Discover Applied Sciences

Research

Predictive modeling of CO_2 capture efficiency using piperazine solutions: a comparative study of white-box algorithms

Fahimeh Hadavimoghaddam[1,2] · Jianguang Wei[1,3,4] · Alexei Rozhenko[5] · Peyman Pourafshary[6] · Abdolhossein Hemmati-Sarapardeh[7,8]

Received: 19 December 2023 / Accepted: 1 October 2024
Published online: 06 November 2024

Abstract
The urgency to mitigate carbon dioxide (CO_2) emissions and combat climate change has spurred the development of effective CO_2 capture technologies. One of the industry's most well-known CO_2 collection processes is CO_2 absorption utilizing amine solvents. However, designing a successful amine scrubbing system in power plants requires precise prediction of CO_2 absorption in aqueous amine solutions under various operating circumstances. Using aqueous piperazine (PZ) solutions for chemical absorption is promising due to the favorable reactivity of PZ with CO_2. In this study, a comprehensive evaluation of PZ solution performance in CO_2 capturing, employing the white-box algorithms, namely, Genetic Programming (GP), Gene Expression Programming (GEP), and Group Method of Data Handling (GMDH), was performed. Through extensive experimentation and data analysis, several correlations were developed with high and acceptable R^2 values, such as 0.933 for GP, 0.949 for GEP, and 0.889 for GMDH, which shows high accuracy and reliability in predicting the CO_2 capture efficiency of PZ solutions under varying operating conditions. The results of sensitivity analysis revealed that CO_2 partial pressure increased CO_2 absorption, while PZ concentration and temperature had negative and decreasing effects. These insights provide essential guidance for optimizing process conditions to enhance the CO_2 capture efficiency. Finally, the leverage method was used to assess the reliability of both experimental and predicted data from white-box algorithms. This analysis identified potential outliers and validated the accuracy of the model's predictions, enhancing the credibility of developed correlations and demonstrating the robustness of the approach.

Article Highlights

- Aqueous piperazine solutions show promise, with white-box algorithms predicting high efficiency (R2 values: GP 0.933, GEP 0.949, GMDH 0.889).
- Insights from sensitivity analysis indicate that increasing CO_2 partial pressure enhances absorption, while higher PZ concentration and temperature have decreasing effects.
- The study's use of white-box algorithms, including Genetic Programming and Group Method of Data Handling, demonstrates a reliable approach for predicting and optimizing CO2 capture in power plants.

✉ Peyman Pourafshary, peyman.pourafshary@nu.edu.kz | [1]Institute of Unconventional Oil and Gas, Northeast Petroleum University, Daqing 163318, Heilongjiang, China. [2]Ufa State Petroleum Technological University, Ufa 450064, Russia. [3]National Key Laboratory of Continental Shale Oil, Northeast Petroleum University, Daqing 163318, China. [4]Key Laboratory of Continental Shale Hydrocarbon Accumulation and Efficient Development (Northeast Petroleum University), Ministry of Education, Northeast Petroleum University, Daqing 163318, China. [5]ITMO University, Saint-Petersburg, Russia 197101. [6]School of Mining and Geosciences, Nazarbayev University, 01000 Astana, Kazakhstan. [7]Department of Petroleum Engineering, Shahid Bahonar University of Kerman, Kerman, Iran. [8]State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum, Beijing, China.

Keywords CO_2 capture · Aqueous piperazine solution · Liquid absorption method · Machine learning algorithms · White-box algorithms · Leverage technique

1 Introduction

Even though carbon dioxide (CO_2) is essential for organisms on Earth, a significant portion of the greenhouse effect is attributable to the release of CO_2 into the atmosphere. In addition to natural CO_2 sources, such as the oceans, humans also contribute to CO_2 emissions. The energy-related industries are the primary contributors to CO_2 emissions [1–3]. This rise may be primarily attributed to the use of fossil fuels as the primary source of CO_2 emissions by humans and industrial activities. As industrialization and population growth have accelerated over the past century, burning fossil fuels, deforestation, and other human activities have significantly increased CO_2 emissions. The concentration of CO_2 in the atmosphere has risen to unprecedented levels, exceeding 415 parts per million (ppm) as of 2021 [1]. This alarming trend has resulted in the greenhouse effect, where CO_2 and other greenhouse gases trap heat in the Earth's atmosphere, leading to rising global temperatures and disruptive climate patterns [2]. The consequences of global warming are becoming increasingly evident, with rising sea levels, extreme weather events, and the loss of biodiversity impacting ecosystems and human livelihoods worldwide. Urgent and coordinated efforts are required to reduce CO_2 emissions and mitigate the adverse effects of global warming on the planet and its inhabitants. The demand for credible energy generation prevents the majority of nations from abandoning fossil fuels. Therefore, it appears that the most effective method of managing CO_2 emissions is to integrate CO_2 capture techniques into producing and refining fossil fuels [4]. Along with its environmental benefits, CO_2 elimination has supplementary positive effects, such as an increase in the calorific value of natural gases [5]. In addition, transportation expenses are reduced [6].

Absorption-based capture of CO_2 is a technique that is commonly applied in practice [7]. Several investigations have examined the possibility of using amine-blend solvent systems for CO_2 elimination due to their reactivity with CO_2 [8–11]. The most prevalent industrial solvents utilized for CO_2 separation from sour streams are alkanolamine solutions such as triethanolamine (TEA), methylidiethanolamine (MDEA), monoethanolamine (MEA), diethanolamine (DEA), and piperazine (PZ) [12–16]. These solvents facilitate the separation of CO_2 from gas streams through chemical reactions that involve the formation of stable carbamate or bicarbonate species. The selective separation of CO_2 from industrial gas emissions is conducted by forming reversible chemical species [17, 18]. The absorption process is characterized by intricate thermodynamics, where parameters like temperature, pressure, solvent concentration, and CO_2 partial pressure influence the overall efficiency of CO_2 capture. Among the mentioned solvents, PZ, as a cyclic amine compound, has garnered attention due to its unique capacity to react with CO_2 faster, often resulting in higher CO_2 absorption. Figure 1 depicts a schematic representation of a standard amine scrubbing setup for CO_2 absorption. In this configuration, the interactions between CO_2 and the amine occur within an absorber unit, while the stripper or regenerator column detaches the CO_2 from the CO_2-rich solvent. This process subsequently allows the CO_2-lean solvent to be reintroduced into the absorber, fostering continued CO_2 absorption and establishing the cyclic pattern.

Forecasting CO_2 absorption with amine solutions is a challenging yet crucial endeavor in the industry, prompting researchers in recent years to employ diverse methods, including machine learning (ML), for predictive modeling. ML has been increasingly used in recent years to predict the CO_2 loading capacity of MEA, DEA, TEA, and PZ solutions [6, 19–21]. Ghiasi and Mohammadi [6] designed a least-squares support vector machine (LSSVM) to calculate CO_2 solubility in a number of amine solutions based on the CO_2 partial pressure, temperature, and concentration of amine. A comparable study was also undertaken using an adaptive neuro-fuzzy inference system (ANFIS) [19]. Daneshvar et al. [20] applied the ANN technique to determine CO_2 loading in triisopropanolamine (TIPA), TIPA/MEA, and TIPA/PZ solvents. In another study [21], by utilizing radial basis function (RBF) and multilayer perceptron (MLP) networks, the absorption capacity of CO_2 in MDEA and DEA was calculated. Moreover, Dashti et al. [22] developed 4 intelligent methods to anticipate CO_2 solubility in 12 amine-based solvents. They concluded that the LSSVM model adjusted by the coupled simulated annealing (CSA) optimization technique could produce the most trustworthy outcomes. Using artificial neural networks (ANN), Salooki et al. [23] endeavored to forecast the outcome variables of a stripper operating in one of the Iranian gas refineries. In addition, the output temperature and flow rate of this stripper were computed using the SVM methodology [24]. ANN was also utilized for predicting steady-state CO_2 capture by MEA aqueous solution [25]. Other sources also report similar research studies [26–29].

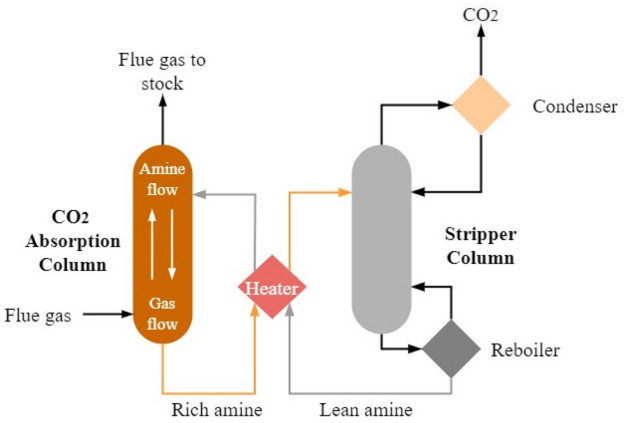

Fig. 1 Schematic of the absorption-stripping system

Considering the capability of PZ aqueous solutions to capture CO_2, numerous scholars have been working on developing accurate methods for estimating the CO_2 absorption capacity of PZ solvents. In 2016, Tatar et al. [30] presented two intelligent techniques, including ANFIS coupled with Conjugate Hybrid-Particle Swarm Optimization (CHPSO-ANFIS) and CSA-LSSVM, to forecast CO_2 solubility in PZ solutions, demonstrating the superior performance of the CHPSO-ANFIS model. Yarveicy et al. [31] undertook comparable research by applying four intelligent methods, including ANN, adaptive boosting classification and regression tree (AdaBoost-CART), ANFIS, and LSSVM. Dashti et al. [32] designed GA-ANFIS and genetic programming (GP) models to forecast CO_2 solubility in PZ solutions as a function of temperature, PZ concentration, and CO_2 partial pressure. Their models were generated by applying a data bank of 390 data points. The average absolute relative deviations (AARDs) for the generated GP and GA-ANFIS models were 5.3% and 9.7%, respectively. To the best of our knowledge, this databank is the most extensive collection of records employed in establishing intelligent models to forecast CO_2 loading in aqueous PZ solutions. The application of ML in CO_2 loading capacity prediction is still in its early stages. However, the results of recent studies suggest that ML has the potential to be a powerful tool for optimizing CO_2 capture processes [21, 32–36]. As indicated by the literature review, numerous decent intelligent models have been introduced for predicting CO_2 loading by PZ during past years; however, they frequently operate as black-boxes, lacking transparency in elucidating distinct relationships between input variables. The adoption of white-box approaches, on the other hand, has the potential to formulate explicit mathematical correlations with precision.

