Fuel 354 (2023) 129263

Contents lists available at ScienceDirect

Fuel

journal homepage: www.elsevier.com/locate/fuel

Full Length Article

Predicting minimum miscible pressure in pure CO_2 flooding using machine learning: Method comparison and sensitivity analysis

Harith F. Al-Khafaji [a,b], Qingbang Meng [a,*], Wakeel Hussain [c], Rudha Khudhair Mohammed [b,d], Fayez Harash [e,f], Salah Alshareef AlFakey [c]

[a] Key Laboratory of Theory and Technology of Petroleum Exploration and Development in Hubei Province, China University of Geosciences, Wuhan 430074, China
[b] Petroleum Research and Development Center, Ministry of Oil, Baghdad, Iraq
[c] School of Geophysics & Geomatics, China University of Geosciences, Wuhan 430074, China
[d] Interdisciplinary Graduate School of Engineering Sciences, Kyushu University, Fukuoka 816-8580, Japan
[e] State Key Laboratory of Geological Processes and Mineral Resources, School of Geophysics & Geomatics, China University of Geosciences, Wuhan 430074, China
[f] Geology Department, Faculty of Sciences, Damascus University, Damascus, Syria

ARTICLE INFO

Keywords:
MMP
ML
Pure CO_2 flooding
Empirical correlations
Computational methods
Sensitivity parameters

ABSTRACT

CO_2 injection for enhanced oil recovery (EOR) is widely recognized as an efficient technique for carbon capture, utilization, and storage (CCUS). This operation has a significant impact on various technical parameters, emphasizing the need to carefully consider and select the optimum approach. Among these factors, the minimum miscible pressure (MMP) plays a crucial role in determining the effectiveness and performance of CO_2 injection. Therefore, this study aims to assess the reliability of machine learning (ML) in predicting the MMP of pure CO_2 and examine the influence of different independent parameters. To achieve this, five ML methods were employed to predict the pure CO_2 MMP, and the results were compared to statistical evaluations based on empirical correlations. In addition, three types of data with different functional input parameters were used in this research. Two types of data were obtained from existing literature, while the third category was collected from the thesis and PVT reports for specific Iraqi oil fields. The ML models were constructed by splitting the dataset into 20% for testing and 80% for training using Python programming. The significance of this study lies in its ability to identify the most efficient approach for forecasting MMP. The results of this work revealed that the K-nearest neighbors (KNN) model indicated the best statistical evaluation among the ML learning algorithms for two types of data (2) and (3) in predicting the MMP for pure CO_2 flooding. This was evidenced by the lowest mean square error and the highest coefficient of determination. Additionally, the findings indicated that the support vector regression (SVR) method is an effective technique for smaller datasets. Moreover, the sensitivity analysis and assessment of the relative impacts of various input parameters revealed that the prediction of MMP is most sensitive to the composition of the injected gas and temperature, accounting for 46% and 28.5% of the variation, respectively. Finally, the presented ML models indicate exceptional accuracy, speed, adaptability in handling diverse conditions, and cost-effectiveness when compared to conventional approaches. These results verify the ability of ML models to provide high-quality predictions.

1. Introduction

Presently high hydrocarbon demand, enhancing total oil recovery from reservoirs is of utmost importance. To achieve this goal, extensive research literature and practical implementations discovered a broad variety of Enhanced Oil Recovery (EOR) technologies [1,2]. Particularly CO_2 flooding stands out as one of the most widely employed and highly impactful EOR methods for enhancing displacement capacity, sweep efficiency, and reservoir pressure [3,4]. Several oil reservoirs have extracted upwards of thirty percent of the oil initially in place (OIIP) through their primary and secondary production processes [5]. Technically, the application of the CO_2 flooding technique is recommended after the primary or secondary production stages [6,7]. There is a positive environmental impact from using CO_2 technology compared with other methods since it can be recycled after being injected into the reservoir, resulting in fewer CO_2 emissions into the atmosphere [8,9] and thereby lessening the greenhouse gas (GHG) issue [10,11]. CO_2

* Corresponding author.
E-mail address: mengqb@cug.edu.cn (Q. Meng).

https://doi.org/10.1016/j.fuel.2023.129263
Received 5 May 2023; Received in revised form 17 June 2023; Accepted 19 July 2023

Nomenclature

EOR	Enhanced Oil Recovery
MMP	Minimum Miscible Pressure
ANN	Artificial Neural Network
MLR	Multiple linear Regression
ML	Machine learning
POS	Particle swarm optimization
RF	Random forest
DT	Decision Tree
KNN	K-Nearest Neighbors
SVR	Support Vector Regression
VIT	Vanishing Interfacial Tension
MWC_{7+}	Molecular weight of component C7+
MWC_{5+}	Molecular weight of component C5+
MWC_{6+}	Molecular weight of component C6+
X_{VOL}	Volatile oil components, including (C_1 &N_2)
X_{INT}	Intermediate oil components, including(C_2-C_4, CO_2, and H_2S)
EOS	Equation of State
LSSVM	Least squares support vector machine
MAE	Mean Absolute Error
MSE	Mean Square Error
MED	Median Absolute Error
R^2	Coefficient of determination
CMG	Computer modelling group
GHG	Greenhouse gases
OIIP	Oil initial in place
MLP	Multilayer perceptron
ABSVR	Adaptive boosting support vector regression
GMDH	Group method of data handling
Psat	Saturation Pressure
T_r	Reservoir Temperature
API	American Petroleum Institute
sp.gr	Oil density
APRE	absolute percent relative error
P_B	Bubble point pressure
Tc	Critical temperature
Pi	Reservoir pressure
RBF	Radial Basis Function
AARD	average absolute relative deviation
$\xi_i, \xi_i^{\wedge}$	Slack parameters
RF	Relevancy factor
PVT	Pressure, Volume, and Temperature
MKF	Mixed kernel function
MMC	Multiple mixing cell
RBFN	Radial Basis Function Networks
TLBO	Teaching learning-based optimization
ANFIS	Adaptive neuro-fuzzy inference system
CART	Classification and regression tree
AI	Artificial Intelligent
BIC	Akaike Information Criterion (AIC)
AIC	Bayesian Information Criterion (BIC)

injection is classified into numerous varieties, including immiscible CO_2 flooding, miscible CO_2 flooding, near miscible CO_2 flooding, huff and puff, and so on,[1] each of which has particular implementation restrictions.

During miscible CO_2 flooding, the minimum miscible pressure (MMP) is an important parameter to identify the mechanism of CO_2 injection operation into the reservoir [12]. Physically, dynamic miscibility occurs at MMP, which is the lowest pressure at which CO_2 is soluble in the reservoir's crude oil [13] and at this point, 80% of OIIP can be recovered at CO_2 breakthrough [14]. Furthermore, the accuracy of the MMP prediction for CO_2 injection is critical to avoiding process failure and ensuring good sweep efficiency [15,16]. Because the CO_2 flooding method is costly, MMP is regarded as one of the most essential screening criteria for determining the accuracy of miscible CO_2 flooding [17]. Due to this, a few experiments have been suggested to detect MMP, including the slim-tube experiment, which is the first conventional method described by Yellig et al. [18]. The rising bubble test is an efficient method suggested by Christiansen & Haines [19] to treat the slow rate issue in the slim tube experiment. Vanishing Interfacial Tension (VIT) is a common experiment introduced by Rao & Lee [20] to predict MMP. These experimental methods have a high degree of accuracy [21]. Nevertheless, they require a long time and a high cost. Also, they are impacted by different experimental factors [22] and may be subject to human error [1].

As mentioned in the literature, plenty of empirical correlations have been introduced in defining the MMP of pure CO_2 flooding. As previously stated, the estimation of pure CO_2 MMP is dependent on several major parameters, including the molecular weight of (C_5+), reservoir temperature, the mole fraction of volatile oil elements, and the mole fraction of intermediate oil elements. The oldest MMP correlation proposed by Holm & Josendal [14] depends on the molecular weight of C_5+ and the reservoir temperature of the crude oil. Lee [23] suggested a correlation between individuals relying on only reservoir temperature to estimate the MMP of pure CO_2. Yellig et al. [18] modified (L.W. Holm & Josendal, 1974)'s empirical relationship for anticipating MMP as a variable of reservoir temperature. Cronquist [24] employed three independent variables of crude oil such as molecular weight (C_5+), reservoir temperature, and volatile oil components (C_1 and N_2) as functions to predict MMP. Using gas purity, reservoir temperature, and pressure, Johnson and Pollin [25] presented a correlation for determining MMP, which applies to various kinds of stock tank oil and live oil [26]. Alston et al. [27] hypothesized a relationship by including extra variables such as reservoir temperature, the molecular weight of (C_5+), the mole fraction of volatile oil components (C_1 and N_2), and intermediate oil components (C_1-C_4). During the same period, Glaso [28] put forward another correlation as a function for three parameters: reservoir temperature and molecular weight of (C_7+), although it was of limited utility for intermediate oil compositions (C_2-C_6) in the fluid of the reservoir. In 1993, Zuo et al. [29] adjusted the relation presented by Johnson and Pollin [25] by utilizing two independent parameters: volatile and light components for reservoir oil. Dong et al. [30] contributed to enhancing the precision of forecasting MMP and proving the influence of the gas solution on CO_2, and the findings confirmed that the gas solution should be taken into consideration. Emera & Sarma [31] utilized a genetic algorithm to develop the correlation of Alston. Applying the alternative conditional expectation (ACE) technique, Shokir [32] and Alomair et al. [17] found a different correlation that could be utilized to compute the MMP of CO_2 injection. Although these correlations and mathematic methods have a quicker and less expensive prediction method (MMP), they are incapable of being applied to a broad variety of conditions and still contain several inadequate, strict assumptions [11,33].

Simultaneously, based on a computational model using the equation of state (EOS), numerical simulation has been employed for MMP prediction of CO_2 injection by utilizing commercial software [17]. Abdullah and Hasan [6] performed research to estimate the effect of miscible CO_2 injection on the recovery factor, which investigated MMP calculation from two equations (Glaso [28] and Alson et al. [27]) versus simulation, and the results confirmed that the calculation of MMP from the Glaso equation was close to the simulation. However, sometimes this

computational method still takes more time to tune the physical properties of the fluid, requires more effort to achieve stability [34], and is regarded as a costly process due to the necessity to get a license [35]. Moreover, computational models are dependent on the amount of precision for physical characteristics [36]. Nonetheless, plenty of approaches have been used to estimate MMP, such as experiments, simulations, and known empirical correlations, but these techniques are affected by various factors and cannot be used in all circumstances.

Contrarily, artificial neural networks (ANN) have been employed for different areas of oil and gas engineering, which has contributed to reducing wasted time and using it for broad operational conditions [15]. One of these areas of application is forecasting the MMP of CO_2 injection. In general, the advantage of machine-learning approaches is that they tend to sidestep the challenges of conventional problem-solving methods and may be used to solve a wide range of issues [37], whereas the first attempt has been made for MMP prediction of the CO_2 technique by using the ANN backpropagation method developed by Huang et al. [33]. Numerous studies have proven the efficiency and accuracy of the ML and ANN to compute the MMP of pure or impure CO_2 injection. Birang et al. [38] constructed an original ANN model that includes a multilayer perceptron (MLP) with two-layer back-propagation for predicting MMP during hydrocarbon injection based on 52 data points. Through comparisons with MMP values obtained from slim-tube experiments and correlations, the average error and the correlation coefficient (R^2) were determined as 18.58% and 0.938, respectively. Dehghani et al. [39] employed a genetic algorithm (GA-ANN) for estimating MMP during gas injection processes by using experimental data of MMP around 46 points. Shokrollahi et al. [40] utilized the least squares support vector machine (LSSVM) for the first time to anticipate the MMP of pure or impure CO_2, achieving an impressive 9.6% overall AARD using 147 experimental databases. Tatar et al.[41] used the same datasets in another investigation to construct another CO_2 MMP approach that relies on the kernel function radial basis function (RBF). In 2014, the fuzzy logic technique has adopted by Ahmadi and Ebadi [42] to define the MMP of gas injection and oil reservoirs. At the same time, Sayyad et al. [43] suggested a new approach (PSO-ANN) for expecting pure and impure CO_2 MMP. In 2016, Zhong & Carr [44] advanced a new mixed kernel function of the SVR model (MKF-SVR) to anticipate the minimum miscible pressure for CO_2 pure and impure based on three independent parameters with the highest R^2 (0.93) and the lowest RMSE (1.9151). In 2017, Karkevandi-Talkhooncheh et al. [45] employed an adaptive neuro-fuzzy inference system (ANFIS) based on large data sets (approximately 270 data points) to create multiple intelligent models for forecasting CO_2 MMP for pure and impure, with a total AARD of 7.53%. Depending on published data around 144 points, Saeedi Dehaghani and Soleimani [46] suggested new models "a hybrid artificial neural network (ANN) and stochastic gradient boosting (SGB)" for CO_2 MMP prediction in 2020. At same year, Dargahi-Zarandi et al. [47] used 270 points of databank stated by Karkevandi-Talkhooncheh et al. creating various smart developed techniques for forecasting CO_2 MMP using GMDH, MLP, and ABSVR. Ghiasi et al. [48] suggested a regression tree and classification improved using AdaBoost (AdaBoostCART) with an ANFIS model to predict CO_2 MMP. Chen et al. [49] assessed the efficacy of numerous ML techniques for forecasting the MMP of CO_2 injection. More recently, Lv et al. [15] carried out comprehensive research in which they employed three models (tree-based, deep learning, and thermodynamic) to anticipate the MMP of CO_2 by using an extensive databank of 310 with an overall AARD of 1.34%, and the parameter sensitivity demonstrated that reservoir temperature has a significant impact on predicting MMP.

The wide range of contributions in MMP modeling suggests that the task of forecasting MMP for CO2 remains challenging, emphasizing the need for more precise and robust predictions. As can be noticed in Table 1, the majority of published research has not studied the impact of other independent variables on MMP expectations in their AI models. In addition, the performance comparison of the computational model and machine learning was not addressed. Furthermore, the main distinction between the current research and prior studies lies in the utilization of novel datasets incorporating several additional parameters. This enables a comprehensive evaluation of the impact of these elements on MMP prediction. Indeed, the main contribution of this study is to forecast the MMP for pure CO_2 with various input parameters and investigate the performance of machine-learning models for predicting the MMP with a wide range of data. Therefore, five machine-learning (ML) methods were employed for this objective in several different scenarios to achieve optimum prediction. In order to verify the reliability models, a comparison between the literature correlations and the computational method with ML techniques was performed. Following that, the effectiveness of these approaches is assessed using a range of statistical and graphical error evaluations. Finally, sensitivity analysis and influence parameters are examined to thoroughly investigate the models' dependability.

2. Theoretical background and methodology

2.1. Machine learning techniques

2.1.1. Multiple linear regression (MLR)

MLR approach is popular among the most widely used supervised ML algorithms for predicting, differentiated by its capacity to analyze data quickly and easily through accommodating out over one independent parameter, in contrast to other linear regression methods[50–52]. MLR is a multivariate linear regression approach used to simulate the linear interconnectivity between many independent parameters (input variables) and one output-dependent parameter (output variable) [53]. Nevertheless, this technology has proven to be an efficient and important method for detecting data structure patterns. Evidently, the MLR technique's approach depends on the predictions that are the existing correlation between the dependent and independent parameters [54]. Based on the numerous factors X, a hypothetical dependent variable Y is forecast mathematically. Furthermore, the MLR paradigm can be expressed using the formula (1) [55]:

$$Y_i = B_0 + B_1\ X_{1i} + B_2\ X_{2i} + \ldots. + B_p X_{pi}\ + \alpha_i \tag{1}$$

where Y_i indicates the dependent variable (output) and p denotes the independent variable (input) (X_1, X_2…, X_p). B_0 represents the intercept term, as well as B_i the coefficient value (slope) determines the contribution for every predicted parameter. α_i is the model's random error item, and i = (1, 2, 3…, n) denotes the total number of samples. Fundamentally, the least square approach is used to create the multiple linear regression model, with the goal of reducing the overall percentage of error between both the observed and anticipated dependent variables [56].

2.1.2. Support vector regression (SVR)

SVR is a common supervised machine learning algorithm that was developed to solve challenges in model production and generalization. In 1995, Vapnik [57] created and developed an SVR model that quickly earned popularity due to its numerous appealing properties. Typically, the SVR can be employed to solve both linear and non-linear regression issues. The primary goal of SVR is to generate a function f(x) that represents the maximum deviation ε from the target Y_i acquired for all training data while remaining as flat as feasible. As a result, the datasets fall between the two margin boundaries, preventing the inclusion of outliers under proper conditions [22], as illustrating in Fig. 1.

The nonlinear-support vector regression technique is always conducted by map in an area of high dimensional features (x_i, i.e.) there is a map ($\varphi : x \rightarrow \varphi \in R$) from which the regression hyperplane is formed as:

$$f(x) = \omega\varphi(x) + b \tag{2}$$

where ω and b denote the weight vector of the hyperplane and hyperplane bias, respectively.

Table 1
A summary of the most literature's proposed models for prediction MMP of CO2-oil.

