that is, there is no repeated effect between the characteristic factors. It shows that the influencing factors of division are reliable and accurate.

Fig. 3. Partial correlation analysis of different characteristic factors

2.3. Machine learning algorithms

2.3.1. Restricted Boltzmann Machine (RBM)

The restricted Boltzmann machine (RBM) originated from the Boltzmann machine. RBM is a special topology of BM, which was first proposed by Smolensky in 1986. BM not only has powerful
unsupervised learning ability, but also can learn complex rules in high-dimensional data. The BM network is composed of a visible layer and a hidden layer. The neurons in the visible layer and in the hidden layer are fully connected, and the neurons in the same layer are connected in pairs. It is this complex structure that makes network training relatively complex and the training time is longer. In order to overcome this problem, an improved BM, namely RBM, was born. The similarities and differences between RBM and BM are: neurons in the same layer of RBM are not connected to each other, and neurons in different layers are fully connected (Roux and Bengio, 2008); at the same time, RBM retains the binary Bernoulli distribution of neurons in BM. The network topology of RBM is shown in Figure 4.

RBM is composed of a visible layer and a hidden layer, $h_1$ to $h_n$ are n real numbers, $v_1$ to $v_m$ are m real numbers, the range of these real numbers is (0,1), they form $h$ vector and $v$ vector respectively; The weight between the neurons of the visible layer and the hidden layer is $w$, because the neurons are fully connected, so there are $n \times m$ weights in total, and $a$ and $b$ are the bias vectors of the visible layer and the hidden layer respectively.

RBM is an energy-based undirected graph structure, and its energy function can be defined as:

$$E(v,h|\varphi) = - \left( \sum_{i=1}^{m} a_i v_i + \sum_{j=1}^{n} b_j h_j + \sum_{i=1}^{m} \sum_{j=1}^{n} v_i w_{ij} h_j \right) \tag{1}$$

Where $\varphi$ is the parameter of the RBM network, $w$ is the weight between neurons, $a$ is the bias of neurons in the visible layer, and $b$ is the bias of neurons in the hidden layer. When the
parameters are determined, the joint probability distribution between the visible layer and the hidden layer is:

\[ P_\varphi(v, h) = \frac{1}{G(\varphi)} e^{-E(v, h|\varphi)} \]  

(2)

Where \( G(\varphi) \) is a normalized coefficient, and its mathematical meaning is to add all the vector pairs of the visible layer and the hidden layer, defined as

\[ G(\varphi) = \sum_{v, h} e^{-E(v, h|\varphi)} \]  

(3)

At the same time, the partial derivatives of \( v \) and \( h \) in formula (2) are calculated separately, and the marginal probability distribution of the visible layer and the hidden layer can be obtained as:

\[ P_\varphi(v) = \frac{1}{G(\varphi)} \sum_{h} e^{-E(v, h|\varphi)} \]  

(4)

\[ P_\varphi(h) = \frac{1}{G(\varphi)} \sum_{v} e^{-E(v, h|\varphi)} \]  

(5)

Since the neurons of the same layer in the RBM network topology are not connected to each other, the independence of the respective states of the neurons is ensured. Therefore, according to the state of neurons in the visible layer, the conditional probability of activation of the \( j \)-th hidden layer neuron can be obtained as:

\[ P(h_j = 1|v) = \beta(b_j + \sum_{i=1}^{m} v_i w_{ij}) \]  

(6)

According to the state of neurons in the hidden layer, the conditional probability of activation of the \( i \)-th visible layer neuron can be obtained as

\[ P(v_i = 1|h) = \beta(a_i + \sum_{j=1}^{n} h_j w_{ji}) \]  

(7)

Where \( \beta \) is the Sigmoid activation function, defined as: \( \beta(x) = \frac{1}{e^{-x} + 1} \).

2.3.2. Deep Belief Network (DBN)

The RBM mentioned above is a shallow network composed of two layers of neurons, but for high-dimensional nonlinear data, its prediction effect is not as good as that of deep learning algorithms. Deep Belief Network (DBN) was first proposed by Hinton in 2006. DBN has a simple
structure, is easy to train, has fast convergence speed, and can be flexibly combined with other methods, and has strong generalization ability. Compared with the traditional machine learning model, DBN has a strong effect on the high-dimensional nonlinear MMP data constructed in this study. Feature extraction capabilities (Yang et al., 2016).

