



OPEN

Modeling solubility of CO₂–N₂ gas mixtures in aqueous electrolyte systems using artificial intelligence techniques and equations of state

Reza Nakhaei-Kohani¹, Ehsan Taslimi-Renani², Fahime Hadavimoghaddam^{3,4},
 Mohammad-Reza Mohammadi⁵ & Abdolhossein Hemmati-Sarapardeh^{5,6}✉

Determining the solubility of non-hydrocarbon gases such as carbon dioxide (CO₂) and nitrogen (N₂) in water and brine is one of the most controversial challenges in the oil and chemical industries. Although many researches have been conducted on solubility of gases in brine and water, very few researches investigated the solubility of power plant flue gases (CO₂–N₂ mixtures) in aqueous solutions. In this study, using six intelligent models, including Random Forest, Decision Tree (DT), Gradient Boosting-Decision Tree (GB-DT), Adaptive Boosting-Decision Tree (AdaBoost-DT), Adaptive Boosting-Support Vector Regression (AdaBoost-SVR), and Gradient Boosting-Support Vector Regression (GB-SVR), the solubility of CO₂–N₂ mixtures in water and brine solutions was predicted, and the results were compared with four equations of state (EOSs), including Peng–Robinson (PR), Soave–Redlich–Kwong (SRK), Valderrama–Patel–Teja (VPT), and Perturbed-Chain Statistical Associating Fluid Theory (PC-SAFT). The results indicate that the Random Forest model with an average absolute percent relative error (AAPRE) value of 2.8% has the best predictions. The GB-SVR and DT models also have good precision with AAPRE values of 6.43% and 7.41%, respectively. For solubility of CO₂ present in gaseous mixtures in aqueous systems, the PC-SAFT model, and for solubility of N₂, the VPT EOS had the best results among the EOSs. Also, the sensitivity analysis of input parameters showed that increasing the mole percent of CO₂ in gaseous phase, temperature, pressure, and decreasing the ionic strength increase the solubility of CO₂–N₂ mixture in water and brine solutions. Another significant issue is that increasing the salinity of brine also has a subtractive effect on the solubility of CO₂–N₂ mixture. Finally, the Leverage method proved that the actual data are of excellent quality and the Random Forest approach is quite reliable for determining the solubility of the CO₂–N₂ gas mixtures in aqueous systems.

Abbreviations

AAPRE	Average absolute percent relative error
SD	Standard deviation
SRK	Soave–Redlich–Kwong
SAFT	Statistical associating fluid theory
RMSE	Root mean square error
PR	Peng–Robinson
VPT	Valderrama–Patel–Teja
PC-SAFT	Perturbed-chain statistical associating fluid theory
EOS	Equation of state
GB	Gradient boosting

¹Department of Chemical & Petroleum Engineering, Shiraz University, Shiraz, Iran. ²Department of Electrical Engineering, University of Malaya (UM), 50603 Kuala Lumpur, Malaysia. ³Ufa State Petroleum Technological University, Ufa 450064, Russia. ⁴Institute of Unconventional Oil & Gas, Northeast Petroleum University, Heilongjiang, Daqing 163318, China. ⁵Department of Petroleum Engineering, Shahid Bahonar University of Kerman, Kerman, Iran. ⁶Key Laboratory of Continental Shale Hydrocarbon Accumulation and Efficient Development, Ministry of Education, Northeast Petroleum University, Daqing 163318, China. ✉email: hemmati@uk.ac.ir

AdaBoost	Adaptive boosting
RF	Random forest
DT	Decision tree
SAFT	Statistical associating fluid theory
SVM	Support vector machine
SVR	Support vector machine for regression
SVC	Support vector machine for classification

In the last decade, one of the most important challenges in the petroleum and chemical industries has been evaluating the solubility of different gases in liquids, including hydrocarbon and non-hydrocarbon gases^{1–3}. The solubility of gases in liquids can be vital in the petroleum and chemical industries for a variety of reasons, including transport operations and the production of hydrates^{1,4}. CO₂ as a greenhouse gas has been considered a serious problem in recent decades^{5–7}. Carbon capture and storage (CCS)^{8,9} is a technique that involves capturing CO₂ from major point sources and storing it in formations^{10,11}. Flue gas storage in saline aquifers, as well as CO₂ extraction and storage using gas hydrates, are considered as potential CCS methods. As a result, information gaps about these methods, such as the solubility of gas mixtures in water and brine, must be filled before commercialization. Due to the high cost of traditional CCS technologies, considerable efforts have been made to improve the efficiency of CCS operations by creating cost-effective and practical CCS approaches; however, there are still a lot of technological and financial roadblocks to overcome^{10,12–14}.

Flue gas or the mixture of CO₂ and N₂ injected within gas hydrate reservoirs have been suggested as a potential alternative for CO₂ underground storage. The thermodynamic mechanism by which CO₂ in flue gas or a CO₂–N₂ mixture is collected as hydrate, on the other hand, is not well recognized¹⁵. CO₂ storage in hydrate reservoirs has expensive obstacles that limit its widespread usage, despite all of the stated benefits. The primary expense in this scenario is CO₂ collection before storage^{15,16}. Injecting CO₂–N₂ mixture within gas hydrate reservoirs rather than pure CO₂ might considerably cut CO₂ separation expenses. Furthermore, an industrial-scale CO₂ substitution experiment on the North Slope of Alaska found that injecting a gas combination of 77/23 ratio of N₂/CO₂ into a hydrate reservoir while recovering methane successfully avoided CO₂ hydrate creation around the injection well. Although the prior studies show that injecting CO₂–N₂ gas mixes into gas hydrate reservoirs might be a cost-effective technique for CCS, a primary concern remain: How can the reservoir circumstances following CO₂–N₂ mixtures or flue gas injection into a gas hydrate reservoir affect the production of CO₂ and CO₂–mixed hydrates¹⁵? Since different thermodynamic conditions affect the injection process of the CO₂–N₂ mixture and make the injection process difficult, the first important step is to evaluate the solubility of the CO₂–N₂ mixture at different thermodynamic conditions. It should be noted that these limitations have also led to limited laboratory data on the solubility of CO₂–N₂ mixture in liquids. Therefore, finding a solution to measure the solubility of the CO₂–N₂ mixture has great importance. As a result of these considerations, assessing the solubility of gases in liquids has become a contentious issue. CO₂ and N₂ have been extensively considered as two frequently used non-hydrocarbon gases in recent studies^{17,18}. The injection of CO₂ into the aquifer and the injection of a mixture of CO₂ and N₂ into oil and gas reservoirs are two examples of these situations, where knowing the degree of solubility of the gas is critical^{10,19}. As a result, a thorough understanding of the physical and chemical interactions between CO₂, N₂, and water is required. For instance, solubility trapping and mineral trapping are the two significant mechanisms that influence the injection of CO₂ into the aquifer. To accurately determine the effect of these mechanisms, it is necessary to conduct a sufficient number of theoretical and experimental studies, which can be time-consuming and costly^{10,20,21}.

In addition to the laboratory experiments, another technique for determining the solubility of CO₂ and N₂ in water is to utilize equations of state (EOSs); however, it should be noted that EOSs are more appropriate for pure fluids but have limitations for pure compounds. Some of these limitations are as follows^{22,23}:

- To determine the solubility using these types of equations, critical characteristics of pure substances are necessary. Many of the chemicals studied, particularly those with complicated chemical structures, break down before meeting critical conditions. As a result, measuring the relevant characteristics does not appear to be feasible.
- To adjust the thermodynamic coefficients of the equation for a more precise estimation of the physical properties of the system, several physicochemical aspects of the system should be evaluated, such as the characteristics of the donor and the acceptor of the hydrogen bond of the molecule.
- Interaction factors setting for solubility data for each model is a time-consuming process.
- Numerical methods are often divergent to solve some equations for pure materials that have low solubility in water.
- The solubility estimations are heavily influenced by the optimization techniques used to get the best values for the thermodynamic model parameters.

As a result, choosing the appropriate optimization technique is another issue to consider. Despite these flaws, thermodynamic techniques have been extensively used to forecast the solubility of CO₂, N₂, and other gases in water, which are often found in the oil and gas industry under a variety of thermodynamic conditions. In the literature, CO₂ solubility in water and aqueous solutions^{24–26} of salts like NaCl, KCl, and CaCl₂ has been thoroughly documented. Also, the solubility of N₂ and CO₂–N₂ mixture in water and brine has been studied^{22,27–29}. Tomoya et al.³⁰ measured CO₂ solubility in aqueous solutions and then correlated the experimental data with the Peng–Robinson–Stryjek–Vera EOS. Yiteng et al.³¹ also needed to know the solubility of CO₂ in brine to estimate CO₂ capturing potential in deep saline aquifers. For this purpose, they utilized the Peng–Robinson

Cubic-Plus-Association (PR-CPA) EOS to calculate the solubility of CO₂ in brine. They represented that good agreement was achieved with laboratory data.

The second group of methods for estimating solubility involves creating correlations, particularly mathematical methods that employ the physical characteristics of the chemicals in a manner that makes these approaches broad and thorough. These techniques may represent/predict the solubility of substances from diverse chemical categories in water in any condition²². Abraham et al.³² suggested a linear solvation energy relationship (LSER) approach. However, the relationship can predict the solubility of ordinary organic substances; the model's properties are challenging to be determined from the compounds' chemical structures. Other researchers have taken the same method^{33,34}.

In the previous studies, a number of experimental data have been reported for the solubility of non-hydrocarbon gases, including CO₂ and N₂ in liquids, especially in water^{18,35,36}. There is a scarcity of experimental results for non-hydrocarbon solubility due to the difficulties and sophistication of measured data of natural gas including gas equilibrium data. As a result, the utilization of laboratory data in new modeling approaches like artificial neural networks has gotten much attention¹. Machine learning techniques have recently found widespread use in forecasting challenges such as hydrate formation³⁷, ammonia solubility in liquids³⁸, simulating asphaltene behavior³⁹, and hydrocarbon-CO₂ interfacial tension⁴⁰. They have received much interest as a result of their captivating performance⁴¹. Samani et al.⁴² proposed different intelligence techniques for estimating the solubility of various gases in aqueous electrolyte systems. Regarding the solubility of non-hydrocarbon gases (i.e., N₂ and CO₂) in aqueous electrolyte systems, their database includes 774 data points, of which only 81 data are related to the N₂-CO₂ gas mixture and the rest are related to the solubility of N₂ and CO₂ pure gases. Their model was based on Coupled Simulated Annealing (CSA) linked to the Least-Squares Support Vector Machine (LSSVM) method. Average absolute relative error and root mean square error (RMSE) values of their proposed CSA-LSSVM model were 10.71% and 0.0011, respectively. Hemmati-Sarapardeh et al.⁴³ investigated the solubility of CO₂ in water at high pressures and temperatures using four powerful machine learning techniques. In this study, Multilayer Perceptron (MLP), Radial Basis Function (RBF), Least-Squares Support Vector Machine (LSSVM), and Gene Expression Programming (GEP) models were developed using temperature and pressure as input data to estimate the solubility of CO₂ in water. The results showed that the LSSVM-FFA model with an RMSE value of 0.3261 had the best performance compared to other models. Nabipour et al.¹ investigated the solubility of CO₂ and N₂ in aqueous solutions using Extreme Learning Machine (ELM) and LSSVM approaches. Their solubility database was similar to Samani et al.'s work⁴² including 774 data points with less than 90 data related to CO₂-N₂ mixture solubility. The results showed that the LSSVM technique with an RMSE value of 0.001 had higher proficiency than the ELM approach in estimating the solubility values of CO₂ and N₂ in aqueous solutions. Temperature, pressure, and composition were the most critical input parameters to the models. Saghafi et al.⁴⁴ investigated the solubility of CO₂ in Monoethanolamine (MEA), Diethanolamine (DEA), Triethanolamine (TEA), and N-Methyldiethanolamine (MDEA) aqueous solutions. In this study, the AdaBoost-Decision Tree method and intelligent neural networks were used. The results showed that AdaBoost-Decision Tree models with RMSE values of 0.005–0.022 obtained the best solutions for different aqueous solutions. Gharagheizi et al.²² estimated the solubility of pure compounds such as CO₂ in water using an Artificial Neural Network-Group Contribution (ANN-GC) technique. The results showed that this model with an RMSE value of 0.4 could have a good performance in estimating the solubility of pure materials in water.