Author	Model	Independent parameters	Limitations
Holm & Josendal 1974 [14]	A graphical model that depends on two variables (MWC5+, reservoir temperature)	Tr, MWC_{5+}	-Temperature range limit (32.2 °C to 82.2 °C) -Pressure range limit (9.65 MPa to 22 MPa). $-180 < MWC5^{+} < 240$
Cronquist 1978 [24]	$MMP = 0.11027 \times (1.8T_r + 32)^{B}$ $whereB = 744206 \times 10^{-6} + (11.038 \times 10^{-4} \times MW_{C5+}) + (15.279 \times 10^{-4} \times X_{VOL})$	Tr, MWC_{5+}, Xvol	-Oil API gravity range from 23.7° to 44°, -Tr range from 21.67 to 120 °C, -MMP range from 7.4 to 34.5 MPa.
Yelling and Mectalfe 1980 [18]	$MMP = 126472 \times 10^{-4} + 1.5531 \times 10^{-2} \times (1.8\ T_r + 32) + 1.24192 \times 10^{-4} (1.8\ T_r + 32)^2 - (716.94/32 + 1.8T_r)$	Tr	-35 °C < TR < 88.9 °C
Glaso 1985 [28]	$MMP = 810 - 3.4MW_{C_{7+}} (0.017 \times 10^{-7} (MW_{C_{7+}})^{1.2785} e^{(786.8MW_{C_{7+}}{}^{-1.058})}) T_r$	Tr, MWC_{7+}	-The range of intermediate components (C_2– C_6) -Specific gravity
Alston et al. 1985 [27]	$MMP = 6056 \times 10^{-9} \times (1.8\ T_r + 32)^{1.06} \times (MW_{C5+}) \times (X_{VOL}/X_{INT})^{0.136}$	Tr, MWC_{5+}, Xvol, Xint	Bubble point pressure
Zuo et al 1993. [29]	Developed models based on previous models and equation of state.	Tr, average molecular weight, °API, Xvol, Xint	-Composition of gas drive -Crude oil composition -Binary interaction coefficient
Dong et al. [30] 2000	Experimental study to investigate the effect of gas composition on prediction MMP of CO2.	/	/
Huang et al 2003 [12].	Developed ANN model to predict CO2-MMP for pure and impure	Tr, MWC_{5+}, Xvol, Xint,	-Temperature unit
Emera & Sarma 2005 [31]	Proposed several equations to predict CO2- Oil MMP based on a genetic algorithm with special limitations for each equation.	Tr, MWC_{5+}, Xvol, Xint	-Bubble point pressure - Percentage of the volatile and intermediate oil fractions.
Shokir [32]	Developed an advanced equation based on the alternating conditional expectation (ACE) algorithm.	Tr, MWC_{5+}, Xvol, Xint, composition of injected gas	
Birang et al 2007 [40]	ANN-backpropagation network model to forecast CO2-MMP.	Tr, MW of C_2-C_5 and MWC_{7+} component, Xvol, $X_{C1\text{-}C5}$, composition of injected gas.	- Hidden neurons
Dehghani et al. 2007 [39]	A hybrid neural genetic model to predict MMP of CO2.	Tr, reservoir fluid composition, and injected gas composition	-Number of hidden layers -learning and momentum coefficients
Shokrollahi et al. 2013 [40]	Evolved Model by utilizing LSSVM	Tr, MWC_{5+}, (Xvol/Xint), composition of injected gas	-Hyper-parameters of LSSVM method
Tatar et al. 2013 [41]	Intelligent developed model based on RBFN	Tr, MWC_{5+}, (Xvol/Xint), composition of injected gas	-Objective function
Sayyad et al. 2013 [43]	Hybrid ANN model by using neural (POS) algorithm for forecasting MMP	Tr, reservoir fluid composition, and injected gas Composition	-Number of hidden layers -Number of neurons -Weight
Ahmadi and Ebadi 2014 [42]	Fuzzy Model for prediction of CO2-MMP.	Tr, MWC_{5+}, (Xvol/Xint), Tc	-Membership functions setting
Zhong & Carr 2016 [44]	Developed mixed new model based on SVR and POS algorithm	Tr, MWC_{5+}, (Xvol/Xint), average Tc	-Objective function -Hyperparameters setting
Karkevandi-Talkhooncheh et al. 2017 [45]	Evolved ANFIS model for prediction MMP of CO2-oil.	Tr, MWC_{5+}, Xvol, Xint, Tc	-Membership functions setting - Hyperparameters selection
Saeedi Dehaghani and Soleimani 2020 [46]	Four developed models (ANN, POS-ANN, ANN-TLBO, and SGB)	Tr, MWC_{5+}, Xvol, Xint, composition of injected gas	-Number of neurons in hidden layers -Transfer function -Learning rate -Trees number
Dargahi-Zarandi et al. 2020 [47]	Three developed models (GMDH, MLP, and ABSVR)	Tr, MWC_{5+}, Xvol, Xint, composition of injected gas	-Kernal function -Parameters optimization -Transfer function -Training algorithm
Ghiasi et al. 2021 [48]	Two developed models (hybrid-ANFIS and AdaBoost-CART)	Tr, MWC_{5+}, (Xvol/ Xint), composition of injected gas, Tc	-Membership functions setting - Epoch number - Number of trees
Chen et al. [49]	Eight ML models	Tr, reservoir fluid composition, and injected gas Composition	-Hyperparameters selection -Cross validation
Lv et al. [15]	Eight intelligent evolved models	Tr, MWC_{5+}, Xvol, Xint	-Data range -Parameter optimization
Present work	Five developed ML models for prediction MMP of pure CO_2.	A number of different parameters as listed in Table 2.	- Hyperparameters selection - Random state setting

Table 2
Statistical data for all dependent and independent parameters for three data types.

DATASETS		INPUT PARAMETERS	MIN	MAX	MEAN	STD
TYPE 1		Temperature (°C)	32.2	137.22	71.26585	26.37
		Mwc+5	136.17	391	208.2946	41.91
		X_{VOL}/X_{INT}	0.14	13.61	2.062827	2.46
		MMP (MPa)	6.89	42.5	17.50536	7.47
TYPE 2		Temperature (°C)	8.95	130	84.74073	25.97
	Composition of injected gas(Mol%)	HX_{N2}	0	80.1	3.361	14.44
		HX_{CO2}	0.59	100	54.049	40.604
		HX_{H2S}	0	50	4.196	10.113
		HX_{C1}	0	85.34	22.584	25.605
		HX_{C2-C6}	0	58.442	15.77165	17.95
		HX_{C7+}	0	0.98	0.020394	0.096
	Component of crude oil (Mol%)	X_{VOL}	4.405	54.98	19.77289	11.61
		X_{INT}	2.63	58.15	26.012	13.107
		X_{C5-C6}	1.909	11.19	6.755	2.30
		X_{C7+}	19.59	80.75	47.45	21.31
		MWC_{7+}	153.9	402.7	227.93	46.94
		MMP (MPa)	6.55	41.47	21.11	8.083
TYPE 3		Temperature (°C)	73.88	148.8	102.95	15.53
		X_{VOL}	17	49	32.42	6.67
		X_{INT}	18	37	24.96	3.83
		Mwc+6	156	450	271.95	77.86
		X_{C6+}	21.9	54.4	38.5	7.5
		API	18.5	45.9	27.73	6.667
		Sp.gr	0.8	0.94	0.889	0.034
		P_b (MPa)	7.49	24.15	16.84	4.24
		MMP (MPa)	22.56	58.94	33.74	9.753

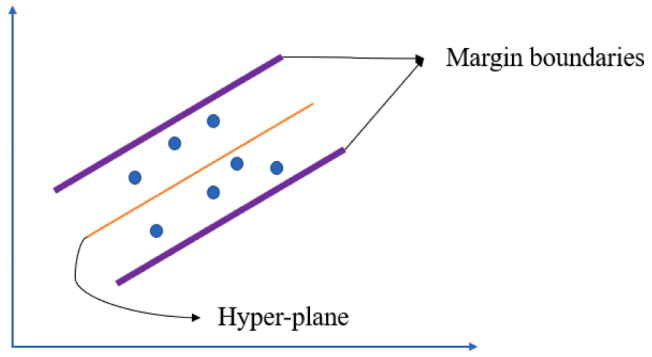

Fig. 1. The principle of the hyper-plane and margin boundaries in SVR.

Afterward, adopting a loss function as being unaffected by the influence of ε, its objective function and minimum restrictions can be described as follows:

$$\frac{1}{2}||\omega||^2 + C\sum_{i=1}^{I}\left(\xi_i + \xi_i^{\wedge}\right) \tag{3}$$

where C is a hyperparameter that determines the trade-off between maximizing the margin and minimizing the classification error that needs to be carefully tuned to achieve the right balance between model complexity and generalization performance [58].

As a result, the SVM model mentioned above is constrained by the following constraints:

$$\begin{cases} Y_i - \omega\varphi(x_i) - b < \varepsilon + \xi_i^{\wedge} \\ \varphi(x_i) + b - Y_i < \varepsilon + \xi_i \\ \xi_i, \xi_i^{\wedge} > 0 \end{cases} \tag{4}$$

where, $\{\xi_i, \xi_i^{\wedge}\}$ represents the slack parameters that measure the output characteristics' divergence from the positive as well as negative classes.

By utilizing the Lagrange function as the SVR's linear situation and picking partial derivatives with regard to the main variables, they employed it to solve Eq. 3. Then, it can set the resultant derivatives to zero [59]. The answer is given by:

$$MAX = -\frac{1}{2}\sum_{i,j=1}^{I}\left(\alpha_i - \alpha_i^{\wedge}\right)\left(\alpha_j - \alpha_j^{\wedge}\right)K(x_i, x_j) - \varepsilon\sum_{i=1}^{I}\left(\alpha_i - \alpha_i^{\wedge}\right) + \sum_{i=1}^{I}\left(\alpha_i - \alpha_i^{\wedge}\right)Y_i \tag{5}$$

where, $K(x_i,x_j)$ represents the Kernel Function that describes of the inner product $\langle\varphi(x_i)|\varphi(x_j)\rangle$. The- Gaussian kernel is regarded as the most proper function of the kernel, which is also called the radial basis function [60]. It is described as:

$$K(xi, xj) = (\varphi(x_i)\varphi(x_j)) = \exp\left(-\gamma||x_i - x_j||^2\right) \tag{6}$$

where $||x_i - x_j||^2$ is expressed as the square of the Euclidean distance that separates both feature vectors, and the Gaussian kernel width variable is denoted by γ. The regression function is generated by solving 3 with the constraint equation:

$$f(x) = \sum_{j=1}^{I}\left(\alpha_i - \alpha_i^{\wedge}\right)K(x_i, x) + \overline{\overline{b}} \tag{7}$$

where the calculation of $\overline{\overline{b}}$ can be simply omitted by pretreatment and centralization of the data, eventually resulting in a bias of zero [61].

2.1.3. Decision trees (DT)

DTs is an effective algorithm that applies to solving classification and regression problems based on data set splitting and was proposed by Breiman et al in 1984 [62]. DTs have been widely utilized in variable selection, data manipulation, missing value management, and prediction because of their simplicity, explainability, capability to provide visual analysis, and low processing cost [61,63]. The DTs algorithm consists of roots, internals, and leaves or nodes connected by branches [64]. Each terminal node, or leaf, has a basic regression model linked to it that only applies to that node. After the induction process is complete, pruning may be used to improve the tree's generalization capability by

eliminating structural complexity. The pruning criterion might be the number of cases in nodes [61]. Typically, the processes of DTs begin at the root node, which is placed at the top of the tree. The root node carries out operations on the input data, while the leaf is responsible for delivering the output data. Data begins to flow from the root node to internal nodes, then to leaf nodes. As a result, the model resembles an upside-down tree [34].

As elaborated by Breiman[62], The major acts in a DT's development phase are splitting, pausing, and pruning. The first operation of DT is splitting, which attempts to supply the optimal splitting for the dependent properties. The advancement phase begins by splitting the training data at the root node. The- splitting progresses to internal nodes. The dividing procedure will continue until the set halting requirements are met. In addition, the pruning strategy tries to reduce the intricacy of the tree, and prevent overfitting [34]. In DT, the goal of optimal splitting is to maximize purity while minimizing impurities.

$$\Delta i(s,t) = i(t) - p_l i(t_L) - - p_r i(t_r) \tag{8}$$

where, s denote the nominee split at node t, and the node t is split by s into the left of the child node t_l with a ratio of p_l, and the right of the child node t_r with a ratio of p_r. $i(t)$ is the measurement of the impurity before splitting, $i(t_l)$ and $i(t_r)$ are the measurement of the impurity after splitting, and $\Delta i(s,t)$ is the measurement of the reduction in impurity from split s.

The most prevlant approximations for computing the impurity is Gini index for measuring i(t), which can be described it by the following Eq. (9) [62]:

$$G_I\left(t_{X_{(x_i)}}\right) = 1 - \sum_{j=1}^{n} f\left(t_{X_{(x_i)}}, j\right)^2 \tag{9}$$

where $f(t_{X_{(x_i)}}, j)$ is the fraction of datasets with the value x_i that belong to leave j as node t. The criterion of decision tree splitting is depending on selecting the feature with the minimum Gini impurity index.

2.1.4. Random forest (RF)

After 17 years of introducing the decision tree approach, Breiman presented an RF as a more powerful model in 2001[65]. RF is a supervised machine learning technique that is commonly utilized in regression and classification issues that incorporates the performance of many DT algorithms to generate classification or prediction models [65,66]. When the RF gets the input vector (S), containing the values of the various evidentiary characteristics examined for a specific training region, it generates N regression trees and mean values of the findings. After growing N such trees {T(S)}, the predictor of the RF regression is expressed by Eq. (10):

$$\widehat{f}_{rf}^{N}(S) = \frac{1}{N}\sum_{n=1}^{N} T(S) \tag{10}$$

RF algorithm theory is constructed on two concepts: random feature selection and bagging[67]. To prevent the links between the various trees, there is an important approach to carry out this process called bagging, which assists in the construction of diverse training data subsets and leads them to develop depending on the original training data. Bagging is a process for creating training data that involves randomized resampling of the existing dataset by replacing without deleting the data picked from the input data set for producing the next subset "{h (x,Θ_N), n = 1,…,N}", where {Θ_N} represents the random variable vectors with the same distribution [61]. As a result, some of the data might be used several times throughout the training, whereas others might never be utilized. Consequently, higher stability is gained, since it renders it more durable in the face of minor deviations in input data, while also increasing forecast accuracy. Another interesting feature is that RF classifier trees develop without pruning, rendering them computationally light [61,65]. To reduce the generalization error and the relation between the trees, the RF chooses input data at random rather than selecting the best data set.

The establishment of the forest tree requires choosing the sub-feature from the original feature haphazardly. Afterward, different splitting ways are carried out selecting the best feature at the root node, and the inside node tests are picked using the same splitting strategy till the leaves are reached. "Out of Bag" (OOB) means the part of the dataset that is excluded from the training, but these data have another function and are utilized to assess the model's performance. Therefore, the positive thing about the RF is that it doesn't require a validation assessment [15,68]. Furthermore, the predicted OOB output for data S is provided below:

$$H^{OOB}(S) = argmax\sum_{n=1}^{N} I(h(S)) = y \tag{11}$$

And the following equation is used to compute the error of the OOB dataset:

$$\beta^{OOB}(S) = \frac{1}{|D|}\sum_{s,y\in D} I\left(H^{OOB}(S) \neq y\right) \tag{12}$$

Finally, the RF algorithm's randomness operation is governed by the variable q, which is defined as $q = \log_2 d$. The following formula is employed to compute the feature importance of the variable Si:

$$I(S_i) = \frac{1}{N}\sum_{t}^{N} \overline{OBB}_{ERR_{t^i}} - OBB_{ERR_t} \tag{13}$$

where S_i represents the ith factor of the vector S, N describes the number of trees in the model, $\overline{OBB}_{ERR_{t^i}}$ signifies the estimated error of the permuted S_i sample's OOB samples in tree t, and the first OOB samples are displayed as the OBB_{ERR_t}, which includes the subset parameters. The permutation importance procedure demonstrates how much a feature is beneficial for the prediction. As a result, a trivial practical characteristic has no or little effect on network forecasting.

2.1.5. K-Nearest Neighbors (KNN)

The KNN approach is known as one of the most basic and non-parametric supervised machine learning techniques, which can be employed for both regression and classification [69]. The input variables in regression and classification comprise the positive integer k nearest training datasets inside a feature space. Typically, the forecasted data sample's output value is calculated by taking the mean of its k closest neighbors[69–71].

$$Y = \frac{1}{k}\sum_{i=1}^{k} Y_i \tag{14}$$

where Y_i is the ith instance in the sample of examples and Y is the query point's expectation (output). However, compared to regression, KNN predictions in classification problems are dependent on a voting mechanism, with the winner used to classify the query. Euclidean distance measuring is widely used in this technique to predict. Therefore, Euclidean distance between the sample instances and the query point must be specified in order to make predictions with KNN, which may be computed as follows [69,72]:

$$D(x,y) = \sqrt{S_i(x_i - y_i)^2} \tag{15}$$

Significantly, the primary advantages of the KKN method are its simplicity in tackling complicated tasks, efficacy, intuitiveness, and a wide variety of applications. Additionally, it is effective with large amounts of training data and can cope with noisy training datasets efficiently [70,71].

2.2. Concept of computational approach

Real pressure, volume, and temperature PVT is crucially required during reservoir modelling. Equations of state (EOS) are considered important techniques for thermodynamic and mathematical modeling of fluid-phase behavior, and PVT results are utilized to tune these equations. Typically, Cubic equation of state is one of the most widely used techniques for determining the fluid-phase behavior of reservoir oil that was developed by Van der Waals in 1873 [26,73,74]. Several equations of state have been proposed by many authors. As stated in the literature, one of the most effective and accurate equations is Peng-Robinson[75], which agrees well with the experimental findings [76], as specified below:

1- The general formulation of cubic EOS can be expressed as follow:

$$f(P, V, T) = 0$$

2- The equation of Peng-Robinson can be described as follow:

$$P = \frac{RT}{V-b} - \frac{a(T)}{V(V+b)+b(V-b)} \tag{16}$$

$$a(T) = a_c\alpha(T) \tag{17}$$

$$a_c = \frac{0.45724R^2T_c^2}{P_C} \tag{18}$$

$$\alpha(T) = \left(1 + m\left(1 - \sqrt{\frac{T}{T_c}}\right)\right)^2 \tag{19}$$

$$b = \frac{0.07780RT_C}{P_C} \tag{20}$$

$$m = 0.37464 + 1.54226\omega - 0.26992\omega^2 \tag{21}$$

where T, P and V indicate temperature, pressure, and volume, respectively. R, Tc, Pc and denote standard gas constant, critical temperature, and pressure, respectively.