**Fig. 5.** Network topology of RBM.

Figure 5 is a network topology of DBN. It is a probability-based deep learning model, which is a hierarchical generative model stacked from bottom to top by multi-layer RBM. The bottom layer is the visible layer, that is, the input layer, which is used to input training data; the visible layer and the first hidden layer $h_1$ constitute the first $RBM_1$; the hidden layer $h_1$ of $RBM_1$ and the hidden layer of the second layer Layer $h_2$ constitutes the second $RBM_2$; and so on to construct $n$ $RBM_n$. The output data of the previous layer is used as the input data of the next layer, and is passed layer by layer until the final layer outputs the result.

The training process of DBN can be divided into two ways: unsupervised training and supervised training. DBN first performs unsupervised training through a layer-by-layer greedy algorithm, which can ensure that the parameters of the single-layer RBM are optimal, and ensure that when the input data is mapped to a high-dimensional space through a certain mapping rule, the characteristics of the
original data are preserved as much as possible; After completing the unsupervised training, DBN enters the supervised reverse fine-tuning stage, calculates the difference between the output result of the last layer and the target result according to the loss function (Baqar et al., 2017), and then reverse fine-tunes each layer through the gradient descent algorithm. The parameters of neurons to achieve the purpose of reducing the loss function.

2.3.3. Support Vector Regression (SVR)

American scholar Vapnik (Vapnik, 1995) proposed a method of nonlinear regression for high-dimensional data, and called it Support Vector Regression (SVR). SVR is a branch of Support Vector Machine (SVM). The difference between them is that SVM divides different types of data, while SVR looks for functions that can accurately predict data distribution. The SVR algorithm takes sample data as input variables, and maps the input variables to a high-dimensional space through a kernel function, establishes a hyperplane as a decision surface, and the distance to the data sample farthest from the hyperplane is the smallest, that is, the data is in the optimal strip area, the sample data is then trained to obtain the optimal prediction model.

2.3.4. Adaptive Boosting (AdaBoost)

Adaptive Boosting (AdaBoost) is an integrated iterative algorithm whose core idea is to train different classifiers (weak classifiers) for the same training set, and then combine these weak classifiers to form a stronger final classification device (strong classifier). The process of the algorithm is: firstly, initialize the weight of the sample dataset; then train the weak classifier, if the sample data is accurately classified, the weight will be reduced, otherwise the weight will be increased. Finally, several weak classifiers are combined into a strong classifier for data prediction.

2.3.5. Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is a better performing ensemble algorithm based on and improved upon Gradient Boosting Decision Trees (GBDT). The difference between them is that the predicted value in GBDT is the weighted sum of the predicted results on all weak classifiers, while the predicted value in XGBoost is the direct sum of the leaf weights on all weak classifiers.
2.3.6. Swarm Intelligence Optimization Algorithm

Meta-heuristic optimization algorithms are popular in related engineering fields, because they are easy to implement, do not require gradient information, and are not easy to fall into local optimum. In general, metaheuristic optimization algorithms can be divided into three categories: evolution-based, physics-based, and population-based. Relevant studies have shown that (Chao et al., 2021), swarm intelligence-based optimization algorithms have more advantages than evolution-based optimization algorithms: swarm intelligence-based optimization algorithms can retain search space information in subsequent iterations, while evolution-based optimization algorithms are emerging when new. Immediately after population, all intermediate information is discarded, so the former can contain fewer operators, which is beneficial to the implementation.

Inspired by the hunting behavior of humpback whales, Australian scholar Mirjalili et al. (Mirjalili et al., 2016) proposed a new swarm intelligence optimization algorithm—Whale Optimization Algorithm (WOA) in 2016. WOA has the advantages of simple operation, less control parameters and avoiding local optimum, and has attracted the attention of scholars in related fields at home and abroad. The process of the algorithm includes three stages: encircling prey, spiral bubble-net attacking, and searching for prey.

1. Encircling prey

Humpback whales can pinpoint the location of their prey and surround it. Since the optimal position is unknown in the search space, the WOA assumes that the optimal candidate solution or near-optimal solution is the target prey. After defining the optimal search agent, other search agents will update their positions, gradually approaching the optimal search agent. Its mathematical model is expressed as follows:

\[ D_1 = |\bar{C} \cdot \bar{X}^*(t) - \bar{X}(t)| \]  
\[ \bar{X}(t+1) = \bar{X}^*(t) - \bar{A} \cdot D_1 \]
Where $D_1$ is the distance between the humpback whale and the prey; $t$ is the number of current iterations; $X^*$ is the position vector of the current optimal solution; $\bar{X}$ is the position vector of the humpback whale; $\bar{A}$ and $\bar{C}$ are coefficient vectors respectively.