Taking the progress of the previous researches into account, the main purpose of the current study is to provide a more universal approach which would not only be accessible to a broad range of specialists, scientists as well as enthusiasts, but may also be used to make forecasts in more varying operational conditions. Therefore, in this study, a pioneering and extensive collection of 517 data points is compiled from the literature to model the CO_2 loading by PZ. Additionally, one of the most modern and efficient white-box algorithms such as Group Method of Data Handling (GMDH), Gene Expression Programming (GEP), and Genetic Programming (GP) are uniquely employed to present simple-to-use correlations. These algorithms are particularly adept at constructing mathematical expressions, thus rendering them highly applicable for precisely predicting CO_2 loading capacity. This approach advances beyond conventional ML methods by predicting outcomes and offering interpretable insights into the underlying mechanisms governing the CO_2 absorption rate. Also, a sensitivity analysis is performed to investigate the impact of key operational parameters on CO_2 absorption in PZ aqueous solutions. Finally, the leverage method was applied to investigate the reliability of both experimental data and predicted data obtained from the white-box algorithms.

2 Data gathering

For the development of intelligent models, an extensive collection of 517 data points was compiled from experimental data in the literature [37–44]. Table 1 provides detailed information of the PZ concentration (M), CO_2 partial pressure (kPa), temperature (K), and, finally, CO_2 loading (mol CO_2/mol PZ). CO_2 loading refers to the maximum amount of CO_2 that a given quality or volume of a solvent can absorb and retain under specific conditions. CO_2 loading capacity is a critical factor in evaluating the performance of solvents used in carbon capture processes.

Table 1 Statistical description of input parameters

	T (K)	PZ concentration (M)	CO_2 partial pressure (kPa)	CO_2 loading (mol CO_2/mol PZ)
Count	517	517	517	517
Mean	317.6	1.3	443.8	1.0
Skew	1.0	1.5	3.9	0.4
Std	15.6	1.4	995.2	0.2
Min	298.0	0.1	0.0	0.7
25%	308.0	0.4	2.3	1.0
50%	313.2	0.8	28.0	1.2
75%	328	1.6	474.6	3.0
Max	373.2	6.2	7399	517

The statistical categories in the left part of a table represent the amount of points (Count), the mean value (Mean), skewness (Skew), standard deviation (Std), minimum and maximum values (Min, Max) as well as data quartiles (25, 50 and 75%).

Figure 2 shows the box plots for all parameters, representing the wide range of data used to develop the models. As can be seen, temperature and PZ concentration are measured at certain levels, while CO_2 partial pressure and CO_2 loading are taken more unsystematically, covering almost every segment from 0.032 to 7399 and from 0.72 to 2.956, accordingly. In general, the data are sufficiently extensive and diverse to develop accurate models to anticipate the target variables.

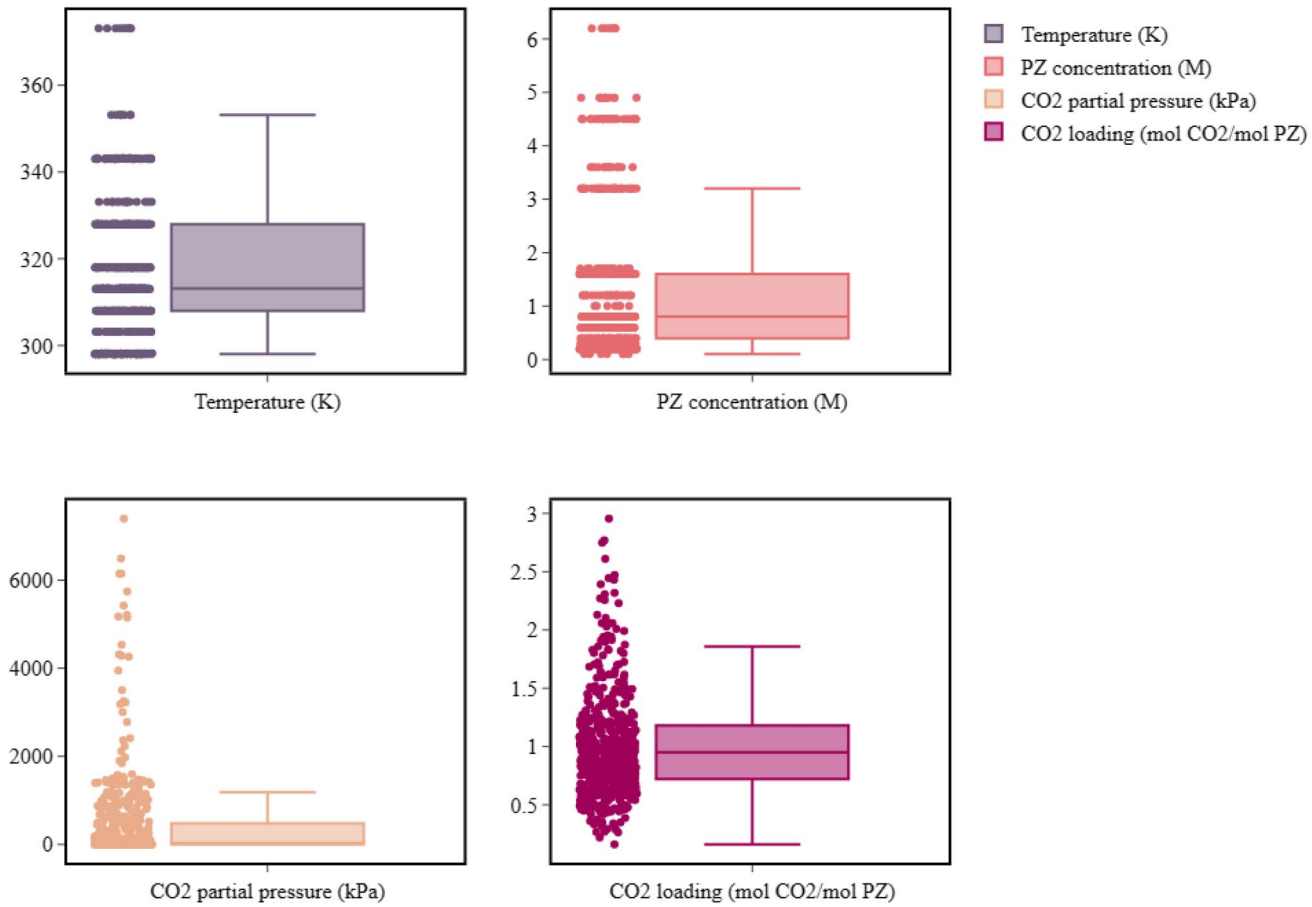

Fig. 2 Graphical analysis of the input parameters

The last part of the initial data analysis was the heatmap presented in Fig. 3. According to the findings, the parameter influencing the CO_2 loading the most is considered to be the CO_2 partial pressure, accounting for 64% of the output behavior. At the same time, temperature as well as piperazine concentration have almost similar impacts on the target parameter, with correlation scores of − 41% and − 39%, respectively. Speaking more specifically about the relationships between input parameters, piperazine concentration and temperature are the only pair that have a significant dependence (43%).

3 Model development

3.1 Genetic programming (GP)

Genetic Programming (GP) is an example of an artificial intelligence (AI) algorithm; it was invented by Cramer [45] and cultivated by Koza [46]. In addition, GP is now a primary technique comprising many sub-techniques that were all founded on the same principles, such as Cartesian Genetic Programming (CGP), Linear genetic programming (LGP), Gene Expression Programming (GEP), Evolutionary Polynomial Regression (EPR), Meta-genetic programming (MGP), and Multi-Gene Genetic Programming (MGGP). GP relies heavily on applying the Genetic Algorithm (GA) method to mathematical equations. GA, in turn, is based on the concepts of selection (determination of the individuals who will be chosen as parents for the subsequent batch), crossover (generation of offsprings from parents), and mutation (random modification of some offspring).

GA is one of the earliest AI techniques. It imitates the process of developing biological organisms by producing random solutions to a problem, evaluating the appropriateness of each solution, selecting the most fitting ones and discarding the rest, merging the selected answers to form the next generation of solutions employing crossover procedures, and repeating the cycle till an acceptable level of fit is reached [47]. In GP, mathematical equations need to be expressed in the form of a binary tree and encoded as genetic forms to make it possible to apply GA operators. The notion of a genetic form and a binary tree is depicted in Fig. 4, which represents a simple equation "$8x + y^2 - 5y$" expressed in a logical structure of GP [48].