Therefore, as mentioned earlier, particular importance and attention to the issue of determining the solubility of CO₂ and N₂ in liquids and especially water with various techniques including laboratory methods⁴⁵, EOSs, mathematical methods, and intelligent neural networks^{46,47} in previous studies has caused further studies in this field and is still of interest to researchers. Although many studies have been done on pure CO₂ and N₂, few studies investigated the solubility of CO₂-N₂ mixtures in water and brine. Only two papers^{1,42} applied intelligent models for CO₂-N₂ mixtures, however, they used less than 90 data points and in limited ranges of operating parameters.

In this study, to estimate the solubility of CO₂-N₂ mixtures in water and aqueous brine solutions, an extensive database containing 289 laboratory is collected from the literature. This paper uses six machine learning approaches, including Random Forest, Decision Tree (DT), Gradient Boosting-Decision Tree (GB-DT), Adaptive Boosting-Decision Tree (AdaBoost-DT), Adaptive Boosting-Support Vector Machine for Regression (AdaBoost-SVR), and Gradient Boosting-Support Vector Machine for Regression (GB-SVR), for determining CO₂-N₂ mixture solubility in aqueous solutions in terms of temperature, pressure, ionic strength of aqueous brine solutions, CO₂ mole percent in gaseous mixture, and finally the index of non-hydrocarbon gases (i.e., N₂ and CO₂) whose solubility is to be estimated. Also, four reputable equations of state, including Peng–Robinson (PR), Soave–Redlich–Kwong (SRK), Valderrama–Patel–Teja (VPT), and Perturbed-Chain Statistical Associating Fluid Theory (PC-SAFT) are utilized to have a comparison with artificial intelligence models. Moreover, the sensitivity analysis of input parameters utilizing the relevancy factor is performed to check their impact on the solubility of CO₂-N₂ gas mixtures in aqueous electrolyte systems. Lastly, the Leverage method is applied to investigate the quality of actual data and the reliability of the best-proposed approaches for determining the solubility of the CO₂-N₂ gas mixtures in aqueous systems.

Data gathering

In this study, to estimate the solubility of CO₂-N₂ mixtures in water and aqueous brine solutions, an extensive database containing 289 laboratory data has been collected from the literature^{10,18}, which is presented in the Supplementary file. Although two studies^{1,42} have been performed to estimate the solubility of CO₂, N₂, and CO₂-N₂ mixture in aqueous electrolyte systems using artificial intelligence models, in these studies, the number of data related to the solubility of CO₂-N₂ mixture in water is much less than the data for the two pure substances (i.e., CO₂, N₂). The number of CO₂-N₂ mixture solubility data of these studies^{1,42} is less than 90 data points. The

	IDX	Temperature (K)	Pressure (MPa)	Ionic strength (M)	CO ₂ (mole %)	Solubility (mole fraction)
Mean	1.505	294.13	11.11	0.8158	31.6114	0.004323
SD	0.5008	15.76	5.83	1.1633	29.1251	0.006132
Min	1	273.25	1.51	0	0	0.0001
Max	2	318.15	21.74	3.99	100	0.025

Table 1. Statistical details of the dataset in this work.

database used in this work has 200 data points of CO₂-N₂ mixture solubility in aqueous electrolyte solutions more than Nabipour et al.¹ and Samani et al.⁴² works. Therefore, what distinguishes this study from other previous studies is the use of a large data bank containing a large number of data related to CO₂-N₂ mixture solubility in aqueous brine solutions. Therefore, the results of the developed models can be more comprehensive and reliable for use in the cases mentioned at the beginning of the introduction. To develop the models, temperature, pressure, ionic strength of aqueous solutions, CO₂ mole percent in gaseous mixture, and the index of non-hydrocarbon gases (IDX: 1 = N₂ and 2 = CO₂) whose solubility is to be estimated, have been used as input parameters. The statistical parameters of inputs and output data are summarized in Table 1.

Models' implementation

Support vector machine for regression (SVR). The Support Vector Machine (SVM) is a type of controlled machine learning system that can be employed for both regression (SVR) and classification (SVC) problems⁴⁸. SVM has been widely used in various research areas due to its superior feature, notably in solving non-linear problems called the kernel trick, mapping the input space into a higher-dimensional space. For the sake of conciseness, this article briefly explains the concept of SVR; however, it has extensively been presented in literature⁴⁹. Let the given dataset be a set of n independent samples, $[(x_1, y_1), \dots, (x_n, y_n)]$, where $x \in R_d$ has d dimension and $y \in R$. The objective of SVR is to identify regression function as below:

$$y = f(x) = w \cdot \phi(x_i) + b \quad (1)$$

here w , b , and $\phi(x)$ denote the weight, bias, and kernel function, respectively.

To get the appropriate values of the weight and bias vectors, Vapnik et al.⁵⁰ suggested the following optimization procedure:

$$\begin{aligned} & \text{minimize } \frac{1}{2} w^T w + C \sum_{j=1}^N (\zeta_j^- + \zeta_j^+) \\ & \begin{cases} (w \cdot \phi(x_i) + b) - y_i \leq \varepsilon + \zeta_j^- \\ y_i - (w \cdot \phi(x_i) + b) \leq \varepsilon + \zeta_j^+ \\ \zeta_j^+, \zeta_j^- \geq 0, i = 1, 2, \dots, m \end{cases} \end{aligned} \quad (2)$$

here w^T indicates the transposed matrix, ε is the toleration of error, ζ_j^+ and ζ_j^- are regarded positive variables reflecting the lower and higher excessive variations, respectively, and C interprets a positive regularization factor determining the deviance from ε . By employing the Lagrange multiplier, Eq. (2) can be converted into a dual optimization problem as follows, which makes it easier to solve⁴⁸.

$$y = f(x) = \sum_{i=1}^n (a_i - a_i^*) \cdot K(x_i, x) + b \quad (3)$$

where $K(x_k, x_l)$ is the kernel function, a_k and a_k^* represent the Lagrange multipliers.

It should be noted that in the present study, the polynomial kernel function was used in the SVR model which was selected by using grid search for the best performance. Weight and bias in Eq. (1) stand for trainable variables of SVR model.

Random forest (RF). Decision Trees, a tree-like structure, are easy to interpret and perform well, notably when the dataset is large. However, the problems of the model are twofold. First, the Decision Trees usually experience low prediction bias and high variance, so-called over-fitted, which means the model picks up even small perturbations and random noises in the training dataset. Furthermore, although the most optimum decision is determined at each step, this greedy model does not consider the global optimum; therefore, the overall decision tree might not be optimal. The abovementioned issues can be mitigated by ensembling methods, integrating the results of multiple trees (weak learners) into the final result (strong learner)^{51,52}. Such ensemble learning algorithm in which each tree is trained in parallel forms a Decision Tree ensemble, which is referred to as Random Forests. The greedy strategy in RF determines the importance of each tree at each stage⁵³. Moreover, RF can measure the feature's importance and retain the most informative input features⁵⁴. To improve the variable selection and diversity of the trees, the RF algorithm employs a technique called bagging or bootstrap aggregation. The model will decide how to split the input data into multiple sub-datasets according to the given trees'

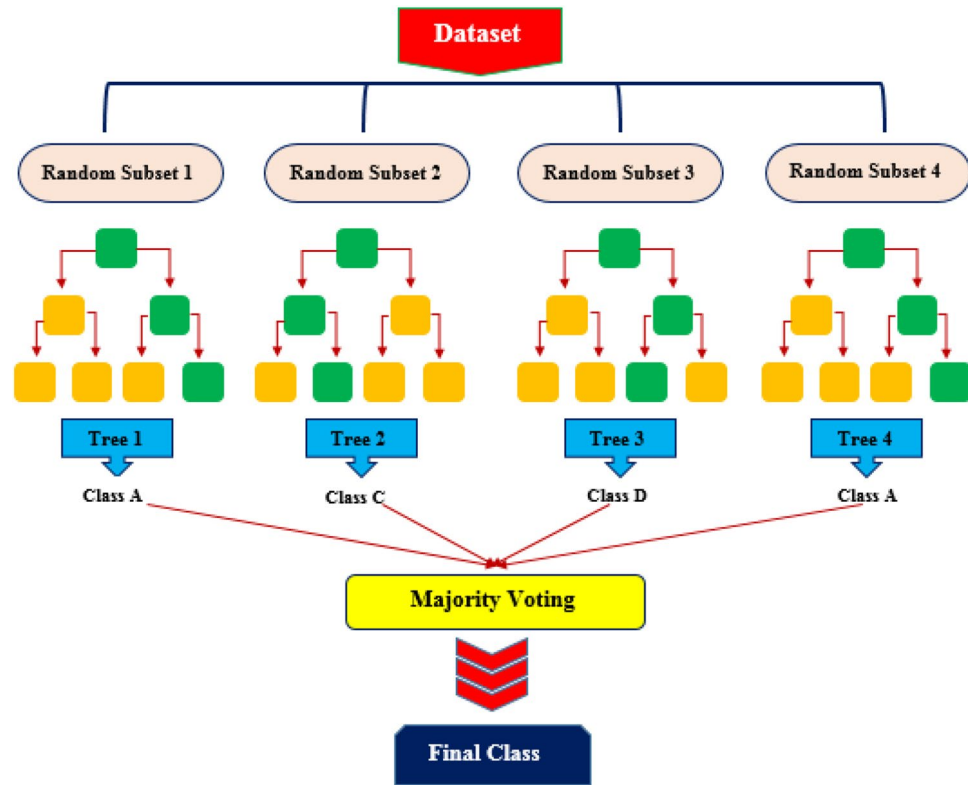


Figure 1. Schematic illustration of random forest algorithm.

population. Bagging, a type of random sampling technique, allocates a third of data for the training purpose of a subtree development process, and the remaining will be left behind, which are referred to as out-of-bag samples. Additionally, the cross-validation technique is unnecessary while using the RF algorithm as multiple bagging in the training process prevents over-fitting⁵⁵. The framework of RF construction is illustrated in Fig. 1.