2. Spiral bubble-net attacking

This stage includes two mechanisms, namely the contraction and wrapping mechanism, and the helical update position mechanism. Humpback whales also follow a spiral path while swimming around their prey in a narrowing circle. In order to express these two mechanisms at the same time, it is assumed that there is a 50% probability of choosing between the two, so the mathematical model is expressed as follows:

$$\bar{X}(t + 1) = \begin{cases} 
\bar{X}^2(t) - \bar{A} \cdot D_1 & p < 0.5 \\
D_2 \cdot e^{bl} \cdot \cos(2\pi l) + \bar{X}^2(t) & p \geq 0.5 
\end{cases} \#(10)$$

Where $D_2$ is the distance between the individual whale and the prey; $l$ is a random number between [-1,1]; $b$ is a constant that determines the shape of the spiral; $p$ is a random number between [0,1].

3. Searching for prey

At this stage, in order to increase the search range of humpback whales and find the optimal solution to maintain the diversity of the population, when $|\bar{A}| \geq 1$, the position of humpback whales is updated according to the randomly selected search agent. The mathematical model is expressed as follows:

$$D_{\text{rand}} = |\bar{C} \cdot X_{\text{rand}} - \bar{X}(t)| \#(11)$$

$$\bar{X}(t + 1) = X_{\text{rand}} - \bar{A} \cdot D_{\text{rand}} \#(12)$$

Where $D_{\text{rand}}$ is the distance from the randomly selected whale to the prey; $X_{\text{rand}}$ is the position vector of the randomly selected whale.

Table 3 presents a comparative analysis between commonly used machine learning models and four proposed smart schemes, which are intended to predict the minimum miscible pressure of CO$_2$-
The purpose of this comparison is to demonstrate the advantages and disadvantages of the method proposed in this study relative to existing machine learning models. In general, all of these methods have own strengths and weaknesses.

<table>
<thead>
<tr>
<th>Models</th>
<th>pros</th>
<th>cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liner Regression</td>
<td>Simple, easy to interpret, works well with small datasets</td>
<td>cannot capture complex nonlinear patterns</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>Easy to interpret, works well with binary classification problems</td>
<td>cannot capture complex nonlinear patterns</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>handle complex samples, easy to interpret, performs well in classification and prediction problems</td>
<td>Prone to overfitting, can create complex trees that are difficult to interpret</td>
</tr>
<tr>
<td>Random Forest</td>
<td>performs well in classification and prediction problems, less prone to overfitting</td>
<td>slow and memory-intensive with large datasets</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>Handles high-dimensional data well, robust to outliers</td>
<td>tuning of hyperparameters</td>
</tr>
<tr>
<td>Support Vector Regression</td>
<td>Strong robustness</td>
<td>Sample insufficient, low precision</td>
</tr>
<tr>
<td>Adaptive Boosting</td>
<td>Solve the problem of insufficient samples and low accuracy</td>
<td>Data imbalance, Sensitive to outliers</td>
</tr>
<tr>
<td>Adaptive Boosting with Support vector regression</td>
<td>Low generalization error rate, high precision</td>
<td>Overall low robustness</td>
</tr>
<tr>
<td>Extreme Gradient Boosting</td>
<td>handle complex samples, performs well in classification and prediction problems, computationally efficient</td>
<td>require tuning of hyperparameters, less interpretable</td>
</tr>
</tbody>
</table>
Deep belief network with Whale optimization algorithm

Excellent performance, fast convergence, high optimization accuracy

require tuning of hyperparameters, less interpretable

3. Results and discussion

Based on the data set mentioned above, in the modeling stage, 75% of the dataset is used as training set, and 25% of the dataset is used as test set, and the aforementioned machine learning model is used to construct the prediction of the minimum miscible pressure of CO$_2$-crude oil model. In view of the fact that previous studies generally used the quality of evaluation indicators as the final test standard for model validation, and did not use other evaluation methods to evaluate model performance, so this study uses three evaluation methods to test the accuracy and robustness of the constructed model.

3.1. Model Accuracy Evaluation

3.1.1. Evaluation based on evaluation indicators

In order to evaluate the accuracy of the model more comprehensively, this study selected five statistical error functions to evaluate the performance of the model, namely mean relative error (MRE), root mean square error (RMSE), mean absolute error (MAE), R-Square ($R^2$), Theil inequality coefficient (TIC), different errors function reflect different evaluation emphases. Among them, the closer the value of the first three error functions is to 0, the higher the accuracy of the model; the closer the value of $R^2$ is to 1, it means that the fitting effect of the model is better; the closer the value of TIC is to 0, it means the model predicts well.

