scientific reports

OPEN

Predictive modeling of CO_2 solubility in piperazine aqueous solutions using boosting algorithms for carbon capture goals

Mohammad-Reza Mohammadi[1]✉, Aydin Larestani[1], Mahin Schaffie[1], Abdolhossein Hemmati-Sarapardeh[1,2]✉ & Mohammad Ranjbar[1,3]

Carbon dioxide (CO_2) is the main greenhouse gas that drives global warming, climate change, and other environmental issues. CO_2 absorption using amine solvents stands out as one of the most well-known industrial technologies of CO_2 capture. However, accurate prediction of CO_2 absorption in aqueous amine solutions under different operating conditions is crucial for designing an efficient amine scrubbing system in power plants. In this work, CO_2 solubility in aqueous piperazine (PZ) solutions was modeled using 517 experimental data points covering a temperature range of 298 to 373 K, PZ concentration of 0.1 to 6.2 mol/L (M), and CO_2 partial pressure of 0.03 to 7399 kPa. To this end, four robust machine learning algorithms, including gradient boosting with categorical features support (CatBoost), light gradient boosting machine (LightGBM), extreme gradient boosting (XGBoost), and adaptive boosting decision trees (AdaBoost-DT) were utilized. Among the developed models, the CatBoost model presented the highest accuracy with an overall determination coefficient (R^2) of 0.9953 and an average absolute relative error of 2.36%. Sensitivity analysis revealed that CO_2 partial pressure had the greatest influence on CO_2 absorption in aqueous PZ solutions, followed by PZ concentration and temperature. Moreover, CO_2 partial pressure positively influenced CO_2 absorption in aqueous PZ solutions, while PZ concentration and temperature exhibited negative effects. Finally, the leverage technique indicated that both the experimental data bank used for modeling and the model's estimates were statistically acceptable and valid showing only 8 points (~1.5% of total data) as possible suspected data.

Keywords CO_2 capture, Aqueous piperazine solution, Liquid absorption method, Machine learning algorithms, CatBoost, Leverage technique

The deterioration of climate change and global warming problems is not hidden from anyone and one of the major concerns in this regard is the anthropogenic carbon dioxide (CO_2) emissions worldwide[1,2]. CO_2 emissions have progressively augmented in recent years[3], mainly due to fossil fuels and industry. Global warming is to be mitigated to prevent agricultural output reduction and extreme weather patterns[4], while industries must meet the energy demand of additional 900 million people by 2035 [5]. Various technologies have been developed so far for capturing CO_2 from flue gas such as cryogenic distillation, membrane separation, chemical/physical absorption, adsorption, and bioremediation[1,2,6–8], among which the use of chemical absorption is the most attractive option for post-combustion CO_2 capture at room pressure and temperature owing to its low cost and easy implementation[2]. Among various organic and inorganic solvents proposed for CO_2 chemical absorption, aqueous solutions of amines containing reactive nitrogen atoms, which can absorb CO_2 in a reversible and selective process, are the most appealing options. Amine structures significantly impact the CO_2 capture process[9]. Also, they are cheap and have low steam pressures[2]. Amine-based aqueous solutions can be potentially applied to extract CO_2 in power plants[10–12]. Specifically, piperazine (PZ) has shown a great potential to absorb CO_2 with respect to its high absorption capacity (almost twice as monoethanolamine (MEA))[2]. Moreover, the product of

[1]Department of Petroleum Engineering, Shahid Bahonar University of Kerman, Kerman, Iran. [2]State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum (Beijing), Beijing, China. [3]Department of Mining Engineering, Shahid Bahonar University of Kerman, Kerman, Iran. ✉email: mohammadi.mrm@eng.uk.ac.ir; mohammmadi.mrm@gmail.com; hemmati@uk.ac.ir; aut.hemmati@gmail.com

its reaction with CO_2 includes PZ carbamate at low loadings and PZ bicarbamate within the concentrated PZ process operational spectrum, thereby enhancing the overall rate of CO_2 absorption under varying operational parameters[13,14].

The CO_2 equilibrium absorption capacity or CO_2 solubility in the amine-based solution is the most significant parameter that directly affects the performance of the solvent in the CO_2 absorption process[15]. This crucial parameter was traditionally obtained through various experimental tests or several thermodynamic models, which are developed on the basis of vapor-liquid equilibrium (VLE) theory[16–19]. Although experimental procedures are reliable methods for evaluating CO_2 loading in amine solutions, they are costly and time-consuming. In contrast, thermodynamic-based methodologies are not capable of assessing CO_2 solubility in broad ranges of operational circumstances[20]. By contrast, recent advancements in computer science have led to the development of powerful and robust machine learning approaches that have been extensively employed in energy and environmental sciences[21–28]. They have also proven their competency in modeling the CO_2 capture process using alkanolamine solvents[29]. Salooki et al.[30] attempted to predict the output variables of a stripper operating in one of the Iranian gas refineries using artificial neural networks (ANN). The output temperature and flow rate of this stripper were also modeled by the support vector machine (SVM) framework[31]. The process of steady-state CO_2 capture in monoethanolamine (MEA) aqueous solution was also modeled using ANN[32] and optimized through statistical methods[33]. More analogous research works related to the CO_2 capture can be found in the literature[34–38].

The CO_2 loading in amine-based aqueous solutions was another point of interest among researchers. Ghiasi and Mohammadi[39] developed a least-squares support vector machine (LSSVM) to estimate CO_2 solubility in a variety of amine solutions with respect to the concentration of amine, temperature, and CO_2 partial pressure. A similar study was then conducted by using an adaptive neuro-fuzzy inference system (ANFIS)[40]. Daneshvar et al.[41] implemented an ANN algorithm to estimate CO_2 loading in triisopropanolamine (TIPA), TIPA/PZ, and TIPA/MEA solvents. In another study, the absorption capacity of CO_2 in diethanolamine (DEA) and methyldiethanolamine (MDEA) was estimated using radial basis function and multilayer perceptron networks[42]. More recently, Dashti et al.[20] implemented four intelligent approaches to forecast CO_2 solubility in twelve amine-based solvents. They concluded that the LSSVM model optimized by coupled simulated annealing (CSA) optimization technique could provide the most reliable results in comparison to the other models.

Given the potential of PZ aqueous solutions in capturing CO_2, many researchers have focused on proposing reliable approaches for accurate estimation of the CO_2 absorption capacity of PZ solvents. Tatar et al.[43] proposed two intelligent approaches, namely CSA-LSSVM and ANFIS coupled with Conjugate Hybrid-Particle Swarm Optimization (CHPSO-ANFIS) to predict CO_2 solubility in PZ solutions and reported the superiority of CHPSO-ANFIS model. A similar study was conducted by Yarveicy et al.[44] using four intelligent approaches including LSSVM, ANFIS, ANN, and adaptive boosting-classification and regression tree. Dashti et al.[45] developed genetic programming (GP) and GA-ANFIS models to predict CO_2 solubility in aqueous solutions of PZ using CO_2 partial pressure, PZ concentration, and temperature as input variables. Their models were developed using a databank gathered from the literature consisting of 390 data points. They reported average absolute relative deviations (AARDs) of 9.7% and 5.3% for the developed GA-ANFIS and GP models, respectively. To the best of our knowledge, this database represents the most extensive collection utilized for developing predictive models of CO_2 loading in aqueous PZ solutions. Furthermore, a thorough literature review indicates that existing models for CO_2 solubility in PZ solvents employ outdated algorithms, highlighting the necessity to enhance their applicability across broader operational conditions and to develop novel intelligent approaches using cutting-edge algorithms for estimating CO_2 loading in PZ aqueous solutions based on an expanded database.

In this work, an extended databank comprising 517 data points gathered from open-source literature is utilized to develop several novel intelligent approaches for estimating CO_2 loading in PZ aqueous solutions. To achieve this goal, four robust machine learning algorithms including, gradient boosting with categorical features support (CatBoost), light gradient boosting machine (LightGBM), extreme gradient boosting (XGBoost), and adaptive boosting decision trees (AdaBoost-DT) are utilized. Then, the performance of the models is evaluated by employing a variety of statistical and graphical assessments. Furthermore, additional trend analyses are conducted to assess the validity of the best-developed model. Also, sensitivity analysis is performed to examine the relationships between inputs and the outcomes of the model. Finally, the Leverage technique is employed to evaluate the credibility and application range of the best-predictive model.

Data gathering

In this work, 517 experimental findings related to the absorption of CO_2 into aqueous PZ solutions were gathered from the literature[46–53]. This data bank has more than 120 data points more than what was used in the studies of Dashti et al.[20,45]. Three independent variables, namely temperature (K), PZ concentration (M), and CO_2 partial pressure (kPa) were considered as inputs to the models, while CO_2 loading (mol CO_2 / mol PZ) is the output. Table 1 reports the statistical description of the data bank used for modeling in this work. As is evident, the solubility of CO_2 in aqueous PZ solutions was modeled using a wide range of influencing parameters including PZ molarities up to 6.2 M, temperatures between 298 and 373 K, and pressures up to about 7400 kPa. A snapshot of the CO_2 solubility changes with the three input parameters was displayed in the 2D contour plots of Fig. 1. A quick glance at the contour plots shows that the higher CO_2 solubility in PZ solutions corresponds to the more elevated CO_2 partial pressures, lower temperatures, and lower PZ concentrations.

Figure 2 shows the correlation matrix between all variables in the gathered data bank in this work. The correlation coefficients shown in the matrix can specify the relationship between two variables, where an absolute value close to 1 is deemed a strong relationship and 0 is neutral. Also, positive and negative values demonstrate direct and inverse relationships between the two variables, respectively[54]. The following formula was used to compute the linear correlation coefficient between two variables[55]:

	Temperature (K)	PZ concentration (M)	CO_2 partial pressure (kPa)	CO_2 loading (mol CO_2 / mol PZ)
Mean	317.61	1.34	443.77	1.01
Median	313.15	0.80	28.01	0.95
Mode	328.00	0.20	0.92	0.98
SD	15.55	1.44	995.23	0.44
Kurtosis	1.14	1.37	17.44	2.48
Skewness	0.96	1.52	3.88	1.29
Minimum	298.00	0.10	0.03	0.16
Maximum	373.15	6.20	7399.00	2.96
Count	517	517	517	517
Variable status	Input	Input	Input	Target

Table 1. Statistical description of the input and target parameters.

$$r\left(x,y\right)=\frac{\sum_{i=1}^{n}(x_i-x_a)(y_i-y_a)}{\left(\sum_{i=1}^{n}\left(x_i-x_a\right)^2\sum_{i=1}^{n}\left(y_i-y_a\right)^2\right)^{0.5}} \tag{1}$$

where, x_i and y_i show the values of the x-variable and y-variable in two sets of data, respectively. Also, x_a and y_a stand for the average of the x-variable and the average of y-variable in the mentioned data sets, respectively.