Suppose D is the training dataset with n number of observations, $D = [(x_1, y_1), (x_2, y_2) \cdots (x_n, y_n)]$, and D_t is the training dataset for the tree h_t , the predicted output corresponding to the out-of-bag dataset of sample x can be expressed as follows⁵⁶:

$$H^{oob}(x) = \operatorname{argmax} \sum_{t=1}^T I(h_t(x)) = y \tag{4}$$

The learning error of the OOB can be obtained by:

$$\varepsilon^{oob}(x) = \frac{1}{|D|} \sum_{(x,y) \in D} I(H^{oob}(x) \neq y) \tag{5}$$

The procedure of RF must be random and this feature can be controlled over a parameter formulated as k^{55} . The significance of a characteristic of a variable X_i could be obtained as follows:

$$I(X_i) = \frac{1}{B} \sum_t^B \widetilde{OOBerr}_i - OOBerr_t \tag{6}$$

where X_i is the i th parameter in vector X , B indicates the current number of trees in the RF, \widetilde{OOBerr}_i denotes the predicted error of the OOB samples for the feature X_i of tree t , and $OOBerr_t$ is the initial OOB samples including permuted variables⁵⁶.

Decision tree (DT). Decision Tree, a nature-inspired supervised learning algorithm, has been widely utilized in the literature and can be used for classification and regression⁵⁷. This algorithm consists of four elements: root node, which is the topmost node in the tree carrying the input data; leaf nodes, which are the final section of the flowchart and denotes the output of the system; internal nodes, which are placed between the root and leaf nodes; branches, which are the connection between nodes. A tree-building process in a decision tree algorithm includes three techniques: splitting, pruning, and stopping⁵⁸. The input data is split into branches and decision nodes starting from the root node. The splitting process carries on till a stopping criterion is convinced. The pruning technique implies removing the low-importance branches⁵⁹. A simple architecture of a DT model is illustrated in Fig. 2.

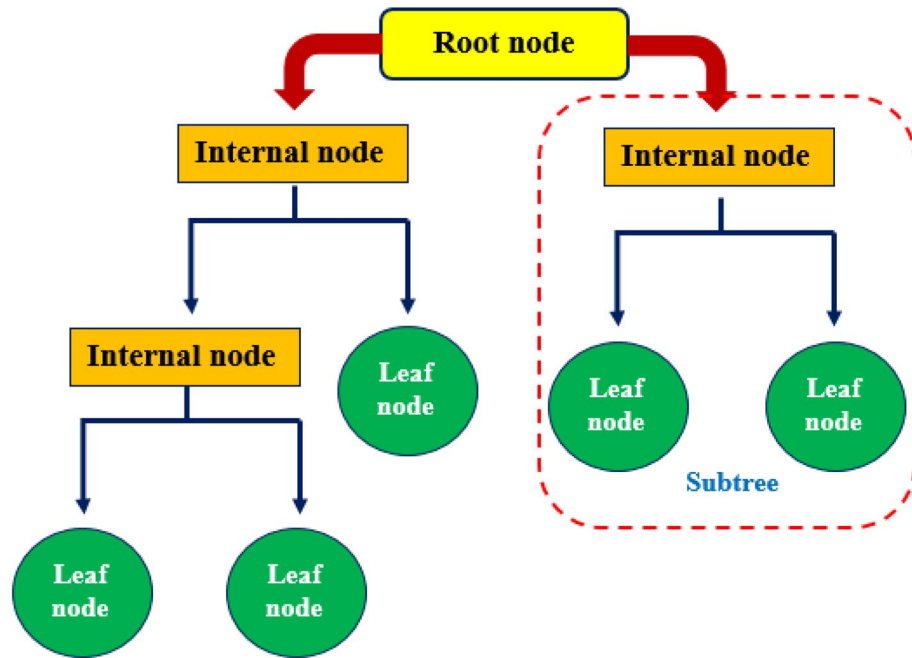


Figure 2. Schematic illustration of a typical decision tree.

Gradient boosting (GB). Gradient Boosting (GB) is an effective machine learning technique that can be used in both regression and classification to reduce bias error or overfitting. Gradient boosting, as functional gradient descent, obtains the residual errors generated from the previous learner, and adds a new learner to it to minimize the loss function of the model at each stage of gradient descent. This technique aims to combine a group of weak learners in a stage-wise manner to build a strong learner and in turn, a more robust model to fit more accurately to the response variable. In other words, the new base-learner must have two conditions: be correlated with the negative gradient of the loss function and also be associated with the whole ensemble. As the idea behind gradient boosting is to minimize the loss function, there is a range of loss functions that can be used. Assume $h(x, \theta)$ is a custom base-learner and $\Psi(y, f)$ a loss function, it is tough to predict the variables and a repetitive model; therefore, is proposed to choose a new function as $h(x, \theta_t)$, where the t enhancement is directed by^{60,61}:

$$g_t(x) = E_y \left[\frac{\partial \Psi(y, f(x))}{\partial f(x)} \middle| x \right]_{f(x)=\tilde{f}^{t-1}(x)} \tag{7}$$

This converts a potential sophisticated optimization problem into a classic least square minimization^{60,62}.

$$(\rho_t, \theta_t) = \underset{\rho, \theta}{\operatorname{argmin}} \sum_{i=1}^N [-g_t(x_i) + \rho h(x_i, \theta)]^2 \tag{8}$$

The following are the steps in the GBDT technique process⁶³:

- Suppose that \hat{f}_0 is a constant
- Evaluate the $g_i(x)$ and training $h(x_i, \theta)$ function
- Obtain parameter ρ_i and modify the function:

$$\hat{f}_i = \hat{f}_{i-1} + \rho_i h(x_i, \theta) \tag{9}$$

The method starts with a single leaf and optimizes the training algorithm for each node and record. Figure 3 shows a schematic example of a conventional GBDT.

Adaptive boosting (AdaBoost). The adaptive boosting algorithm presented by Freund and Schapire⁶⁴ aims to combine weak classifiers and learn from their mistake to create a strong classifier. In other words, it selects the training dataset iteratively to combine the multiple classifiers and assign the appropriate weight to each classifier based on the accuracy of each classifier so that higher weights are assigned to the misclassified/mislabeled samples⁶⁵. The following are the general stages of the AdaBoost technique^{66,67}:

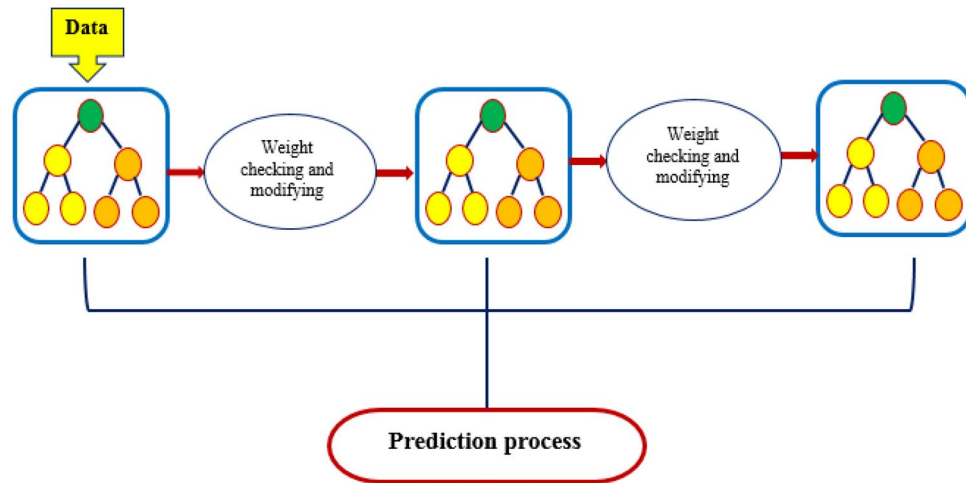


Figure 3. Schematic illustration of a typical GBDT.

- Weights definition: $w_j = \frac{1}{n}$, $j = 1, 2, \dots, n$
- Apply the training data to a weak learner $Wl_i(x)$, weights, and obtain the weighted error for each i .

$$I(x) = \begin{cases} 0 & \text{if } x = \text{false} \\ 1 & \text{if } x = \text{true} \end{cases} \quad (10)$$

$$Err_i = \frac{\sum_{j=1}^n w_j I(t_j \neq w_l_i(x))}{\sum_{j=1}^n w_j} \quad (11)$$

- Determine the weights for predictors for each i as follows:

$$\beta_i = \log\left(\frac{1 - Err_i}{Err_i}\right) \quad (12)$$

- Update the sample weights for each i to N (where N is the learner's number)
- Assign a weak learner to the data test (x) as a result.

Support vector regressors (SVR) and Decision Trees (DT) have been used as weak learners in AdaBoost systems in this study.

In this paper, we have applied ensemble models such as Adaboost-DT, Adaboost-DT, and GB-SVR. To discover the functionality and different possibilities of regression methods, AdaBoost and Gradient boosting as varieties of clustering methods have been executed to enhance the conventional weak regressors by incorporating the outcome of the weak regressors into a weighted combination that determines the best output of the enhanced powerful regressor and also the outcome of the weak regressors is distorted in pursuit of incorrectly estimated samples autonomously.

More details are as follows:

Linear SVR indistinguishability is achieved by using a nonlinear imaging approach to convert features with linearly unidentifiable low-dimensional input space into a high-dimensional feature space. This allows the non-linear features of the samples to be analyzed linearly using a linear algorithm in a high-dimensional feature space. However, the choice of kernel functions and parameters has a significant impact on its performance⁶⁸. The AdaBoost method trains many base learners, and the sample generalization could be further improved by combining techniques to produce the final strong learner. Anomaly samples are susceptible to the AdaBoost method, and anomalous samples may obtain greater weights in iterations, affecting the prediction accuracy of strong learners. Furthermore, the decision tree is widely used as a basic learning method, but it is inadequate in dealing with nonlinear issues, and prediction accuracy varies substantially⁶⁹. The AdaBoost method, on the other hand, is sensitive to anomalous data, and anomalous samples may obtain greater weights in iterations, affecting strong learners' prediction accuracy.