MRE represents the ratio between the true value of each sample and the error value, the calculation formula is as follows:

\[
MRE = \frac{1}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{y_i}\tag{13}
\]

RMSE represents the arithmetic square root of the square and mean of the difference between the true value and the predicted value of each sample, and the formula is as follows:
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (14)

MAE refers to the average of the absolute errors of each measurement. It can avoid the problem that errors cancel each other, so it can accurately reflect the actual size of prediction errors.

MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (15)

\( R^2 \) represents the ratio of the regression sum of squares to the total sum of squares, the value range is between (0,1), and the calculation formula is as follows:

\[ R^2 = 1 - \frac{S_{\text{model}}}{S_{\text{wal}}} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (16) \]

The value range of TIC is (0,1). The smaller the value of TIC, the more accurate the model prediction is. The calculation formula is as follows:

\[ \text{TIC} = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} y_i^2} + \sqrt{\frac{1}{n} \sum_{i=1}^{n} \hat{y}_i^2}} \quad (17) \]

Where \( y_i \) is the corresponding observation, \( \hat{y}_i \) is the prediction of the \( i \)-th sample and \( n \) is the number of samples; \( S_{\text{model}} \) is the error generated by using the predicted value and \( S_{\text{wal}} \) is the used data The error produced by the mean forecast.

The comprehensive statistical results of each model on the train and test set are displayed in Table 4. It can be seen from the table that the prediction accuracy of the four models is relatively high, and the scores of all aspects of the model are also relatively excellent. The MRE is less than 0.1; the RMSE value fluctuates around 2; the MAE value is close to 1; the values of \( R^2 \) are greater than 0.9 and close to 1; the values of TIC are generally close to 0. The above values indicate that the predictive ability of the models is strong. Considering the error in the real test, the accuracy of the four models is already very high.
By comparing the errors of the training dataset and test dataset, it can be seen that although the prediction accuracy of the XGBoost model is relatively high overall, the prediction errors of the training dataset and the dataset are quite different. Among them, the difference of MRE reached 0.0371, the difference of RMSE reached 0.5528, and the difference of MAE reached 0.4207. From the analysis of error precision, it can be seen that the XGBoost model may have the problem of overfitting.

Whether WOA-DBN is in the training dataset or the test dataset, the values of the five evaluation indicators are all superior. The $R^2$ values of WOA-DBN on the training set and the test set are close to 1, which are 0.9822 and 0.9792 respectively; the TIC are both close to 0, which are 0.0218 and 0.0266 respectively; the MRE on the training set and other indicators are very small as 0.0365. It can be concluded that the WOA-DBN model has higher accuracy, stronger stability and stronger robustness than other models. XGBoost also performs very well, second only to WOA-DBN, followed by AdaBoost-SVR, and AdaBoost has the lowest performance in comparison.

Because DBN has a special network structure and training method, it has a simple structure, is easy to train, has fast convergence speed, and has strong generalization ability. Compared with other machine learning models, DBN has stronger feature extraction ability for high-dimensional nonlinear data. And in this study, the dataset uses full-component data, considering the impact of each component on the target value, therefore DBN has good training and prediction performance. Swarm intelligence optimization algorithm--WOA has simple operation, few control parameters, and can optimize the hyperparameters in the DBN very well. The combination of the two shows amazing predictive ability.

<table>
<thead>
<tr>
<th>Data evaluation indicators</th>
<th>AdaBoost</th>
<th>AdaBoost-SVR</th>
<th>XGBoost</th>
<th>WOA-DBN</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRE</td>
<td>0.096</td>
<td>0.077</td>
<td>0.0452</td>
<td>0.0365</td>
</tr>
<tr>
<td>RMSE</td>
<td>2.0967</td>
<td>2.0564</td>
<td>2.018</td>
<td>1.8629</td>
</tr>
<tr>
<td>MAE</td>
<td>1.8364</td>
<td>1.5858</td>
<td>0.9281</td>
<td>0.8624</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.9301</td>
<td>0.9328</td>
<td>0.9756</td>
<td>0.9822</td>
</tr>
</tbody>
</table>

Table 4 Train and Test results on each model used for comparative studies.
### 3.1.2. Evaluation based on Visualization

Using the visualization-based evaluation method can more intuitively show the accuracy of each prediction model. Fig. 6 is a prediction effect diagram of training set and test set of four prediction models, the abscissa represents the actual value of the data, and the ordinate represents the predicted value of the data. The smaller the difference between real value and predicted value, that is, the higher the prediction accuracy, the closer data points are to the 45° line.