Frequently, the variables of the equation of state (EOS) must be adjusted (tuned) prior to producing useful reservoir predictions. During calibration, the variables of the EOS adjust to ensure that the forecasts correspond to a wide range of experimental data. Furthermore, WinProp is one of the common software components in the CMG program that will be used to construct a PVT model by using EOS and thermodynamics in order to predict MMP after achieving optimal matching. Based on that, the most common techniques for calculating MMP are mixing-cell approaches, which are employed in a variety of commercial products[15,77]. In this study, three-parameter Peng-Robinson (PR) employed to estimate MMP by using multiple mixing cell approaches.

2.2.1. Cell-to-cell (multiple mixing cell) concept

This technique suggested by Ahmadi and Johns [78] for MMP prediction is based on the concept of separating the fluid system into many mixing cells, each reflecting a distinct step of the miscibility process, which depends on one of the EOS types. Typically, in this method two cells employes at first calculation of MMP for CO_2. The fluid is considered to flow successively through these cells, with mass transfer occurring between them to achieve phase equilibrium. Generally, each cell's fluid composition is estimated using mass balance and phase equilibrium formula $Z = X^o + \alpha(Y^G - X^o)$. Iterations are performed till a specific convergence threshold is fulfilled. The pressure and composition of the fluid in the initial cell are modified to represent the injection of CO_2 into the reservoir. Once the fluid composition and pressure in the first cell are known, they can be used to determine the pressure and fluid composition in the next cell. By comparing the fluid composition at the final stage of the steps to the starting composition, the miscibility of the oil and injection gas is calculated. The fluids are deemed miscible when the variance between their beginning and final compositions is less than a particular threshold, and their corresponding pressure represents the minimum miscible pressure, as implied in Fig. 2. In this work, CMG software has been employed to estimate the minimum miscible pressure for CO_2 injection.

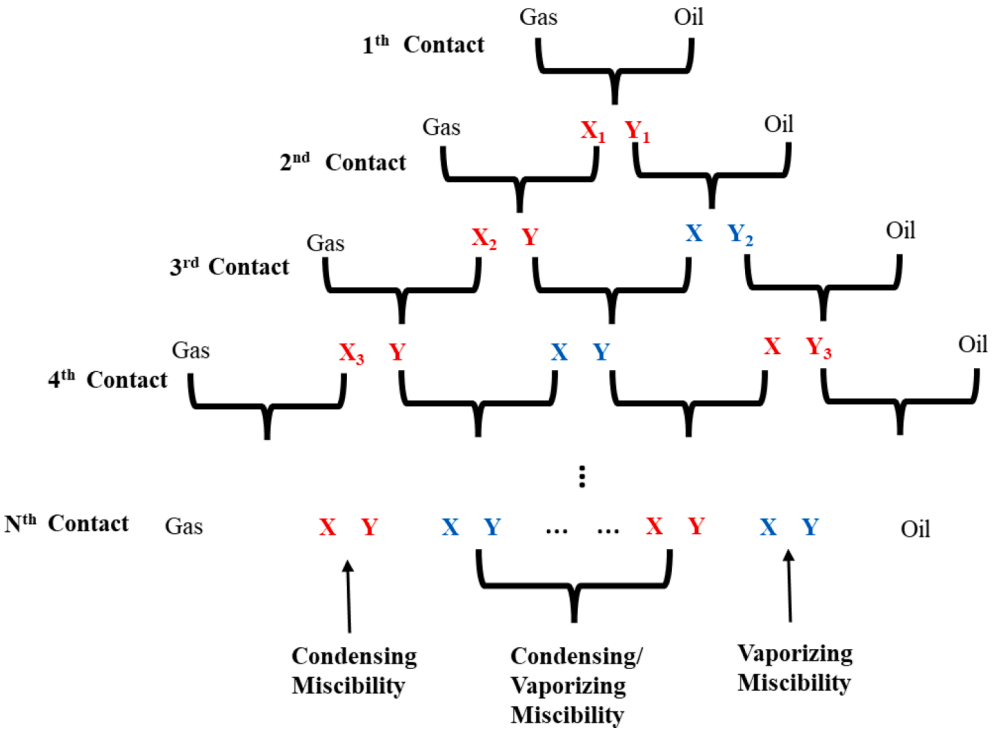

Fig. 2. Demonstration of the stages of MMP calculation by the multiple mixing method.

2.3. Model assessment techniques

Various mathematical parameters were employed to evaluate the competency and accuracy of the created models. To assess the performance of the established models, the most important variables have been utilized to evaluate the models of prediction MMP, such as absolute percent relative error, mean absolute error, mean square error, coefficient of determination (R^2), and median. In order to achieve further evaluation, kernal density estimation (KDE), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) have been used to provide a trade-off between model complexity and goodness of fit, which are shown below in the following equations [61,70,79]:

- Absolute Percent Relative Error (APRE)

$$\text{APRE} = \frac{1}{V}\sum_{i=1}^{v}\left|\left(\frac{MMP_{i,EXP} - MMP_{i,PRED}}{MMP_{i,EXP}}\right)\right| \times 100 \tag{22}$$

- Mean Absolute Error (MAE)

$$MAE = \frac{1}{V}\sum_{i=1}^{v}|MMP_{i,\,EXP} - MMP_{i,\,PRED}| \tag{23}$$

- Root Mean Square Error (RMSE)

$$\text{RMSE} = \sqrt{MSE} = \frac{1}{V}\sum_{i=1}^{v}(MMP_{i,EXP} - MMP_{i,PRED})^2 \tag{24}$$

- Coefficient of determination (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{V}(MMP_{i,EXP} - MMP_{i,PRED})^2}{\sum_{i=1}^{v}(MMP_{i,EXP} - \widehat{MMP})^2} \tag{25}$$

- Median (MED)

$$\text{MED} = \text{median}(|MMP_{i,EXP} - MMP_{i,PRED}| \tag{26}$$

- Akaike Information Criterion (AIC)

$$AIC = -2\ln(likelihood) + 2L_N \tag{27}$$

- Bayesian Information Criterion (BIC)

$$BIC = -2\ln(likelihood) + [\ln(n)\,]L_N \tag{28}$$

- Kernal density estimation

KDE is a non-parametric method of estimating the probability density that may be used to both analyze data and draw conclusions about a sample or larger population. [80].

Where $MMP_{i,EXP}$ *and* $MMP_{i,PRED}$ represent experiment and predicted values of MMP, respectively, $\widehat{MMP}$ represents the average of MMP, while V signifies the total number of points in data. Likelihood is a measure of how well the model fits the data in both AIC and BIC, with L_N indicates the number of free parameters in the model, which often includes coefficients, intercepts, and other model-specific factors, and n is the number of samples in the dataset.

2.4. Data normalization

As stated in literature [44,81,82], appropriate normalization of input database before training process may minimize error rates and training durian. As a result, database normalization is a necessary stage in preparing data. In this investigation, an absolute scale is employed. The following is the normalized formula:

$$X_i^{norm} = \left[\frac{X_i^{original} - X_{min}}{X_{max} - X_{min}}\right] \tag{29}$$

where X_i^{norm} represents normalized input values, X_{max} and X_{min} describe the maximum and minimum for input database, $X_i^{original}$ indicates to the original input database. The value of normalized input data is ranging from 0 to 1. The goal of normalization is to reduce the divergence of the data, which leads to reduce error estimation.

3. Database processing

According to previous studies, the MMP of CO_2 injection is controlled by some major parameters. The most important effect parameters that have been discussed in the literature include the molecular weight of the C_5+ component, the light oil components (X_{VOL} are composed of C_1 and N_2), and the intermediate oil components (X_{INT} are composed of $C_{2\sim4}$, CO_2, and H_2S). Whereas this study is an extension of past research in this field. In this work, three various sorts of data were employed to execute the pure CO_2 MMP prediction utilizing certain machine-learning techniques. Two kinds of data with experimental MMP values have been gathered from the literature, containing around 147 and 197 points, respectively [18,22,79,83]. The third category has been collected from different Iraqi fields to detect the influence of other parameters on forecasting MMP. The first type of data includes three independent variables (molecular weight of the C_5+ component, the ratio of light oil components (X_{VOL} are comprising of C_1 and N_2) to intermediate oil components (X_{INT} are comprising of $C_{2\sim4}$, CO_2, and H_2S) as input data, as documented in Table.1 the distribution of the data, which contains around 147 points. The second kind of data includes the composition of gas injection, light oil components, intermediate oil components, and the molecular weight of C_7+. The reason behind using the second category of data is to reveal the impact of the composition of the injected gas on predicting MMP. The third sort of data, with a total of 28 points and containing other new independent parameters (input parameters) such as API, specific gravity, and molecular weight of C_{6+}, and bubble point pressure (Pb), was gathered from reports of certain Iraqi fields and some theses by Hameed A. and Jani G. [84,85], that were published in the libraries of the University of Baghdad and the University of Technology, Iraq. Typically, 26 points have MMP values that were computed by Hameed A. and Jani G. [84,85] using computational software (Eclipse- PVTi) because these data don't have experiment MMP values. However, two points of the report data don't have MMP values, which these were determined by creating a PVT model for each report by choosing the equation of state ("EOS-Peng-Robinson 1978") and using computational software from Computer Modelling Group Ltd. (CMG) (WINPROP) [86]. The significance of including this type of data and taking plenty of variables into account is to figure out the influence of these factors on the prediction of MMP, as well as to compare the performance estimation of MMP from the computational model with machine learning techniques. As shown in Figs. 3 and 4, the relationship coefficient evaluation between the independent parameters and the output parameter (MMP). Obviously, Fig. 5, Fig. S1, and Fig. S2 display the more correlated parameter with output (MMP) for each database.

As a matter of fact, the objective of using multiple kinds of data is to investigate a broad spectrum of effect features on prediction MMP and compare the accuracy of ML with diverse methodologies such as experiments, empirical correlations, and computational modeling. After data gathering, five reliable machine learning algorithms were applied to the MMP estimation of pure CO_2. Finally, this work has been implemented by Python Language Programming V3.11.1 by using the Spyder Platform. Additionally, a sensitivity study has been performed to identify the factors that influence MMP expectations.

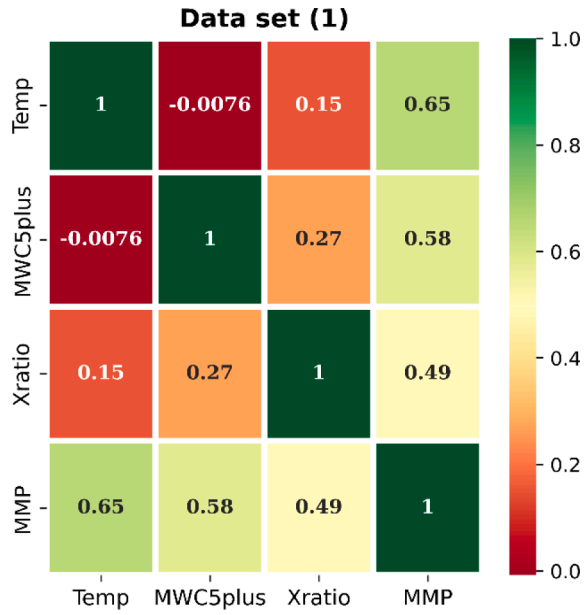

Fig. 3. Heat map implying the correlation between input and output variables for dataset type (1).

4. Results and discussion

In this study, five machine-learning techniques were employed with several types of data to reveal the performance of ML techniques compared with other approaches and to investigate the mechanism impact of important parameters on MMP estimation. Moreover, the assessment of these models was carried out by comparing them with various results of correlation and either computational model, and the major goal of these models is to have the lowest mean square error (MSE), AIC, BIC and mean absolute error (MAE) as well as the highest (R^2). The input parameters of each data type have different independent parameters (functional parameters), as listed in Table 3. The MMP was projected as a function of all input factors of the data. Consequently, numerous runs with different hyperparameters were tested in order to achieve optimal results for SVR, KNN, DT, and RF.

4.1. Models development

In this work, there are two kinds of MMP in the data values that have been employed: the first is an experiment value, and the second is computed by a computational model, as illustrated in the following:

4.1.1. ML and PVT model development

Typically, several kinds of Equation of state (EOS) proposed by some authors that used to make a PVT model [26]. In this work, Equation of state (PR-EOS- "Peng-Robsonin"[75]) was employed to compute MMP for two points by using computational program (CMG-WINPRONP[86]) for data type (3), which the aim to perform this section is to compare the effectiveness of ML with computational methods, as demonstrated in Fig. 6. After importing the required data from the differential liberation (DL) laboratory and other fluid physical properties to software, numerous trials have been carried out in order to discover an appropriate fit with the observed vales of the DL by modifying and tuning some important main parameters of PR-EOS, such as P_c,V_c, Tc, acentric factor, Volume shift, and molecular weight as well as changing the weight percent of some parameters to fulfill optimal matching. Figs. 7 and 8 shows the difference of calculated results before and after regression processes for PVT experiments data. Fig. 9 demonstrates how the MMP is calculated through the utilization of the multiple mixing cell technique. This method indicates the point at which CO_2 becomes miscible with oil, causing the calculation process to halt, and providing the value of the MMP at that stage. In reality, the process of adjusting variables during regression in EOS takes a long time to yield satisfactory results. In comparison to computational approaches, ML needs a short time (around 15 s) to anticipate MMP.

4.1.2. ML development processes

4.1.2.1. Data normalization and splitting. Data normalization and splitting are crucial processes before carrying out ML calculations. In this study, before carrying out the model run, data normalization has been utilized to eliminate the divergence values within the data and make all the values converge on each other's in order to reduce the error estimation and avoid overfitting during model training. Consequently, splitting the data is necessary before running the model to confirm its accuracy, and the data was divided into 20% for testing and 80% for training, as implied in Fig. 6.

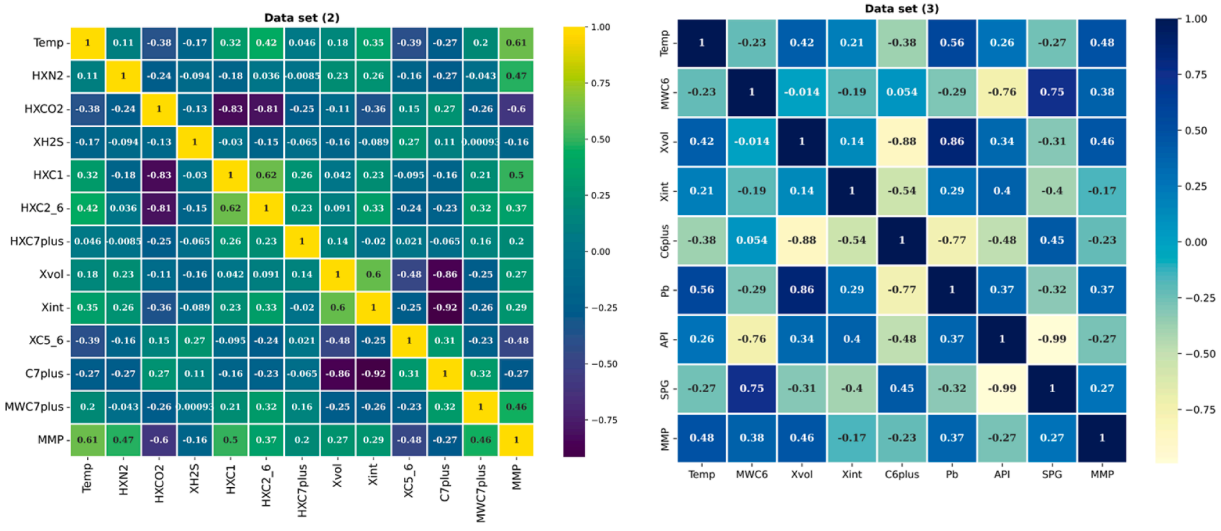

Fig. 4. Heat map implying the correlation between input and output variables for datasets (2) and (3).