When using SVR for sample learning, the model's performance is determined by the kernel function and kernel parameters. Using SVR as the AdaBoost base learner, on the other hand, lowers the influence of the SVR algorithm's kernel functions and parameters. It also overcomes AdaBoost's standard algorithm's inability to address nonlinear issues. This makes the AdaBoost-SVR method appropriate for dealing with nonlinear feature data prediction while also ensuring the model's generalizability⁷⁰. We combined GB and SVR algorithms⁷¹. The combined GB and SVR algorithm into a single predictive model is another meta-algorithm applied in this paper in order to enhance the overall performance. Gradient Boosting as part of an ensemble technique attempts to create a strong regressor from several weak regressors.

EOS	PVT relation	Parameters	Reference
SRK	$P = \frac{RT}{v-b} - \frac{a_c \alpha}{v(v+b)}$	$a_c = 0.42747 \frac{R^2 T_c^2}{P_c}$ $b = 0.08664 \frac{RT_c}{P_c}$ $m = 0.480 + 1.574\omega - 0.176\omega^2$ $\alpha = [1 + m(1 - \sqrt{T_r})]^2$	76
PR	$P = \frac{RT}{v-b} - \frac{\alpha}{v(v+b)+b(v-b)}$	$\alpha = a_c \alpha(T)$ $\alpha(T) = [1 + m(1 - \sqrt{T_r})]^2$ $a_c = 0.45724 \frac{(RT_c)^2}{P_c}$ $m = 0.37464 + 1.54226\omega - 0.26992\omega^2$ $b = 0.07780 \frac{RT_c}{P_c}$	76
VPT	$P = \frac{RT}{v-b} - \frac{a(T)}{v(v+b)+c(v-b)}$	$a(T) = a^C + a^A$ $a^C = \sum_i \sum_j x_i x_j \sqrt{a_i a_j} (1 - k_{ij})$ $a^A = \sum_p x_p^2 \sum_i x_i a_{pi} l_{pi}$ $a_{pi} = \sqrt{a_p a_i}$ $l_{pi} = l_{pi}^0 - l_{pi}^1 (T - 273.15)$	77-79
PC-SAFT	$\tilde{a} = \frac{A}{kT} = \tilde{a}^{hc} + \tilde{a}^{id} + \tilde{a}^{disp} + \tilde{a}^{assoc}$	$\tilde{a}^{hc} = \overline{m} \tilde{a}^{hs} + \tilde{a}^{chain} = \overline{m} \tilde{a}^{hs} - \sum_i x_i (m_i - 1) \ln g_{ij}^{hs}$ $\overline{m} = \sum_i x_i m_i$ $\tilde{a}^{hs} = \frac{1}{\zeta_0} \left[\frac{3\zeta_1 \zeta_2}{1 - \zeta_3} + \frac{3\zeta_2^3}{\zeta_3(1 - \zeta_3)^2} + \left(\frac{\zeta_2^3}{\zeta_3^2} - \zeta_0 \right) \ln(1 - \zeta_3) \right]$ $\zeta_n = \frac{\pi}{6} \rho \sum_i x_i m_i d_i^n, n \in \{0, 1, 2, 3\}, \eta = \zeta_3$ $d_i = \sigma_i \left[1 - 0.12 \exp\left(-3 \frac{\varepsilon_i}{kT}\right) \right]$ $g_{ij}^{hs} = \frac{1}{1 - \zeta_3} + \left(\frac{d_i d_j}{d_i + d_j} \right) \frac{2\zeta_2}{(1 - \zeta_3)^2} + \left(\frac{d_i d_j}{d_i + d_j} \right)^2 \frac{2\zeta_2^2}{(1 - \zeta_3)^2}$ $\tilde{a}^{dis} = -2\pi \rho I_1(\eta, \overline{m}) \overline{m}^2 \varepsilon \sigma^3 - \pi \rho \overline{m} C_1(\eta, \overline{m}) I_2(\eta, \overline{m}) \overline{m}^2 \varepsilon^2 \sigma^3$ $I_1(\eta, \overline{m}) = \sum_{i=0}^6 a_i(\overline{m}) \eta^i, I_2(\eta, \overline{m}) = \sum_{i=0}^6 b_i(\overline{m}) \eta^i$ where a_i and b_i depend on the chain length as given in Gross and Sadowski ⁸⁰ $C_1 = \left[1 + \overline{m} \frac{8\eta - 2\eta^2}{(1 - \eta)^4} + (1 - \overline{m}) \frac{20\eta - 27\eta^2 + 12\eta^3 - 2\eta^4}{[(1 - \eta)(2 - \eta)]^2} \right]$ $\overline{m}^2 \varepsilon \sigma^3 = \sum_i \sum_j x_i x_j m_i m_j \left(\frac{\varepsilon_{ij}}{kT} \right) \sigma_{ij}^3$ $\overline{m}^2 \varepsilon^2 \sigma^3 = \sum_i \sum_j x_i x_j m_i m_j \left(\frac{\varepsilon_{ij}}{kT} \right)^2 \sigma_{ij}^3$ $\varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j} (1 - k_{ij})$ $\sigma_{ij} = \frac{(\sigma_i + \sigma_j)}{2}$ The formulation for the contributions from the dispersion and ideal gas are similar to those of Gross and Sadowski ⁸⁰	80,81

Table 2. EOSs relationships and parameters.

Equations of state (EOS). An EOS is a mathematical representation that connects system parameters to represent the state of a material under a range of predefined circumstances, including pressure, temperature, or volume⁷². These thermodynamic models can characterize the thermal characteristics and volumetric behavior of mixtures and pure materials⁷³. During the last few decades, cubic EOSs have been widely employed. New EOSs, like various forms of the Statistical Associating Fluid Theory (SAFT), have been applied successfully in the past

Substance	M_w (g/mol)	m	σ (Å)	ϵ/k (K)	Reference
N ₂	28.013	1.2053	3.313	90.96	⁸⁰
CO ₂	44.01	2.0729	2.7852	169.21	⁸⁰
H ₂ O	18.015	2	2.3533	207.84	⁸²

Table 3. PC-SAFT EOS factors for the substances utilized in this paper.

Substance	P_c (MPa)	T_c (K)	Z_c	ω
N ₂	3.394	126.10	0.2917	0.0403
CO ₂	7.382	304.19	0.2744	0.2276
H ₂ O	22.055	647.13	0.2294	0.3449

Table 4. Critical properties and acentric factors utilized in the EOSs for the substances used in this paper⁷⁹.

few years. To explain the interactions between the molecules in a system, the SAFT EOSs were constructed using statistical mechanics^{74,75}. SAFT EOSs are designed to depict molecules as chains of spherical particles that engage with others via long-range attraction, short-range repulsion, and hydrogen bonding at particular places. In this study, four equations of state such as PR, SRK, VPT, and PC-SAFT, have been used. The PVT relationships and parameters of the respective equations of state are reported in Tables 2 and 3. The critical properties and acentric coefficients of the materials utilized in this study are summarized in Table 4.

Performance analysis of models

The mathematical description of the statistical parameters employed in this study are summarized below^{72,83}:

- Average absolute percent relative error (AAPRE)

$$AAPRE = \frac{1}{N} \sum_{i=1}^N |(S_{iEXP} - S_{iPRED})/S_{iEXP}| \times 100 \quad (13)$$

- Standard deviation (SD)

$$SD = \sqrt{\frac{1}{N-1} \sum_{i=1}^N \left(\frac{S_{iEXP} - S_{iPRED}}{S_{iEXP}} \right)^2} \quad (14)$$

- Coefficient of determination (R²)

$$R^2 = 1 - \frac{\sum_{i=1}^N (S_{iEXP} - S_{iPRED})^2}{\sum_{i=1}^N (S_{iEXP} - \overline{S_{iEXP}})^2} \quad (15)$$

- Root mean square error (RMSE)

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (S_{iEXP} - S_{iPRED})^2} \quad (16)$$

In the above equations, S_{iEXP} , S_{iPRED} , $\overline{S_{iEXP}}$, and N refer to experimental solubility, predicted solubility, mean experimental solubility, and the total number of data points, respectively.

Also, several graphical analyses, namely, cross-plot, relative error distribution diagram, cumulative frequency plot, and trend plot were utilized to visually evaluate the developed models. Descriptions of these analyses can be found elsewhere⁷².

Results and discussion

Statistical evaluation of models. The models discussed in the previous sections have been developed to predict the solubility of CO₂-N₂ mixtures in water utilizing 289 laboratory data. In this study, we have employed six algorithms, which were rarely used, to estimate CO₂-N₂ gas mixture solubility in water. The structure of the models was modified and also the grid search algorithm was used to optimize the hyperparameters of the models to avoid overfitting in this particular problem. The hyperparameters obtained by the grid search are different for each model. It is based on the importance of the hyperparameters according to theoretical and practical aspects. Total data has been divided randomly to 80/20 for the training and testing phase. It should be noted that experimental data and predictions of different models are presented in the Supplementary file. The calculated

Statistical criteria		RMSE	SD	R ²	AAPRE (%)
DT	Train	0.000297	0.1044	0.9978	6.1904
	Test	0.000290	0.3172	0.9965	12.3069
	Total	0.000295	0.1721	0.9977	7.4179
GB-DT	Train	0.000166	0.2324	0.9993	10.5323
	Test	0.000155	0.3973	0.9991	15.3978
	Total	0.000164	0.2745	0.9992	11.5088
AdaBoost-DT	Train	0.000217	0.2331	0.9988	12.5086
	Test	0.000204	0.2901	0.9985	13.7655
	Total	0.000214	0.2457	0.9987	12.7609
AdaBoost-SVR	Train	0.000161	0.2220	0.9993	9.5464
	Test	0.000147	0.2076	0.9992	9.6933
	Total	0.000159	0.2192	0.9993	9.5759
GB-SVR	Train	0.000300	0.1120	0.9977	6.7068
	Test	0.000290	0.0716	0.9970	5.3403
	Total	0.000298	0.1051	0.9976	6.4326
Random forest	Train	0.000132	0.0740	0.9995	2.9086
	Test	0.000131	0.0608	0.9994	2.5999
	Total	0.000132	0.0716	0.9995	2.8466

Table 5. Calculated statistical criteria for the proposed models.

statistical parameters for the represented models are summarized in Table 5. In this table, different statistical parameters such as RMSE, R², SD, and AAPRE are reported. GB-SVR outperforms other models except for Random Forest because SVR is more like a soft fabric that can bend and fold in whatever way we need to better fit our data. This gives more degrees of freedom and flexibility so that a more accurate model can be achieved. Moreover, SVR can capture the non-linear relationships between variables. The performance of the model is further improved by tuning the hyperparameters. These are the main reasons that GB-SVR has shown a higher accuracy. Random Forest proved the highest accuracy in this study even higher than SVR-GB. Random Forest is built for multiclass issues, whereas SVM is for two-class problems. In SVM, in the case of a multiclass problem, the problem must be broken down into numerous binary classification tasks. With a combination of numerical and categorical variables, Random Forest performs well. Also, in classification problems, it is not necessary to do normalization or scaling in Random Forest. SVM seeks to maximize the "margin," relying on the idea of "distance" between points. It is up to us to decide if "distance" is significant. As a consequence, one-hot encoding for categorical features is a must-do. Further, min-max or other scaling is highly recommended in preprocessing step. Random forests are good for a specific set of issue types when given a specific set of data, but they do not act well for many others. We should mention that random forests are unexpectedly effective for a wide range of issues because they are built on trees, the variables cannot be scaled. A tree inherently captures any monotonic alteration of a single variable, and in random forest built-in feature selection is automated⁸⁴.