Based on Fig. 2, CO_2 partial pressure has a direct relationship with CO_2 loading, and on the other hand, PZ concentration and temperature have an inverse one with it. It is important to remember that correlation coefficients measure the strength and direction of a linear relationship but do not imply causation. For example, a correlation coefficient of 0.33 between PZ concentration and temperature suggests a weak positive relationship between these two variables. However, the presence of a correlation does not mean that changes in PZ concentration cause changes in temperature or vice versa. The correlation coefficient simply indicates the degree to which the two variables move together in a linear fashion. This analysis only provides an overview of the correlation coefficient matrix for the data collected in this research, focusing on the linear relationships between inputs (temperature, PZ concentration, and CO_2 partial pressure) and the target variable (CO_2 loading). No general conclusions about causation or trends are drawn at this stage. Further analysis, including trend analysis and other statistical methods, will be presented in the continuation of the manuscript to provide more comprehensive insights and conclusions.

Model development

In this study, four powerful tree-based machine learning algorithms are implemented to predict the CO_2 solubility in PZ aqueous solutions accurately considering CO_2 partial pressure, temperature, and PZ concentration using a databank comprised of 517 data points. The theoretical concepts behind these intelligent models are described in what follows.

Extreme gradient boosting (XGBoost)

This algorithm is proposed as a supervised machine learning approach on the basis of the tree-boosting method and is capable of solving regression tasks as well as ranking and classification problems[56,57]. XGBoost operates based on the Newton-Raphson method. Analogous to the structure of a decision tree (DT), XGBoost consists different types of node[58]. In the initial step of model training, the entire databank is divided into *k* datasets and then they form two distinct internal nodes followed by leaf nodes after the last classification[59,60]. When the model structure completes, the model outputs will be calculated as follows:

$$\widehat{y}_i=\sum_{k=1}^{N}f_k\left(X_i\right),f_k\,\mathrm{inf} \tag{2}$$

and

$$f=\left\{f\left(X\right)=\omega_{h(x)}\right\},\left(h:m\rightarrow T,\omega\in T\right) \tag{3}$$

where *h(x)* is determined by mapping example *X* and denotes binary leaf index, *f* represents the regression tree's space, *T* stands for the leaves of the tree, f_k exhibits the *k*th tree and ω means the weight of the tree[59]. Afterwards, the objective function (*L*) is to be iteratively minimized for each leaf[59]:

$$L=\sum_{i=1}^{n}l\left(\widehat{y}_i,y_i\right)+\sum_{k=1}^{N}\left(f_k\right) \tag{4}$$

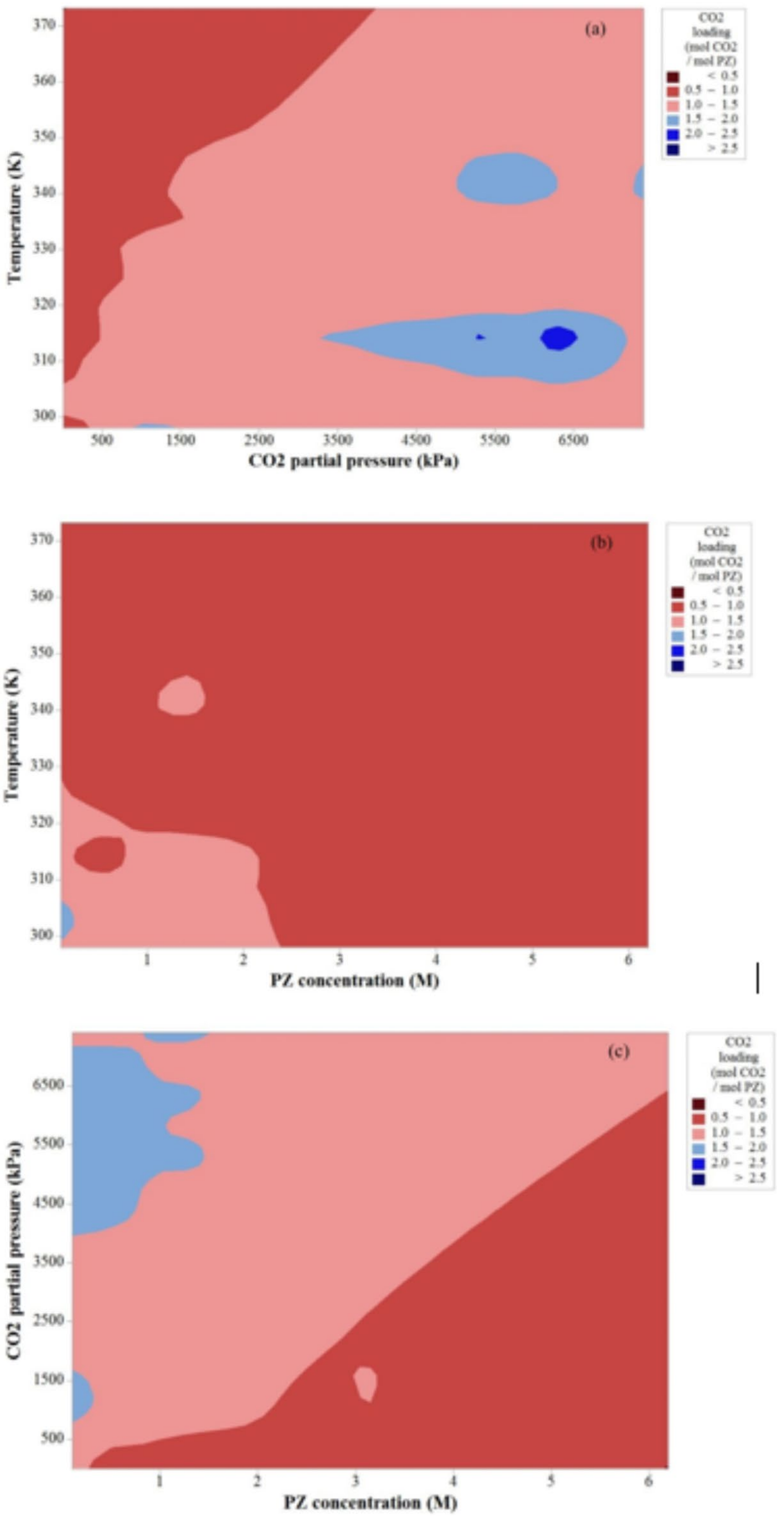

Fig. 1. 2D contour plots of changes in CO_2 solubility in aqueous PZ solutions with the inputs; (**a**) Temperature and CO_2 partial pressure; (**b**) Temperature and PZ concentration; (**c**) CO_2 partial pressure and PZ concentration.

where the regularization and loss functions are respectively denoted by Ω and l, λ signifies the regulation coefficient, and γ shows the minimum loss. The model uses parameters γ and λ to control its variance and avoid overfitting. Figure 3 represents a representation of the XGBoost algorithm.

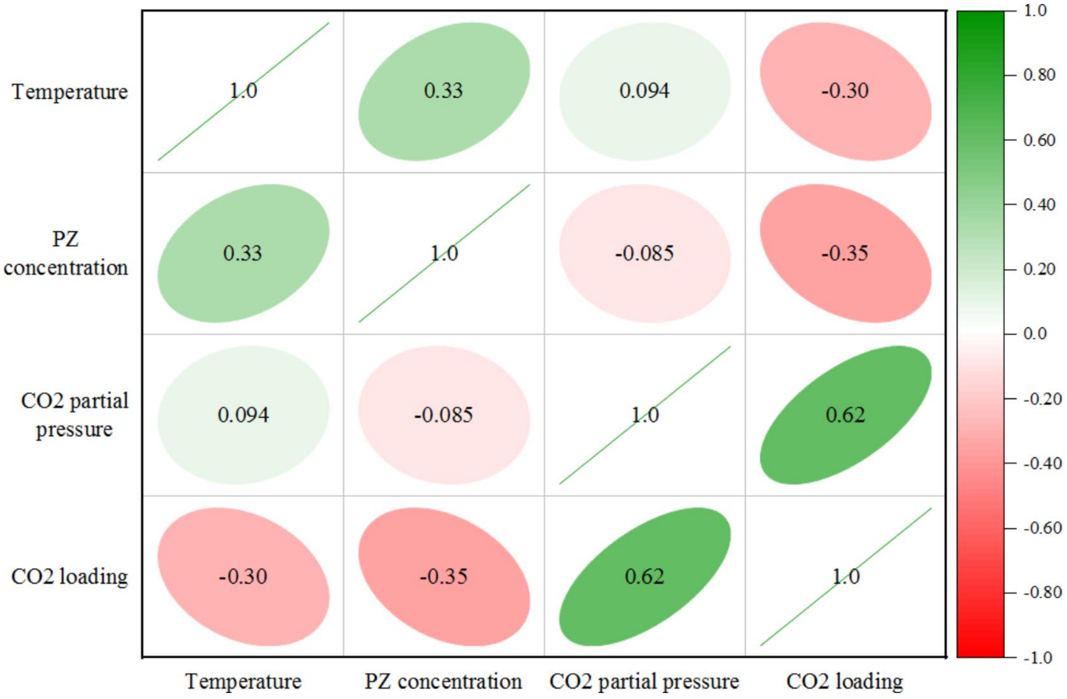

Fig. 2. The correlation coefficient matrix for the data bank gathered in this research.

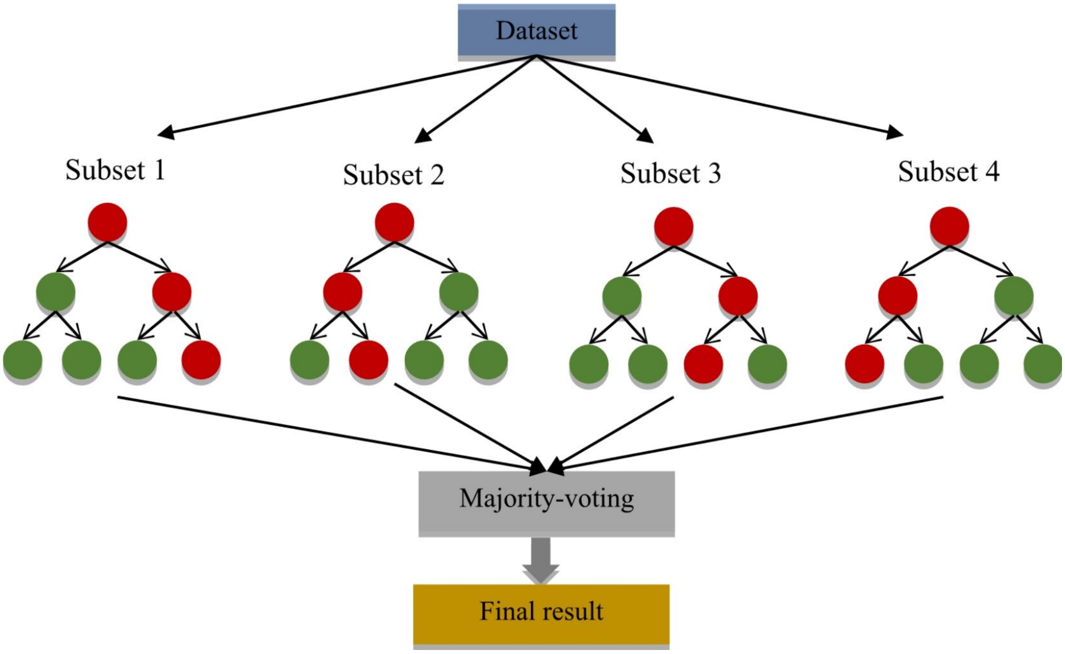

Fig. 3. An illustration of XGBoost model.