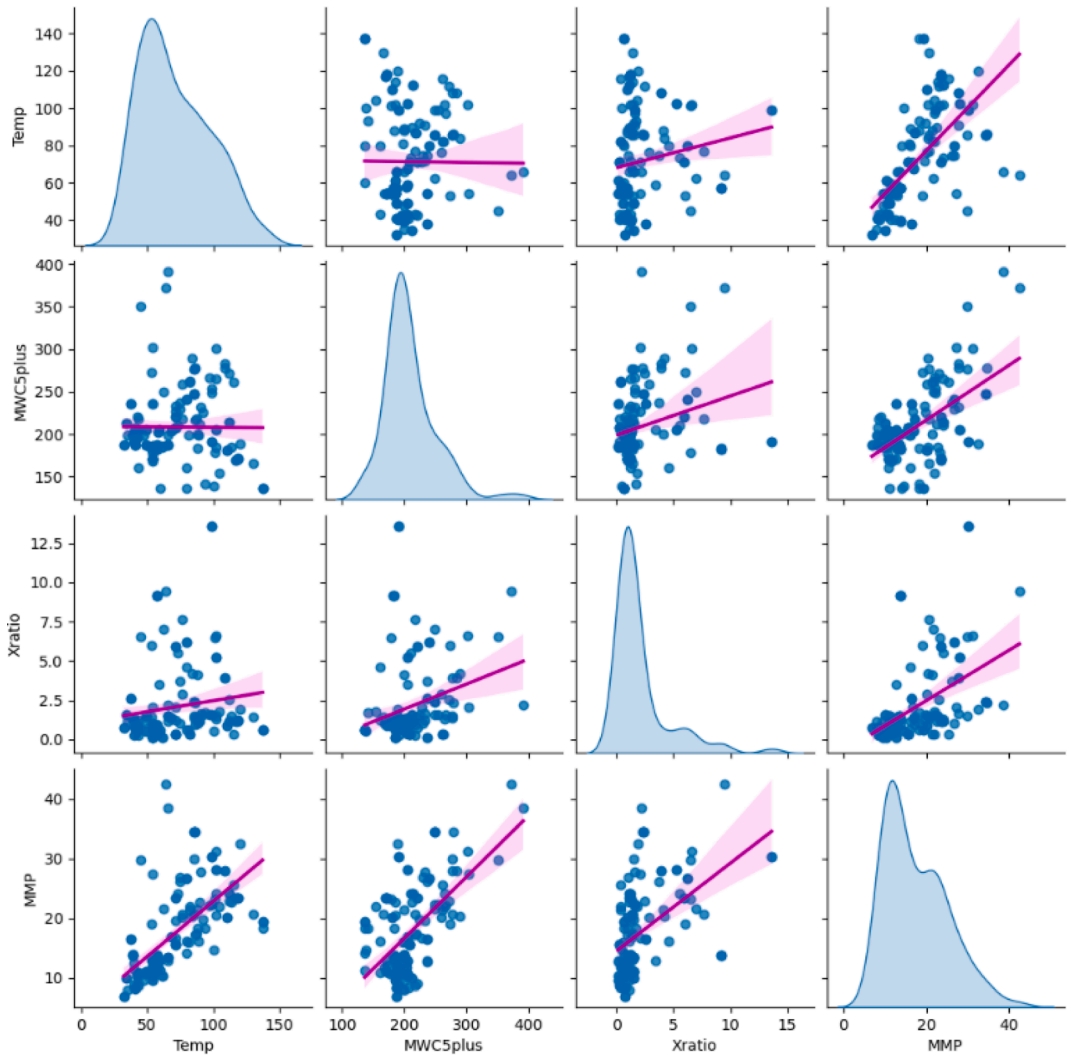

Fig. 5. A pair plot showing the clear regression correlation between input and output variables for dataset (1).

Table 3
Independent parameters for each data type for predicting MMP.

Data type	Independent variables (input parameters)
TYPE 1	*T, MWC_5+, X_{VOL}/X_{INT}*
TYPE 2	Injected gas composition, T, MWC_7+, C_7+, X_{VOL}, X_{INT}
TYPE 3	T, X_{VOL}, X_{INT}, MWC_6+, Sp.gr, API, P_b

4.1.3. Hyperparameters setting

During carrying out the run of ML models, several trials have been executed to acquire the optimal choice of hyperparameters. As highlighted in Table 4, the hyperparameter settings change for each data set, implying that they are not identical for each data set. As is clearly noted, the accuracy of SVR technique outcomes is largely dependent on the appropriate selection of hyperparameters such as C, gamma (γ), and epsilon (ε). Significantly, the most popular sort of kernel function in SVR that produced superior results was the radial basis function (RBF).

4.2. Models comparison

4.2.1. Case 1: Comparison of ML models with (Experimental and empirical correlations)

In this part, data types (1) and (2) have been taken to contrast the performance approaches for the prediction MMP of pure CO_2. In order to compare the results of ML with other approaches, two existed correlations [18,31] were employed to estimate the MMP for pure CO_2 injection. The visual graphs, as shown in Fig. 10, exhibit the assessment of the anticipated MMP findings for each ML approach and two existing correlations for each data types. Based on these figures, the proposed models can be evaluated visually by observing the scatter points that are closest to a 45^0 (X = Y) line. Furthermore, the closer scatter points for any approach to line 45^0 indicate the robustness of the MMP predictive model. Visibly, Fig. 10 a and b depict the cross plots of training and testing between projected ML models and experiments for data type (1), demonstrating that the DT and SVR techniques are the most two effective methods for obtaining closer points and meeting satisfactory outcomes. At the same data type, Fig. 10 c displays the cross plot for the whole data points between two empirical correlations and experiments, where the majority of the points are not aggregated around line 45^0. For data type (2) with different input parameters, Fig. 10 d and e imply cross plots of training and testing between anticipated ML models and experiments, and the results show that RF and KNN were the top two approaches among ML methods that generated closer points to line 45^0. Simultaneously, the outcomes of prediction MMP by literature correlations demonstrate that a large number of the points are located distant from the diagonal line, as implied in Fig. 10 f Based on the graphical plots aforementioned, it is possible to infer that all ML techniques perform proficiently with low error accuracy for estimating MMP when compared to various correlations.

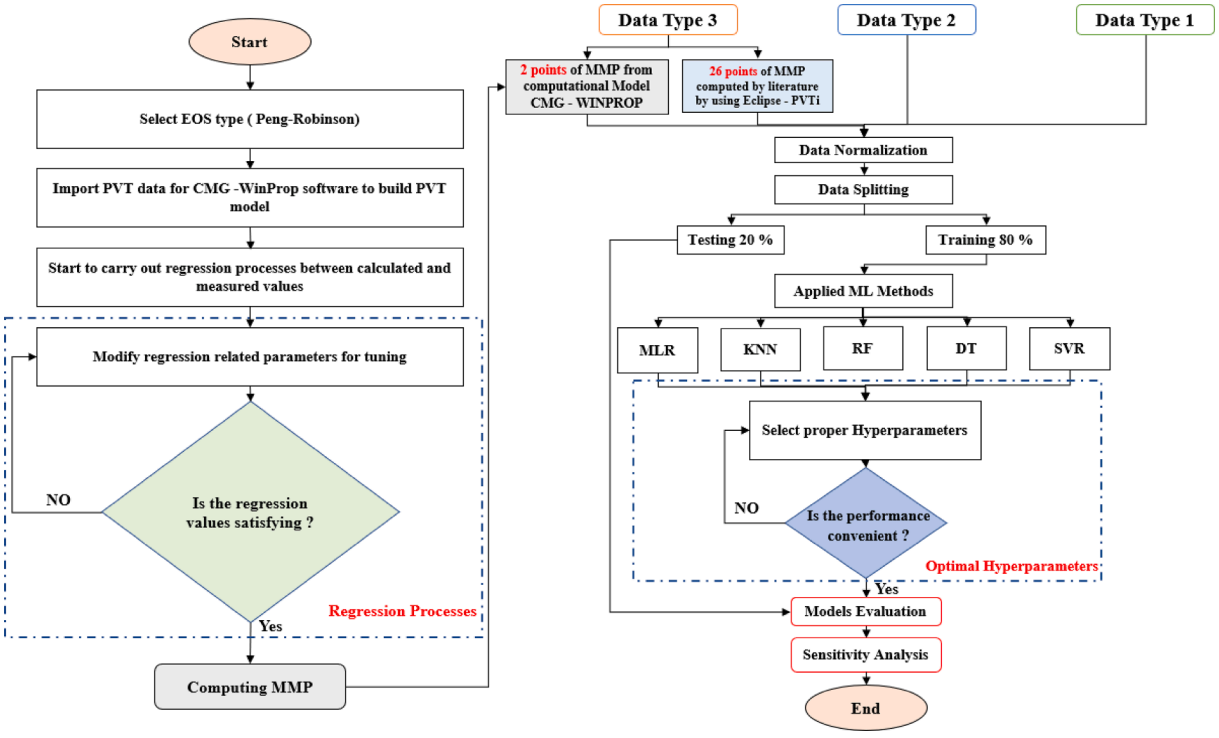

Fig. 6. Shows the flowchart of ML and computational model processes.

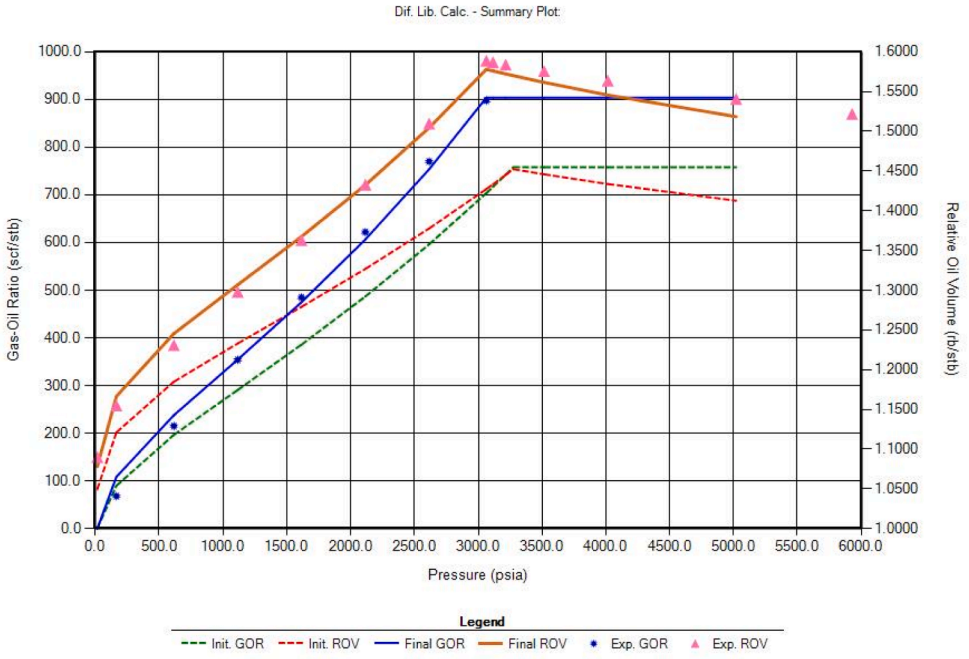

Fig. 7. Shows the tuning regression between observed and calculated values.

4.2.2. Case 2: Comparison of ML models with (Computational and empirical correlations)

This part was included in this research to compare the efficacy of ML methods for MMP prediction with a computational methodology. Furthermore, data type (3) lacks experiment values for MMP because the experiment test is costly; consequently, it is preferable to execute a computational model because it might represent a real fluids condition rather than an experiment. Even though the computational processes might require a long time to obtain the matching, it is necessary to examine the effectiveness of ML for forecasting MMP and compare it with the computational method because ML processes require a short time, thus it can be argued that it is not expensive. Fig. 11 highlights a comparison between predicted ML models and computational model. As implied in Fig. 11 a and b, the two highest-ranking ML algorithms to anticipate MMP for testing and training data that has accumulating points on the diagonal line are KNN and SVR. Nonetheless, Fig. 11 c depicts forecasting MMP using correlations against a computational model, where the vast majority of the points are far from the diagonal

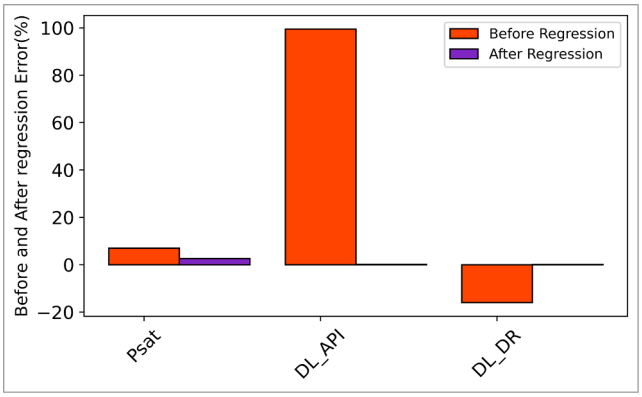

Fig. 8. Implies the error reduction after performing regression processes.

line, indicating that correlations may have poor accuracy in predicting CO_2-oil MMP in certain circumstances. Regarding these findings, it is reasonable to state that ML techniques are particularly appropriate for estimating MMP at low cost and in a short period of time.

4.3. Performance of ML models

Following the completion of the training operations, the anticipated regression models were created. The well-trained models will be evaluated with the testing data set (20% of the data) that was not included during the training procedures in order to validate the model's potential generalization and reliability. For further clear assessment, the histogram of error distribution was used to explore the range of precision of testing ML- models to determine MMP of pure CO_2. Based on that, Fig. 12 presents the error distributions of data type (1), which indicate that DT, RF and SVR have the best distributions because the majority of their values are closer to zero and their lowest error margin range is around (-0.2–0.3). For data type (2), Fig. 13 represents the accepted error of ML algorithms for three methods, including SVR, RF, and KNN. The best ML technique was SVR because virtually all of its points are centered around zero with an error margin of (-0.2–0.25), whereas RF and KNN have error margins of approximately (-0.4–0.25). As can be observed in Fig. 14, KNN and SVR are the most efficient ML approaches for data type (3) that fulfill the error distribution conditions that were mentioned before.

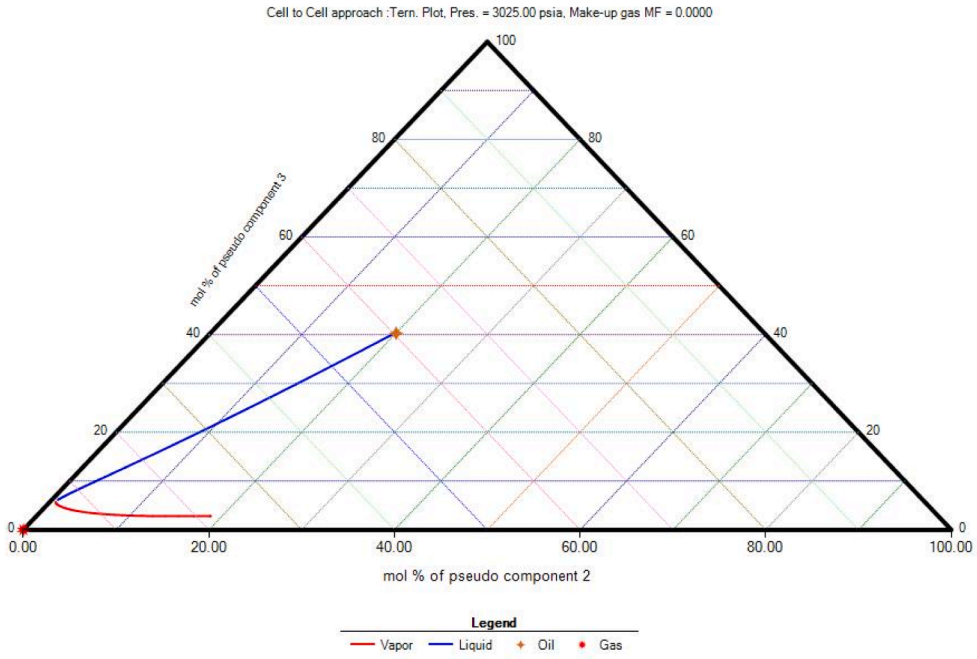

Fig. 9. MMP calculation via using multiple mixing cell (MMC) by CMG- WinProp software.

Table 4
Optimal Hyperparameters for some ML methods of each datasets.

Data Type	Method	Optimal setting Hyperparameters			
DATA TYPE 1	RF	N - estimators	Random state	N -jobs	
		1000	20	−1	
	DT	Max depth	Random state		
		150	80		
	SVR	Kernel Function	Gamma (ϒ)	Epsilon (ε)	C
		RBF	0.1	0.21	250
DATA TYPE 2	DT	Max depth	Random state		
		130	80		
	SVR	Kernel Function	Gamma (ϒ)	Epsilon (ε)	C
		RBF	0.00001	0.1	212.14
DATA TYPE 3	DT	Max depth	Random state		
		90	20		
	SVR	Kernel Function	Gamma (ϒ)	Epsilon (ε)	C
		RBF	0.1	0.09	780

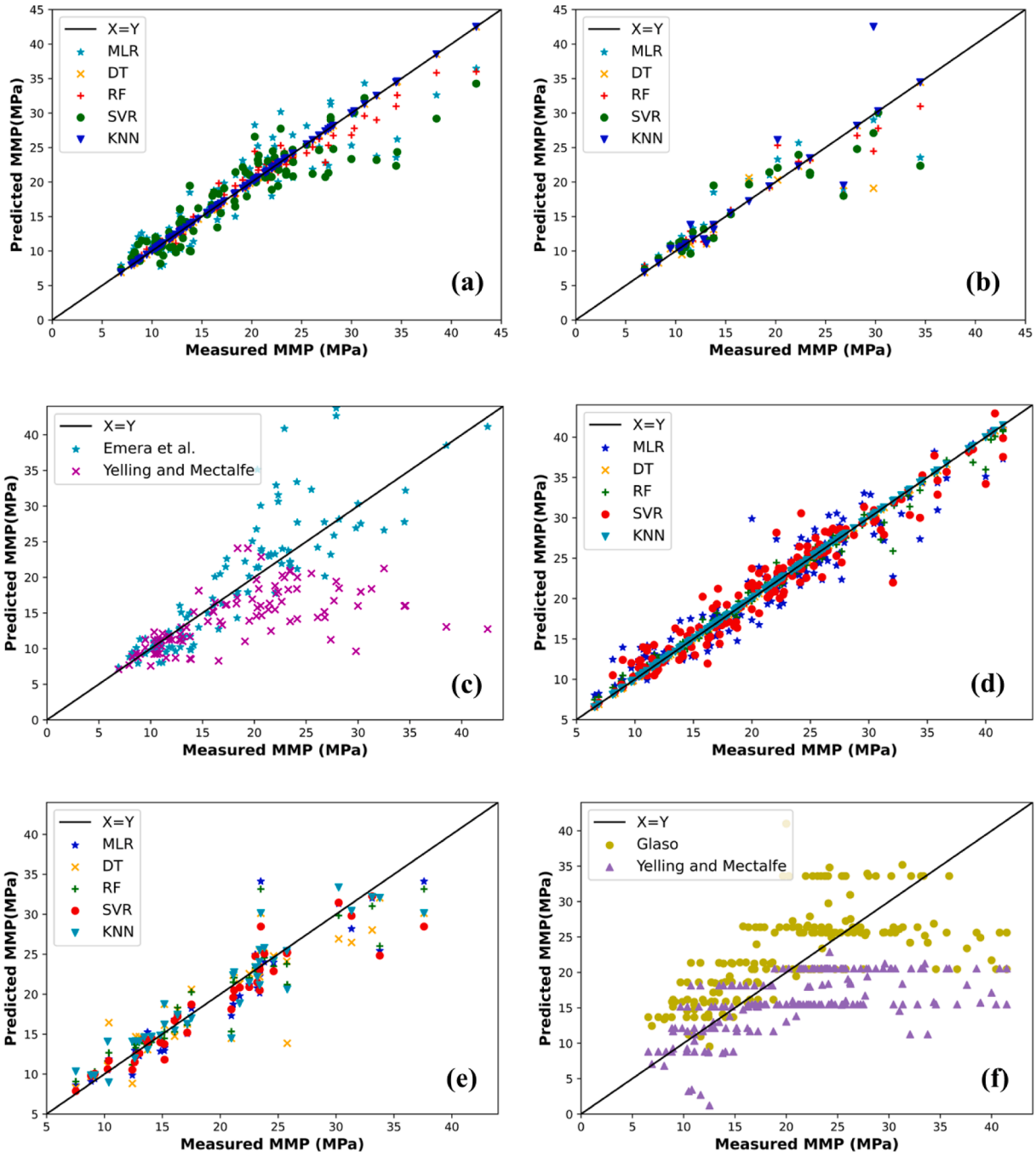

Fig. 10. Shows the results of prediction MMP for (a) ML models vs. measured as function of data type (1) for training, (b) ML models vs. measured as function of data type (1) for testing, (c) Empirical correlation vs. measured as function of data type (1), (d) ML models vs. measured as function of data type (2) for training, (e) ML models vs. measured as function of data type (2) for testing, (f) Empirical correlation vs. measured as function of data type (2).