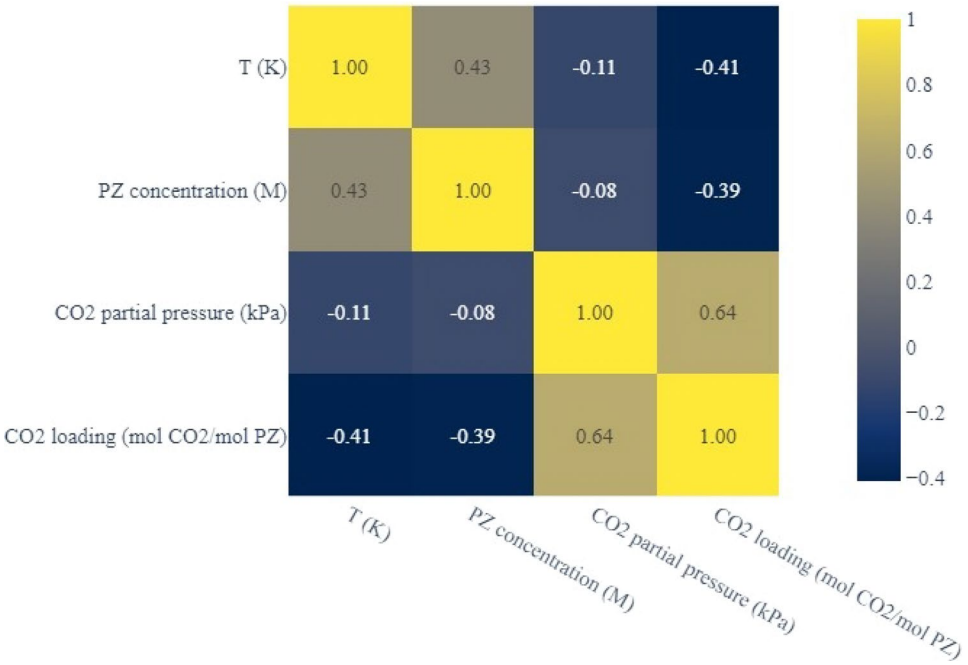

Fig. 3 Heatmap of parameters used in this study

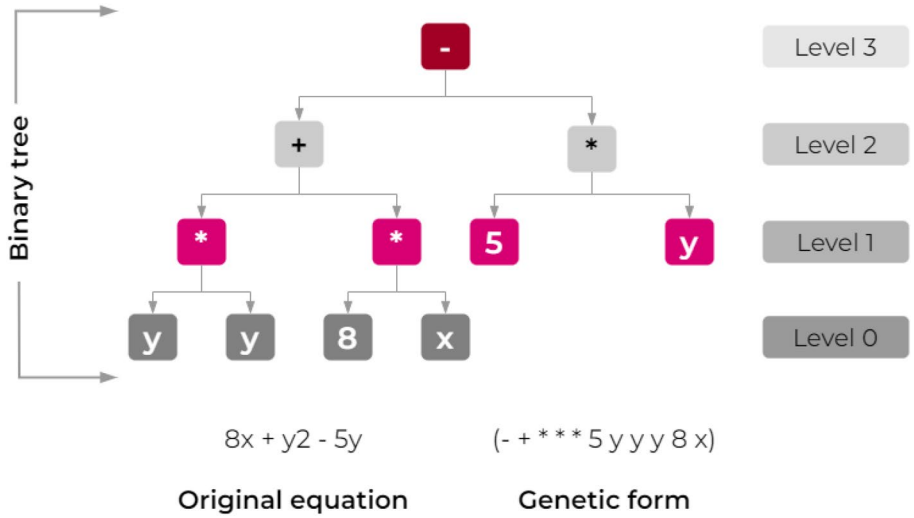

Fig. 4 Example of binary tree and genetic form for an equation

3.2 Group Method of Data Handling (GMDH)

GMDH is a technique for anticipation and recognition initially proposed by Ivakhnenko [49]. It was created to address complicated structures with massive dimensions that are difficult to formulate when the data series is brief or when the data relationships exhibit multicollinearity [50]. The primary objective of the GMDH is to inductively learn and progressively construct a generalized analytic function with the appropriate level of complexity [51], as presented in Fig. 5.

The GMDH technique constructs successive layers of connections. The first layer is the input layer, comprised of numerous nodes that refer to the predictor variables. The nodes are connected pairwise via a quadratic polynomial to generate candidate nodes for the subsequent layer. The relationship between the output y_i and the input variables $x_{i,j,\ldots,k}$ is constructed by employing a nonlinear function known as Volterra Kolmogorov Gabor presented in Eq. (1) [52, 53]:

$$y_i = a_0 + \sum_{i=1}^{d} a_i x_i + \sum_{i=1}^{d}\sum_{j=1}^{d} a_{ij} x_i x_j + \ldots + \sum_{i=1}^{d} \times \sum_{j=1}^{d} \ldots \sum_{k=1}^{d} a_{ijm} x_i x_j \ldots x_m \tag{1}$$

where a_o, a_i, a_{ij}, …, a_{ijm} represent coefficients or weights of polynomial variables, and d represents the number of input variables.

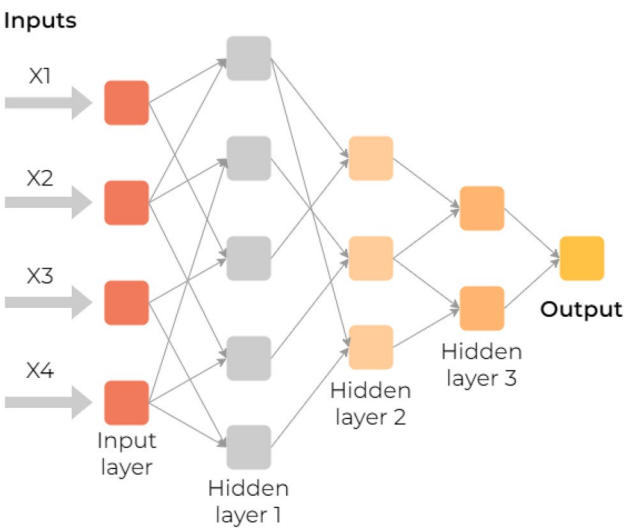

Fig. 5 Schematic of GMDH approach

3.3 Genetic expression programming (GEP)

Ferreira [54] proposed the GEP as a novel GA founded on genotypes and expressions. Similar to GA and GP, it replicates the natural selection of populations by the environment and the processes of heredity and mutation among members of populations [55–57]. It employs fitness to pick individuals within populations and a number of genetic operators to modify the chosen ones to produce new populations while promoting evolution.

The flowchart in Fig. 6 depicts the fundamental phases of GEP. This process continues until an effective solution is found. GEP employs genes of fixed length, but by linking functions, such genes may correspond to expression trees (ET) of varying sizes and structures [58]. These linking functions permit the gene combinations to generate chromosomes.

4 Evaluation of models

The precision of the suggested models was evaluated by employing multiple statistical indicators. The definition of these parameters is shown as below [59]:

5 Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_i'\right)^2} \tag{2}$$

6 Standard Deviation (SD):

$$\mathrm{SD} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{y_i - y_i'}{y_i}\right)^2}. \tag{3}$$

7 The Coefficient of determination (R^2):

$$R^2 = 1 - \frac{\sum_{t=1}^{T}\left(y_i - \mathrm{f}_i\right)^2}{\sum_{t=1}^{T}\left(y_i - y_i'\right)^2} \tag{4}$$

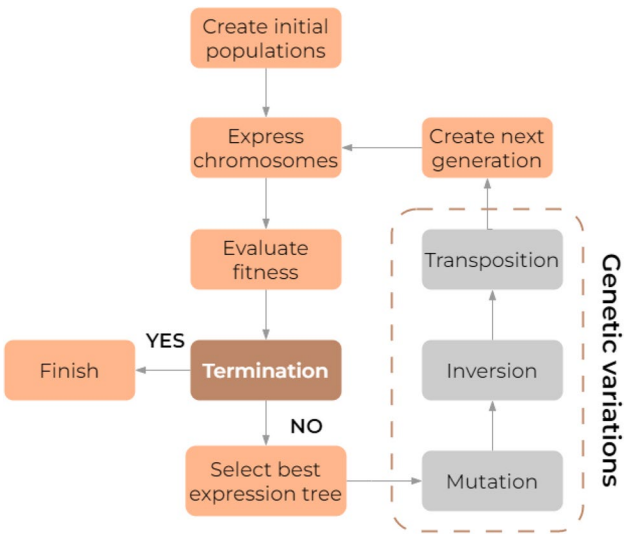

Fig. 6 Schematic of GEP approach

8 Mean Absolute Percentage Error (MAPE%):

$$E_r = \frac{1}{n}\sum_{i=1}^{n} abs\left(\left[\frac{y_i - y_i'}{y_i}\right]\right) \times 100 \tag{5}$$

9 Mean Absolute Value (MAE):

This parameter reflects a measure of risk corresponding to the predicted absolute error loss figure. The mean absolute error (MAE) over n_{samples} is calculated using Eq. (6) if y_i' is the anticipated value of the i-th sample while y_i is the actual number:

$$MAE(y, y') = \frac{1}{n_{samples}} \sum_{i=0}^{N_{samples}-1} |y_i - y_i'|. \tag{6}$$

10 Mean Bias Error (MBE):

This indicator calculates the mean prediction error in the following manner depicted in Eq. (7):

$$MBE = \frac{1}{n}\sum_{i=1}^{n} (y_i' - y_i) \tag{7}$$

In addition, to visually introduce the outcomes of the present study, 4 types of graphical materials have been utilized, mainly scatter plot, residual plot, histogram, and cumulative frequency plot.

11 Results and discussion

11.1 Developed correlations

GP, GEP, and GMDH have been employed to anticipate CO_2 loading using PZ concentration, CO_2 partial pressure, and temperature as input parameters. Before modeling, the input data was randomly shuffled and divided into 2 datasets used to train and test the resulting models, respectively.