![Fitting results of four models](https://ssrn.com/abstract=4520436)
From the visualization results of the above four models, the data points are densely distributed on both sides of the 45° line, indicating that the prediction accuracy of each model is high. It can be clearly seen that the distribution of data points of the WOA-DBN model is relatively balanced, and there are few abnormal points, which is the model with the highest prediction accuracy. The second is the XGBoost. The data in the training set can be better predicted, and the distribution of the test set is relatively scattered, but within the scope of engineering error. In contrast, the data points of the AdaBoost and AdaBoost-SVR models have a higher degree of dispersion. It shows that the prediction accuracy is slightly lower, but within the acceptable range.

3.2. Learning Curve Evaluation

In order to more fully evaluate the performance of the model constructed in this study, the learning curve of each model in the modeling process is drawn. Learning curve is the score change curve of the model on the training set and test set according to the size of different training sets, which can clearly show the state of the model, that is, overfitting and underfitting. For this reason, in this study, the data set is divided into 75% of the training set and 25% of the test set. In the verification process, a 10-fold cross-validation method is used to obtain the training error and test error under different data volumes. The specific results are shown in Fig. 7.
Fig. 7. Learning Curve results of four models

By analyzing the learning curves of the four models, the status of the models can be seen clearly and intuitively. As the training data continues to increase, the mean square error of the test data of each model decreases continuously, and the mean square error of the training data fluctuates within a small range and remains at a low level. It can be shown that the four models constructed in this study can predict the data set very well, and can reach an acceptable level.

Further analysis shows that the mean square error of the training data of the AdaBoost, AdaBoost-SVR and XGBoost models is quite different from the test data, and their difference is generally around 20. The above models have obvious overfitting phenomenon, that is, for the data is over-learned, breaking away from the original law of the data.

In contrast, the difference between the mean square error of the training data and the test data of the WOA-DBN model gradually decreases with the increase of the training data. When the amount of data increases to the maximum, the difference between the two is stable at 2 nearby, it can be seen
that the WOA-DBN model has carried out reasonable learning on the data set, reflecting the real law
between the data, and achieving more accurate prediction results.

3.3. Analysis of single factor control variables evaluation

From the comprehensive analysis of the above evaluation results, it can be seen that the WOA-
DBN model has excellent performance. In order to gain a deeper understanding of the impact of input
parameters on the output value, the WOA-DBN model is used to predict the variation of MMP value
with each factor. The analysis was based on a single factor control variable approach, whereby one
study variable was gradually varied while keeping the other variables constant. In this study, the
effects of reservoir temperature, crude oil composition and injected gas composition on MMP were
selected. Due to the relatively small content of some components in the data set, this part is not
considered for the time being. Fig 8 shows the change curve of MMP relative to each influencing
factor. It can be seen from the figure that the change of MMP is not stable, showing a certain law
with the change of various factors.

(1) In the study of the influence of reservoir temperature on MMP, keeping the average value
of other variables constant, change the temperature gradually, and observe its influence on MMP. Fig
8(a) shows that MMP increases with increasing temperature. MMP increased sharply after 30°C, a
trend consistent with experimental measurements. It shows that the WOA-DBN model can accurately
predict the MMP value and reveal the law of physical changes more accurately.

(2) In the study of the influence of crude oil components on MMP, keeping other variables
unchanged, change N₂-C₁-CO₂、C₂-C₃-C₄-C₆ and C₅ in crude oil one by one, and observe their
influence on MMP. Figure 8(b) shows that MMP increases with the increase of volatile component
content in crude oil, decreases with the increase of intermediate light hydrocarbon content, and
increases with the increase of C₅ content. Likewise, the WOA-DBN model conforms to certain
physical laws.
(3) In the study of the influence of the molar mass of C_{7+} in crude oil on MMP, keeping other variables constant and change the molecular weight of C_{7+} one by one. Figure 8(c) shows that the value of MMP increases with the increase of the molecular weight of C_{7+}. The result is more obvious, and WOA-DBN still conforms to the physical law.

(4) In the study of the influence of the components in the injected gas on MMP, the mole fractions of N_{2}-C_{1}, C_{2}-C_{3}-C_{4}-H_{2}S in the injected gas were changed one by one. Figure 8(d) shows that the value of MMP increases with the content of N_{2}-C_{1} and increased, and decreased with the increase of C_{2}-C_{3}-C_{4}-H_{2}S content.