4.4. Statistical evaluation

Numerous ML algorithms were used in this research to anticipate the MMP for pure CO_2 injection based on experimentation data in order to examine the reliability of ML approaches for MMP prediction with different conditions. Additionally, the existing correlations were used to validate or compare the competence of ML models with other techniques. Moreover, to further assess and compare the efficacy of various MMP forecasting techniques for three groups of data, a number of statistical evaluation variables have been utilized. According to the results in Table 5, the findings of the average statistical assessment parameters indicate that the best two approaches to ML with an ideal value of regression evaluation for data type (1) are DT and SVR, which have the highest coefficient of determination (R^2) of 0.95 and 0.94 respectively, and the lowest MSE of 3.12 and 3.53 respectively. For data type (1), the following order shows that MLR provides the lowest accuracy values among ML methods: DT > SVR > RF > KNN > MLR. Based on functional group for data type (2), the top two techniques of ML that provide the best precision are KNN and RF with highest coefficient of determination (R^2) of 0.93 and 0.92 respectively, and the lowest MSE of 3.36 and 3.91 respectively as shown in the following arrangement: KNN > RF > SVR > DT > MLR. The KNN and SVR have the best statistical characteristics for data type (3), as demonstrated in the following sequence: KNN > SVR > MLR > RF > DT. The KNN and SVR also have the greatest coefficient of

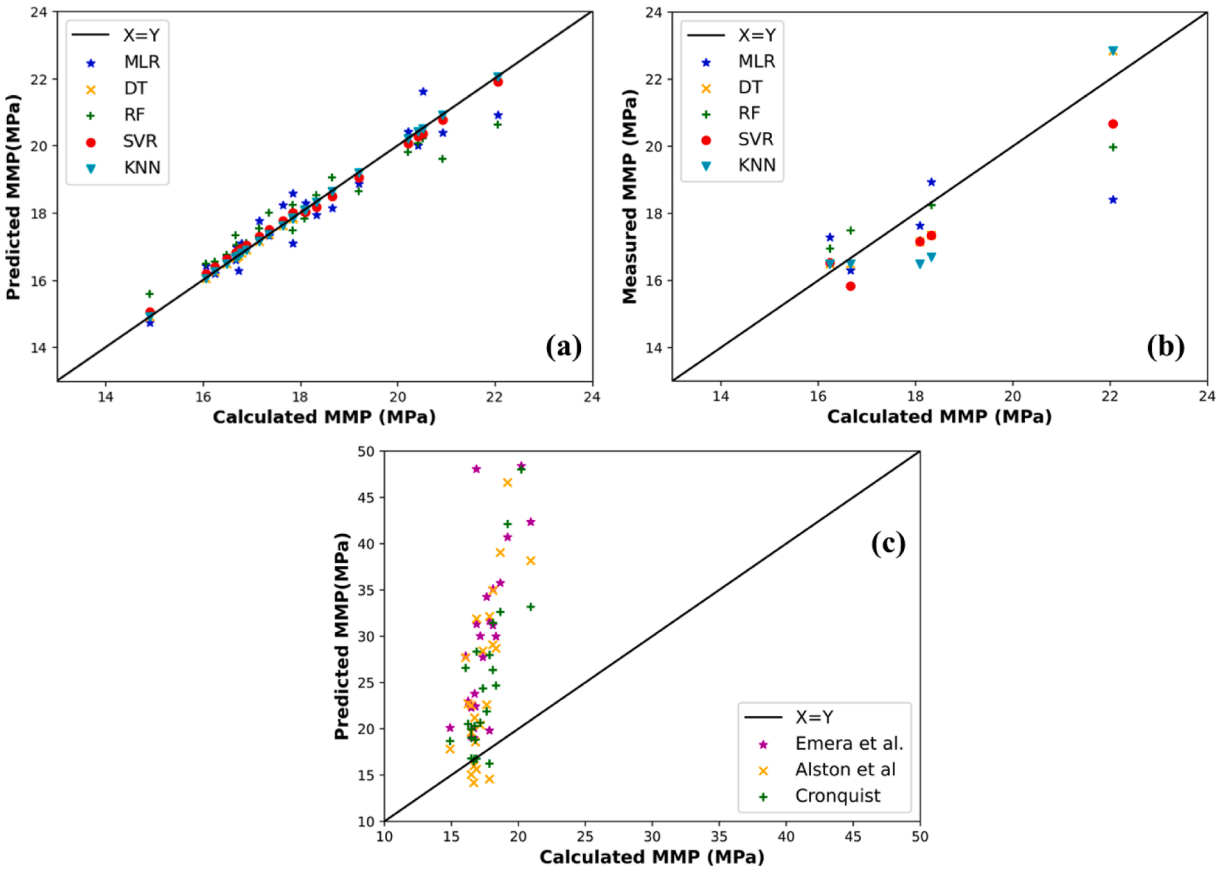

Fig. 11. Shows the results of prediction MMP for (a) ML models vs. computational model as function of data type (3) for training, (b) ML models vs. computational model as function of data type (3) for testing (c) Empirical correlation vs. computational model as function of data type (3).

determination (R^2) of 0.99 and 0.95, respectively, and the lowest MSE of 0.09 and 0.62, respectively. Nevertheless, for data type (3), the accuracy assessment for training data was greatest and the testing was lowest, leading to a low total accuracy evaluation, suggesting that the values are not enough to generate the best regression, despite DT having the highest degree of accuracy appraisal based on the functional groupings for data type (1). It has been observed that SVR was creating an acceptable level of precision for data type (3) on account of its effectiveness for small data, as is mentioned in the literature [87].

Depending on the overall visualization for absolute percent relative error (APRE), as shown in Fig. 15, the evaluation outcomes showed the effectiveness of three advanced ML methods: KNN, SVR, and DT for prediction MMP. As illustrated in Fig. 16, the kernel density estimation of all used models, of which most produce satisfactory outcomes when compared to the real test data. Practically, SVR and KNN achieve a better match with the KDE of the real data set for all three types of databases. For further evaluation, AIC and BIC have been employed in this study to investigate the compatibility of the models, with lower values implying better model fit and lesser complexity. Furthermore, Fig. 17 a and b demonstrate that KNN has the lowest AIC and BIC for all datasets, indicating KNN has an appropriate fit and is less complex compared with other models.

On the other hand, the findings of statistical assessment for MMP estimation using empirical correlations demonstrated the poor efficiency of some correlations to compute MMP due to their impact on specific variables and their incapacity to compute under a wide range condition, as noted in Table 6. Based on that, it can be observed that the parameter assessment for Yelling and Mectalfe's [18] correlation provided adequate MAE, MSE, MED, and R^2 values for data type (1). However, the appraisal variables for data type (2) demonstrated the inefficiency of Yelling and Mectalfe's [18] correlation to estimate MMP for a broad range of circumstances. Therefore, it might be concluded that the most significant correlations can only be applied to specific instances and not to various situations. In general, all graphical analyses and statistical evaluations confirmed the efficiency of some ML methods without limitations in comparison to other methods. Thus, it can be argued that ML approaches are appropriate for anticipating MMP with acceptable accuracy and without restrictions.

4.5. Predictability of models

To evaluate the predictability of machine learning (ML) models for MMP prediction across a wide pressure range, dataset type (2) was selected. This dataset consists of a broad pressure range, which was further divided into three pressure intervals: (6–15) MPa, (15–25) MPa, and (25–41) MPa. The root mean square error (RMSE) was employed as a metric for comparison. According to the results in Fig. 18, the ML models showed the lowest average RMSE within the pressure range of 6–15 MPa. This finding suggests that as MMP increases, the accuracy of ML models tends to slightly decline. On the other hand, as remarked in certain literature, it has been observed that there may be variations in findings when utilizing temperatures in Celsius and Fahrenheit measurements. However, it is crucial to emphasize that this possible variance was carefully explored in the research. Consequently, the results have confirmed that there is no significant difference between the obtained outcomes.

4.6. Sensitivity analysis

4.6.1. Relevancy factor

To analyze the effect of each input parameter on the projected MMP value, a variable impact study was performed using the relevance factor

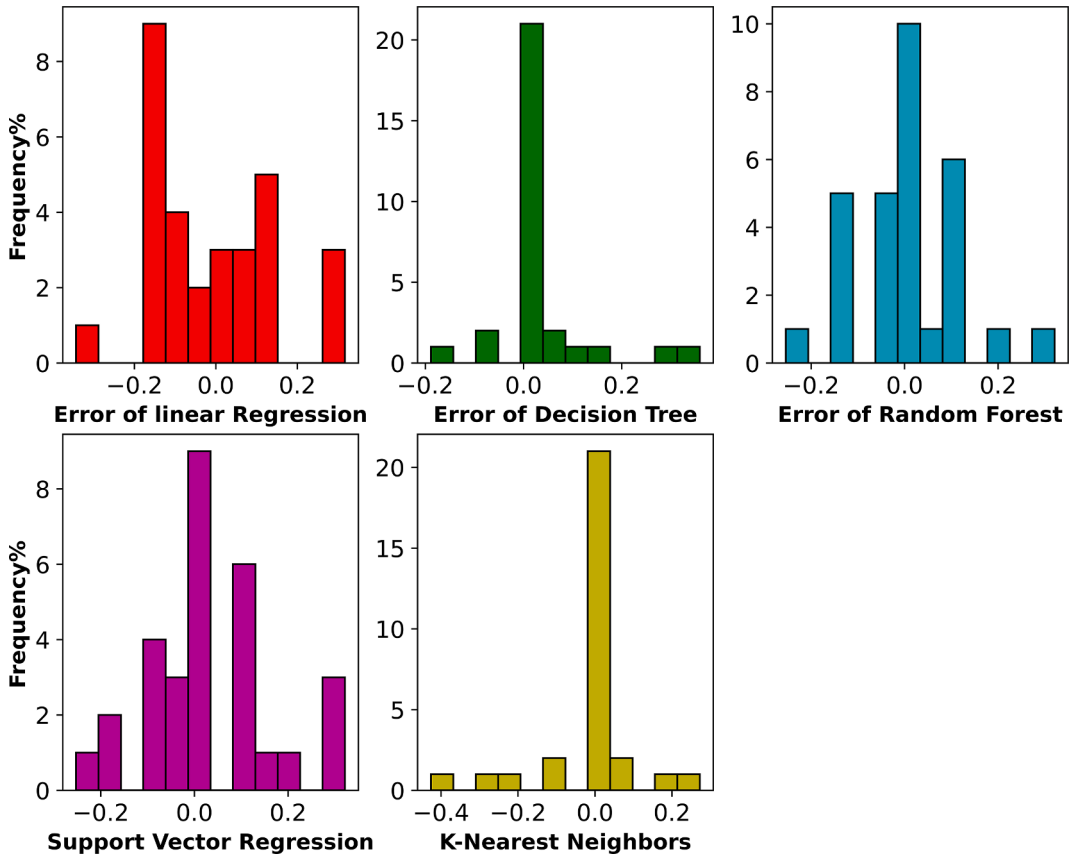

Fig. 12. Error distribution of the ML approaches as function of data type (1) for testing.

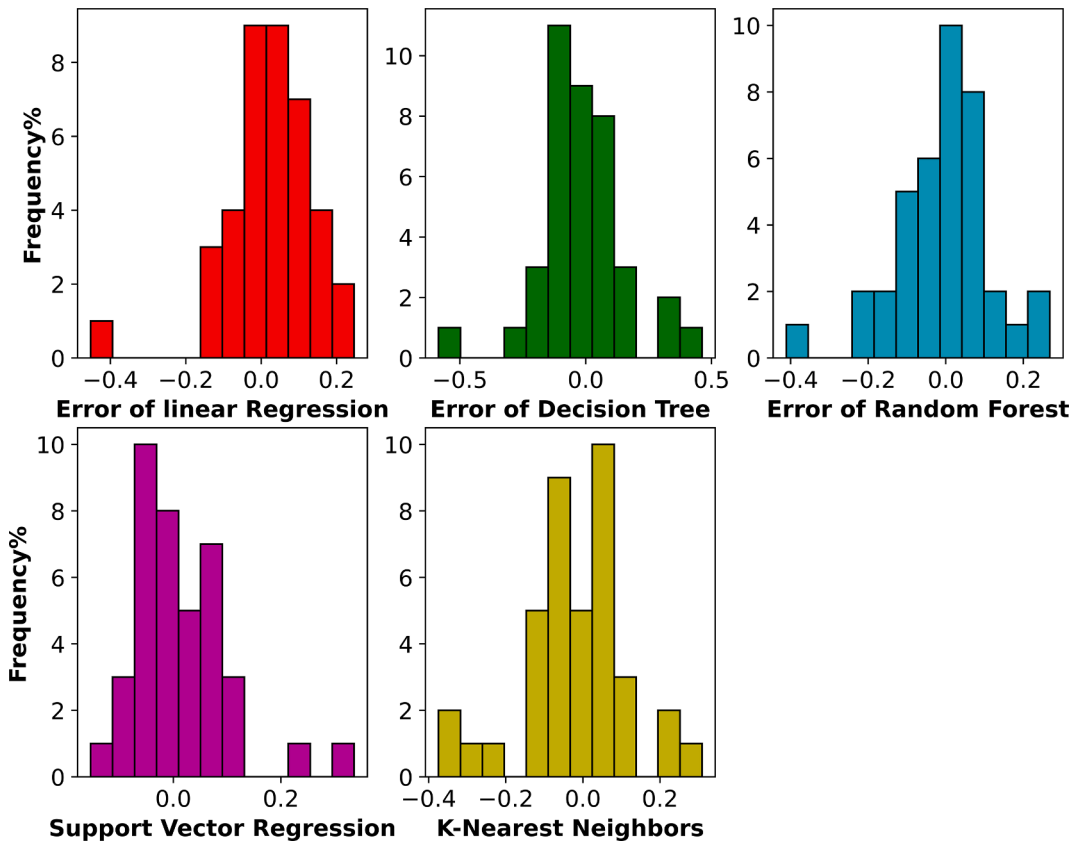

Fig. 13. Error distribution the ML approaches as function of data type (2) for testing.

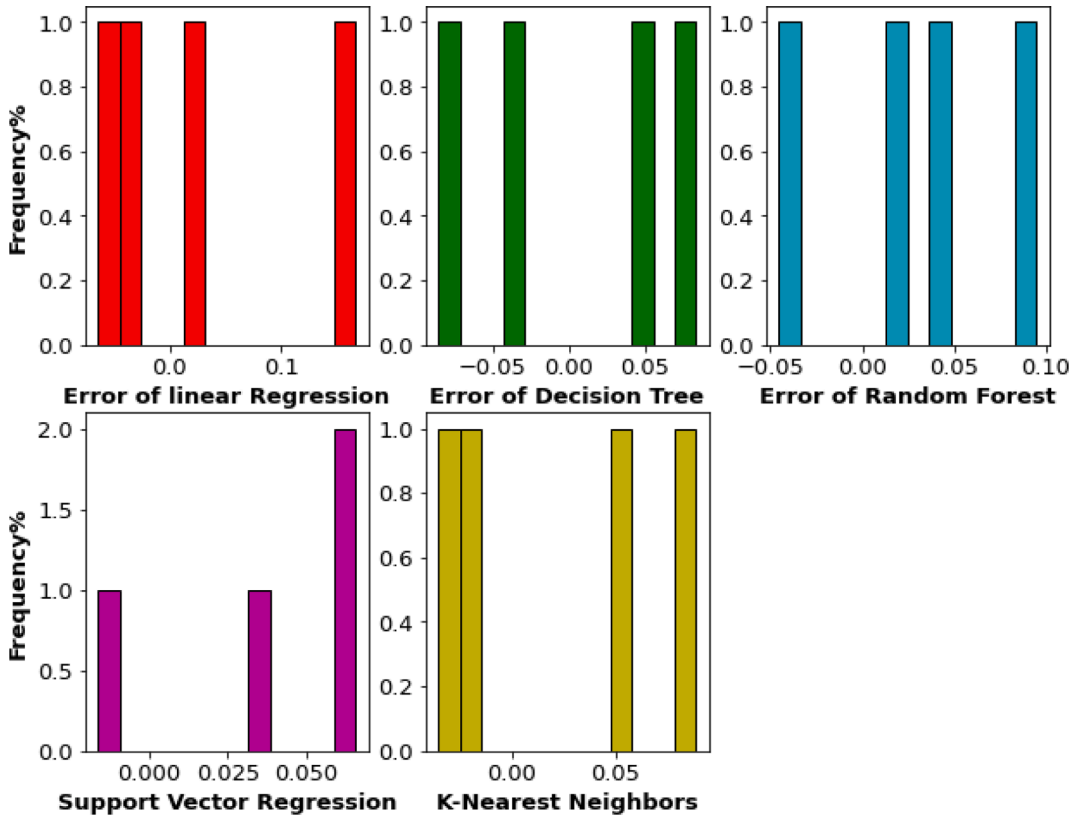

Fig. 14. Error distribution the ML approaches as function of data type (3) for testing.