- For the GP algorithm, the following correlation has been proposed:

$$CO_2 \text{ loading} = \left(c_0 T + \frac{c_1 PP}{c_2 T} + \frac{1}{c_3 PP} + \log\left(c_4 PP\right) * c_5 + \frac{\log\left(c_6 PP\right)}{c_8 PZ} + c_9\right),$$

$$c_0 = -0.007,\ c_1 = 0.064,\ c_2 = 1.342,\ c_3 = -82.455,\ c_4 = 1.565,\ c_5 = 0.029,$$
$$c_6 = 1.957,\ c_7 = 0.029,\ c_8 = 1.303,\ c_9 = 2.902.$$

- For GEP algorithm, the following correlation has been proposed:

$$CO_2 \text{ loading} = \left(c_0 PP + c_1 T + \log\left(c_2 PP\right) \times c_3 + \log\left(c_4 PP\right) \times c_5 + \frac{\log\left(c_6 PP\right) \times c_7}{c_8 PP} + \frac{\log\left(c_9 PP\right) \times c_{10}}{c_{11} PZ} + c_{12}\right).$$

$c_0 = 0.0001446; c_1 = -0.0073282; c_2 = 0.89973; c_3 = -0.0032597; c_4 = 2.0512; c_5 = 0.035897;$
$c_6 = 0.6008; c_7 = 0.0032597; c_8 = 1.2568; c_9 = 1.8354; c_{10} = 0.0032597; c_{11} = 0.14814; c_{12} = 2.9684.$

- For GMDH algorithm, the following correlation has been proposed:

$$CO_2 \text{ loading} = -0.0260363 - \text{PP} \times 8.99617\text{e}-05 + (\text{PP})^2 \times 1.98022\text{e}-08 + \text{N2} \times 1.04222.$$

$$\text{N}_1 = -0.407217 + \text{N}_4 \times 1.70701 - \text{N}_4 \times \text{N}_2 \times 1.1815 + N_2^2 \times 0.886315;$$

$$\text{N}_2 = 1.1632 - \text{N8} \times 2.07024 + \text{N8} \times \text{N10} \times 2.92341 - \text{N}_5^2 \times 1.04949;$$

$$\text{N}_3 = 0.921387 - \text{PZ} \times 0.0649256 - \text{PZ} \times \text{PP} \times 0.000115387 + \text{PP} \times 0.000697744 - (\text{PP})^2 \times 6.97729\text{e}-08;$$

$$\text{N}_4 = 15.467 - \text{T*}0.0819317 - \text{T} \times \text{PP} \times 1.8041\text{e}-06 + (\text{T (K)})^2 \times 0.000112539 + \text{PP*}0.00113445 - (\text{PP})^2 \times 5.4595\text{e}-08;$$

$$\text{N}_5 = 2.94259 - \text{T*}0.00573336 - \text{PZ} \times 0.0859164;$$

where CO_2 partial pressure (kPa) = PP, T(K) = T and PZ concentration (M) = PZ.

11.2 Statistical evaluation of models

According to Table 2, the most accurate models developed are GEP and GP, with the estimates far more precise when compared to GMDH. The RMSE for both techniques for all datasets mentioned is less than 0.13, R^2 is more than 0.89, MAPE values are not greater than 8.4, MAE does not exceed 0.08, and SD is lower than 0.12. Moreover, the distinctions between the training outcomes and data testing sets are insignificant. This fact indicates the absence of overfitting or underfitting in 2 approaches observed. The above evidence represents the robustness and reliability of GP and GEP models in predicting CO_2 loading based on PZ concertation, CO_2 partial pressure, and temperature.

GMDH is the least precise technique among the methods utilized. RMSE, MAE, SD, and MAPE of the "train" and "all" datasets are approximately twice (and in some cases even three times) higher than other algorithms, and the difference between testing and training R^2 is equal to 0.22, which indicates the substantial inconsistency of results. R^2 of the "all" and "train" sets are less than 0.74 and about 0.18 less than the corresponding values of GP and GEP. Hence, the GMDH model could be considered an example of an untrustworthy one that is too inaccurate to predict the target parameter precisely.

Table 2 Statistical results of models developed

	Dataset	RMSE	SD	R^2	MAPE	MBE	MAE
GEP	All	0.120	0.116	0.924	8.153	0.000	0.080
	Train	0.119	0.115	0.929	8.105	0.003	0.080
	Test	0.124	0.115	0.890	8.345	−0.011	0.079
GP	All	0.118	0.111	0.927	7.818	0.000	0.077
	Train	0.117	0.111	0.932	7.825	0.003	0.078
	Test	0.122	0.110	0.894	7.787	−0.010	0.075
GMDH	All	0.228	0.354	0.725	20.358	0.000	0.161
	Train	0.225	0.356	0.735	20.260	−0.003	0.160
	Test	0.091	0.107	0.955	7.690	−0.003	0.067

11.3 Graphical evaluation of models

Figure 7 shows scatter plots for 3 developed models. GEP and GP are the most accurate techniques, with most observations placed around the 0% error line and within the borders of ∓10% line. It should be mentioned that the GP and GEP graphs are almost similar. GMDH is the least precise algorithm with the most significant error terms. The fact worth mentioning is that the noticeable outliers belong to the training group, and the points of the testing group are all located close to the 45° lines.

Figure 8 shows the residual plots. The main data portion is located within the scope of – 0.2 and 0.2 for both GEP and GP, and the overall spread is from approximately – 0.9 to 0.4. It again should be underlined that the graphs look practically identical.

GMDH has the highest number of significant outliers. Despite having the most observations in the interval between – 0.5 and 0.5, the general data range is from almost – 1.5 to around 0.6. That, in turn, substantially decreases the model's performance, making it the least precise among all techniques. Furthermore, this figure provides a clear explanation for the observed discrepancy between the testing and training results, as well as the entire dataset. The