Light gradient boosting machine (LightGBM)

Alike XGBoost, LightGBM is applicable in a variety of machine learning tasks as another tree-based learning model[59,61]. LightGBM applies a histogram by splitting eigenvalues into '*P*' distinct bins so as to reduce memory consumption and speed up the model's development steps[59]. This algorithm reduces memory consumption even more by keeping values in an eight-bit integer[62]. LightGBM is trained through a leaf-wise process which is more effective than the traditional level-wise method[63,64]. It is also possible to minimize the error by recognizing the leaves with the maximum branching gain. However, this process makes a deeper and more complex model that is more prone to overfitting, which should be prevented by defining an upper limit on the depth of the leaf top[59,65]. A schematic of LightGBM is depicted in Fig. 4.

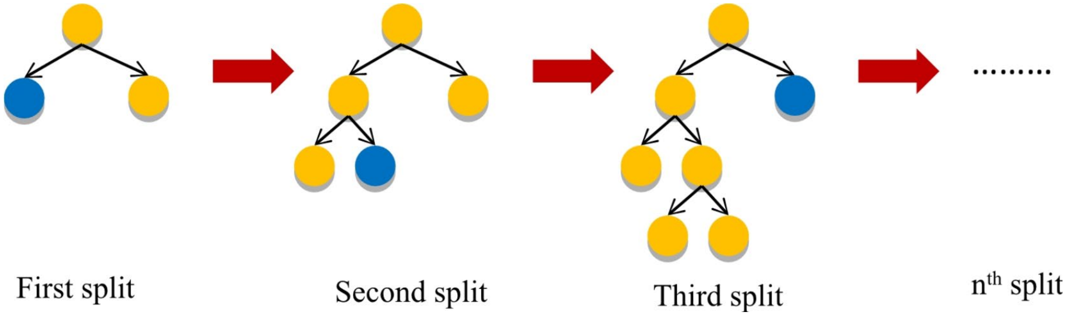

Fig. 4. A schematic image of LightGBM model.

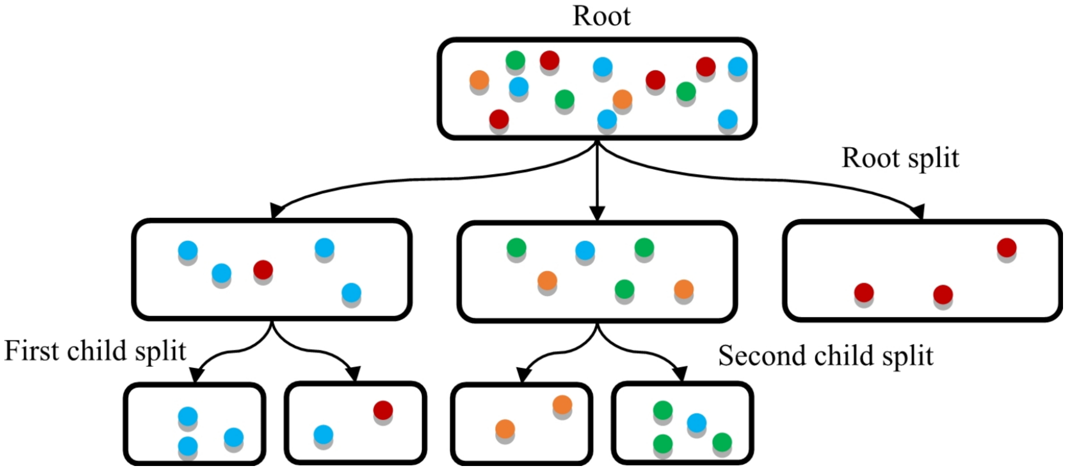

Fig. 5. A schematic illustration of CatBoost method.

Categorical boosting (CatBoost)

As another variation of gradient boosting techniques, CatBoost applies categorical columns to take advantage of target-based statistics and one_hot_max_size (OHMS) features[66–68]. The algorithm employs a greedy method to split a tree and find the exponential evolution of the feature combination[66]. If a feature possesses more category compared to OHMS, the following steps are applied in the algorithm:

1. Making random subsets from the available records.
2. Converting labels into integers.
3. Using the equation below to transform categorical features to numeric ones[66]:

$$avgTarget = \frac{CountInClass + prior}{totalCount + 1}$$

where *CountInClass* and *totalCount* denote the number of targets and the number of preceding objects, respectively, while *prior* is specified by the starting parameters to count objects[66]. CatBoost prevents overfitting through ordered boosting, regularization, and early stopping, ensuring effective handling of categorical features and robust model performance. This algorithm is schematically illustrated in Fig. 5.

Adaptive boosting decision tree (AdaBoost-DT)

AdaBoost was first introduced by Freund and Schapire[69] as a powerful tool that is capable of learning the mistakes of weak learners and executing a strong classifier/regressor. In this algorithm, an initial group of learners is developed based on weighted datasets, and different weights are assigned to each learner with respect to its accuracy[70]. The less accurate learners get higher weights so that new learners will affect them the most. The algorithm typically follows the steps below[71]:

1. Allocating initial weights: $w_j = 1/n, j = 1, 2, \ldots, n$.
2. Developing weak learners based on training data and obtaining weighted errors of each learner.
3. Assigning weights to each learner.

4. Updating the weight of the training samples.
5. Testing the learners with testing data.

In this study, decision trees (DTs) were employed as weak learners. A schematic of the AdaBoost-DT algorithm is shown in Fig. 6.

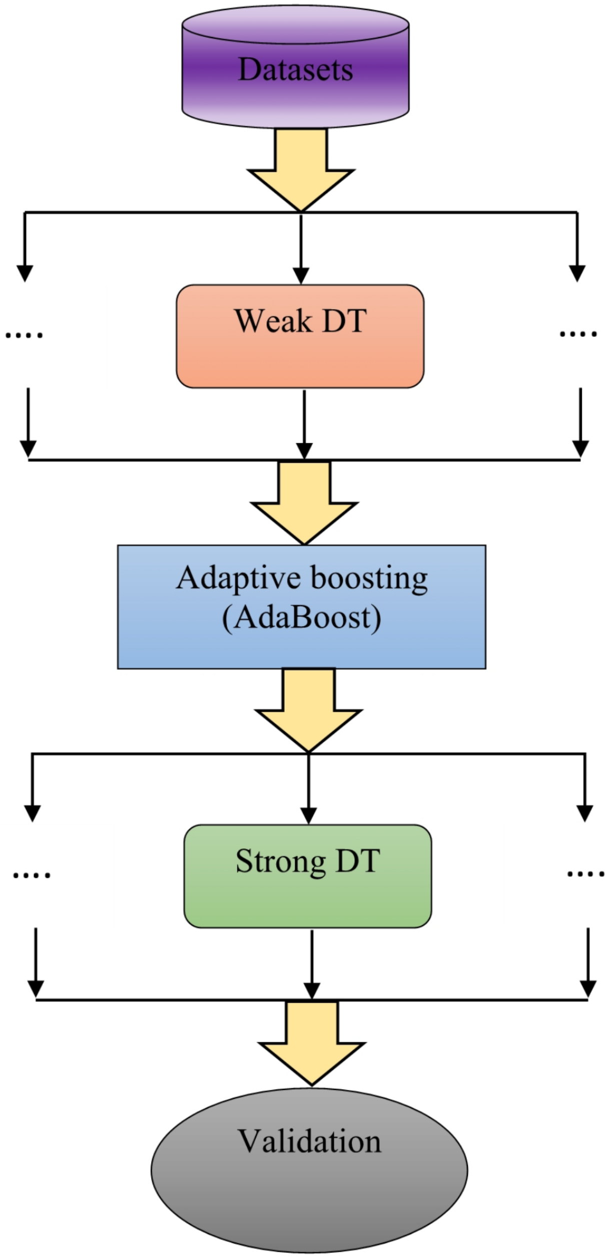

Fig. 6. A schematic diagram of AdaBoost-DT model.

Assessment of models

Using five statistical indicators, namely determination coefficient (R^2), average absolute percent relative error (AAPRE), average percent relative error (APRE), standard deviation (SD), and root mean square error (RMSE), the accuracy of the proposed models was assessed. These statistical criteria are listed below[72]:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(Y_{i,\exp} - Y_{i,pred}\right)^2}{\sum_{i=1}^{N}\left(Y_{i,\exp} - \overline{Y_{\exp}}\right)^2} \tag{5}$$

$$APRE = \frac{100}{N}\sum_{i=1}^{N}\left(\frac{Y_{i,\exp} - Y_{i,pred}}{Y_{i,\exp}}\right) \tag{6}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Y_{i,\exp} - Y_{i,pred}\right)}^{2} \tag{7}$$

$$AAPRE = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{Y_{i,\exp} - Y_{i,pred}}{Y_{i,\exp}}\right| \tag{8}$$

$$SD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{Y_{i,\exp} - Y_{i,pred}}{Y_{i,\exp}}\right)}^{2} \tag{9}$$

In these formulas, $Y_{i,\ exp}$, $Y_{i,\ pred}$, and N show the experimental CO_2 solubility data, the predicted CO_2 solubility data by the proposed models, and the number of data, respectively.

In tandem with statistical analysis, this work incorporates graphical evaluation of model outcomes, outlined succinctly as follows:

Cross-plot: This analysis allows the cross-plotting of two sets of data (experimental and modeling data). The more data concentrated around the unit-slope line, the better the estimates of the model.

Trend plot: In this analysis, the validity of the model is assessed by plotting both experimental and modeling data according to the inputs.

Error distribution graph: In this analysis, the error distribution around the zero error line is assessed to specify the possible error trend of the model. In this graphical analysis, the percent relative error (E_i) values are used, which can be calculated according to the following formula:

$$\mathrm{E_i} = \left[\frac{Y_{i,\exp} - Y_{i,pred}}{Y_{i,\exp}}\right] \times 100\mathrm{i} \ \ = 1, 2, 3, \ldots, n \tag{10}$$

Cumulative frequency plot: This analysis allows checking the accuracy of models by plotting the absolute relative error (E_a), as calculated using the following formula, versus the proportion of the data.

$$\mathrm{E}_a = \left|\frac{Y_{i,\mathrm{e}} - Y_{i,p}}{Y_{i,\mathrm{e}}}\right| \times 100\mathrm{i} \ \ = 1, 2, 3, \ldots, n \tag{11}$$

Results and discussion

Developed models

In this work, CO_2 absorption in aqueous PZ solutions was modeled using robust boosting machine learning algorithms. In this regard, 517 experimental findings were used in the modeling process considering temperature, PZ concentration, and CO_2 partial pressure as input parameters. Here, 80% of the data was utilized for model training, while the remaining 20% served as the test subset. To evaluate model performance and ensure unbiased predictions, a widely used approach in machine learning is the 10-fold cross-validation method used in this work. This technique involves partitioning the dataset into ten equal segments, referred to as "folds." In each iteration, one fold is set aside for validation, while the other nine are used for training the model[73]. This process repeats ten times, with each fold serving as the validation set once. Consequently, the model undergoes nine training phases before each validation, cycling through all folds to provide a comprehensive assessment. A grid search was employed for optimizing the hyperparameters of each model throughout the modeling process. Grid search is a method for optimizing hyperparameters by exhaustively evaluating all possible combinations within a defined range, using cross-validation to assess model performance. It systematically trains and evaluates the model for each combination to identify the best-performing parameters. The search range and tuned values of the principal hyperparameters obtained in the modeling process were reported in Table 2. Hyperparameter tuning plays a crucial role in minimizing prediction errors in machine learning models[74]. In addition to grid search, metaheuristic optimization methods like genetic algorithms, particle swarm optimization, and grey wolf optimization can effectively navigate vast hyperparameter spaces to quickly discover optimal solutions, as demonstrated in the literature[73].