Table 5
Statistical assessment of developed ML models for predicting MMP with various data.

DATASET	METHOD	Average Statistical Parameters between (Training set and Testing set)			
		MAE	**MSE**	**MED**	**R^2**
TYPE 1	MLR	2.35	11.20	1.69	0.81
	DT	0.47	3.12	0.005	0.95
	SVR	0.72	3.53	0.21	0.94
	RF	1.15	3.88	0.602	0.93
	KNN	0.55	4.71	0.005	0.92
TYPE 2	MLR	1.80	6.58	1.25	0.88
	DT	1.15	5.82	0.65	0.89
	SVR	1.55	5.36	1.10	0.90
	RF	1.11	3.91	0.69	0.92
	KNN	0.93	3.36	0.55	0.93
TYPE 3	MLR	0.57	0.53	0.44	0.78
	DT	0.59	0.92	0.76	0.64
	SVR	0.29	0.12	0.28	0.95
	RF	0.62	0.56	0.67	0.76
	KNN	0.09	0.02	0.11	0.99

(RF). The following equation [15,88] computes the relevance factor for every parameter:

$$RF(X_J, Y) = \frac{\sum_{i=1}^{n}(X_{J,i} - \overline{X}_J)(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_{J,i} - \overline{X}_J)^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \quad (30)$$

where $X_{J,i}$ and $\overline{X}_J$ signify the i-th and average number of input J, respectively, while Y_i and $\overline{Y}$ indicate the i-th and average number of MMP output, respectively. For any input parameter, this method produces a value ranging from −1 to 1. Negative and positive numbers represent the inverse and direct connections between the input and output variables, respectively. The maximum absolute value indicates the greatest importance of an input parameter. A sensitivity study was performed to further detect the relation between the independent factors and the prediction of MMP. Throughout this investigation, the impact of independent parameters on forecasting MMP by using ML techniques has been verified. The findings of sensitivity analysis of dependent variables for each type of data are implied in Figs. 18-20. As demonstrated in Fig. 19, for data types (1), reservoir temperature, molecular weight of C_{5+}, and the ratio of volatile and intermediate components have an obvious correlation with MMP, and among other factors, temperature has a significant influence on model predictions,

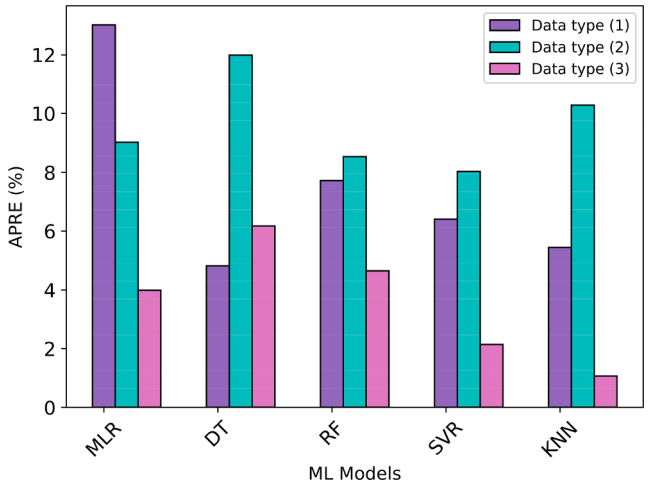

Fig. 15. Total comparison between ML methods for each data type.

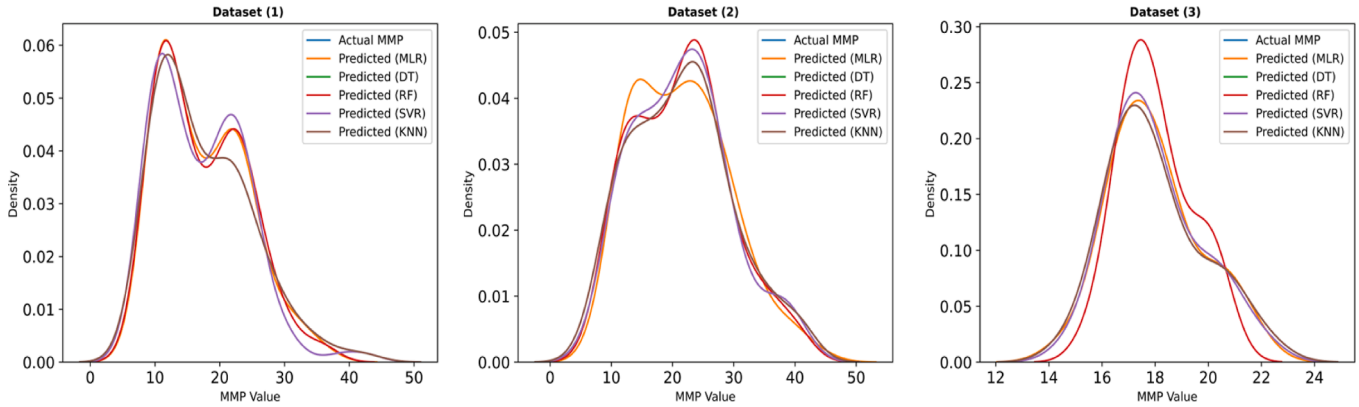

Fig. 16. Comparison of all models' kernel density estimation performance between real data and expected outcomes for each dataset.

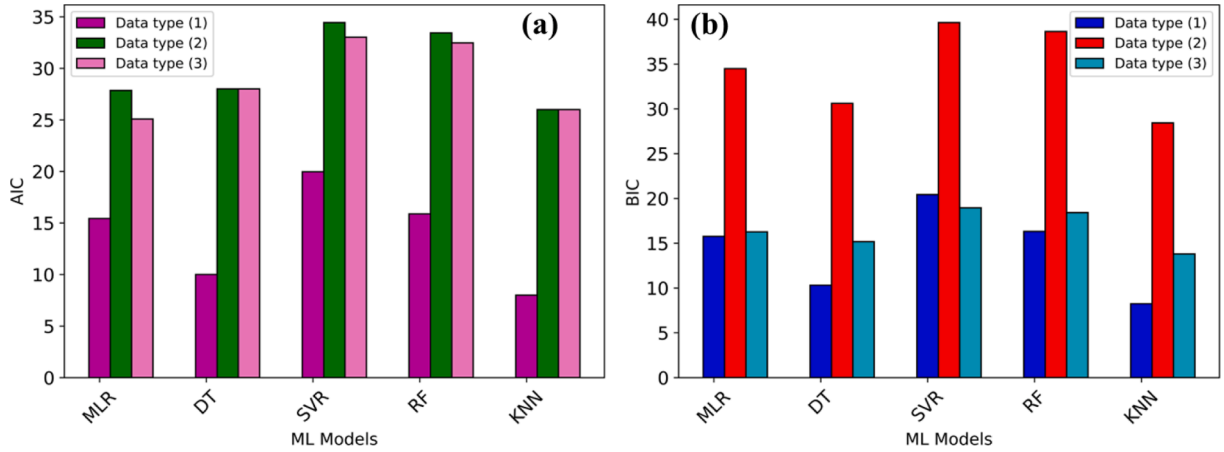

Fig. 17. Illustration of the performance of ML models depending on (a) AIC and (b) BIC.

Table 6
Statistical assessment of empirical correlations for predicting MMP with various datasets.

DATASET	METHOD	MAE	MSE	MED	R^2
TYPE 1	Alston et al.[27]	3.75	43.1	1.61	0.73
	Emera et al.[31]	2.48	17.19	1.28	0.78
	Yelling and Mectalfe [18]	4.32	46.77	2.09	0.91
TYPE 2	Glaso[28]	5.2	45.57	4.32	0.92
	Yelling and Mectalfe	6.19	70.51	4.61	0.65
TYPE 3	Alston et al.	13.09	382.71	8.39	0.67
	Emera et al.	9.58	173.87	7.05	0.76
	Cronquist[24]	10.26	266.77	5.31	0.63

with a relevance value of 0.74.

Consequently, as shown in Fig. 20 for database (2), all compositions (HX_{N2}, HX_{H2S}, HX_{C1}, $HX_{C2\sim C6}$, HX_{C7+}) in injected gas, temperature, MWC_{7+}, volatile components (X_{VOL}) and intermediate components (X_{INT}) in crude oil have a positive relationship with MMP, while $X_{C5\sim C6}$ and X_{C7+} in crude oil have a negative relationship with MMP. Among these parameters, some compositions of injected gas, HX_{C1} and $HX_{C2\sim C6}$ have the greatest effect on MMP, followed by the impact of temperature, and the influence of HX_{N2} and HX_{H2S} is small. For additional clarification, the effects of MWC_{7+}, volatile fraction (X_{VOL}) and intermediate components (X_{INT}) in crude are less significant than the gas injection's chemical composition. In the parameter interval, every impacting factor's level of effect on MMP is listed in descending sequence: HX_{C1} > $HX_{C2\sim C6}$ > temperature > X_{VOL} > MWC_{7+} > HX_{C7+} > X_{INT} > HX_{H2S} > HX_{N2}.

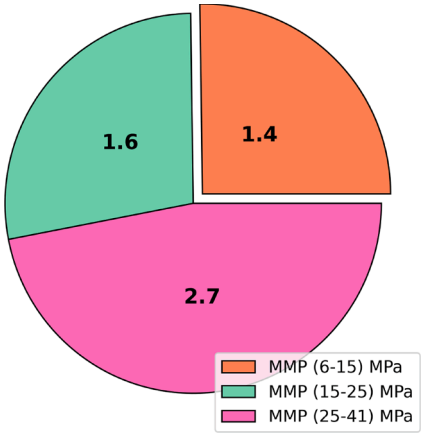

Fig. 18. Performance comparison for the ML models between different ranges of MMP.

Another sensitivity analysis for database type (3) included new functional parameter groups such as sp.gr, P_b, API, viscosity, C_{6+}, and MWC_{6+}, as illustrated in Fig. 21, which demonstrates these factors (temperature, sp.gr, viscosity, X_{VOL}, X_{C6+}, and MWC_{6+}) have a positive effect on prediction of MMP. In contrast, API, P_b, X_{VOL}, and X_{INT} have a negative impact on MMP estimation. The following descending order shows the importance degree for independent parameters: Temperature

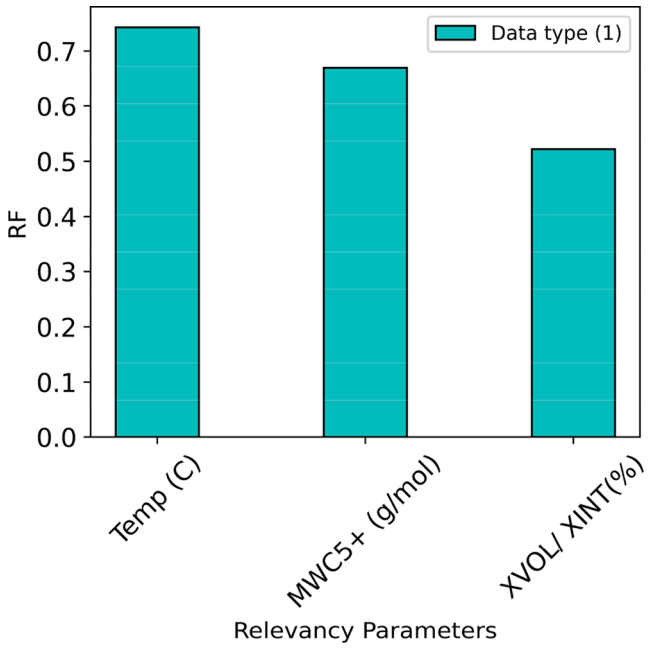

Fig. 19. Sensitivity analysis for the impact of independent variables on MMP prediction for data type (1).

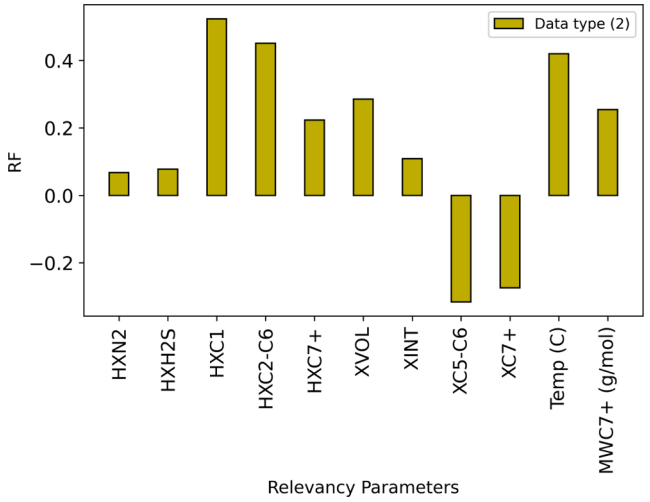

Fig. 20. Sensitivity analysis for the impact of independent variables on MMP prediction for data type (2).

> sp.gr > MWC_{6+} > viscosity > X_{VOL} > X_{C6+} > P_b > X_{INT} > API.

4.6.2. Shapely explanation plot (SHAP)

The Shapley graph is one of the most valuable tools to define or interpret the influence of each attribute parameter on the output of a machine learning model. The plot's y-axis displays the relevance of each feature; the features at the top have the most effect on the output, while those at the bottom have less influence. Each characteristic is represented in the plot by a horizontal bar. The length of the bar represents the amount of the feature's influence on the model's output. Positive Shapley values (red) imply that the feature enhances output, while negative Shapley values (blue) indicate that the feature reduces output. Features of importance are listed on the y-axis of the plot. This can reveal which properties contribute the most impact to the model's predictions. Because the KNN approach involves a parametric algorithm that is unable to apply in a shape plot, the SVR model was selected in this part to analyze the influence of parameters on the model's predicting.

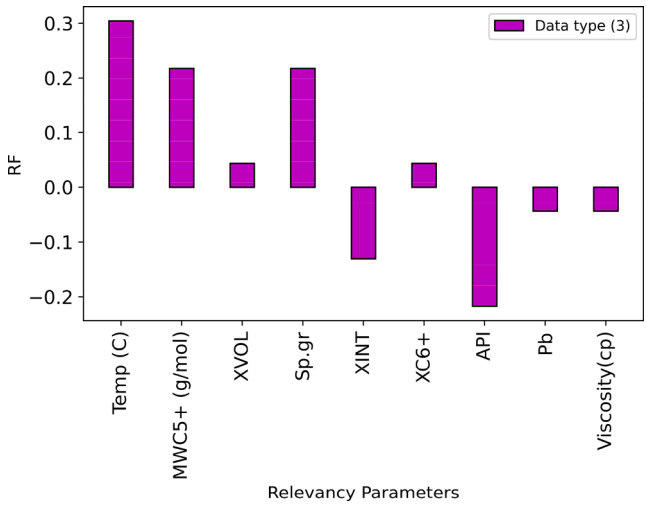

Fig. 21. Sensitivity analysis for the impact of independent variables on MMP prediction for data type (3).

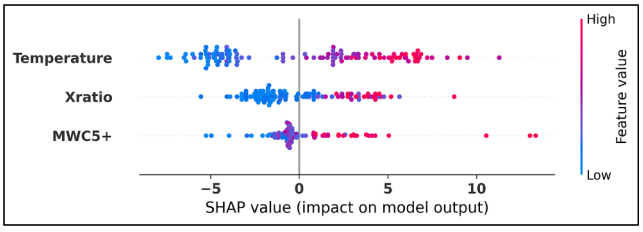

Fig. 22. Shapely plot shows the summary of the input features on output of SVR model for dataset (1).

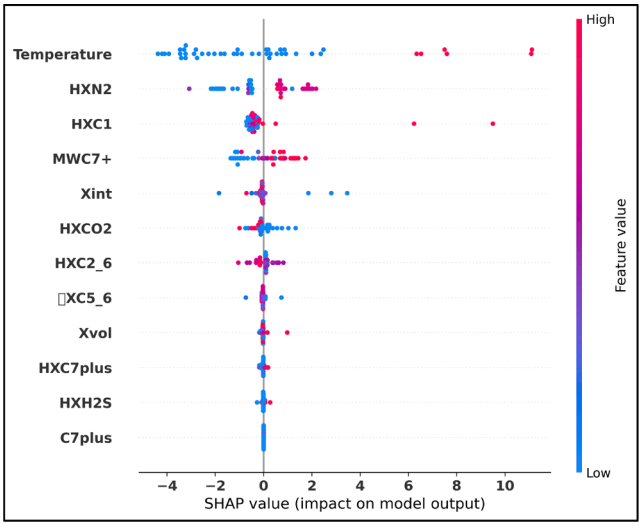

Fig. 23. Shapely plot shows the summary of the input features on output of SVR model for dataset (2).