According to Table 5, it can be seen that the Random Forest model with an AAPRE value of 2.84% has the most accurate prediction for the solubility of CO₂-N₂ mixtures in water. The GB-SVR and DT models with AAPRE values of 6.43% and 7.41%, respectively, have the closest prediction to the Random Forest model compared to other models. However, it should be noted that other models also have relatively good results. Another noteworthy point is that sometimes the high accuracy of a model in predicting outputs may be due to over-training. In order to ensure that this does not happen, the results of training and test data should be compared with each other. If the difference between the statistical parameters of the training and test data is significant, the model may be over-trained. If the results of the training and test data are close to each other, it can be stated that over-training has not happened. As the results show, the statistical parameters for the training and test data are very close.

To evaluate the performance of artificial intelligence methods in comparison with mathematical methods, four equations of state such as SRK, PR, VPT, and PC-SAFT, have been used. For this purpose, the solubility of CO₂ and N₂ in different CO₂ + N₂ + H₂O (brine) systems was calculated using 24 laboratory data points extracted from the literature¹⁰, and the results are reported in Table 6 and Table 7. As shown in Tables 6 and 7, the value of AAPRE obtained for the SRK and PR equations of state is much higher than VPT and PC-SAFT equations of state and the intelligent models. For solubility of CO₂ in aqueous solutions, the Random Forest approach outperforms the other intelligent techniques with an AAPRE value of 1.16%, and the PC-SAFT model has the best results among the EOSs with an AAPRE value of 3.35%. For solubility of N₂ in aqueous solutions, the Random Forest technique has the best results among the intelligent approaches with an AAPRE value of 4.13%, and the VPT model has the best results among the EOSs with an AAPRE value of 5.71%.

Graphical analysis of models. Figure 4 shows the cross-plot diagrams for the six models presented in this study. In this graph, where the predicted results are plotted against actual values, the higher the compaction of data around the Y = X line indicates that the estimated values are closer to the actual values; therefore, the model

Solubility system	Data no.	P (MPa)	CO ₂ solubility (mole fraction)										
			Exp	DT	GB-DT	AdaBoost-DT	AdaBoost-SVR	GB-SVR	Random Forest	SRK	PR	VPT	PC-SAFT
CO ₂ (14.6%) + N ₂ (85.4%) + H ₂ O, at 303.05 K	1	1.98	0.0008	0.000791	0.000800	0.000900	0.000900	0.000816	0.000800	0.0010	0.0013	0.0009	0.0008
	2	5.63	0.0022	0.002101	0.002200	0.002150	0.002200	0.002095	0.002200	0.0023	0.0032	0.0023	0.0021
	3	9.35	0.0031	0.003384	0.003350	0.003542	0.003460	0.003258	0.003050	0.0033	0.0041	0.0033	0.0031
	4	13.17	0.0039	0.003888	0.003900	0.003975	0.003900	0.003794	0.003750	0.0039	0.0049	0.0043	0.0038
	5	16.97	0.0045	0.004417	0.004500	0.004500	0.004500	0.004444	0.004650	0.0045	0.0061	0.0048	0.0043
	6	20.75	0.0048	0.004717	0.004575	0.004560	0.004575	0.004718	0.004650	0.0051	0.0066	0.0053	0.0046
CO ₂ (3%) + N ₂ (97%) + H ₂ O, at 283.15 K	7	2.05	0.0003	0.000405	0.000400	0.000780	0.000400	0.000410	0.000300	0.0002	0.0003	0.0003	0.0003
	8	5.74	0.0008	0.001142	0.000800	0.000900	0.000800	0.001135	0.000800	0.0005	0.0007	0.0008	0.0008
	9	9.84	0.0012	0.001258	0.001400	0.001400	0.001400	0.001241	0.001200	0.0007	0.0010	0.0013	0.0012
	10	13.58	0.0014	0.001501	0.001600	0.001700	0.001500	0.001477	0.001400	0.0008	0.0012	0.0015	0.0014
	11	18.06	0.0017	0.001739	0.001700	0.001750	0.001800	0.001739	0.001700	0.0010	0.0014	0.0018	0.0016
	12	21.5	0.0018	0.001843	0.001800	0.001800	0.001800	0.001828	0.001800	0.0011	0.0016	0.0021	0.0017
CO ₂ (61%) + N ₂ (39%) + H ₂ O, at 303.05 K	13	1.92	0.0024	0.003120	0.002400	0.002400	0.002400	0.003257	0.002400	0.0032	0.0057	0.0026	0.0024
	14	5.59	0.0064	0.007550	0.006400	0.006400	0.006400	0.008086	0.006400	0.0093	0.0099	0.0065	0.0062
	15	9.06	0.0091	0.009971	0.009100	0.009100	0.009100	0.009878	0.009100	0.0113	0.0124	0.0091	0.0087
	16	13.3	0.0112	0.011830	0.011200	0.011400	0.011467	0.011521	0.011200	0.0133	0.0144	0.0113	0.0107
	17	17.03	0.012	0.012651	0.012000	0.012000	0.012000	0.012598	0.012250	0.0142	0.0178	0.0125	0.0117
	18	20.99	0.0125	0.013007	0.012500	0.012500	0.012500	0.012988	0.012250	0.0156	0.0201	0.0133	0.0123
CO ₂ (14.6%) + N ₂ (85.4%) + brine (10 wt.% NaCl), at 273.25 K	19	2	0.0013	0.001176	0.001300	0.001200	0.001250	0.001186	0.001300	0.0011	0.0016	0.0013	0.0013
	20	5.59	0.0033	0.003190	0.003225	0.003175	0.003300	0.003184	0.003300	0.0025	0.0036	0.0032	0.0031
	21	9.42	0.0047	0.004564	0.004665	0.004462	0.004733	0.004561	0.004700	0.0034	0.0048	0.0046	0.0044
	22	13.26	0.0056	0.005460	0.005600	0.005428	0.005600	0.005432	0.005900	0.0041	0.0055	0.0055	0.0052
	23	17.12	0.0061	0.005996	0.006000	0.005800	0.006002	0.005914	0.006500	0.0053	0.0058	0.0063	0.0056
	24	21.08	0.0065	0.006276	0.006079	0.005900	0.006129	0.006212	0.006500	0.0058	0.0061	0.0066	0.0059
AAPRE, %	–	–	8.71	3.67	11.87	4.42	9.03	1.16	23.34	30.18	5.08	3.35	

Table 6. Predictions of EOSs and smart models for CO₂ solubility in different CO₂ + N₂ + H₂O (brine) systems.

is more accurate. In addition, R² value for this dataset will close to 1. As shown in Fig. 4, the Random Forest model is in a better position than the other models, which also confirms the results reported in Table 5.

Figure 5 shows the error distribution diagram for the developed models. This diagram shows the relative error on the Y-axis and the experimental data on the X-axis. The closer and the more compaction of the points around the zero line, the less the predicted data error. On the other hand, according to this diagram, the relative error range for experimental data can be visually observed. For example, it can be seen how the relative error will change as the value of experimental data increases. As shown in Fig. 5, it can be observed that the Random Forest model is in a better condition and shows relatively lower errors than other models.

A cumulative frequency graph is one of the most important diagrams that can be used to compare the performance of several models simultaneously. Figure 6 shows a cumulative frequency diagram for different models. In this diagram, which is a cumulative frequency of the number of data in terms of absolute relative error, the higher the curve of one model than the curve of other models, the higher the accuracy. In other words, if a model's curve is higher than another model's curve in a constant AAPRE value, it means that a higher percentage of the data in that model has a lower absolute relative error than another model. The higher the curve of one model at small absolute relative errors (close to 1), the higher the percentage of that data, the lower the absolute relative error, and the more accurate the model. Therefore, according to Fig. 6 and what is said, the Random Forest model is in a better situation than other models and has a higher accuracy, which also confirms the results presented in Table 5.

Trend analysis. Investigating the trend of solubility changes in terms of different parameters can give us a better understanding of the solubility of CO₂-N₂ mixture in water and brine solutions. On the other hand, the validity of the developed models can be investigated by comparing the trend of measured changes with laboratory data, equations of state, and predicted data. For example, when an input parameter shows an increasing trend in experimental data, the developed models should show the same trend. In this case, the validity of the developed model will be more. In the following, we examined the trend analysis of various parameters.

Figure 7 shows the effect of pressure on the solubility of CO₂ and N₂ in an aqueous system consisting of 39% N₂ and 61% CO₂ at 283 K. In this figure, the changes in solubility in terms of pressure using laboratory and predicted data in the Random Forest model as the best model and equations of state were investigated. According to Fig. 7a and b, all methods show an incremental trend. What is debatable in this figure is the degree to which the models are overestimated and underestimated. Another noteworthy point is the perfect agreement of the Random Forest model data with the experimental data, which confirms the efficiency of the intelligent models. As shown in Fig. 7a, the curves related to the equations of state are generally in a higher position than the curve of