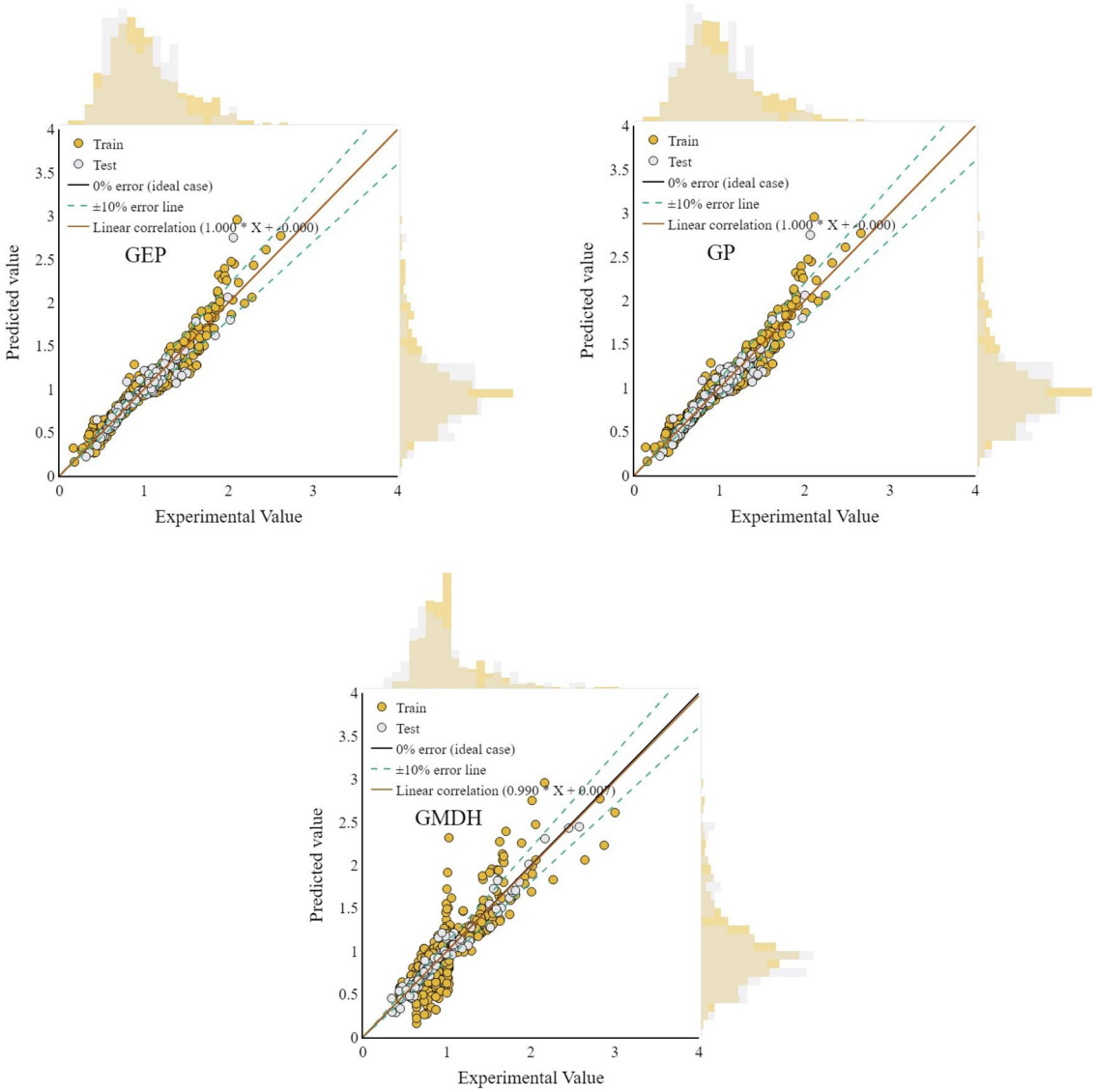

Fig. 7 Scatter plots for developed models

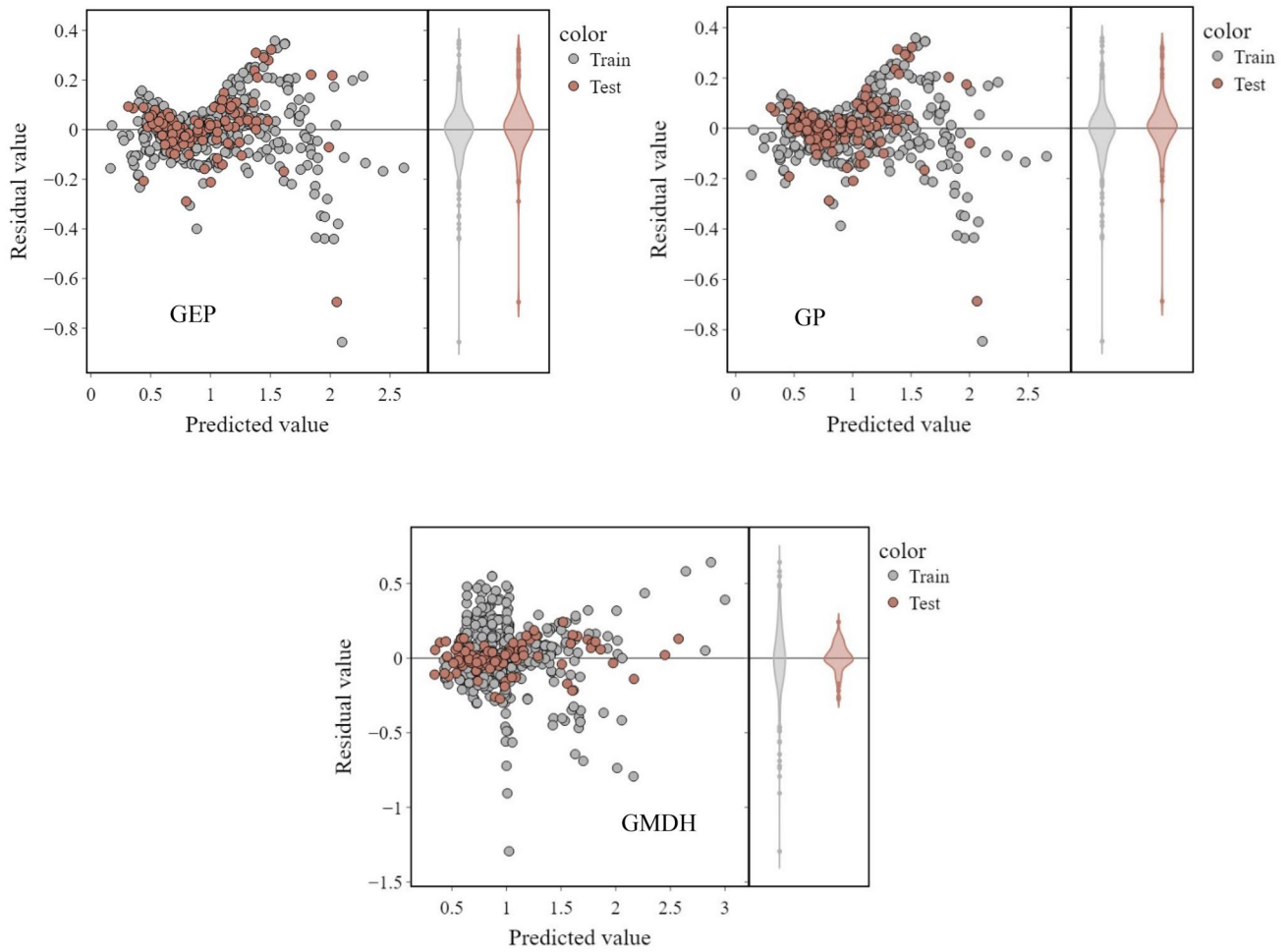

Fig. 8 Residual plots for developed models

testing set was randomly chosen so that all points were placed within the approximate range of 0.3 and – 0.3, and no errors were taken, whereas all the outliers were in the training group.

The histograms presented in Fig. 9 prove the patterns observed in the last figures. GEP and GP have the minor deviation borders of approximately (– 0.6 to 0.5) and (– 0.5 to 0.6), respectively. Furthermore, the largest share of observations is around the zero point, enabling the algorithms to show remarkable results.

GMDH has most of its data centered around the zero-point, but the substantial deviations of up to – 3 make its performance the most inaccurate among all the approaches employed.

Figure 10 displays the cumulative frequency plot. GEP and GP are more precise than GMDH, having almost identical performance levels. Their angles are steeper, indicating that with more observations added, the relative error increase is relatively lower than GMDH. Moreover, about 90% of points in GEP and GP have an error of 25%, whereas only 80% of GMDH data accounts for a considerable 40% error.

11.4 Trend analysis

It is essential to evaluate the performance of the GP model to predict the physical trend of CO_2 absorption in aqueous PZ solutions, taking into account influential variables. First, as studied experimentally in the literature [35], the investigation focused on examining the GP model's forecast of the CO_2 solubility in a 0.2 M PZ solution at different partial pressures of CO_2 and temperatures. As observed in Fig. 11, the absorption values of CO_2 exhibited an upward trend as the CO_2 partial pressure rose. This trend may be attributed to the increased driving force for absorption at higher levels of CO_2 partial pressure.