Model	Parameter	Search range	Value
XGBoost	learning rate	0.01–0.9	0.1
	Booster	[gbtree, gblinear, dart]	gbtree
	Max depth	1–14	4
	n_estimator	1–1000	200
	reg_alpha	0–1	0.11
	sub sample	0.1–1	0.15
LightGBM	learning rate	0.01–0.9	0.2
	num leaves	2–20	6
	Max depth	1–14	5
	min data in leaf	[2, 5, 10, 15]	2
CatBoost	learning rate	0.01–0.9	0.036
	Max depth	1–14	4
	loss function	[RMSE, MAE]	RMSE
AdaBoost-DT	learning rate	0.01–0.9	0.29
	Max depth	1–14	6
	min sample split	[2, 5, 10, 20]	2
	min sample leaf	[1, 2, 5, 10]	1
	n_estimator	1–1000	111

Table 2. The search range and tuned hyperparameters for the developed models.

Statistical factor	Status	XGBoost	LightGBM	CatBoost	AdaBoost-DT
APRE (%)	Train	-0.15	-0.16	-0.14	-1.78
	Test	-0.22	-0.05	-0.13	-2.58
	Total	-0.17	-0.14	-0.14	-1.94
AAPRE (%)	Train	3.38	2.53	2.06	3.71
	Test	4.71	3.75	3.54	5.53
	Total	3.64	2.77	2.36	4.08
RMSE	Train	0.047	0.030	0.025	0.039
	Test	0.069	0.051	0.046	0.063
	Total	0.053	0.035	0.030	0.045
SD	Train	0.051	0.033	0.030	0.059
	Test	0.067	0.049	0.048	0.093
	Total	0.055	0.037	0.035	0.067
R^2	Train	0.9885	0.9951	0.9968	0.9931
	Test	0.9759	0.9885	0.9892	0.9709
	Total	0.9858	0.9935	0.9953	0.9901

Table 3. Statistical evaluation of the developed models.

Statistical and graphical evaluation of models

Considering evaluating the accuracy of the proposed models, Table 3 summarizes the values of R^2, RMSE, APRE, AAPRE, and SD. According to statistical principles, the closer the R^2 of a model is to 1 and the lower the values of RMSE, AAPRE, APRE, and SD in the modeling process, the more accurate and valid that model is. As shown in Table 3, the CatBoost model represents AAPRE values of 2.36%, 2.06%, and 3.54% for the total, train, and test collections, which are the lowest error values among the four models developed in this work. Furthermore, this model shows the highest overall R^2 value of 0.9953 along with the lowest values of APRE, RMSE, and SD compared to the remaining three models. Hence, the CatBoost model can be considered the most accurate model developed in this study for predicting CO_2 absorption in aqueous PZ solutions. Summing up the statistical analyses, CatBoost, LightGBM, XGBoost, and AdaBoost-DT models are classified from the best performance to the weakest, respectively.

Moreover, the performance of the suggested models was compared using graphical error analyses. First, Fig. 7 illustrates cross-plots of the predicted data by the developed models versus the experimental data. As is evident, all the boosting models show good performance having most of the data points around the unit slope line, however, the CatBoost model delivers the closest cloud of data to this line suggesting that the estimations of this model match the experimental values better than the rest.

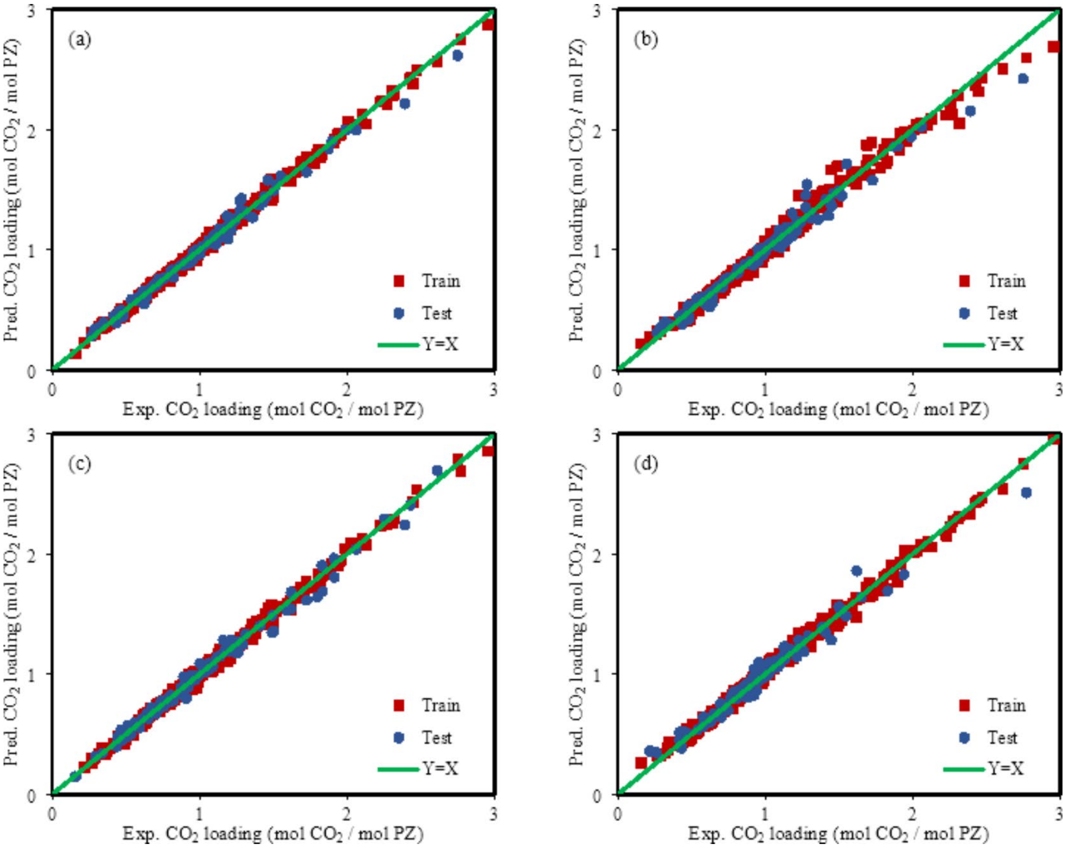

Fig. 7. Cross plots of different models developed in this study; (**a**) CatBoost; (**b**) XGBoost; (**c**) LightGBM; (**d**) AdaBoost-DT.

In the subsequent stage, the distributions of the CO_2 solubility estimation errors applying developed models against the experimental data were plotted in Fig. 8. As can be seen, the predictions of the models developed in this work show relative errors close to zero, which confirms their accuracy and reliability. However, again, the CatBoost model delivers relatively lower errors than others, and the formed cloud of errors is more concentrated near the zero error line.

Next, Fig. 9 depicts the cumulative frequency of the absolute relative error for different models when applied to the whole data bank. If the yellow horizontal dashed line that defines 70% of the data in the figure is considered, it can be seen that the CatBoost, LightGBM, XGBoost, and AdaBoost-DT models show absolute relative errors of 2.7%, 3.2%, 3.9%, and 4.9%, respectively, which means that the error of the models for predicting 70% of the data is less than the mentioned values. Similarly, about 90% of the estimated values by the CatBoost model had an absolute relative error of less than 5%, while the error values of other models are more than this. These observations along with other statistical and graphical analyses prove that the CatBoost model is highly accurate for predicting CO_2 absorption into aqueous PZ solutions.

Trend analysis

At this stage, it is time to check how the CatBoost model predicts the physical trend of CO_2 absorption in aqueous PZ solutions based on influencing variables. First, the prediction of the CatBoost model related to the solubility of CO_2 in 0.2 M PZ solution, as studied experimentally in the literature[48], was investigated with respect to temperature and partial pressure of CO_2. As illustrated in Fig. 10, CO_2 absorption values increased with increasing CO_2 partial pressure. This behavior is due to the more driving force for absorption at higher CO_2 partial pressure. Experimental studies showed that when CO_2 or a sour gas is added to an aqueous PZ solution, since the gas is mainly dissolved in non-volatile and ionic form, the total pressure initially rises very slightly with a raising extent of gas in the liquid. For higher gas loadings, the total pressure and of course CO_2 partial pressure increase steeply when PZ has been spent in the liquid phase by chemical reactions. This means that more sour gas can no longer be absorbed chemically but must be dissolved physically[75,76]. This is while the temperature has a destructive effect on the CO_2 solubility in PZ aqueous solution, and with the increase in temperature, the amount of CO_2 loading has decreased significantly. Actually, CO_2 absorption in aqueous PZ (amine) solutions decreases at higher temperatures due to the nature of the exothermic mass transfer process of chemisorption. Moreover, lower temperatures raise the viscosity of the liquid phase; thus lower CO_2 diffusion coefficients and consequently decrease CO_2 solubility[50,77]. Considering the modeling results illustrated in the

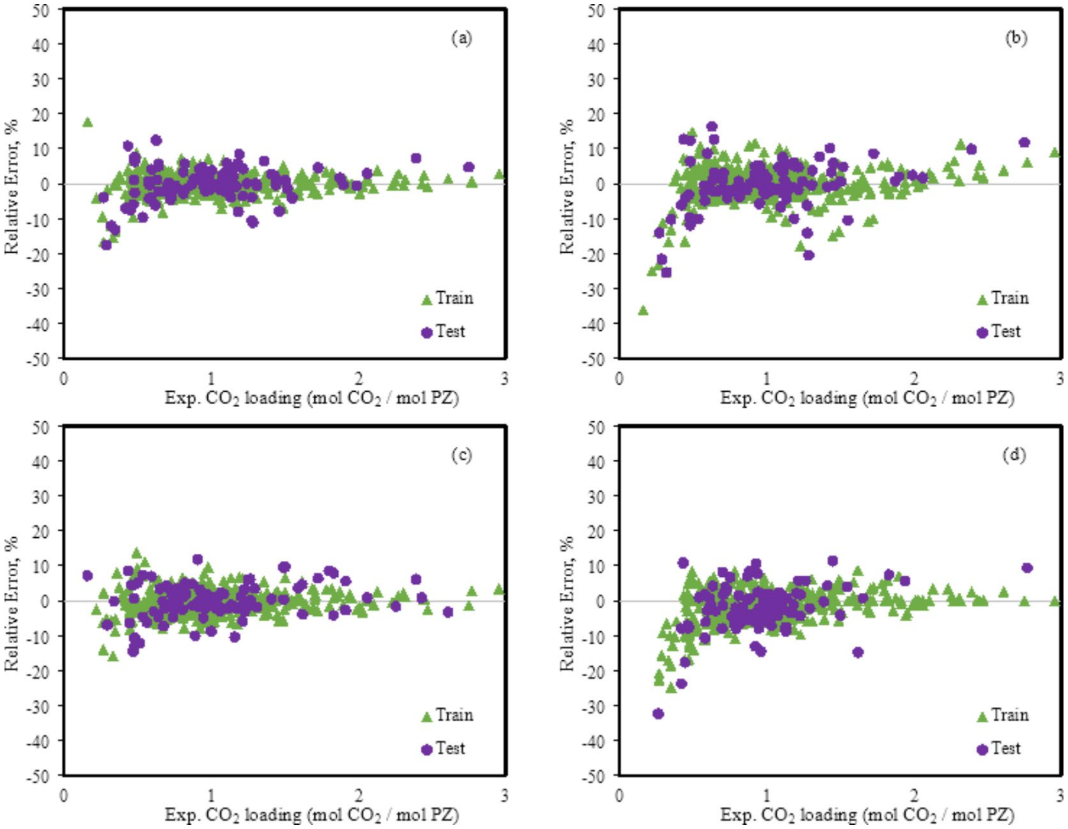

Fig. 8. Error distribution plots corresponding to the proposed models in this study; (**a**) CatBoost; (**b**) XGBoost; (**c**) LightGBM; (**d**) AdaBoost-DT.