Solubility system	Data no.	P (MPa)	N ₂ solubility (mole fraction)											
			Exp	DT	GB-DT	AdaBoost-DT	AdaBoost-SVR	GB-SVR	Random Forest	SRK	PR	VPT	PC-SAFT	
CO ₂ (14.6%) + N ₂ (85.4%) + H ₂ O, at 303.05 K	1	1.98	0.0002	0.000196	0.000300	0.000300	0.000300	0.000300	0.000195	0.000200	0.0002	0.0003	0.0002	0.0002
	2	5.63	0.0005	0.000498	0.000611	0.000600	0.000500	0.000500	0.000486	0.000660	0.0005	0.0007	0.0005	0.0005
	3	9.35	0.0008	0.000802	0.000900	0.000900	0.000900	0.000900	0.000819	0.000750	0.0007	0.0011	0.0008	0.0007
	4	13.17	0.0011	0.001034	0.001200	0.001100	0.001100	0.001100	0.001044	0.001000	0.0010	0.0013	0.0011	0.0009
	5	16.97	0.0013	0.001269	0.001350	0.001367	0.001300	0.001300	0.001223	0.001300	0.0013	0.0016	0.0014	0.0011
	6	20.75	0.0015	0.001471	0.001500	0.001500	0.001500	0.001500	0.001485	0.001500	0.0015	0.0019	0.0016	0.0013
CO ₂ (3%) + N ₂ (97%) + H ₂ O, at 283.15 K	7	2.05	0.0003	0.000331	0.000400	0.000320	0.000300	0.000300	0.000338	0.000300	0.0002	0.0004	0.0003	0.0003
	8	5.74	0.0008	0.000790	0.000800	0.000750	0.000800	0.000793	0.000748	0.0004	0.0009	0.0008	0.0007	0.0007
	9	9.84	0.0012	0.001235	0.001213	0.001200	0.001200	0.001200	0.001247	0.001200	0.0007	0.0012	0.0012	0.0011
	10	13.58	0.0016	0.001632	0.001600	0.001600	0.001600	0.001600	0.001625	0.001600	0.0009	0.0018	0.0016	0.0014
	11	18.06	0.002	0.002035	0.002000	0.002000	0.002000	0.002000	0.002050	0.002000	0.0011	0.0024	0.002	0.0018
	12	21.5	0.0023	0.002332	0.002300	0.002100	0.002300	0.002300	0.002329	0.002300	0.0013	0.0027	0.0023	0.0020
CO ₂ (61%) + N ₂ (39%) + H ₂ O, at 303.05 K	13	1.92	0.0001	0.000157	0.000307	0.000200	0.000200	0.000200	0.000163	0.000100	0.0001	0.0002	0.0002	0.0001
	14	5.59	0.0004	0.000430	0.000600	0.000600	0.000600	0.000440	0.000433	0.000354	0.0003	0.0005	0.0004	0.0004
	15	9.06	0.0006	0.000613	0.000600	0.000600	0.000600	0.000600	0.000640	0.000560	0.0004	0.0008	0.0006	0.0006
	16	13.3	0.0008	0.000816	0.000800	0.000800	0.000800	0.000800	0.000842	0.000727	0.0006	0.0011	0.0009	0.0008
	17	17.03	0.001	0.001079	0.001000	0.001000	0.001000	0.001000	0.001019	0.001000	0.0008	0.0013	0.0011	0.0009
	18	20.99	0.0012	0.001268	0.001200	0.001200	0.001200	0.001200	0.001265	0.001200	0.0010	0.0017	0.0012	0.0011
CO ₂ (14.6%) + N ₂ (85.4%) + brine (10 wt.% NaCl), at 273.25 K	19	2	0.0002	0.000172	0.000200	0.000200	0.000200	0.000200	0.000171	0.000200	0.0001	0.0003	0.0002	0.0001
	20	5.59	0.0004	0.000420	0.000500	0.000484	0.000400	0.000424	0.000328	0.0003	0.0007	0.0004	0.0004	0.0004
	21	9.42	0.0006	0.000615	0.000650	0.000800	0.000600	0.000608	0.000600	0.0005	0.0009	0.0006	0.0006	0.0006
	22	13.26	0.0008	0.000776	0.000800	0.000850	0.000800	0.000776	0.000800	0.0007	0.0011	0.0008	0.0008	0.0008
	23	17.12	0.001	0.000977	0.001000	0.001100	0.001000	0.000979	0.001000	0.0009	0.0014	0.001	0.0009	0.0009
	24	21.08	0.0012	0.001134	0.001200	0.001200	0.001200	0.001158	0.001200	0.0011	0.0016	0.0012	0.0012	0.0010
AAPRE, %	-	-	-	6.13	17.59	13.74	7.18	6.81	4.13	21.72	35.18	5.71	8.78	

Table 7. Predictions of EOSs and smart models for N₂ solubility in different CO₂ + N₂ + H₂O (brine) systems.

the experimental data, and this indicates that these equations overestimate the solubility of CO₂ in the mentioned system. Figure 7b also shows the conformity of the data curve predicted by the Random Forest model with the experimental data, but the different point is that the PR EOS overestimates the solubility of N₂ in the mentioned system and other models underestimate although the degree of agreement of the VPT EOS to the experimental data is significant. Again, for solubility of CO₂ present in gaseous mixtures in aqueous systems, the PC-SAFT model, and for solubility of N₂, the VPT model had the best results among the EOSs.

Figure 8 shows the effect of CO₂ content in the gas mixture for the solubility of CO₂ and N₂ in an aqueous system containing CO₂ and N₂ at a temperature of 308 K and pressure of 8 MPa, as experimentally investigated in the literature¹⁸. As expected, increasing the amount of CO₂ in the gas mixture reduces the solubility of N₂ in the system and, conversely, increases the solubility of CO₂ at constant temperature and pressure. As it is clear, the solubility of N₂ in water is less than that of CO₂.

Figure 9 shows the effect of pressure on the solubility of CO₂ and N₂ in a system containing 85.4% N₂ and 14.6% CO₂ in water at 303 K for the Random Forest model and laboratory data¹⁰. As shown in Fig. 9, increasing the pressure can have a positive effect on increasing the solubility of both CO₂ and N₂ in the system, although this effect is more significant for CO₂.

Figure 10 shows the effect of pressure on the solubility of CO₂ and N₂ in aqueous systems with different salinity (pure water, 5% NaCl brine, and 15% NaCl brine). What can be seen in both Fig. 10a and b is the effect of salinity on system performance. For both CO₂ and N₂ gases, increasing the pressure increases the solubility, but it is noteworthy that increasing the salinity decreases the solubility of CO₂ and N₂. Therefore, increasing the concentration of NaCl in water, or in other words, an increase in the ionic strength of the solution, reduces the solubility of CO₂ and N₂. The salting-out phenomenon causes a reduction in CO₂ and N₂ solubility in water. The electrolytes influence water to dissolve less gas in this process. As salinity increases, more water molecules are attracted to the salt ions, reducing the amount of H⁺ and O₂⁻ ions available to gather and separate the gas molecules, lowering CO₂ and N₂ solubility in the water⁸⁵.

Input parameters impact analysis. To study the influence of input parameters on the output of the model, a parameter called Relevancy factor was used. Relevancy factor is calculated as follows⁸⁶:

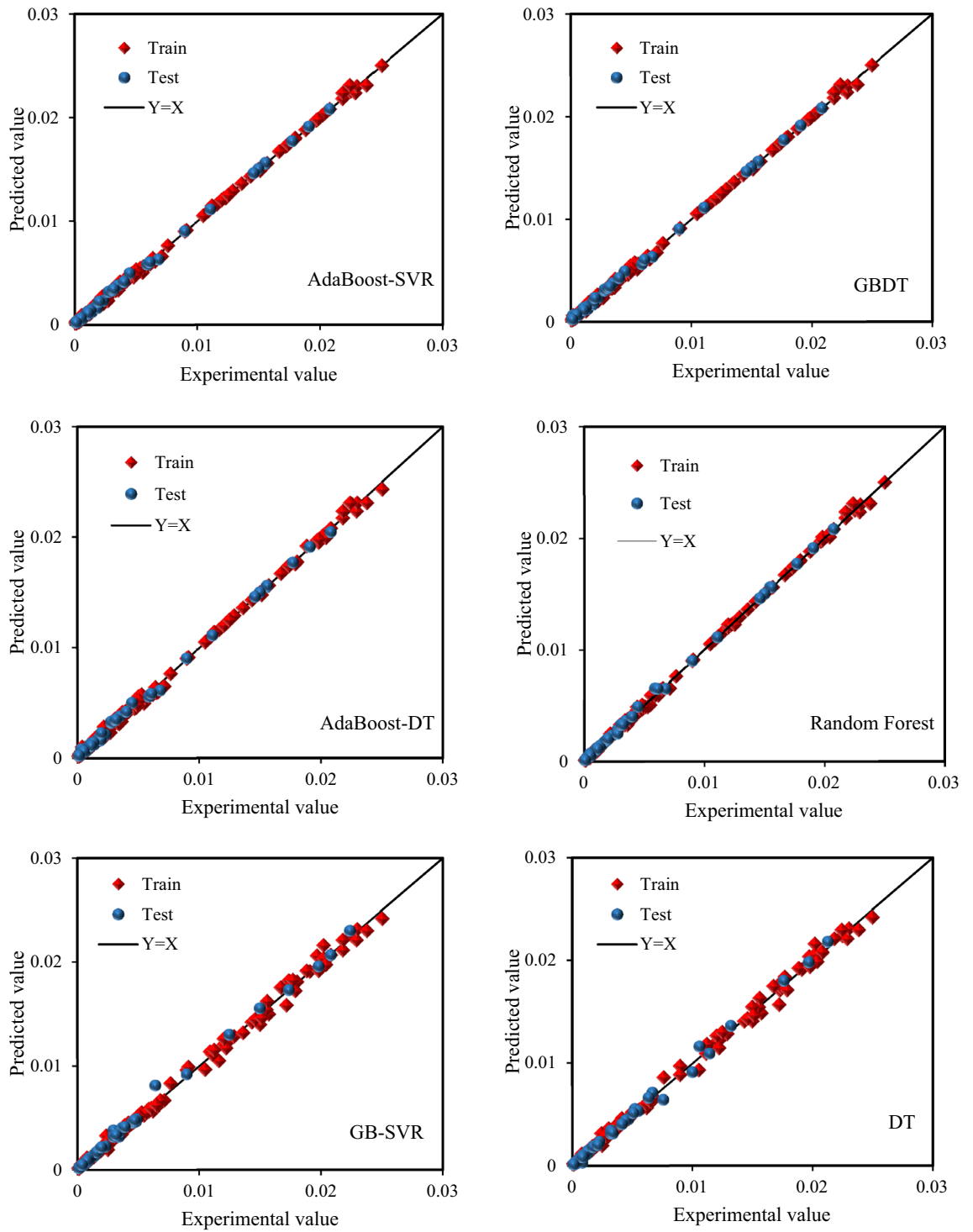


Figure 4. Cross plots of the developed models in this study.

$$r(inp_i, O) = \frac{\sum_{j=1}^n (inp_{i,j} - inp_{m,i})(O_j - O_m)}{\sqrt{\sum_{j=1}^n (inp_{i,j} - inp_{m,i})^2 \sum_{j=1}^n (O_j - O_m)^2}} \quad (17)$$

Here, $inp_{m,i}$ and $inp_{i,j}$ indicate the average value, and the j th value of the i th input, respectively O_j refers to the j th value of predicted output, and O_m is the average of output data.