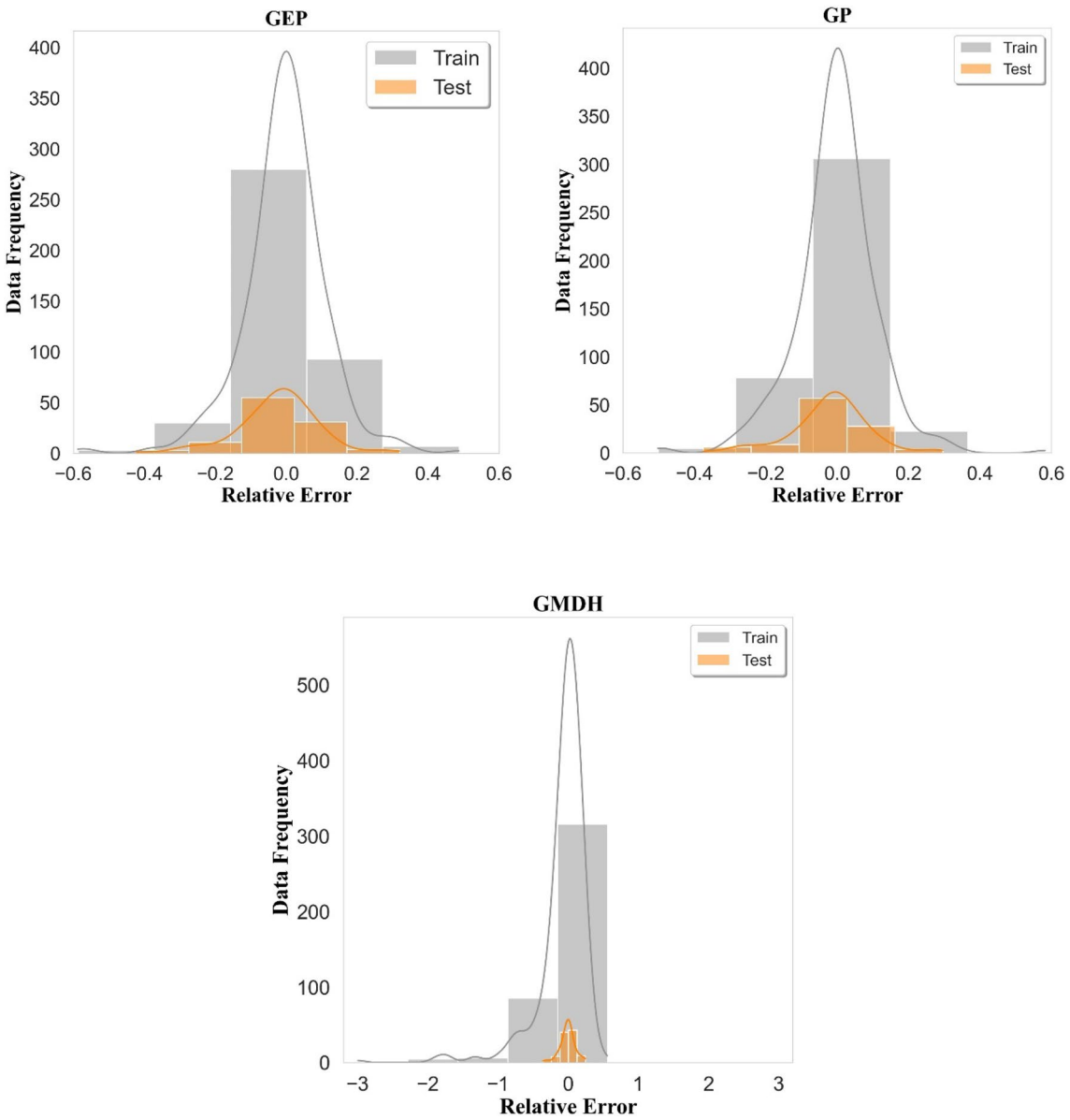

Fig. 9 Histograms of developed models

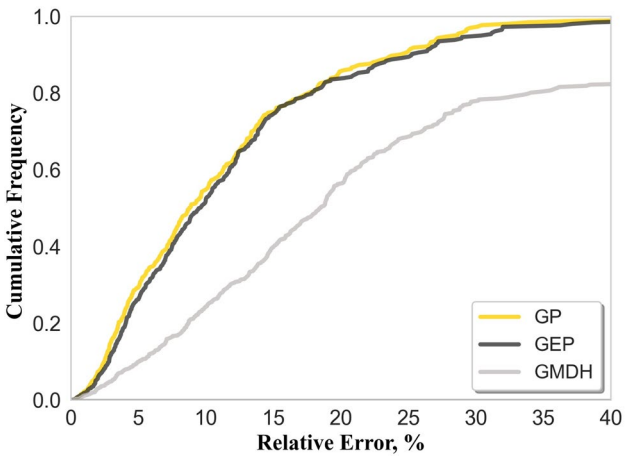

Fig. 10 Cumulative frequency plot for developed models

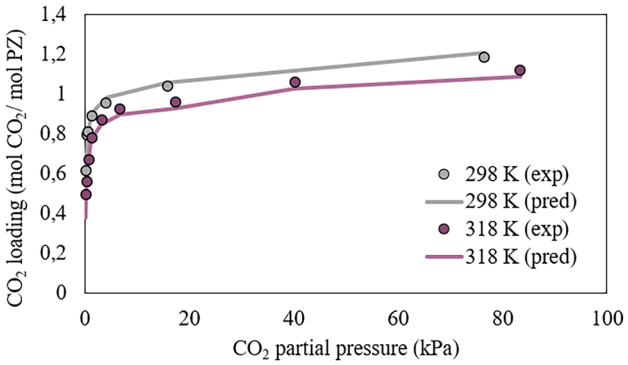

Fig. 11 The impact of temperature on CO_2 solubility in 0.2 M PZ solution; experimental findings [39] and forecasts based on the GP model

Experimental examinations have shown that introducing CO_2 or a sour gas into PZ solution results in the dissolution of the gas, mainly in an ionic and non-volatile form. Consequently, the overall pressure exhibits a marginal increase as the gas content in the liquid phase progressively elevates. Once PZ has been utilized in the liquid phase by chemical processes, the CO_2 partial and total pressure rise substantially with higher gas loadings. Hence, the capacity for chemical absorption of more sour gas has reached its limit, necessitating physical dissolution methods [60, 61]. This happens when the solubility of CO_2 in PZ aqueous solution is adversely affected by temperature, leading to a substantial reduction in CO_2 loading at high temperatures. The decline of adsorption at higher temperatures is attributed to the characteristics of the exothermic mass transfer mechanism involved in chemisorption. In addition, it should be noted that the liquid phase viscosity is higher at lower temperatures, resulting in a lower diffusion coefficient for CO_2 and a drop in the CO_2 solubility [41, 62]. Hence, the GP model effectively identified the pattern of gas absorption and provided correct predictions of the CO_2 loading in an aqueous PZ solution under different pressure and temperature conditions.

Subsequently, an examination was conducted to assess the influence of PZ concentration on the solubility of CO_2 in aqueous PZ solutions. This investigation was carried out at a constant temperature of 318 K, as previously measured in the literature [44]. The GP model forecast and experimental data are shown in Fig. 12. The concentration of PZ has an inverse relationship with the CO_2 loading at constant pressure and temperature. The concentration of free amines has a significant role in the mass transfer of CO_2. The viscosity of the liquid phase also increases at higher PZ concentrations. Consequently, there is a modest reduction in the CO_2 diffusion coefficient, resulting in a decline in the solubility of CO_2 [41].

When the CO_2 partial pressure is increased in a less concentrated solution, the solubility of the gas grows. The loading capacity of a more potent PZ solution is limited due to the little presence of physically absorbed CO_2 compared to chemically absorbed CO_2 [44]. Hence, these figures demonstrate the exceptional capabilities of the GP model in estimating the solubility of CO_2 in PZ solutions across various concentrations and pressures.

The CO_2 solubility is decreased with increasing temperature as a result of the exothermic absorption reaction. Applying Le Chatelier's principle, at higher temperatures, the system would favor the release of CO2, hence reducing the solubility [63]. At higher PZ concentrations, even though the CO_2 absorption enhancement initially takes place because of more reactive sites available, increasing viscosity at high concentrations may reduce mass transfer, thus cutting down overall

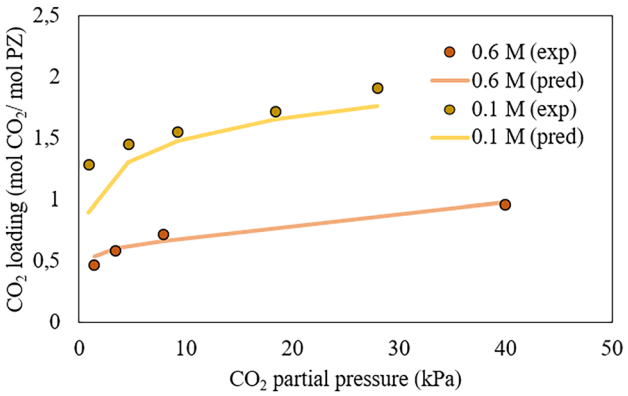

Fig. 12 The effect of PZ concentration on CO_2 solubility in aqueous PZ solutions; experimental data [44] and GP model predictions

absorption efficiency. These kinetic and thermodynamic interactions provide a comprehensive explanation of the trends observed in our experimental results and sensitivity analysis.

11.5 Sensitivity analysis

This study aimed to assess the influence of three variables, namely PZ content, temperature, and CO_2 partial pressure, on the solubility of CO_2 in aqueous PZ solutions. Hence, the Pearson correlation coefficient was applied to calculate the impact of the input parameters, as shown afterwards in Eq. (8) [64, 65] as:

$$r(z_i, y) = \frac{\sum_{j=1}^{n} (z_{i,j} - z_{a,i})(y_j - y_a)}{\left(\sum_{j=1}^{n} (z_{i,j} - z_{a,i})^2 \sum_{j=1}^{n} (y_j - y_a)^2\right)^{0.5}} \tag{8}$$

where $z_{i,j}$ and $z_{a,i}$ stand for the j-th and average values of the i-th input parameter, accordingly. In addition, z represents PZ concentration, temperature, and CO_2 partial pressure. Furthermore, y_a and y_j represent the average and the j-th values of calculated CO_2 solubility in aqueous PZ solutions. The relevance factor (r) varies from − 1 to 1. The greater the absolute value of a parameter, the more dominant its influence is on the model's outcome [66]. The sign of r shows the enhancing or reducing effect of the parameter [67]. The r values for all inputs estimated utilizing the GP model, which is the most trustworthy approach, are depicted in Fig. 13. CO_2 partial pressure shows the most significant impact on the absorption of CO_2 in aqueous PZ solutions. CO_2 partial pressure has a positive and escalating impact, while temperature and PZ concentration have diminishing and negative effects on CO_2 absorption. It is worth mentioning that these findings align with the most recent studies [32, 33].

In fact, the above mentioned results obtained from a sensitivity analysis have implications not only at a theoretical level; for example, the negative correlation between the temperature and the CO_2 solubility would also imply that controlling or mitigating a temperature increase is one of the potential operational strategies to enhance effectiveness in CO_2 capture. Moreover, the effect of PZ concentration on solubility also reflects the requirement for optimal solvent utilization with the aim of balancing a maximum absorption and operational costs.

11.6 Leverage approach

In this study, the leverage approach [68–70] was used to assess the validity range of the proposed GP model and identify any questionable data points. The disparities between the estimates produced by the model and the empirical data are referred to as standardized residuals (SR). If MSE is the mean square of error, H_i is the i-th Leverage value, and e_i is the error value, SR values could be depicted as in Eq. (9) [71, 72]:

$$SR_i = \frac{e_i}{\left[MSE\left(1 - H_i\right)\right]^2} \tag{9}$$

The Hat matrix incorporates standardized residuals. Moreover, hat indexes refer to the elements located on the major diagonal of the Hat matrix. Let T represent the transpose matrix of X, which is a matrix of dimensions ($k \times l$), where k denotes the number of rows (data points) and l denotes the number of columns (input parameters). Consequently, the Hat indexes are generated in accordance with the Hat matrix, which, as shown in Eq. (10), is [71]:

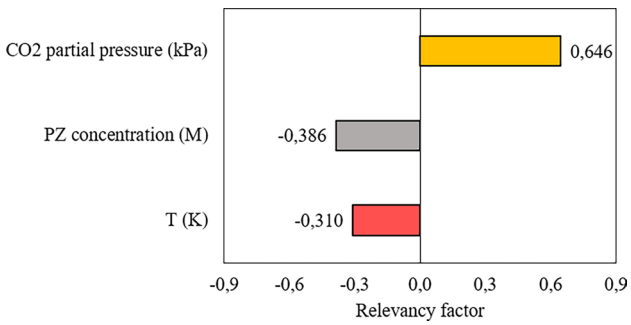

Fig. 13 The relative impacts of input parameters on the CO_2 solubility in aqueous PZ solutions as the GP model output

$$H = X(X^T X)^{-1} X^T \tag{10}$$

Moreover, critical leverage (H^*) represents a fixed value for a certain data bank and could be estimated as in Eq. (11) [70, 73]:

$$H^* = \frac{3 \times (l+1)}{k} \tag{11}$$

The plot by Williams is often used as a visual tool to depict the degree of applicability of a model and the presence of uncertain data within the data bank, as shown in Fig. 14 for the GP model. High leverage points exhibiting SR values exceeding 3 or falling below – 3, irrespective of their Hat values, are considered unfavorable. According to the data displayed in Fig. 13, 8 data points, accounting for about 1.5% of the dataset, were classified as suspicious. These data points were recognized as possible laboratory errors. Furthermore, the data points with SR between – 3 and 3 and a Hat value beyond 0.0232 are acceptable high-leverage points. According to Williams's plot, 20 data points have been recognized as potential outliers. This indicates that, given the model's good prediction, these particular data points fall beyond the realm of application and deviate significantly from most of the dataset. In summary, the experimental data collection used for the purpose of modeling, as well as the estimates generated by the model, demonstrated statistical acceptability and validity.