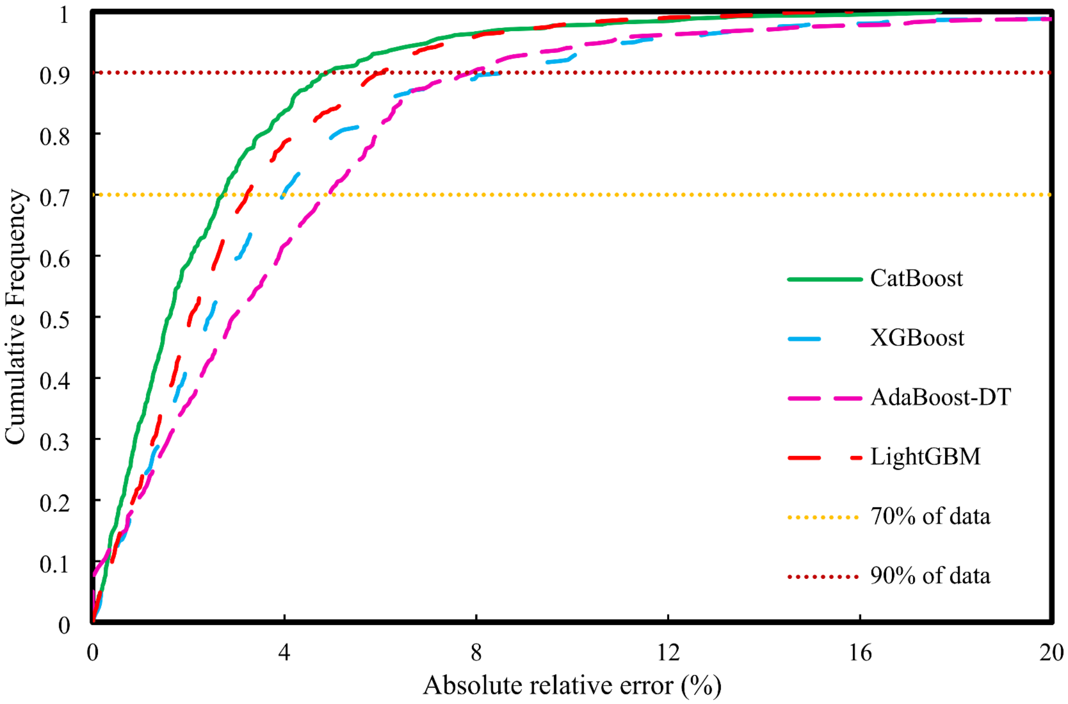

Fig. 9. The cumulative frequency plot for the proposed models.

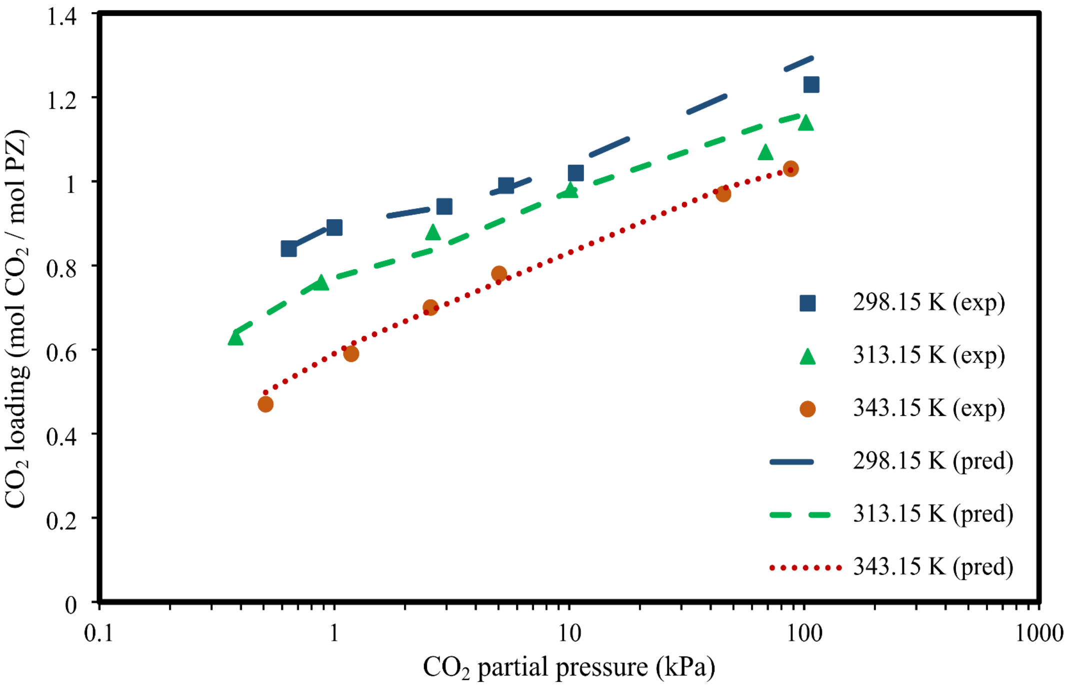

Fig. 10. The effect of temperature on CO_2 solubility in 0.2 M PZ solution; experimental data[48] and CatBoost model predictions.

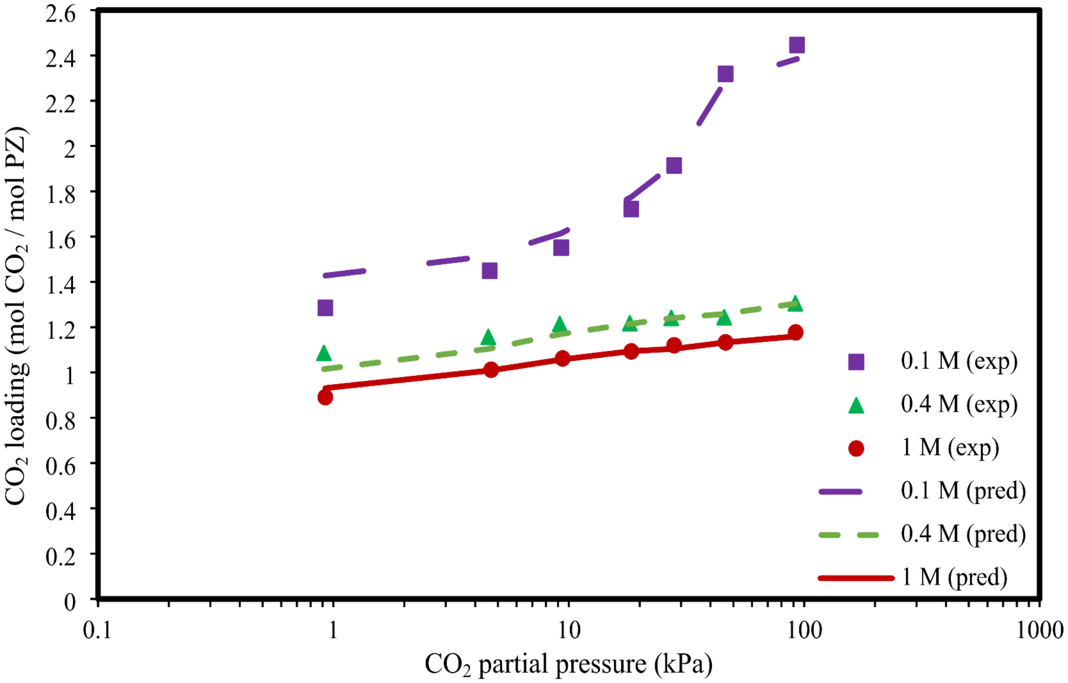

Fig. 11. The effect of PZ concentration on CO_2 solubility in aqueous PZ solutions; experimental data[53] and CatBoost model predictions.

figure, the proposed CatBoost model accurately recognized the absorption trend of gas and forecasts the CO_2 loading in aqueous PZ solution at various temperatures and pressures.

Next, the impact of PZ concentration on CO_2 solubility in PZ solutions was investigated at a fixed temperature of 303.15 K with respect to the partial pressure of CO_2, as experimentally studied in the literature[53], and compared with CatBoost model predictions in Fig. 11. As shown in Fig. 11, increasing the PZ concentration decreases the CO_2 loading at constant temperature and pressure. The free amine concentration, being a component of the mass transfer coefficient, has the potential to influence CO_2 mass transfer. With the increase of PZ concentration, the viscosity of the liquid phase increases, and therefore the CO_2 diffusion coefficient decreases slightly, leading to a decrease in the solubility of CO_2 at constant temperature and pressure[50]. Moreover, at higher CO_2 partial pressures for a more dilute solution, more physical absorption of the gas can be observed, which can ultimately lead to greater solubility of the gas compared to lower CO_2 partial pressures. Actually, a stronger PZ solution cannot be loaded to a high extent since the physically absorbed CO_2 is negligible in comparison to the chemically absorbed CO_2[53]. Again, the modeling results shown in the figure exhibit that the CatBoost model has an outstanding prediction for CO_2 solubility in PZ solutions with different concentrations and at different pressures.

Sensitivity analysis

In this survey, the Pearson and Spearman correlation coefficients were calculated to check the impact of three inputs, namely temperature, CO_2 partial pressure, and, PZ concentration on the output of the CatBoost model (i.e. CO_2 solubility in aqueous PZ solutions). For the Pearson correlation coefficient, the formula used to compute the linear effect of the input parameters is given below[55,78]:

$$r\left(z_i, y\right) = \frac{\sum_{j=1}^{n}(z_{i,j} - z_{a,i})(y_j - y_a)}{\left(\sum_{j=1}^{n}\left(z_{i,j} - z_{a,i}\right)^2 \sum_{j=1}^{n}\left(y_j - y_a\right)^2\right)^{0.5}} \tag{12}$$

here, $z_{i,j}$ and $z_{a,i}$ stand for the *j*-th and average values of *i*-th input parameter, respectively. Moreover, i could be temperature, CO_2 partial pressure, and, PZ concentration. In addition, $y_a$and $y_j$show the average and the *j*-th values of estimated CO_2 solubility in aqueous PZ solutions.