This parameter, which is between 1 and -1, shows the effect of inputs on the output of the model as follows⁸⁷:

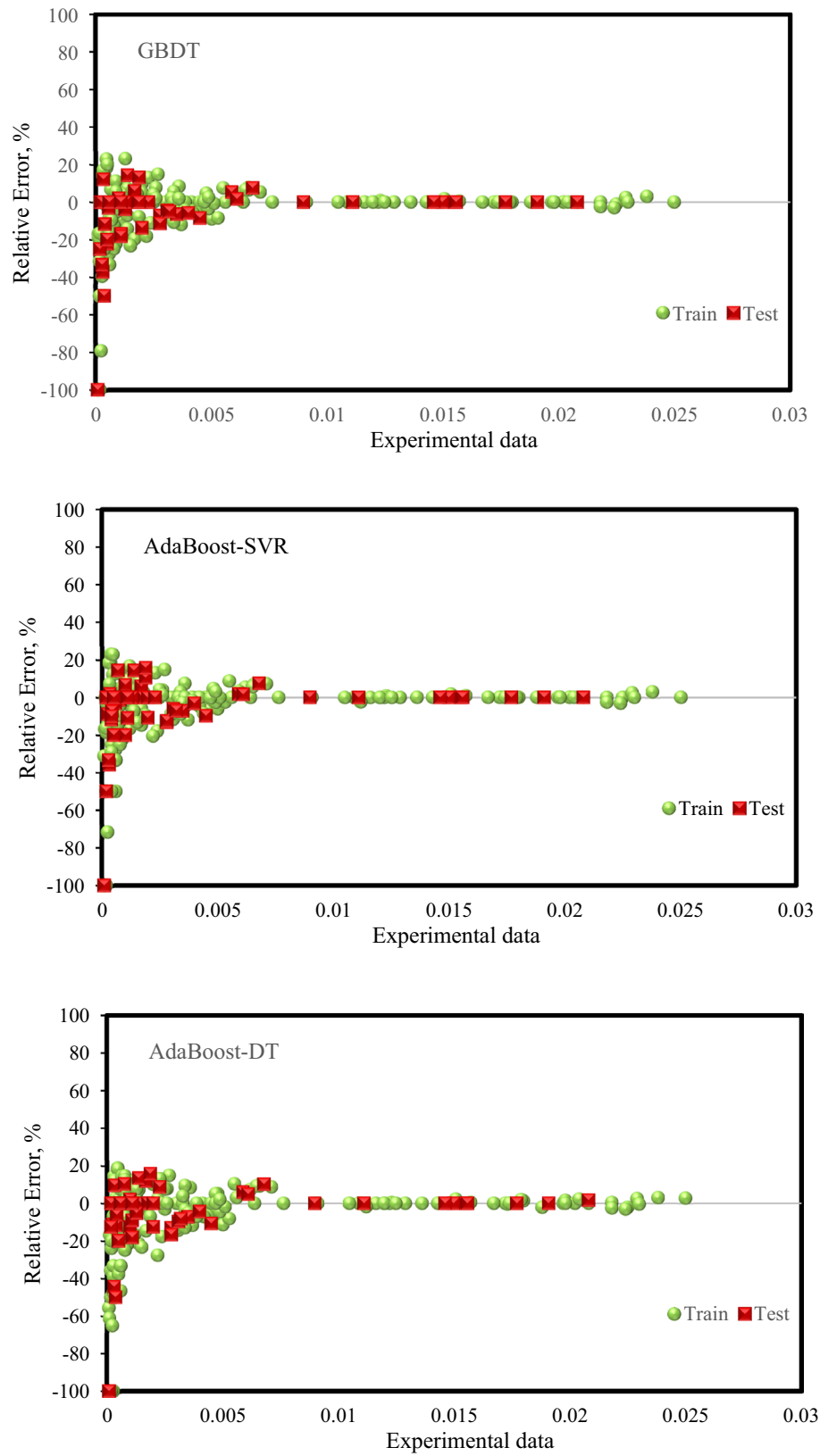


Figure 5. Error distribution plots of the developed models for training and test sets.

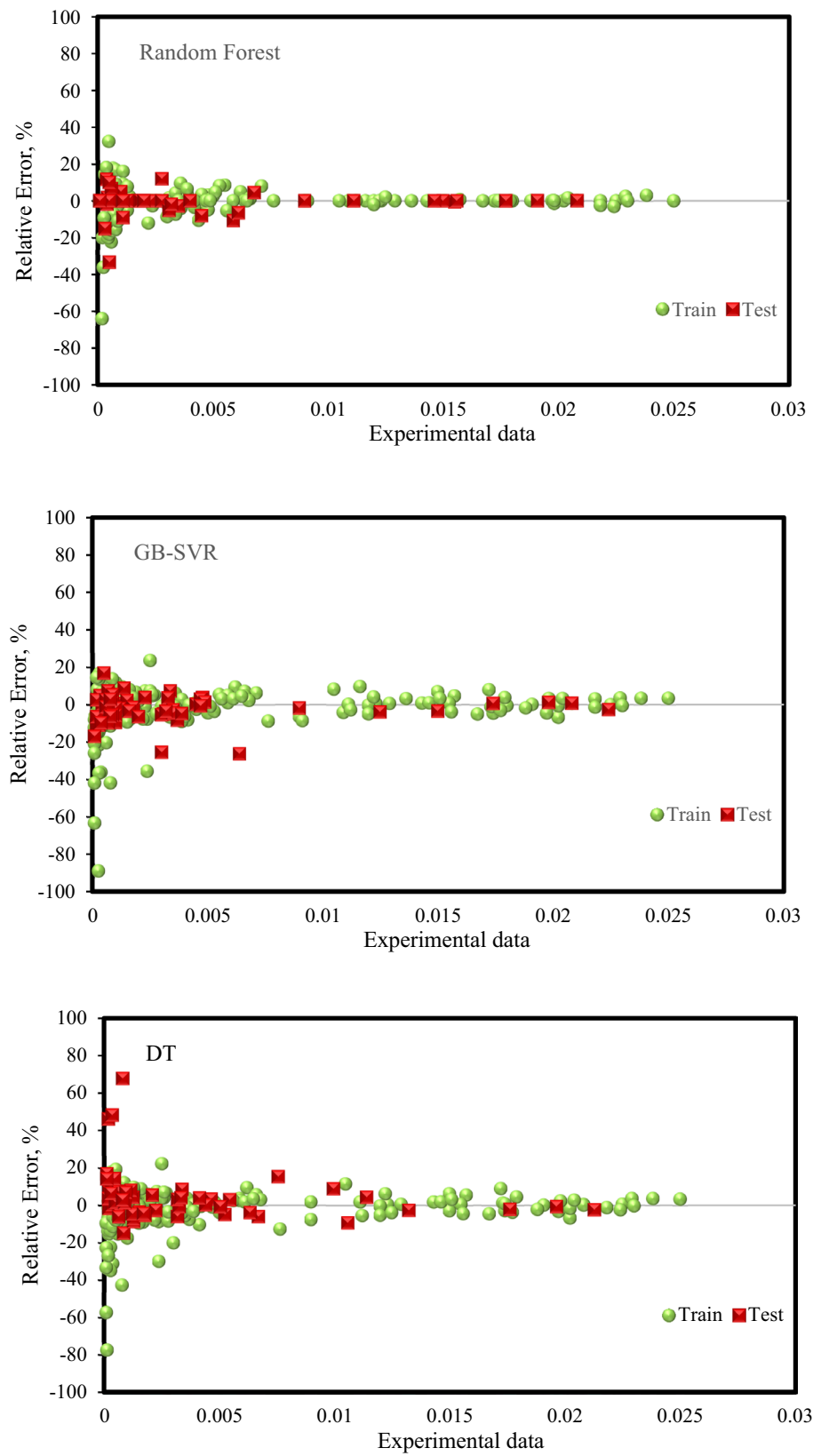


Figure 5. (continued)

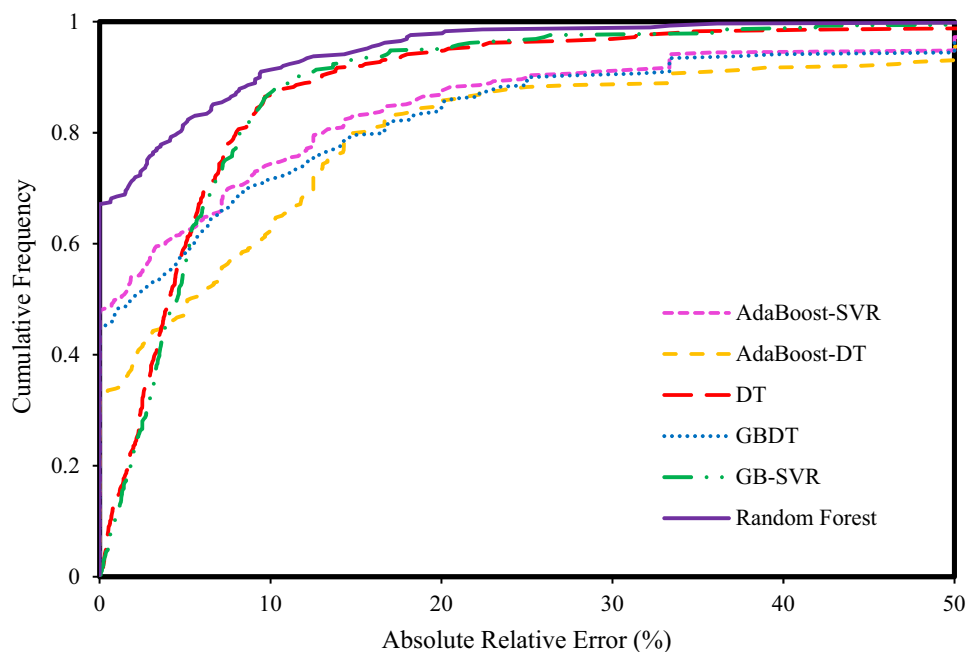


Figure 6. The cumulative frequency plot for the developed predictive models.

- If the relevancy factor < 0 , the impact of the input parameter on the output is decreasing. In other words, by increasing the desired parameter, the value of the output parameter decreases. On the other hand, the closer the relevancy factor to -1 , the greater the influence.
- If the relevancy factor $= 0$, there is no relation between the input parameter and output or this relation is not monotonic.
- If the relevancy factor > 0 , the impact of the input parameter on the output data is incremental. In other words, by increasing the desired parameter, the value of the output parameter also increases. Therefore, the closer the relevancy factor to 1 , the greater the influence.

Figure 11 shows the relevancy factor value for the input parameters of the Random Forest model as the best model. According to this figure, the impact of temperature, pressure, and mole percent of CO_2 in gaseous phase on the solubility of CO_2 - N_2 mixture in aqueous solutions is increasing, and the impact of ionic strength is decreasing. Among the parameters whose relevancy factor values are positive, the mole percent of CO_2 in gaseous phase with a relevancy factor of 0.61 has the most significant impact. Therefore, with increasing temperature, pressure, and the mole percent of CO_2 in gaseous phase, the solubility of CO_2 - N_2 mixture in water and brine solutions increases, and with increasing ionic strength, the solubility decreases.

Implementation of Leverage method. The Leverage method^{88–90} was used to determine the applicability domain of the constructed Random Forest model and to identify any data that is suspect. The Leverage method, which is well-established analytically and visually through Williams' plot, is one of the most important approaches in outlier diagnosis. Standardized residuals (R), which reflect the differences of model's outcomes from experimental observations, and Leverage values, which are the diagonal components of the hat matrix, are determined in this method. The following is the definition of the hat matrix⁸³:

$$H = X(X^T X)^{-1} X^T \quad (18)$$

here, X^T denotes the transpose of the matrix X , which is an $(m \times n)$ matrix, and m and n denote the number of data points and model input variables, respectively. In addition, the critical leverage (H^*) is determined to be $3(n+1)/m$.