Based on that, it can be observed in Fig. 22, Fig. 23, and Fig. 24 that the most important parameter that has a direct impact on the model's output for all datasets is temperature.

4.6.3. Physical parameter analysis

In this section, dataset type (3) has been used to detect the behavior of independent parameters during the training of the model to predict MMP because this dataset has new parameters that are included in the

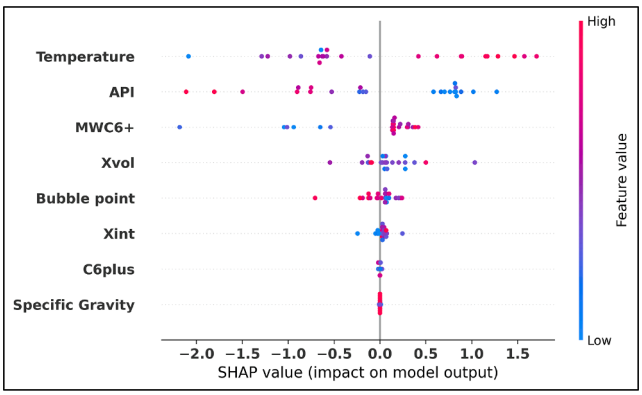

Fig. 24. Shapely plot shows the summary of the input features on output of SVR model for dataset (3).

developed models. According to the pre-evaluation of the models, a KNN model has efficient reliability and good compatibility; hence, it was chosen for this purpose. As can be seen in Fig. 25, the majority of the physical parameters of the created model correspond to the actual data. For deep analysis, during CO_2 injection for increased oil recovery, MMP increases when some parameters (reservoir temperature, molecular weight of hexane plus, volatile percentage and specific gravity) increase. Particularly, crude oil specific gravity has an impact on MMP; denser oils have greater MMPs, whereas less dense oils with lower specific gravity values have lower MMPs. In contrast, raising API, X_{INT}, and C_6 plus lead to a decrease in MMP. Typically, greater API values tend to achieve miscibility easily between oil and gas easily because higher API crude oils have lower viscosities. These tendencies are supported by the experimental trends of data points, which are shown for each figure. In addition, the findings are matched the concept of physical analysis in the literature [12,30,40,44,45,48].

For further analysis, the SHAP dependence plot was also utilized to investigate the physical trends and interactions between each parameter and the MMP using the SVR model based on dataset (3). As depicted in Fig. 26, it is observed that all the physical parameters in the SVR model show a similar trend as those in the KNN method. Moreover, these parameters effectively capture the well-documented physical trends observed in the literature. Fig. 27 clearly implies the mean absolute significant impact of each variable, highlighting temperature as the parameter with the highest effect on predicting MMP.

4.6.4. Screening main impact factors

This procedure involves progressively excluding one of the variables while maintaining the other parameters without modifying them during implementing the run and investigating the effect of this parameter on the accuracy of MMP prediction using ML. The coefficient of determination (R^2) is used for evaluating the effect degree of the important parameters in estimating MMP. The following formulas determine the percentage of impact:

$$Parameter\ \ Impact = R^2\ \ without\ \ removing - R^2\ \ after\ \ removing \tag{31}$$

$$Influencing\ \ Percentage(\%) = \frac{Parameter\ \ Impact}{Total\ \ Parameter\ \ Impact} \times 100\% \tag{32}$$

The essential objective of this part is to detect the amount of influence of each main parameter on the precision of MMP prediction for all data sets. As can be observed in Fig. 28, The composition of the injected gas has the strongest influence among the independent factors on the prediction of MMP, at approximately 46%, followed by reservoir temperature, molecular weight of C_{6+}, molecular weight of C_{5+}, molecular weight of C_{7+}, and the volatile and intermediate components. Simultaneously, specific gravity factor was having a little positive effect on predicting MMP.

5. Conclusion

The determination of the minimum miscible pressure (MMP) is crucial for understanding the complex mechanics involved in CO_2 injection. Therefore, the primary objective of this study is to assess the effectiveness and reliability of machine learning (ML) approaches in predicting MMP for pure CO_2 using a wide range of datasets and various parameters. To accurately address this issue, five ML models have been developed. The main contribution of this work is to investigate the impact of other parameters on the developed ML models and utilize unique evaluation methods for comparison. Based on the comprehensive evaluation, the study's conclusions can be briefly described as follows:

1. The research investigation showed that the DT model produced the optimum paradigm for estimating MMP with the lowest error metrics and highest determination coefficient (MSE = 3.12 and R^2 = 0.95), followed by SVR, RF, and KNN models based on the dataset (1).
2. Based on dataset (2), the KNN model provided efficient accuracy for forecasting MMP relying on the statistical evaluation with MSE = 3.36 and R^2 = 0.93. In addition, the KNN model demonstrated strong

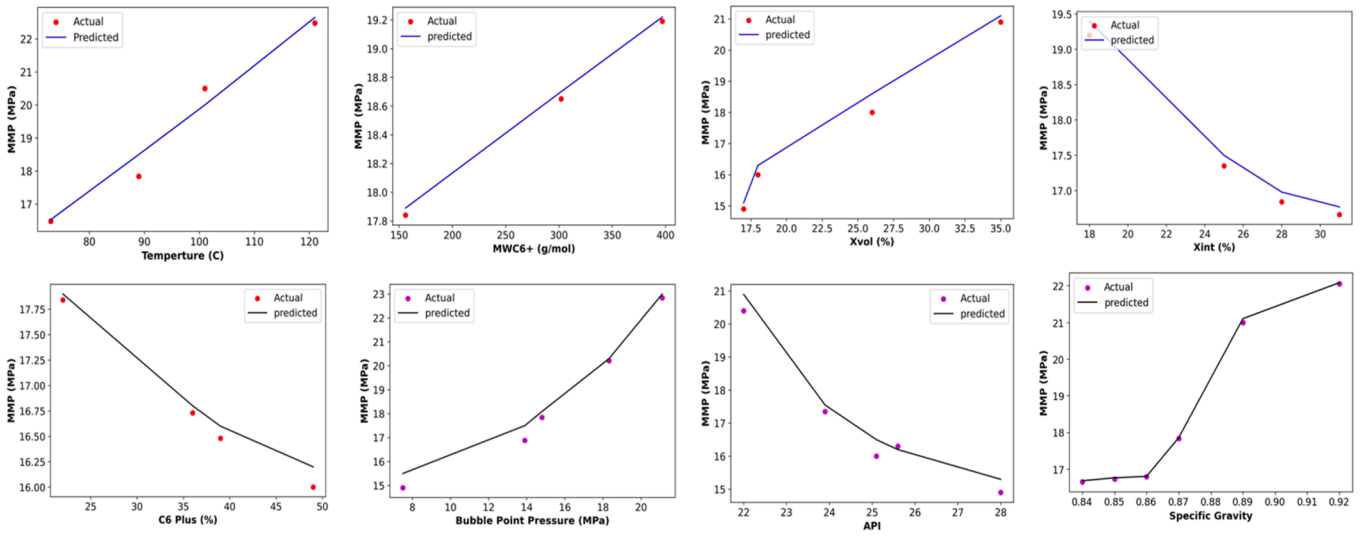

Fig. 25. Physical analysis for all input parameter with actual and predicted MMP by KNN for dataset (3).

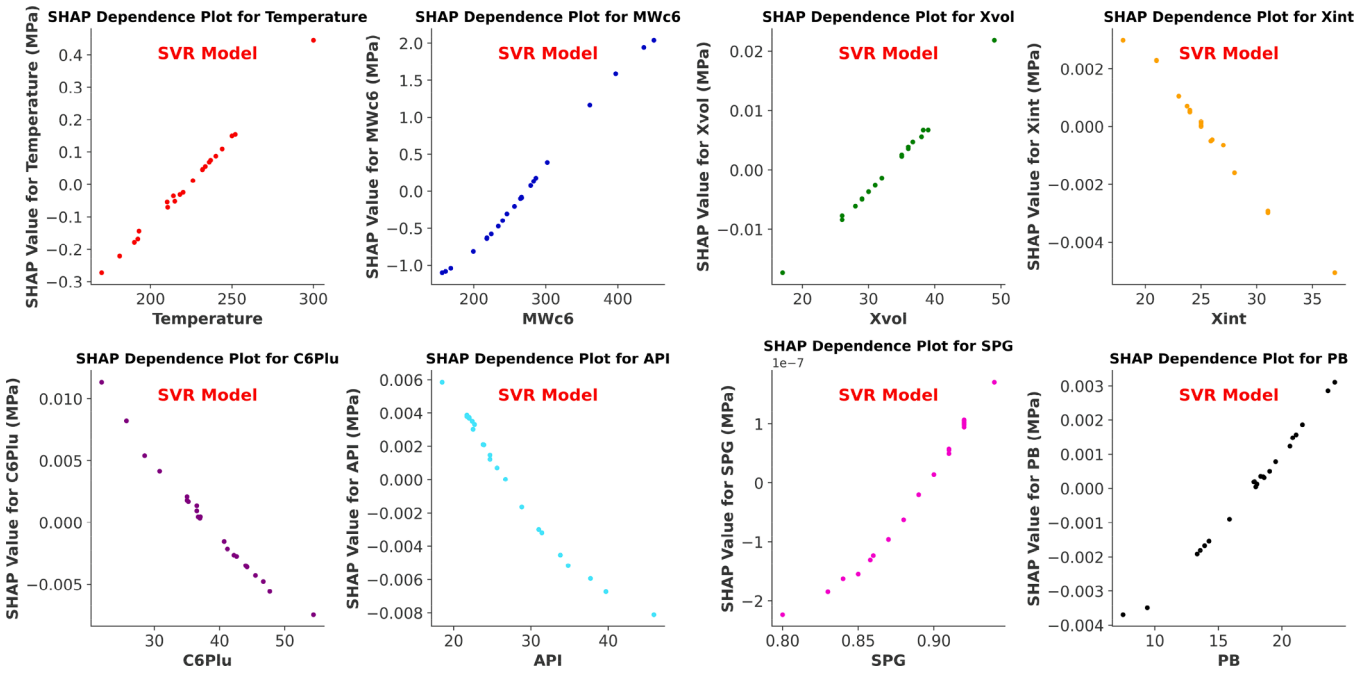

Fig. 26. SHAP dependence plot for each input parameter for the SVR model based on dataset (3).

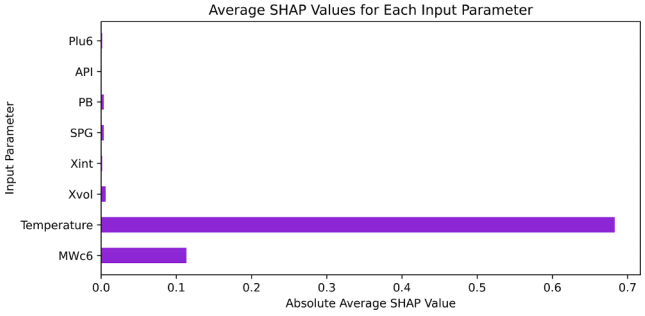

Fig. 27. The mean absolute SHAP values for dataset (3)'s input variables.

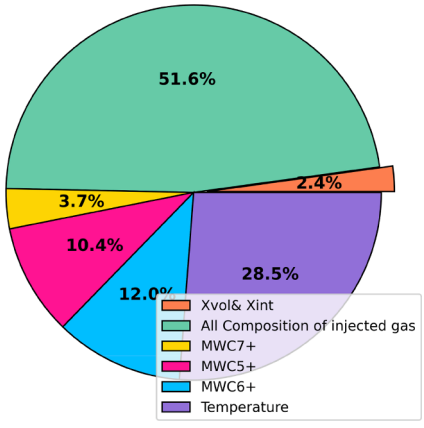

Fig. 28. Screening of main effect parameters on prediction MMP.

prediction as a function of the dataset (3) with MSE = 0.02 and R^2 = 0.99.

3. Depending on the findings of AIC, BIC, and KDE, the KNN model has the lowest values among ML methods, indicating KNN has low complexity and is an efficient fit for all trained models.
4. To assess the predictability of ML models, the dataset was divided into multiple pressure ranges. The results of this analysis revealed that MLR (Multiple Linear Regression) technique showed lower accuracy for the high-pressure range.
5. The results of the physical sensitivity parameters illustrated that the ML model has captured the- physical standard compared with real data. As shown in the Shapely plot, the most impactful parameter that has a direct effect on MMP prediction for three datasets is temperature.
6. The influence of relevant parameters on MMP prediction was verified, focusing on the entire composition of the injected gas, temperature, molecular weight of C6+, and molecular weight of C5+. The investigation revealed that the gas composition parameters had the most significant impact, accounting for approximately 46% of the total effect on MMP prediction. However, further analysis indicated that the inclusion of new components, namely Pb and API, as independent input parameters, had a negative impact on the prediction efficiency of MMP. These results indicate that the gas composition parameters play a significant role; however, the addition of Pb and API as independent parameters does not improve the accuracy of MMP prediction.
7. Hyper-parameters of each developed ML model are typically considered the limitations of these models to achieve the optimum accuracy, whereas tuning these parameters before training the models can have a significant impact on their performance and optimization.
8. Overall, predictive models for precisely assessing MMP have the potential to significantly aid reservoir engineers and EOR practitioners in improving CO_2 injection operations. Moreover, future studies may focus on integrating additional features and using ensemble approaches to improve the accuracy and generalization capabilities of machine learning-based MMP prediction models. Particularly, the computing time needed to train all kinds of ML models stays under 15 s, making them more efficient than other conventional approaches that take longer to complete the same operation.

CRediT authorship contribution statement

Harith F. Al-Khafaji: Writing – review & editing, Methodology. **Qingbang Meng:** Supervision, Resources, Project administration. **Wakeel Hussain:** Writing – review & editing, Validation, Data curation. **Rudha Khudhair Mohammed:** Writing – review & editing, Visualization, Data curation. **Fayez Harash:** Methodology, Conceptualization, Formal analysis. **Salah Alshareef AlFakey:** Software, Methodology, Formal analysis.

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.

Data availability

Data will be made available on request.

Acknowledgement

The project was supported by the National Natural Science Foundation of China (Grant NO.: 51804284) and the Chinese Government Scholarship. The authors would like to acknowledge Computer Modeling Group Ltd. for providing the CMG software for this study. Additionally, the authors express their sincere appreciation to the staff at the Petroleum Research and Development Center (PRDC) of the Iraqi Ministry of Oil, Prof. Zhong and Mr. Wahib Ali Yahya for their valuable criticisms and helpful advice during this research.

Appendix A

Model application

To make any developed model an accessible database for any new or developer user, in this study, five models of ML have been saved as digital models with the PKL extension that can be downloaded by any user. These models will improve the accuracy of the newly developed model based on the saved database, whereas the pre-evolved models will not need to be trained again; it is just a sequential step. The following steps are illustrated the procedure for using the currently developed model for any new user with new data sets:

1. Download the PKL extension.
2. Make sure to import the joblib module from scikit-learn.
3. Load the stored model using the joblib. load (pre-developed model.pkl) method.

```
# Load the saved model
Imported_model = joblib.load('pre-developed.pkl.pkl')
```

4 Once imported, the loaded_model can be used to make predictions on new data.

```
# Use the Imported model for predictions
New_data = [[...]]  # write "Your New Data"
New_predicted_model = Imported_model.predict(New_data)
```

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fuel.2023.129263.

References

[1] Shakeel M, Khan MR, Kalam S, Khan RA, Patil S, Dar UA. Machine Learning for Prediction of CO2 Minimum Miscibility Pressure. SPE J Sci Eng 2023. https://doi.org/10.2118/213322-ms.
[2] Zhang K, Liu L, Huang G. Nanoconfined Water Effect on CO2 Utilization and Geological Storage. Geophys Res Lett 2020;47:15. https://doi.org/10.1029/2020GL087999.
[3] Sheng J. Critical review of field EOR projects in shale and tight reservoirs. J Pet Sci Eng 2017;159:654–65.
[4] Wang Z, Li S, Jin Z, Li Z, Liu Q, Zhang K. Oil and gas pathway to net-zero: Review and outlook. Energ Strat Rev 2023;45:101048. https://doi.org/10.1016/j.esr.2022.101048.
[5] Zerpa L, Queipo N, Pintos S, Salager J. An optimization methodology of alkaline–surfactant–polymer flooding processes using field scale numerical simulation and multiple surrogates. J Pet Sci Eng 2005;47(3–4):197–208.
[6] Abdullah N, Hasan N. Effects of miscible CO2 injection on production recovery. J Pet Explor Prod Technol 2021;11(9):3543–57.
[7] Hoteit H, Fahs M, Soltanian M. Assessment of CO2 injectivity during sequestration in depleted gas reservoirs. Geosciences (Basel) 2019;9(5):199.
[8] Zhang K, Jin Z, Li G, Liu Q, Tian L. Gas adsorptions of geological carbon storage with enhanced gas recovery. Sep Purif Technol 2023;311:123260. https://doi.org/10.1016/j.seppur.2023.123260.
[9] Zhang K, Jin Z, Li S. Coupled miscible carbon utilization-storage processes in fractured shales. Chem Eng J 2022;441:135987. https://doi.org/10.1016/j.cej.2022.135987.
[10] Mavar K, Gaurina-Međimurec N, Hrnčević L. Significance of enhanced oil recovery in carbon dioxide emission reduction. Sustainability 2021;13(4):1800.
[11] Li D, Li X, Zhang Y, Sun L, Yuan S. Four Methods to Estimate Minimum Miscibility Pressure of CO2-Oil Based on Machine Learning. Chin J Chem 2019;37(12):1271–8.
[12] Choubineh A, Helalizadeh A, Wood DA. Estimation of minimum miscibility pressure of varied gas compositions and reservoir crude oil over a wide range of conditions using an artificial neural network model. Advances in Geo-Energy Research 2019; 3: 1(52–66). 10.26804/ager.2019.01.04.
[13] Wang H, Tian L, Zhang K, Liu Z, Huang C, Jiang L, et al. How Is Ultrasonic-Assisted CO2 EOR to Unlock Oils from Unconventional Reservoirs? Sustainability 2021;13(18):10010. https://doi.org/10.3390/su131810010.
[14] Holm LW, Josendal VA. Mechanisms of oil displacement by carbon dioxide. SPE J Petrol Technol 1974;26(12):1427–38. https://doi.org/10.2118/4736-PA.
[15] Lv Q, Zheng R, Guo X, Larestani A, Hadavimoghaddam F, Riazi M, et al. Modelling minimum miscibility pressure of CO2-crude oil systems using deep learning, tree-based, and thermodynamic models: Application to CO2 sequestration and enhanced oil recovery. Sep Purif Technol 2023;213:123086. https://doi.org/10.1016/j.seppur.2022.123086.
[16] Hemmati-Sarapardeh A, Ghazanfari M, Ayatollahi S, Masihi M. Accurate determination of the CO2-crude oil minimum miscibility pressure of pure and impure CO2 streams: a robust modelling approach. Can J Chem Eng 2016;94(2):253–61. https://doi.org/10.1002/cjce.22387.
[17] Alomair O, Malallah A, Elsharkawy A, Iqbal M. Predicting CO2 minimum miscibility pressure (MMP) using alternating conditional expectation (ACE) algorithm. Oil & Gas Sci Technol-Revue d'IFP Energies Nouvelles 2015;70(6):967–82. https://doi.org/10.2516/ogst/2012097.