The Spearman correlation coefficient measures the association between the rankings of two variables using a monotonic function, enabling detection of non-linear relationships. It is robust against sample data distribution, unlike parametric methods, and uses a specific formula given below for rank correlation analysis[79]:

$$\rho\left(z, y\right) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(R(z_i) - R_a(z)\right)\left(R(y_i) - R_a(y)\right)}{\left(\left(\frac{1}{n}\sum_{i=1}^{n}\left(R\left(z_i\right) - R_a\left(z\right)\right)^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}\left(R\left(y_i\right) - R_a\left(y\right)\right)^2\right)\right)^{0.5}} \tag{13}$$

here, *n* denotes the count of data, where *ρ* is the Spearman rank correlation coefficient. Also, *R(z)* stands for the rank of variable *z*, while $R_a(z)$ represents its average rank. Moreover, *R(y)* shows the rank of variable *y*, and $R_a(y)$ is its average rank.

The correlation coefficients range from − 1 to 1, while the higher the absolute value of a parameter, the greater its impact on the output of the model[80]. Positive or negative values of correlation coefficients for a parameter indicate the increasing or decreasing effect of that parameter on the model's output, respectively[81,82]. The Pearson and Spearman correlation coefficients for all inputs calculated using the results of the CatBoost model, as the best paradigm developed in this work, are shown in Fig. 12. Among the input parameters, CO_2 partial pressure had the greatest influence on CO_2 absorption in aqueous PZ solutions. After that, PZ concentration and temperature respectively have shown the greatest effect with a slight difference. Both temperature and PZ concentration exhibit inverse relationships with CO_2 solubility, as indicated by negative Pearson coefficients (-0.299 and − 0.355, respectively) and even stronger negative Spearman coefficients (-0.361 and − 0.383, respectively), suggesting the presence of non-linear elements in these relationships. Conversely, CO_2 partial pressure shows a strong positive correlation with CO_2 solubility, with a Pearson coefficient of 0.621 and an even higher Spearman coefficient of 0.862, highlighting significant non-linear dynamics. In summary, while temperature and PZ concentration negatively influence CO_2 solubility with some non-linear effects, CO_2 partial pressure positively affects solubility, predominantly through non-linear effects.

Leverage approach

To appraise the validity region of the proposed CatBoost model and to discern any dubious data, the leverage technique[83–85] was utilized in this survey. In this approach, the differences between the model's estimates and experimental data are dubbed standardized residuals (*SR*). Taking H_i as the *i*th Leverage value, e_i as the error value, and *MSE* as the mean square of error, *SR* values are represented below[86,87]:

$$SR_i = \frac{e_i}{\left[MSE\left(1 - H_i\right)\right]^{0.5}} \tag{14}$$

Standardized residuals are incorporated in a Hat matrix. Also, hat indexes are elements on the main diagonal of the Hat matrix. Considering *T* as the transpose matrix of *X* as a (*k* × *l*) matrix incorporating *k* rows (data points), *l* columns (input parameters), the Hat indexes are determined according to the Hat matrix presented as follows[86]:

$$H = X\left(X^T X\right)^{-1} X^T \tag{15}$$

In addition, critical leverage (H^*) is a fixed value for a given data bank and can be computed as follows[85,88]:

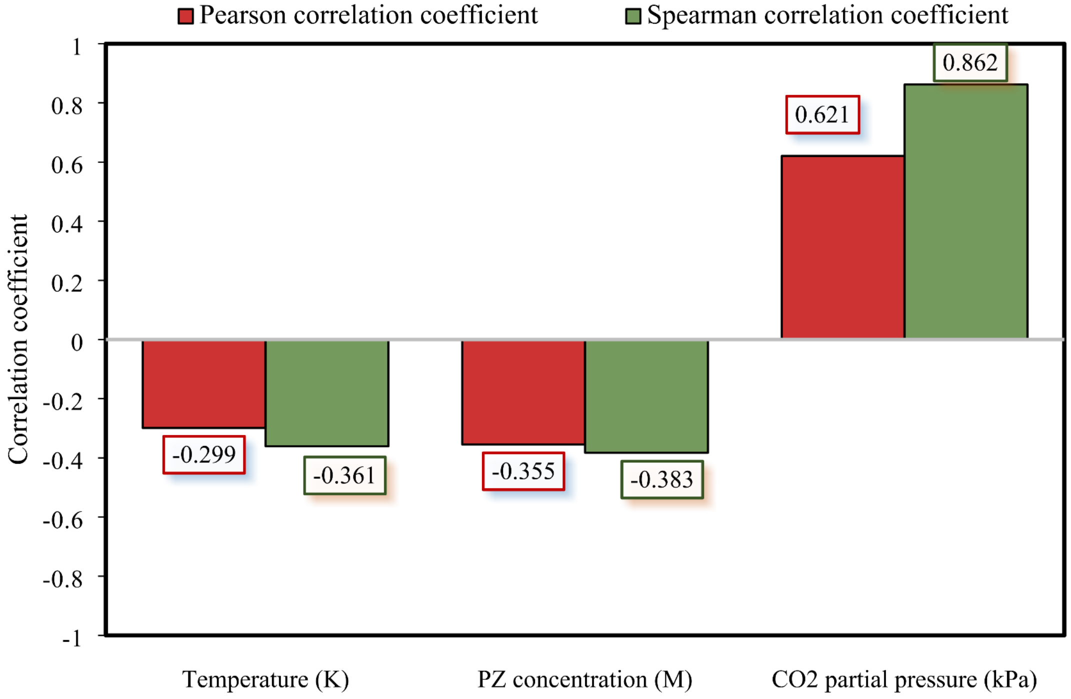

Fig. 12. The relative impacts of input parameters on the CO_2 solubility in aqueous PZ solutions as the CatBoost model output.

$$H^* = \frac{3 \times (l+1)}{k} \tag{16}$$

Williams's plot is usually drawn for a visual representation of the applicability scope of a model and doubtful data existing in the data bank, as shown in Fig. 13 for the CatBoost model developed in this survey. Here, bad high leverage points are those having *SR* values of more than 3 and less than -3 regardless of their Hat values. As shown in Fig. 13, only 8 data points (~1.5% of data) were identified as suspected data, which means that these data were laboratory suspects. Moreover, the data points having *SR* values between -3 and 3 with a Hat value higher than H^* (0.0232) are named good high leverage. As Williams's plot shows, 20 data points were identified as probable outliers, which means that despite the accurate estimation, these data were beyond the applicability scope of the model and are different from most of the data. In conclusion, both the experimental data bank utilized for modeling and the model's estimates were statistically acceptable and valid. Table 4 provides a list of the suspected data along with outliers identified for the proposed CatBoost model using the leverage technique.

In reviewing the literature[43–45], various models such as GP, GA-ANFIS, LSSVM, ANFIS, AdaBoost-CART, CHPSO-ANFIS, and CSA-LSSVM have been effectively utilized to address similar problems. These models have shown considerable success in their respective applications. The present work introduces the application of tree-based boosting algorithms to this domain, which have proven to be highly effective in regression problems but have not been previously applied to this specific subject of study. Through meticulous hyperparameter tuning using grid search and cross-validation, significant improvements in prediction accuracy were achieved, underscoring the potential of these algorithms in this context. For future work, incorporating new datasets to further validate and enhance the model's robustness is proposed. Additionally, exploring advanced metaheuristic optimization techniques and developing novel algorithms could offer further performance gains, ensuring the models remain at the forefront of predictive accuracy and reliability.

Conclusions

In this study, CO_2 solubility in aqueous PZ solutions was modeled using 517 experimental data points and four robust machine learning algorithms, namely CatBoost, LightGBM, XGBoost, and AdaBoost-DT. The CatBoost model represented the lowest error values among the four models developed in this work having AAPRE values of 2.36%, 2.06%, and 3.54% for the total, train, and test collections. Moreover, LightGBM, XGBoost, and AdaBoost-DT models were classified from the best performance to the weakest after the CatBoost model, respectively. Among the input parameters, CO_2 partial pressure had the greatest influence on CO_2 absorption in aqueous PZ solutions based on sensitivity analysis. After that, PZ concentration and temperature respectively demonstrated the greatest effect with a slight difference. Furthermore, both temperature and PZ concentration exhibited inverse relationships with CO_2 solubility, as indicated by negative Pearson and even stronger negative Spearman coefficients, suggesting the presence of non-linear elements. Conversely, CO_2 partial pressure showed a strong positive correlation with CO_2 solubility, with higher Spearman coefficient highlighting significant non-linear dynamics. Eventually, data assessment using the Leverage approach exhibited that 20 data points were

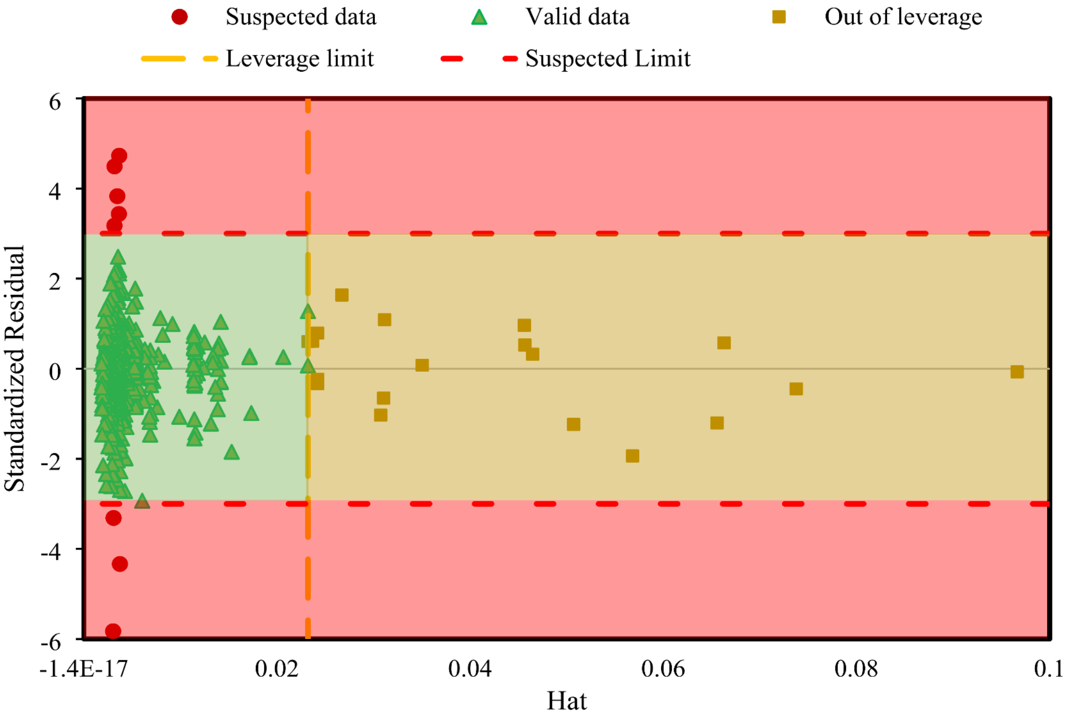

Fig. 13. The Williams plot of the entire data bank for the CatBoost model.

probable outliers, which means that despite the accurate estimation, these data were beyond the applicability scope of the model and were statistically different from most of the data. Moreover, both the experimental data bank used for modeling and the model's estimates were statistically acceptable and valid showing only 8 points (~1.5% of total data) as possible suspected data.