12 Conclusions

This study presented a comprehensive evaluation of aqueous PZ solutions' performance in CO_2 capture using white-box algorithms, namely, GP, GEP, and GMDH. The developed models demonstrated high accuracy, as indicated by the R^2 values of 0.933 for GP, 0.949 for GEP, and 0.889 for GMDH. Also, RMSE values further validated the models' predictive capability, with RMSE of 0.098 for GEP, 0.110 for GP, and 0.228 for GMDH. Notably, the models successfully captured the intricate relationships between input parameters and CO_2 capture performance. The sensitive analysis revealed that CO_2 partial pressure had a positive impact with a sensitivity coefficient of 0.646, indicating its significance in enhancing CO_2 absorption efficiency. On the other hand, PZ concentration and temperature showed negative impacts with sensitivity coefficients of – 0.386 and – 0.310, respectively. These findings underscore the importance of controlling these parameters to optimize CO_2 capture processes using PZ solutions. Additionally, the leverage method was applied to assess the reliability of the experimental and predicted data. This analysis identified 18 data points as probable outliers, suggesting that these data points fell beyond the applicability scope of the models and exhibited statistical differences from most of the dataset. However, the models demonstrated accurate estimation capabilities for most of the data, ensuring the robustness of our approach. In conclusion, this research provides valuable insights into optimizing CO_2 capture processes using PZ solutions. The effective use of white-box algorithms and the sensitivity analysis of input parameters offer a comprehensive understanding of the factors influencing CO_2 absorption efficiency. By enhancing the accuracy of predictions and accounting for outliers through the leverage method, this study contributes to the advancement of

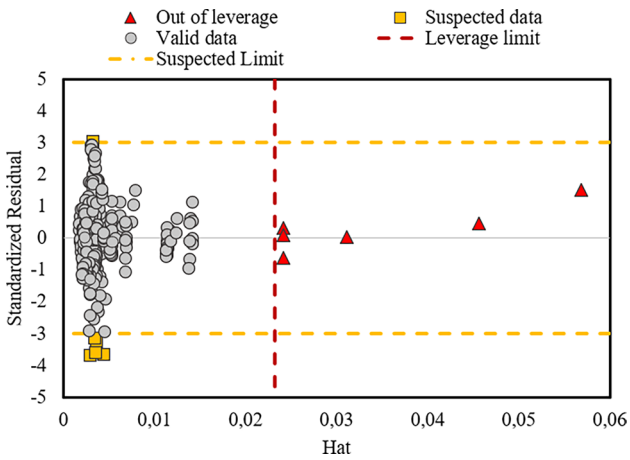

Fig. 14 The Williams plot of the entire data bank for the GP model

efficient and sustainable CO_2 capture technologies, supporting global efforts to mitigate CO_2 emissions and combat the impact of climate change.

Furthermore, despite the trustworthiness of the suggested models, this study has several limitations, including the accuracy lower than of the black box counterparts, a complicated mathematical equation as well as a dataset which is large enough though, but does not cover all of the operational conditions possible. Nevertheless, there are still potential directions of future research which may concern the broader datasets for universal model training as well as the development of frameworks simplifying the calculation process from pure mathematics to automated web-based prediction approach.

Author contributions F.H, J,W, and P.P collected data and prepared the model. F.H and A.R verified the model and prepared the discussion section. F.H. and A.H. supervised the research. All authors reviewed the manuscript.

Funding The authors have not disclosed any funding. There is no specific funding for this project.

Data availability The datasets used during the current study available from the corresponding author on reasonable request.

Declarations

Competing interests The authors declare no competing interests.

References

1. Deja J, Uliasz-Bochenczyk A, Mokrzycki E. CO2 emissions from Polish cement industry. Int J Greenhouse Gas Control. 2010;4(4):583–8.
2. Wen Z, Li H. Analysis of potential energy conservation and CO2 emissions reduction in China's non-ferrous metals industry from a technology perspective. Int J Greenhouse Gas Control. 2014;28:45–56.
3. Friedlingstein P, et al. Global carbon budget 2022. Earth Syst Sci Data Discuss. 2022;2022:1–159.
4. Saghafi H, Ghiasi MM, Mohammadi AH. CO2 capture with aqueous solution of sodium glycinate: modeling using an ensemble method. Int J Greenhouse Gas Control. 2017;62:23–30.
5. Datta AK, Sen PK. Optimization of membrane unit for removing carbon dioxide from natural gas. J Membr Sci. 2006;283(1–2):291–300.
6. Ghiasi MM, Mohammadi AH. Rigorous modeling of CO2 equilibrium absorption in MEA, DEA, and TEA aqueous solutions. J Nat Gas Sci Eng. 2014;18:39–46.
7. Saghafi H, Ghiasi MM, Mohammadi AH. Analyzing the experimental data of CO2 equilibrium absorption in the aqueous solution of DEA+ MDEA with random forest and leverage method. Int J Greenhouse Gas Control. 2017;63:329–37.
8. Adeosun A, Abu-Zahra MR. Evaluation of amine-blend solvent systems for CO2 post-combustion capture applications. Energy procedia. 2013;37:211–8.
9. Choi W-J, Seo J-B, Jang S-Y, Jung J-H, Oh K-J. Removal characteristics of CO2 using aqueous MEA/AMP solutions in the absorption and regeneration process. J Environ Sci. 2009;21(7):907–13.
10. Muchan P, Saiwan C, Narku-Tetteh J, Idem R, Supap T, Tontiwachwuthikul P. Screening tests of aqueous alkanolamine solutions based on primary, secondary, and tertiary structure for blended aqueous amine solution selection in post combustion CO2 capture. Chem Eng Sci. 2017;170:574–82.
11. Narku-Tetteh J, Muchan P, Saiwan C, Supap T, Idem R. Selection of components for formulation of amine blends for post combustion CO2 capture based on the side chain structure of primary, secondary and tertiary amines. Chem Eng Sci. 2017;170:542–60.
12. Gabrielsen J, Michelsen ML, Stenby EH, Kontogeorgis GM. A model for estimating CO2 solubility in aqueous alkanolamines. Ind Eng Chem Res. 2005;44(9):3348–54.
13. Chung P-Y, Soriano AN, Leron RB, Li M-H. Equilibrium solubility of carbon dioxide in the amine solvent system of (triethanolamine+ piperazine+ water). J Chem Thermodyn. 2010;42(6):802–7.
14. Ma'mun S, Nilsen R, Svendsen HF, Juliussen O. Solubility of carbon dioxide in 30 mass% monoethanolamine and 50 mass% methyldiethanolamine solutions,". J Chem Eng Data. 2005;50(2):630–4.
15. Park SH, Lee KB, Hyun JC, Kim SH. Correlation and prediction of the solubility of carbon dioxide in aqueous alkanolamine and mixed alkanolamine solutions. Ind Eng Chem Res. 2002;41(6):1658–65.