[18] Yellig WF, Metcalfe RS. Determination and Prediction of CO2 Minimum Miscibility Pressures (includes associated paper 8876). SPE J Petrol Technol 1980;32(01): 160–8. https://doi.org/10.2118/7477-PA.

[19] Christiansen RL, Haines HK. Rapid measurement of minimum miscibility pressure with the rising-bubble apparatus. SPE Reserv Eng 1987;2(04):523–7. https://doi.org/10.2118/13114-PA.

[20] Rao DN, Lee JI. Application of the new vanishing interfacial tension technique to evaluate miscibility conditions for the Terra Nova Offshore Project. J Petrol Sci Eng 2002;35(3–4):247–62. https://doi.org/10.1016/S0920-4105(02)00246-2.

[21] Wu RS, Batycky JP. Evaluation of miscibility from slim tube tests. J Can Petrol Technol 1990;29(06). https://doi.org/10.2118/90-06-06.

[22] Chemmakh A, Merzoug A, Ouadi H, Ladmia A, Rasouli V. Machine Learning Predictive Models to Estimate the Minimum Miscibility Pressure of CO2-Oil System. In: In: SPE Abu Dhabi International Petrol Exhibition & Conference; 2021. https://doi.org/10.2118/207865-MS.

[23] Lee J. Effectiveness of carbon dioxide displacement under miscible and immiscible conditions. Report RR-40 Calgary: Petroleum Recovery Inst 1979.

[24] Cronquist C. Carbon dioxide dynamic miscibility with light reservoir oils. In: US DOE annual symposium, Tulsa. Springer, 1978; 28–30.

[25] Johnson JP, Pollin JS. Measurement and correlation of CO2 miscibility pressures. In: SPE/DOE enhanced oil recovery symposium;1981.

[26] Azin R, Izadpanahi A, Osfouri S. Fundamentals and Practical Aspects of Gas Injection. Springer 2022;2:23–72. https://doi.org/10.1007/978-3-030-77200-0.

[27] Alston RB, Kokolis GP, James CF. CO2 minimum miscibility pressure: a correlation for impure CO2 streams and live oil systems. Soc Petrol Eng J 1985;25(02):268–74. https://doi.org/10.2118/11959-PA.

[28] Glaso O. Generalized minimum miscibility pressure correlation. Soc Petrol Eng J 1985;25(6):927–34. https://doi.org/10.2118/12893-PA.

[29] Zuo Y, Chu J, Ke S, Guo T. A study on the minimum miscibility pressure for miscible flooding systems. J Petrol Sci Eng 1993;8(4):315–28. https://doi.org/10.1016/0920-4105(93)90008-3.

[30] Dong M, Huang S, Srivastava R. Effect of solution gas in oil on CO2 minimum miscibility pressure. J Can Petrol Technol 2000;39(11). https://doi.org/10.2118/99-47.

[31] Emera MK, Lu J. Genetic algorithm (GA)-based correlations offer more reliable prediction of minimum miscibility pressures (MMP) between the reservoir oil and CO2 or flue gas. SPE In: Can International Petrol Conference 2005. https://doi.org/10.2118/2005-003.

[32] Shokir E-M-E-M. CO2-oil minimum miscibility pressure model for impure and pure CO2 streams. J Petrol Sci Eng 2007;58(173–85). https://doi.org/10.1016/j.petrol.2006.12.001.

[33] Huang YF, Huang GH, Dong MZ, Feng GM. Development of an artificial neural network model for predicting minimum miscibility pressure in CO2 flooding. J Petrol Sci Eng 2003;37(1–2):83–95. https://doi.org/10.1016/S0920-4105(02)00312-1.

[34] Larestani A, Hemmati-Sarapardeh A, Samari Z, Ostadhassan M. Compositional Modeling of the Oil Formation Volume Factor of Crude Oil Systems: Application of Intelligent Models and Equations of State". ACS Omega 2022;7(28):24256–73. https://doi.org/10.1021/acsomega.2c01466.

[35] Akpobi ED, Oboh EP. Algorithm to Compute the Minimum Miscibility Pressure (MMP) for Gases in Gas Flooding Process. In: SPE Nigeria Annual International Conference and Exhibition; 2022. 10.2118/211973-MS.

[36] Sinha U, Dindoruk B, Soliman M. Prediction of CO2 minimum miscibility pressure MMP using machine learning techniques. In: SPE Improved Oil Recovery Conference 2020. https://doi.org/10.2118/200326-MS.

[37] Simovici D. Intelligent data analysis techniques—machine learning and data mining. Springer 2015;1–51. 10.1007/978-3-319-16531-8_1.

[38] Birang Y, Dinarvand N, Shariatpanahi S F, Edalat M. Development of a new artificial-neural-network model for predicting minimum miscibility pressure in hydrocarbon gas injection. In: SPE Middle East Oil and Gas Show and Conference 2007. 10.2118/105407-MS.

[39] Dehghani SAM, Sefti MV, Ameri A, Kaveh NS. Minimum miscibility pressure prediction based on a hybrid neural genetic algorithm. Chem Eng Res Design 2008;86(2):173–85. https://doi.org/10.1016/j.cherd.2007.10.011.

[40] Shokrollahi A, Arabloo M, Gharagheizi F, Mohammadi AH. Intelligent model for prediction of CO2 – Reservoir oil minimum miscibility pressure. Fuel 2013;112: 375–84. https://doi.org/10.1016/j.fuel.2013.04.036.

[41] Tatar A, Shokrollahi A, Mesbah M, Rashid S, Arabloo M, Bahadori A. Implementing Radial Basis Function Networks for modeling CO2-reservoir oil minimum miscibility pressure. J Nat Gas Sci Eng 2013;15:82–92. https://doi.org/10.1016/j.jngse.2013.09.008.

[42] Ahmadi MA, Ebadi M. Fuzzy modeling and experimental investigation of minimum miscible pressure in gas injection process. Fluid Phase Equilib 2014;378:1–12. https://doi.org/10.1016/j.fluid.2014.06.022.

[43] Sayyad H, Manshad AK, Rostami H. Application of hybrid neural particle swarm optimization algorithm for prediction of MMP. Fuel 2014;116:625–33. https://doi.org/10.1016/j.fuel.2013.08.076.

[44] Zhong Z, Carr TR. Application of mixed kernels function (MKF) based support vector regression model (SVR) for CO2 – Reservoir oil minimum miscibility pressure prediction. Fuel 2016;184:590–603. https://10.1016/j.fuel.2016.07.030.

[45] Karkevandi-Talkhooncheh A, Hajirezaie S, Hemmati-Sarapardeh A, Husein MM, Karan K, Sharifi M. Application of adaptive neuro fuzzy interface system optimized with evolutionary algorithms for modeling CO2-crude oil minimum miscibility pressure. Fuel 2017;205:34–45. https://doi.org/10.1016/j.fuel.2017.05.026.

[46] Saeedi Dehaghani AH, Soleimani R. Prediction of CO2-Oil Minimum Miscibility Pressure Using Soft Computing Methods. Chem Eng Technol 2020;43(7):1361–71. https://doi.org/10.1002/ceat.201900411.

[47] Dargahi-Zarandi A, Hemmati-Sarapardeh A, Shateri M, Menad NA, Ahmadi M. Modeling minimum miscibility pressure of pure/impure CO2-crude oil systems using adaptive boosting support vector regression: Application to gas injection processes. J Pet Sci Eng 2020;184:106499. https://doi.org/10.1016/j.petrol.2019.106499.

[48] Ghiasi MM, Mohammadi AH, Zendehboudi S. Use of hybrid-ANFIS and ensemble methods to calculate minimum miscibility pressure of CO2 - reservoir oil system in miscible flooding process. J Mol Liq 2021;331:115369. https://doi.org/10.1016/j.molliq.2021.115369.

[49] Chen H, Zhang C, Yu H, Wang Z, Duncan I, Zhou X, et al. Application of machine learning to evaluating and remediating models for energy and environmental engineering. Appl Energy 2022;320:119286. https://doi.org/10.1016/j.apenergy.2022.119286.

[50] Nathans LL, Oswald FL, Nimon K. Interpreting multiple linear regression: a guidebook of variable importance. Pract Assess Res Eval 2012;17(9):9.

[51] Deumah SS, Yahya WA, Al-khudafi AM, Ba-Jaalah KS, Al-Absi WT. Prediction of Gas Viscosity of Yemeni Gas Fields Using Machine Learning Techniques. In: SPE Symposium: Artificial Intelligence-Towards a Resilient and Efficient Energy Industry, 2021. 10.2118/208667-MS.

[52] Oladeinde MH, Ohwo AO, Oladeinde CA. A mathematical model for predicting output in an oilfield in the niger delta area of nigeria. Nigerian J Technol 2015;34 (4):768–72. https://doi.org/10.4314/njt.v34i4.14.

[53] Tranmer M, Elliot M. Multiple linear regression. The Cathie Marsh Centre for Census and Survey Research (CCSR) 2008;5(5):1–5.

[54] Guo L, Deng X. Application of improved multiple linear regression method in oilfield output forecasting. In: International conference on information management, innovation management and industrial engineering, IEEE 2009: 133–136. 10.1109/ICIII.2009.39.

[55] Cunningham CF, Cooley L, Wozniak G, Pancake J. Using multiple linear regression to model EURs of horizontal Marcellus shale wells. SPE Eastern Regional Meeting 2012. https://doi.org/10.2118/161343-MS.

[56] Ciulla G, D'Amico A. Building energy performance forecasting: A multiple linear regression approach. Appl Energy 2019;253:113500. https://doi.org/10.1016/j.apenergy.2019.113500.

[57] Vapnik V. The nature of statistical learning theory. Springer science & business media; 1999.

[58] Hastie T, Friedman J, Tibshirani R. Support Vector Machines and Flexible Discriminants. In: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer 2001:371–409. 10.1007/978-0-387-21606-5_12.

[59] Smola AJ, Schölkopf B. A tutorial on support vector regression. Springer Stat Comput 2004;14(3):199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88.

[60] Shawe-Taylor J, Cristianini N. Kernel methods for pattern analysis. Cambridge University Press; 2004.

[61] Rodriguez-Galiano V, Sanchez-Castillo M, Chica-Olmo M, Chica-Rivas M. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geol Rev 2015;71:804–18. https://doi.org/10.1016/j.oregeorev.2015.01.001.

[62] Breiman L, Friedman J, Olshen R, Stone C. Classification and regression trees. Florida Boca Raton: CRC Press; 1984.

[63] Song YY, Ying LU. Decision tree methods: applications for classification and prediction. Shanghai Arch. Psychiatry 2015;27(2):130. https://doi.org/10.11919/j.issn.1002-0829.215044.

[64] Sabah M, Talebkeikhah M, Agin F, Talebkeikhah F, Hasheminasab E. Application of decision tree, artificial neural networks, and adaptive neuro-fuzzy inference system on predicting lost circulation: A case study from Marun oil field. J Pet Sci Eng 2019;177:236–49. https://doi.org/10.1016/j.petrol.2019.02.045.

[65] L. Random forests. Mach Learn 2001;45: 5–32. 10.1023/A:1010933404324.

[66] Rodriguez-Galiano VF, Ghimire B, Rogan J, Chica-Olmo M, Rigol-Sanchez JP. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J Photogrammetry Remote Sensing 2012;67:93–104. https://doi.org/10.1016/j.isprsjprs.2011.11.002.

[67] Basith S, Manavalan B, Shin TH, Lee G. iGHBP: computational identification of growth hormone binding proteins from sequences using extremely randomised tree. Comput Struct Biotechnol J 2018;16:412–20. https://doi.org/10.1016/j.csbj.2018.10.007.

[68] Abdi J, Hadavimoghaddam F, Hadipoor M, Hemmati-Sarapardeh A. Modeling of CO2 adsorption capacity by porous metal organic frameworks using advanced decision tree-based models. Sci Rep 2021;11(1):24468.

[69] Ge Z, Song Z, Ding SX, Huang B. Data mining and analytics in the process industry: The role of machine learning. IEEE Access 2017;5:20590–616. https://doi.org/10.1109/ACCESS.2017.2756872.

[70] Sayyad Amin J, Bahadori A, Hosseini Nia B, Rafiee S, Kheilnezhad N. Prediction of hydrate equilibrium conditions using k-nearest neighbor algorithm to CO2 capture. Pet Sci Technol 2017;35(11):1070–7. https://doi.org/10.1080/10916466.2017.1302475.

[71] Imandoust SB, Bolandraftar M. Application of K-nearest neighbor (KNN) approach for predicting economic events theoretical background. Int J Eng Res Appl 2013;3: 605–10.

[72] Kramer O. Dimensionality reduction with unsupervised nearest neighbors. Springer; 2013. p. 51.

[73] Danesh A. PVT and phase behaviour of petroleum reservoir fluids. Elsevier; 1998.

[74] Merrill RC, Hartman KJ. A comparison of equation of state tuning methods. In: SPE Annual Technical Conference and Exhibition; 1994.
[75] Peng DY, Robinson DB. A new two-constant equation of state. Ind Eng Chem Fundam 1976;15(1):59–64.
[76] Ghorbani D, Kharrat R. Fluid characterization of an Iranian carbonate oil reservoir using different PVT packages. In: SPE Asia Pacific Oil and Gas Conference and Exhibition; 2001. 10.2118/68745-MS.
[77] Jensen F, Michelsen ML. Calculation of first contract and multiple contact minimum miscibility pressures. In: Situ;(USA)1990; 14(1).
[78] Ahmadi K, Johns RT. Multiple-mixing-cell method for MMP calculations. SPE J 2011;16(4):733–42. https://doi.org/10.2118/116823-PA.
[79] Huang C, Tian L, Jiang L, Xu W, Wang J. Prediction of Minimum Miscibility Pressure (MMP) of CO2-Crude Oil Systems Considering the Differences of MMP in Different Experiments Based on Artificial Neural Network and Bayesian Optimization Algorithm. Energy 2022;2004:2965.
[80] Kalam S, Yousuf U, Abu-Khamsin SA, Waheed UBin, Khan RA. An ANN model to predict oil recovery from a 5-spot waterflood of a heterogeneous reservoir. J Pet Sci Eng 2022; 210: 110012. 10.1016/j.petrol.2021.110012.
[81] Ali M, Zhu P, Huolin M, Pan H, Abbas K, Ashraf U, et al. A Novel Machine Learning Approach for Detecting Outliers, Rebuilding Well Logs, and Enhancing Reservoir Characterization. Nat Resour Res 2023:1–20. https://doi.org/10.1007/s11053-023-10184-6.
[82] Kalam S, Abu-Khamsin SA, Al-Yousef HY, Gajbhiye R. A novel empirical correlation for waterflooding performance prediction in stratified reservoirs using artificial intelligence. Neural Comput Appl 2021;33(7):2497–514. https://doi.org/10.1007/s00521-020-05158-1.
[83] Huang C, Tian L, Wu J, Li M, Li Z, Li J, et al. Prediction of minimum miscibility pressure (MMP) of the crude oil-CO2 systems within a unified and consistent machine learning framework. Fuel 2023;337:127194. https://doi.org/10.1016/j.fuel.2022.127194.
[84] Hameed A. A New Correlation of Minimum Miscibility Pressure for CO2 Flood. Baghdad: Oil and Gas Engineering Department, University of Technology; 2017 [MS Thesis] [in English].
[85] Jani G. Enhanced oil recovery methods for extraction of crude oil in Noor field [MS Thesis]. Baghdad: Petroleum Engineering Department, University of Baghdad; 2013 [in English].
[86] CMG-WinProp. CMG Software | WinProp Fluid Property Characterization Tool. https://www.cmgl.ca/winprop (accessed Apr. 20, 2023).
[87] Chen H, Zhang C, Jia N, Duncan I, Yang S, Yang Y. A machine learning model for predicting the minimum miscibility pressure of CO2 and crude oil system based on a support vector machine algorithm approach. Fuel 2021;290:120048. https://doi.org/10.1016/j.fuel.2020.120048.
[88] Chen G, Fu K, Liang Z, Sema T, Chen Li, Tontiwachwuthikul P, et al. The genetic algorithm based back propagation neural network for MMP prediction in CO2-EOR process. Fuel 2014;126:202–12. https://doi.org/10.1016/j.fuel.2014.02.034.