No.	Temperature (K)	PZ concentration (M)	CO$_2$ partial pressure (kPa)	Exp. CO$_2$ loading (mol CO$_2$/ mol PZ)	Pred. CO$_2$ loading (mol CO$_2$/ mol PZ)	H	SR	Status	Ref.
1	298	0.2	817.3	2.392	2.216	0.00303	-5.828	Suspected	47
2	298	0.2	1171	2.75	2.619	0.00374	-4.336	Suspected	47
3	328	0.4	385.3	1.191	1.091	0.00308	-3.314	Suspected	47
4	298	0.2	76.51	1.185	1.281	0.00317	3.177	Suspected	47
5	328	0.2	550.7	1.463	1.579	0.00347	3.831	Suspected	47
6	308	0.2	238.5	1.275	1.410	0.00318	4.492	Suspected	47
7	303.15	0.1	0.924	1.285	1.428	0.00367	4.730	Suspected	53
8	313	0.3	1108	1.49	1.594	0.00363	3.437	Suspected	52
9	313	0.3	6489	2.77	2.756	0.07375	-0.448	Outlier	52
10	343	0.3	5745	2.06	2.000	0.05682	-1.937	Outlier	52
11	313	0.3	5421	2.61	2.572	0.05070	-1.236	Outlier	52
12	343	0.3	5214	1.99	2.000	0.04647	0.323	Outlier	52
13	313	0.3	4536	2.43	2.432	0.03503	0.077	Outlier	52
14	343	0.3	4285	1.86	1.840	0.03105	-0.652	Outlier	52
15	343	1.2	7399	1.8	1.798	0.09663	-0.070	Outlier	52
16	313	1.2	6150	2.03	2.048	0.06629	0.574	Outlier	52
17	343	1.2	6150	1.7	1.662	0.06557	-1.204	Outlier	52
18	313	1.2	5149	1.94	1.956	0.04565	0.527	Outlier	52
19	343	1.2	5174	1.59	1.620	0.04560	0.963	Outlier	52
20	343	1.2	4314	1.51	1.543	0.03113	1.088	Outlier	52
21	313	1.2	4265	1.83	1.799	0.03074	-1.030	Outlier	52
22	333.15	6.2	0.331	0.462	0.455	0.02420	-0.235	Outlier	49
23	333.15	6.2	1.865	0.578	0.602	0.02420	0.791	Outlier	49
24	333.15	6.2	6.791	0.708	0.698	0.02419	-0.332	Outlier	49
25	353.15	6.2	2.115	0.444	0.463	0.02370	0.612	Outlier	50
26	353.15	6.2	9.141	0.58	0.571	0.02369	-0.295	Outlier	50
27	373.15	6.2	7.871	0.444	0.462	0.02322	0.603	Outlier	50
28	373.15	6.2	33.652	0.58	0.582	0.02320	0.070	Outlier	50

Table 4. Identified suspected data and outliers for the proposed CatBoost model using the leverage technique.

Data availability

The databank utilized during this research is available from the corresponding author on reasonable request.

Received: 10 June 2024; Accepted: 13 September 2024
Published online: 27 September 2024

References

1. Aghel, B., Behaein, S., Wongwises, S. & Shadloo, M. S. A review of recent progress in biogas upgrading: With emphasis on carbon capture. *Biomass Bioenergy***160**, 106422 (2022).
2. Aghel, B., Janati, S., Wongwises, S. & Shadloo, M. S. Review on CO$_2$ capture by blended amine solutions. *Int. J. Greenh. Gas Control***119**, 103715 (2022).
3. Friedlingstein, P. et al. Global carbon budget 2022. In *Earth System Science Data Discussions* 1–159 (2022).
4. Gelles, T., Lawson, S., Rownaghi, A. A. & Rezaei, F. Recent advances in development of amine functionalized adsorbents for CO$_2$ capture. *Adsorption***26**, 5–50 (2020).
5. Zhang, F., Zhao, P., Niu, M. & Maddy, J. The survey of key technologies in hydrogen energy storage. *Int. J. Hydrog. Energy***41**, 14535–14552 (2016).
6. Chen, P. C., Cho, H. H., Jhuang, J. H. & Ku, C. H. Selection of mixed amines in the CO$_2$ capture process. *Carbon*. **7**, 25 (2021).
7. Wu, S. Y., Liu, Y. F., Chu, C. Y., Li, Y. C. & Liu, C. M. Optimal absorbent evaluation for the CO$_2$ separating process by absorption loading, desorption efficiency, cost, and environmental tolerance. *Int. J. Green Energy***12**, 1025–1030 (2015).
8. Olabi, A. et al. Membrane-based carbon capture: Recent progress, challenges, and their role in achieving the sustainable development goals. *Chemosphere***320**, 137996 (2023).
9. Dai, N. & Mitch, W. A. Influence of amine structural characteristics on N-nitrosamine formation potential relevant to postcombustion CO$_2$ capture systems. *Environ. Sci. Technol.***47**, 13175–13183 (2013).
10. Bui, M. et al. Carbon capture and storage (CCS): The way forward. *Energy Environ. Sci.***11**, 1062–1176 (2018).
11. Liang, Z. H. et al. Recent progress and new developments in post-combustion carbon-capture technology with amine based solvents. *Int. J. Greenh. Gas Control***40**, 26–54 (2015).
12. Wang, Z., Zhang, Z. & Mitch, W. A. Role of absorber and desorber units and operational conditions for N-nitrosamine formation during amine-based carbon capture. *Water Res.***170**, 115299 (2020).
13. Aghel, B., Sahraie, S., Heidaryan, E. & Varmira, K. Experimental study of carbon dioxide absorption by mixed aqueous solutions of methyl diethanolamine (MDEA) and piperazine (PZ) in a microreactor. *Process Saf. Environ. Prot.***131**, 152–159 (2019).