16. Porcheron F, Gibert A, Mougin P, Wender A. High throughput screening of CO2 solubility in aqueous monoamine solutions. Environ Sci Technol. 2011;45(6):2486–92.
17. Rochelle GT. Amine scrubbing for CO2 capture. Science. 2009;325(5948):1652–4.
18. Olajire AA. CO2 capture and separation technologies for end-of-pipe applications–a review. Energy. 2010;35(6):2610–28.
19. Ghiasi MM, Arabloo M, Mohammadi AH, Barghi T. Application of ANFIS soft computing technique in modeling the CO2 capture with MEA, DEA, and TEA aqueous solutions. Int J Greenhouse Gas Control. 2016;49:47–54.
20. Daneshvar N, Moattar MZ, Abdi MA, Aber S. Carbon dioxide equilibrium absorption in the multi-component systems of CO2+ TIPA+ MEA+ H2O, CO2+ TIPA+ Pz+ H2O and CO2+ TIPA+ H2O at low CO2 partial pressures: experimental solubility data, corrosion study and modeling with artificial neural network. Sep Purif Technol. 2004;37(2):135–47.
21. Shahsavand A, Fard FD, Sotoudeh F. Application of artificial neural networks for simulation of experimental CO2 absorption data in a packed column. J Nat Gas Sci Eng. 2011;3(3):518–29.
22. Dashti A, Raji M, Alivand MS, Mohammadi AH. Estimation of CO2 equilibrium absorption in aqueous solutions of commonly used amines using different computational schemes. Fuel. 2020;264: 116616.
23. Salooki MK, Abedini R, Adib H, Koolivand H. Design of neural network for manipulating gas refinery sweetening regenerator column outputs. Sep Purif Technol. 2011;82:1–9.
24. Adib H, Sharifi F, Mehranbod N, Kazerooni NM, Koolivand M. Support Vector Machine based modeling of an industrial natural gas sweetening plant. J Nat Gas Sci Eng. 2013;14:121–31.
25. Sipöcz N, Tobiesen FA, Assadi M. The use of artificial neural network models for CO2 capture plants. Appl Energy. 2011;88(7):2368–76.
26. Wu Y, Chan CW. Analysis of data for the carbon dioxide capture domain. Eng Appl Artif Intell. 2011;24(1):154–63.
27. Zhou Q, Chan CW, Tontiwachwuthikul P, Idem R, Gelowitz D. Application of neuro-fuzzy modeling technique for operational problem solving in a CO2 capture process system. Int J Greenhouse Gas Control. 2013;15:32–41.
28. Zhou Q, Wu Y, Chan CW, Tontiwachwuthikul P. From neural network to neuro-fuzzy modeling: applications to the carbon dioxide capture process. Energy Procedia. 2011;4:2066–73.
29. Zhou Q, Wu Y, Chan CW, Tontiwachwuthikul P. Modeling of the carbon dioxide capture process system using machine intelligence approaches. Eng Appl Artif Intell. 2011;24(4):673–85.
30. Tatar A, et al. Comparison of two soft computing approaches for predicting CO2 solubility in aqueous solution of piperazine. Int J Greenhouse Gas Control. 2016;53:85–97.
31. Yarveicy H, Ghiasi MM, Mohammadi AH. Performance evaluation of the machine learning approaches in modeling of CO2 equilibrium absorption in Piperazine aqueous solution. J Mol Liq. 2018;255:375–83.
32. Dashti A, Raji M, Razmi A, Rezaei N, Zendehboudi S, Asghari M. Efficient hybrid modeling of CO2 absorption in aqueous solution of piperazine: applications to energy and environment. Chem Eng Res Des. 2019;144:405–17.
33. Khoshraftar Z, Ghaemi A. Modeling of CO2 solubility in piperazine (PZ) and diethanolamine (DEA) solution via machine learning approach and response surface methodology. Case Stud Chem Environ Eng. 2023;8: 100457.
34. Zafari P, Ghaemi A. Mixed MDEA-PZ amine solutions for CO2 capture: modeling and optimization using RSM and ANN approaches. Case Stud Chem Environ Eng. 2023;8: 100509.
35. Zafari P, Ghaemi A. Modeling and optimization of CO2 capture into mixed MEA-PZ amine solutions using machine learning based on ANN and RSM models. Results Eng. 2023;19: 101279.
36. Shokri A, Ghaemi A. Developing artificial neural networks and response surface methodology for evaluating CO2 absorption into K2CO3/piperazine solution. Case Stud Chem Environ Eng. 2024;9: 100725.
37. Bishnoi S, Rochelle GT. Absorption of carbon dioxide into aqueous piperazine: reaction kinetics, mass transfer and solubility. Chem Eng Sci. 2000;55(22):5531–43. https://doi.org/10.1016/S0009-2509(00)00182-2.
38. Dash SK, Samanta A, Samanta AN, Bandyopadhyay SS. Vapour liquid equilibria of carbon dioxide in dilute and concentrated aqueous solutions of piperazine at low to high pressure. Fluid Phase Equilib. 2011;300(1):145–54. https://doi.org/10.1016/j.fluid.2010.11.004.
39. Derks P, Dijkstra H, Hogendoorn J, Versteeg G. Solubility of carbon dioxide in aqueous piperazine solutions. AIChE J. 2005;51(8):2311–27.
40. Dugas R, Rochelle G. Absorption and desorption rates of carbon dioxide with monoethanolamine and piperazine. Energy Procedia. 2009;1(1):1163–9. https://doi.org/10.1016/j.egypro.2009.01.153.
41. R. E. Dugas, *Carbon dioxide absorption, desorption, and diffusion in aqueous piperazine and monoethanolamine*. The University of Texas at Austin, 2009.
42. Haghtalab A, Eghbali H, Shojaeian A. Experiment and modeling solubility of CO2 in aqueous solutions of Diisopropanolamine+2-amino-2-methyl-1-propanol+Piperazine at high pressures. J Chem Thermodyn. 2014;71:71–83. https://doi.org/10.1016/j.jct.2013.11.025.
43. Kadiwala S, Rayer AV, Henni A. High pressure solubility of carbon dioxide (CO2) in aqueous piperazine solutions. Fluid Phase Equilib. 2010;292(1):20–8. https://doi.org/10.1016/j.fluid.2010.01.009.
44. Aroua MK, Mohd Salleh R. Solubility of CO2 in aqueous piperazine and its modeling using the Kent-Eisenberg approach. Chem Eng Technol: Ind Chem-Plant Equip-Process Eng-Biotechnol. 2004;27(1):65–70.
45. N. L. Cramer, "A representation for the adaptive generation of simple sequential programs," in *proceedings of the first international conference on genetic algorithms and their applications*, 2014: Psychology Press, pp 183–187
46. Koza JR. Genetic programming II: automatic discovery of reusable programs. Cambridge: MIT press; 1994.
47. Hadavimoghaddam F, Mohammadi M-R, Atashrouz S, Bostani A, Hemmati-Sarapardeh A, Mohaddespour A. Modeling hydrogen solubility in alcohols using group method of data handling and genetic programming. Int J Hydrogen Energy. 2023;48(7):2689–704.
48. Ebid AM. 35 Years of (AI) in geotechnical engineering: state of the art. Geotech Geol Eng. 2021;39(2):637–90.
49. Ivakhnenko AG. The group method of data handling A rival of stochastic approximation. Soviet Autom Control. 1968;13:43–55.
50. Nishikawa T, Shimizu S. Identification and forecasting in management systems using the GMDH method. Appl Math Model. 1982;6(1):7–15.
51. Youcefi MR, Hadjadj A, Boukredera FS. New model for standpipe pressure prediction while drilling using Group Method of Data Handling. Petroleum. 2022;8(2):210–8.
52. Farlow SJ. Self-organizing methods in modeling: GMDH type algorithms. Boca Raton: CrC Press; 2020.

53. Ghazanfari N, Gholami S, Emad A, Shekarchi M. Evaluation of GMDH and MLP networks for prediction of compressive strength and workability of concrete. Bull Soc Roy Sci Liège. 2017;86:855–68.
54. C. Ferreira, Gene expression programming: a new adaptive algorithm for solving problems, *arXiv preprint cs/0102027,* 2001
55. B. Z. Laskar and S. Majumder, "Gene expression programming," in *Bio-Inspired Computing for Information Retrieval Applications*: IGI Global, 2017, pp 269–292
56. Lawal AI, Kwon S, Hammed OS, Idris MA. Blast-induced ground vibration prediction in granite quarries: an application of gene expression programming, ANFIS, and sine cosine algorithm optimized ANN. Int J Min Sci Technol. 2021;31(2):265–77.
57. Onifade M, et al. Development of multiple soft computing models for estimating organic and inorganic constituents in coal. Int J Min Sci Technol. 2021;31(3):483–94.
58. Althoey F, et al. Machine learning based computational approach for crack width detection of self-healing concrete. Case Stud Constr Mater. 2022;17: e01610.
59. Mohammadi M-R, Hadavimoghaddam F, Atashrouz S, Abedi A, Hemmati-Sarapardeh A, Mohaddespour A. Modeling the solubility of light hydrocarbon gases and their mixture in brine with machine learning and equations of state. Sci Rep. 2022;12(1):14943. https://doi.org/10.1038/s41598-022-18983-2.
60. Kamps ÁPS, Xia J, Maurer G. Solubility of CO2 in (H2O+ piperazine) and in (H2O+ MDEA+ piperazine). AIChE J. 2003;49(10):2662–70.
61. Ermatchkov V, Pérez-Salado Kamps Á, Speyer D, Maurer G. Solubility of carbon dioxide in aqueous solutions of piperazine in the low gas loading region. J Chem Eng Data. 2006;51(5):1788–96.
62. Jahangiri A, Nabipoor Hassankiadeh M. Effects of piperazine concentration and operating conditions on the solubility of CO2 in AMP solution at low CO2 partial pressure. Sep Sci Technol. 2019;54(6):1067–78.
63. Lin W, Murphy CJ. A demonstration of Le Chatelier's principle on the nanoscale. ACS Cent Sci. 2017;3(10):1096–102.
64. Mohammadi M-R, Hemmati-Sarapardeh A, Schaffie M, Husein MM, Ranjbar M. Application of cascade forward neural network and group method of data handling to modeling crude oil pyrolysis during thermal enhanced oil recovery. J Petrol Sci Eng. 2021;205: 108836.
65. Chen G, et al. The genetic algorithm based back propagation neural network for MMP prediction in CO2-EOR process. Fuel. 2014;126:202–12.
66. Ansari S, et al. Experimental measurement and modeling of asphaltene adsorption onto iron oxide and lime nanoparticles in the presence and absence of water. Sci Rep. 2023;13(1):122.
67. Mohammadi M-R, Hemmati-Sarapardeh A, Schaffie M, Husein MM, Karimian M, Ranjbar M. On the evaluation of crude oil oxidation during thermogravimetry by generalised regression neural network and gene expression programming: application to thermal enhanced oil recovery. Combust Theor Model. 2021;25(7):1268–95.
68. A. M. Leroy and P. J. Rousseeuw, "Robust regression and outlier detection," *rrod,* 1987.
69. C. R. Goodall, "13 Computation using the QR decomposition," 1993.
70. Gramatica P. Principles of QSAR models validation: internal and external. QSAR Comb Sci. 2007;26(5):694–701.
71. Rousseeuw PJ, Leroy AM. Robust regression and outlier detection. Amsterdam: John wiley & sons; 2005.
72. Hadavimoghaddam F, Mohammadi M-R, Atashrouz S, Nedeljkovic D, Hemmati-Sarapardeh A, Mohaddespour A. Data-driven modeling of H2 solubility in hydrocarbons using white-box approaches. Int J Hydrogen Energy. 2022;47(78):33224–38.
73. Ansari S, Safaei-Farouji M, Atashrouz S, Abedi A, Hemmati-Sarapardeh A, Mohaddespour A. Prediction of hydrogen solubility in aqueous solutions: comparison of equations of state and advanced machine learning-metaheuristic approaches. Int J Hydrogen Energy. 2022;47(89):37724–41.