It can be shown that simply selecting a certain component to study MMP is one-sided, and the influence of all components on MMP cannot be ignored, and the contribution of each component should be considered separately.

**Fig. 8.** Sensitivity analysis based on WOA-DBN model

4. Conclusion
In this study, we collected and sorted out the full component data of 180 groups of fluids at home and abroad, and used the method based on domain knowledge and four correlation ranking methods to screen and evaluate the 21 influencing factors of the sample data set, and divided 10 groups of characteristic factors in total, and use partial correlation analysis to study whether there are characteristic factors that have the same impact on MMP. Afterwards, four kinds of intelligent algorithm models were established, 10 groups of characteristic factors were used as the input variables of the models, and the MMP value was used as output variable. In view of the one-sidedness of the previous research evaluation methods, the evaluation based on model accuracy, learning curve and single-factor control variable were proposed. After evaluating the three evaluation methods, the following conclusions can be drawn.

(1) Using the method of domain knowledge to assist mathematical statistical results, analyze the influence degree of each characteristic factor on MMP one by one from the perspective of petroleum and natural gas engineering, and divide the 21 full-component data in the data set into 10 groups of characteristic factors, which are reservoir temperature (T_R), the molecular fraction of all components of crude oil (O_{N2-C1-CO2}、O_{C2-C3-C4-C6}、O_{C5}、O_{C7+}), the molecular fraction of all components of injected gas (G_{CO2}、G_{C1-N2}、G_{H2S-C2-C3-C4}、G_{C5-C6-C7+}) and the molecular weight of C_{7+} in crude oil (MW_{C7+}). Among them, for the intermediate heavy hydrocarbons (C_5-C_6) in crude oil, the traditional cognitive division method is broken, and the two are considered separately, which is conducive to improving the accuracy of the prediction results.

(2) The four artificial intelligence models constructed in this study, AdaBoost, AdaBoost-SVR, XGBoost and WOA-DBN, no matter which evaluation method is used, the final prediction results are within the accepted error range.

(3) Whether WOA-DBN is in the training set or the test set, the values of the five evaluation indicators are all superior. The $R^2$ values of WOA-DBN on the training set and the test set are close to 1, which are 0.9822 and 0.9792 respectively; the TIC on the training set and the test set are both...
close to 0, which are 0.0218 and 0.0266 respectively; The MRE on the training set and other indicators are very small as 0.0365. This shows that WOA-DBN as a deep learning algorithm exhibits amazing predictive effects. For high-dimensional nonlinear sample data sets, it is easier to obtain the expected effect by using deep learning algorithms first.

(4) The learning curve analysis shows that with the continuous increase of training data, the mean square error of test data of each model decreases continuously, and the mean square error of training data fluctuates within a small range and remains at a low level. Among them, the error between the training data and the test data of the WOA-DBN model decreases gradually with the increase of the number of samples, and finally stabilizes at around 2. It shows that the deep learning model using the swarm intelligence optimization algorithm can conduct reasonable learning on the sample data set, and explain the laws between the data more accurately, so as to obtain accurate prediction results.

(5) According to the results of single factor control variables, within a certain range, the MMP value predicted by the WOA-DBN model can accurately reveal the law of physical changes. Machine learning has broad application prospects in the field of oil and gas engineering. However, most of the current research has many disadvantages, such as incomplete consideration of characteristic factors, lack of parameter screening, single model evaluation method, low prediction accuracy of intelligent models, and underfitting or overfitting. It is recommended to use the modeling and evaluation process in this paper when dealing with high-dimensional nonlinear data in related fields, in order to obtain reliable and accurate results and guide on-site production and related theoretical research.
CRediT authorship contribution statement

References

Amine, H. and Mohammed, M.S., 2023. Application of artificial neural networks (ANN) and gray relational analysis (GRA) to modeling and optimization of the material ratio curve parameters when turning hard steel J The International Journal of Advanced Manufacturing Technology. 124(10).
Can, H. et al., 2023. Prediction of minimum miscibility pressure (MMP) of the crude oil- CO\textsubscript{2} systems within a unified and consistent machine learning framework J Fuel. 337.


Kun, S.U. et al., 2013. CO$_2$-oil Minimum Miscible Pressure Prediction by Genetic Algorithm Method.


Mirjalili, Seyedali, Lewis and software, A.J.A.i.e., 2016. The Whale Optimization Algorithm.