14. Kim, Y. E., Choi, J. H., Nam, S. C. & Yoon, Y. I. CO_2 absorption characteristics in aqueous K_2CO_3/piperazine solution by NMR spectroscopy. *Ind. Eng. Chem. Res.***50**, 9306–9313 (2011).
15. Rochelle, G. T. Amine scrubbing for CO_2 capture. *Science***325**, 1652–1654 (2009).
16. Aronu, U. E. et al. Solubility of CO_2 in 15, 30, 45 and 60 mass% MEA from 40 to 120 °C and model representation using the extended UNIQUAC framework. *Chem. Eng. Sci.***66**, 6393–6406 (2011).
17. Chen, C. C. & Evans, L. B. A local composition model for the excess Gibbs energy of aqueous electrolyte systems. *AlChE J.***32**, 444–454 (1986).
18. Fouad, W. A. & Berrouk, A. S. Prediction of H_2S and CO_2 solubilities in aqueous triethanolamine solutions using a simple model of Kent–Eisenberg type. *Ind. Eng. Chem. Res.***51**, 6591–6597 (2012).
19. Haghtalab, A. & Dehghani Tafti, M. Electrolyte UNIQUAC – NRF model to study the solubility of acid gases in alkanolamines. *Ind. Eng. Chem. Res.***46**, 6053–6060 (2007).
20. Dashti, A., Raji, M., Alivand, M. S. & Mohammadi, A. H. Estimation of CO_2 equilibrium absorption in aqueous solutions of commonly used amines using different computational schemes. *Fuel***264**, 116616 (2020).
21. Song, Z., Shi, H., Zhang, X. & Zhou, T. Prediction of CO_2 solubility in ionic liquids using machine learning methods. *Chem. Eng. Sci.***223**, 115752 (2020).
22. Amar, M. N., Larestani, A., Lv, Q., Zhou, T. & Hemmati-Sarapardeh, A. Modeling of methane adsorption capacity in shale gas formations using white-box supervised machine learning techniques. *J. Pet. Sci. Eng.* 109226 (2021).
23. Naghizadeh, A., Larestani, A., Amar, M. N. & Hemmati-Sarapardeh, A. Predicting viscosity of CO_2–N_2 gaseous mixtures using advanced intelligent schemes. *J. Pet. Sci. Eng.* 109359 (2021).
24. Hashemizadeh, A., Maaref, A., Shateri, M., Larestani, A. & Hemmati-Sarapardeh, A. Experimental measurement and modeling of water-based drilling mud density using adaptive boosting decision tree, support vector machine, and K-nearest neighbors: A case study from the South pars gas field. *J. Pet. Sci. Eng.* 109132 (2021).
25. Larestani, A., Hemmati-Sarapardeh, A. & Naseri, A. Experimental measurement and compositional modeling of bubble point pressure in crude oil systems: Soft computing approaches, correlations, and equations of state. *J. Pet. Sci. Eng.* 110271 (2022).
26. Lv, Q. et al. Modelling minimum miscibility pressure of CO_2-crude oil systems using deep learning, tree-based, and thermodynamic models: Application to CO_2 sequestration and enhanced oil recovery. *Sep. Purif. Technol.* 123086 (2023).
27. Tian, Y., Wang, X., Liu, Y. & Hu, W. Prediction of nitrogen solubility in ionic liquids by machine learning methods based on COSMO-derived descriptors. *Chem. Eng. Sci.***284**, 119482 (2024).
28. Wang, C. et al. Integrating experimental study and intelligent modeling of pore evolution in the Bakken during simulated thermal progression for CO_2 storage goals. *Appl. Energy***359**, 122693 (2024).
29. Saghafi, H. & Arabloo, M. Modeling of CO_2 solubility in MEA, DEA, TEA, and MDEA aqueous solutions using AdaBoost-Decision Tree and Artificial neural network. *Int. J. Greenh. Gas Control***58**, 256–265 (2017).
30. Salooki, M. K., Abedini, R., Adib, H. & Koolivand, H. Design of neural network for manipulating gas refinery sweetening regenerator column outputs. *Sep. Purif. Technol.***82**, 1–9 (2011).
31. Adib, H., Sharifi, F., Mehranbod, N., Kazerooni, N. M. & Koolivand, M. Support vector machine based modeling of an industrial natural gas sweetening plant. *J. Nat. Gas Sci. Eng.***14**, 121–131 (2013).
32. Sipöcz, N., Tobiesen, F. A. & Assadi, M. The use of artificial neural network models for CO_2 capture plants. *Appl. Energy***88**, 2368–2376 (2011).
33. Sahraie, S., Rashidi, H. & Valeh-e-Sheyda, P. An optimization framework to investigate the CO_2 capture performance by MEA: Experimental and statistical studies using Box-Behnken design. *Process. Saf. Environ. Prot.***122**, 161–168 (2019).
34. Wu, Y. & Chan, C. W. Analysis of data for the carbon dioxide capture domain. *Eng. Appl. Artif. Intell.***24**, 154–163 (2011).
35. Zhou, Q., Chan, C. W., Tontiwachwuthikul, P., Idem, R. & Gelowitz, D. Application of neuro-fuzzy modeling technique for operational problem solving in a CO_2 capture process system. *Int. J. Greenh. Gas Control***15**, 32–41 (2013).
36. Zhou, Q., Wu, Y., Chan, C. W. & Tontiwachwuthikul, P. Modeling of the carbon dioxide capture process system using machine intelligence approaches. *Eng. Appl. Artif. Intell.***24**, 673–685 (2011).
37. Hsiao, Y. D. & Chang, C. T. Expandable neural networks for efficient modeling of various amine scrubbing configurations for CO_2 capture. *Chem. Eng. Sci.***281**, 119191 (2023).
38. Wang, X., Chan, C. W. & Li, T. High accuracy prediction of the Post-combustion Carbon capture process parameters using the decision Forest Approach. *Chem. Eng. Sci.* 119878 (2024).
39. Ghiasi, M. M. & Mohammadi, A. H. Rigorous modeling of CO_2 equilibrium absorption in MEA, DEA, and TEA aqueous solutions. *J. Nat. Gas Sci. Eng.***18**, 39–46 (2014).
40. Ghiasi, M. M., Arabloo, M., Mohammadi, A. H. & Barghi, T. Application of ANFIS soft computing technique in modeling the CO_2 capture with MEA, DEA, and TEA aqueous solutions. *Int. J. Greenh. Gas Control***49**, 47–54 (2016).
41. Daneshvar, N., Moattar, M. Z., Abdi, M. A. & Aber, S. Carbon dioxide equilibrium absorption in the multi-component systems of $CO_2 + TIPA + MEA + H_2O$, $CO_2 + TIPA + pz + H_2O$ and $CO_2 + TIPA + H_2O$ at low CO_2 partial pressures: Experimental solubility data, corrosion study and modeling with artificial neural network. *Sep. Purif. Technol.***37**, 135–147 (2004).
42. Shahsavand, A., Fard, F. D. & Sotoudeh, F. Application of artificial neural networks for simulation of experimental CO_2 absorption data in a packed column. *J. Nat. Gas Sci. Eng.***3**, 518–529 (2011).
43. Tatar, A. et al. Comparison of two soft computing approaches for predicting CO_2 solubility in aqueous solution of piperazine. *Int. J. Greenh. Gas Control***53**, 85–97 (2016).
44. Yarveicy, H., Ghiasi, M. M. & Mohammadi, A. H. Performance evaluation of the machine learning approaches in modeling of CO_2 equilibrium absorption in Piperazine aqueous solution. *J. Mol. Liq.***255**, 375–383 (2018).
45. Dashti, A. et al. Efficient hybrid modeling of CO_2 absorption in aqueous solution of piperazine: Applications to energy and environment. *Chem. Eng. Res. Des.***144**, 405–417 (2019).
46. Bishnoi, S. & Rochelle, G. T. Absorption of carbon dioxide into aqueous piperazine: Reaction kinetics, mass transfer and solubility. *Chem. Eng. Sci.***55**, 5531–5543 (2000).
47. Dash, S. K., Samanta, A., Samanta, A. N. & Bandyopadhyay, S. S. Vapour liquid equilibria of carbon dioxide in dilute and concentrated aqueous solutions of piperazine at low to high pressure. *Fluid Phase Equilibria***300**, 145–154 (2011).
48. Derks, P., Dijkstra, H., Hogendoorn, J. & Versteeg, G. Solubility of carbon dioxide in aqueous piperazine solutions. *AIChE J.***51**, 2311–2327 (2005).
49. Dugas, R. & Rochelle, G. Absorption and desorption rates of carbon dioxide with monoethanolamine and piperazine. *Energy Procedia***1**, 1163–1169 (2009).
50. Dugas, R. E. *Carbon Dioxide Absorption, Desorption, and Diffusion in Aqueous Piperazine and Monoethanolamine* (The University of Texas at Austin, 2009).
51. Haghtalab, A., Eghbali, H. & Shojaeian, A. Experiment and modeling solubility of CO_2 in aqueous solutions of diisopropanolamine + 2-amino-2-methyl-1-propanol + piperazine at high pressures. *J. Chem. Thermodyn.***71**, 71–83 (2014).
52. Kadiwala, S., Rayer, A. V. & Henni, A. High pressure solubility of carbon dioxide (CO_2) in aqueous piperazine solutions. *Fluid. Phase. Equilibria***292**, 20–28 (2010).
53. Aroua, M. K. & Mohd Salleh, R. Solubility of CO_2 in aqueous piperazine and its modeling using the Kent-Eisenberg approach. *Chem. Eng. Technol. Ind. Chem. Plant Equip. Process. Eng. Biotechnol.***27**, 65–70 (2004).
54. Hadavimoghaddam, F. et al. Modeling crude oil pyrolysis process using advanced white-box and black-box machine learning techniques. *Sci. Rep.***13**, 22649 (2023).

55. Chen, G. et al. The genetic algorithm based back propagation neural network for MMP prediction in CO_2-EOR process. *Fuel***126**, 202–212 (2014).
56. Chen, T. & Guestrin, C. In *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining* 785–794.
57. Chen, T. et al. *Xgboost: Extreme gradient boosting*. R package version 0.4-2 1, 1–4 (2015).
58. Nakhaei-Kohani, R. et al. Machine learning assisted structure-based models for predicting electrical conductivity of Ionic liquids. *J. Mol. Liq.* 119509 (2022).
59. Abdi, J., Hadavimoghaddam, F., Hadipoor, M. & Hemmati-Sarapardeh, A. Modeling of CO_2 adsorption capacity by porous metal organic frameworks using advanced decision tree-based models. *Sci. Rep.***11**, 1–14 (2021).
60. Mohammadi, M. R. et al. Modeling hydrogen solubility in hydrocarbons using extreme gradient boosting and equations of state. *Sci. Rep.***11**, 17911 (2021).
61. Sun, X., Liu, M. & Sima, Z. A novel cryptocurrency price trend forecasting model based on LightGBM. *Finance Res. Lett.***32**, 101084 (2020).
62. Yang, X., Dindoruk, B. & Lu, L. A comparative analysis of bubble point pressure prediction using advanced machine learning algorithms and classical correlations. *J. Petrol. Sci. Eng.***185**, 106598 (2020).
63. Gu, Y. et al. Data-driven estimation for permeability of simplex pore-throat reservoirs via an improved light gradient boosting machine: A demonstration of sand-mud profile, Ordos Basin, northern China. *J. Petrol. Sci. Eng.* 110909 (2022).
64. Mahmoudzadeh, A. et al. Modeling CO_2 solubility in water using gradient boosting and light gradient boosting machine. *Sci. Rep.***14**, 13511 (2024).
65. Qi, M. *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*. Neural Inform. Process. Syst. Curran Associates Inc (2017).
66. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. CatBoost: Unbiased boosting with categorical features. *Adv. Neural. Inf. Process. Syst.* 31 (2018).
67. Morozov, A. D. et al. Data-driven model for hydraulic fracturing design optimization: Focus on building digital database and production forecast. *J. Petrol. Sci. Eng.***194**, 107504 (2020).
68. Duplyakov, V. et al. Data-driven model for hydraulic fracturing design optimization. Part II: Inverse problem. *J. Petrol. Sci. Eng.***208**, 109303 (2022).
69. Freund, Y. & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.***55**, 119–139 (1997).
70. Dargahi-Zarandi, A., Hemmati-Sarapardeh, A., Shateri, M., Menad, N. A. & Ahmadi, M. Modeling minimum miscibility pressure of pure/impure CO_2-crude oil systems using adaptive boosting support vector regression: Application to gas injection processes. *J. Petrol. Sci. Eng.***184**, 106499 (2020).
71. Zerrouki, N., Harrou, F., Sun, Y. & Houacine, A. Vision-based human action classification using adaptive boosting algorithm. *IEEE Sens. J.***18**, 5115–5121 (2018).
72. Mohammadi, M. R. et al. Modeling the solubility of light hydrocarbon gases and their mixture in brine with machine learning and equations of state. *Sci. Rep.***12**, 14943 (2022).
73. Nair, P. et al. AI-driven digital twin model for reliable lithium-ion battery discharge capacity predictions. *Int. J. Intell. Syst.* 8185044 (2024).
74. Shawki, N., Nunez, R. R., Obeid, I. & Picone, J. In *2021 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)* 1–7 (IEEE).
75. Kamps, Á. P. S., Xia, J. & Maurer, G. Solubility of CO_2 in (H_2O + piperazine) and in (H_2O + MDEA + piperazine). *AIChE J.***49**, 2662–2670 (2003).
76. 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***51**, 1788–1796 (2006).
77. Jahangiri, A. & Nabipoor Hassankiadeh, M. Effects of piperazine concentration and operating conditions on the solubility of CO_2 in AMP solution at low CO_2 partial pressure. *Sep. Sci. Technol.***54**, 1067–1078 (2019).
78. Mohammadi, M. R., Hemmati-Sarapardeh, A., Schaffie, M., Husein, M. M. & 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.***205**, 108836 (2021).
79. Xu, M., Wong, T. C. & Chin, K. S. Modeling daily patient arrivals at Emergency Department and quantifying the relative importance of contributing variables using artificial neural network. *Decis. Support Syst.***54**, 1488–1498 (2013).
80. 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.***13**, 122 (2023).
81. Mohammadi, M. R. et al. 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.***25**, 1268–1295 (2021).
82. Salehi, E. et al. Modeling interfacial tension of N_2/CO_2 mixture + n-alkanes with machine learning methods: Application to eor in conventional and unconventional reservoirs by flue gas injection. *Minerals***12**, 252 (2022).
83. Leroy, A. M. & Rousseeuw, P. J. *Robust Regression and Outlier Detection*. rrod (1987).
84. Goodall, C. R. *13 Computation Using the QR Decomposition* (1993).
85. Gramatica, P. Principles of QSAR models validation: Internal and external. *QSAR Comb. Sci.***26**, 694–701 (2007).
86. Rousseeuw, P. J. & Leroy, A. M. *Robust Regression and Outlier Detection* (Wiley, 2005).
87. Hadavimoghaddam, F. et al. Data-driven modeling of H_2 solubility in hydrocarbons using white-box approaches. *Int. J. Hydrog. Energy***47**, 33224–33238 (2022).
88. Ansari, S. et al. Prediction of hydrogen solubility in aqueous solutions: Comparison of equations of state and advanced machine learning-metaheuristic approaches. *Int. J. Hydrog. Energy***47**, 37724–37741 (2022).

Author contributions

M-R.M: Investigation, data curation, modeling, visualization, writing-original draft; A.L: Investigation, conceptualization, validation, modeling, writing-original draft; M.S: Writing-Review and Editing, Methodology, Validation, Supervision; A.H-S: Methodology, validation, supervision, writing-review and editing; M.R: Writing-Review and Editing, Methodology, Validation, Supervision.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.-R.M. or A.H.-S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.