The proposed model's applicability domain is then graphically evaluated by displaying the standardized residuals versus leverage values. If most of the data points were located in the limits of $0 \leq H \leq H^*$, and $-3 \leq R \leq 3$, the created model is considered trustworthy and its estimations are made in the applicability domain⁹¹.

Following that, as shown in Fig. 12, William's plot is utilized to determine the Random Forest model's applicable scope and outliers. As shown in Fig. 12, the majority of data falls between $0 \leq H \leq 0.062$, and $-3 \leq R \leq 3$, indicating that the experimental results are of excellent quality and the Random Forest model is quite reliable. Suspicious data are data points with $R > 3$ or $R < -3$, linked with a high level of doubt. Out of Leverage data are data points with $H > 0.062$, and $-3 \leq R \leq 3$ beyond the Random Forest model's applicability range. Only nine data points were identified to be as suspected data and one outlier exists in the solubility databank, which proves the high validity of the experimental databank used for modelling. Eight suspected data points along with one

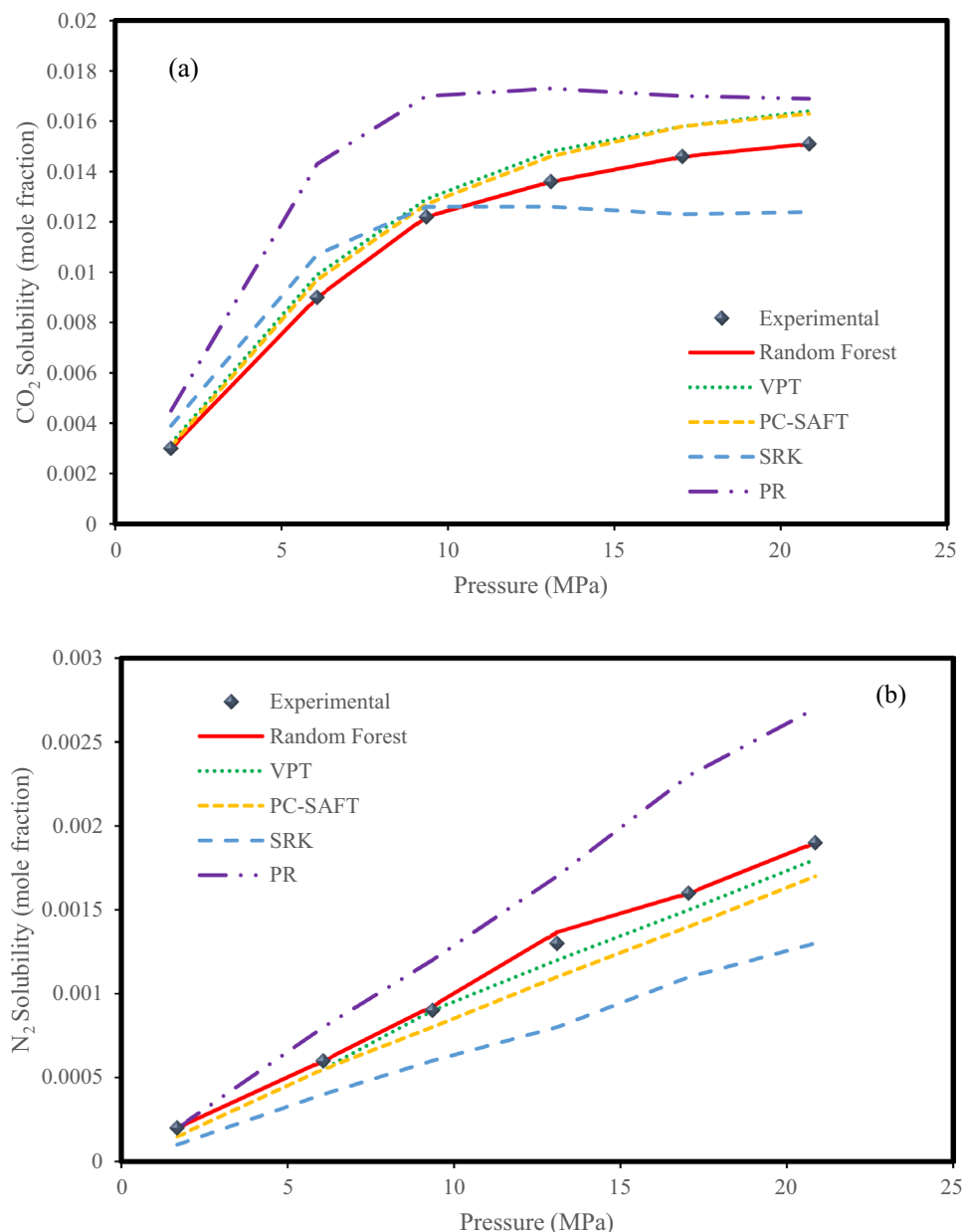


Figure 7. Experimental values¹⁰ with predictions of the EOSs and Random Forest model for the (a) CO₂ solubility and (b) N₂ solubility in the N₂ (39%) + CO₂ (61%) + H₂O system at a temperature of 283 K.

outlier belong to the training subset and one suspected data point belongs to the test subset, which is specified in the Supplementary file.

Conclusions

In this study, using 289 laboratory data and six intelligent models including DT, GBDT, AdaBoost-DT, AdaBoost-SVR, GB-SVR, and Random Forest, the solubility of CO₂ and N₂ in the systems of CO₂-N₂ mixture and aqueous solutions was predicted and comparing their results with thermodynamic models such as SRK, PR, VPT, and PC-SAFT led to the following conclusions:

1. Among the presented models, the Random Forest model with an AAPRE value of 2.84% has the best results. GB-SVR and DT models then have the closest predictions with AAPRE values of 6.43% and 7.41%, respectively. After these models, AdaBoost-SVR, GB-DT, and AdaBoost-DT are ranked in terms of good predictions, respectively. Therefore, intelligent models are very efficient and reliable compared to equations of state.
2. Generally, the equations of state used in this work overestimate the solubility of CO₂ in the aqueous system by increasing the pressure; however, this is the opposite for N₂ except for the PR equation of state for all other models.

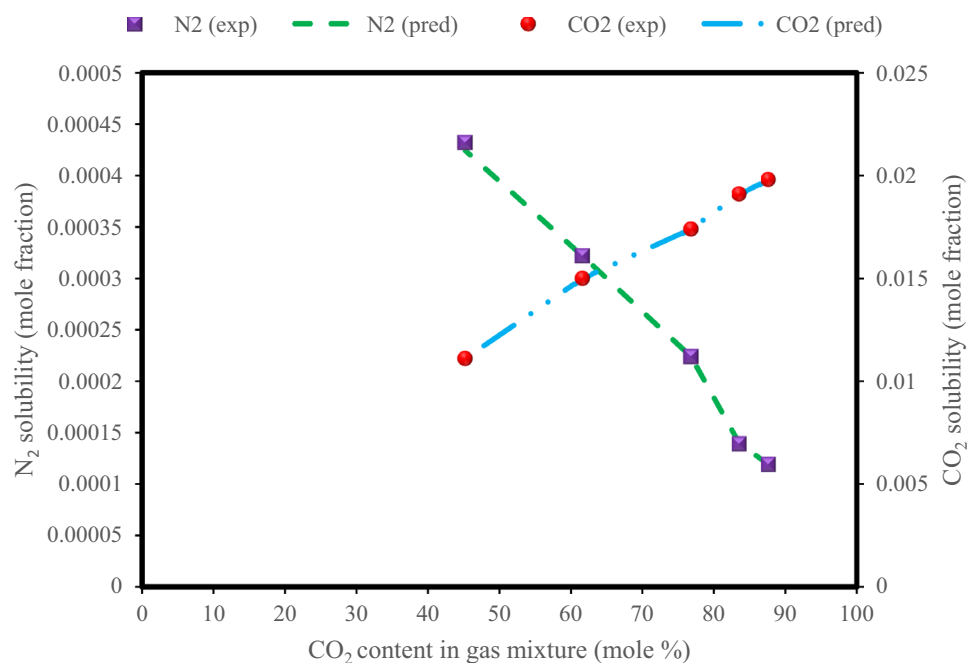


Figure 8. Dependence of CO₂ and N₂ solubilities on the mole percent of CO₂ in the gaseous phase in the N₂ + CO₂ + H₂O system at a temperature of 308 K and pressure of 8.0 MPa.

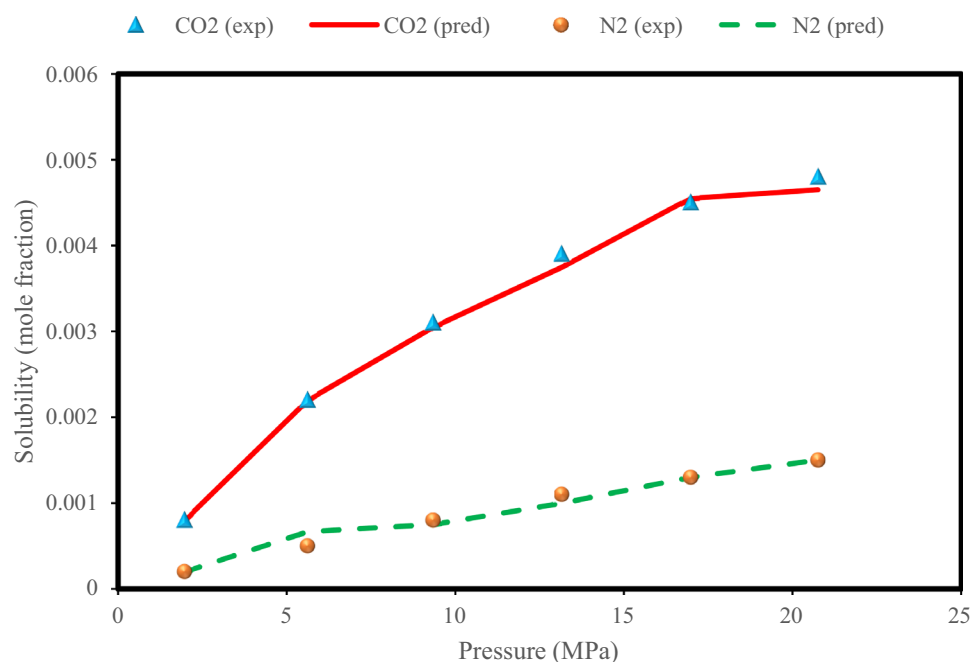


Figure 9. Experimental values¹⁰ of CO₂ and N₂ solubilities in the N₂ (85.4%) + CO₂ (14.6%) + H₂O system at a temperature of 303 K with predictions of the Random Forest model.

3. For solubility of CO₂ present in gaseous mixtures in aqueous systems, the PC-SAFT model, and for solubility of N₂, the VPT model had the best results among the equations of state.
4. Increasing the CO₂ content in the gas mixture increases the solubility of CO₂ in the system and, conversely, decreases the solubility of N₂ at constant temperature and pressure.

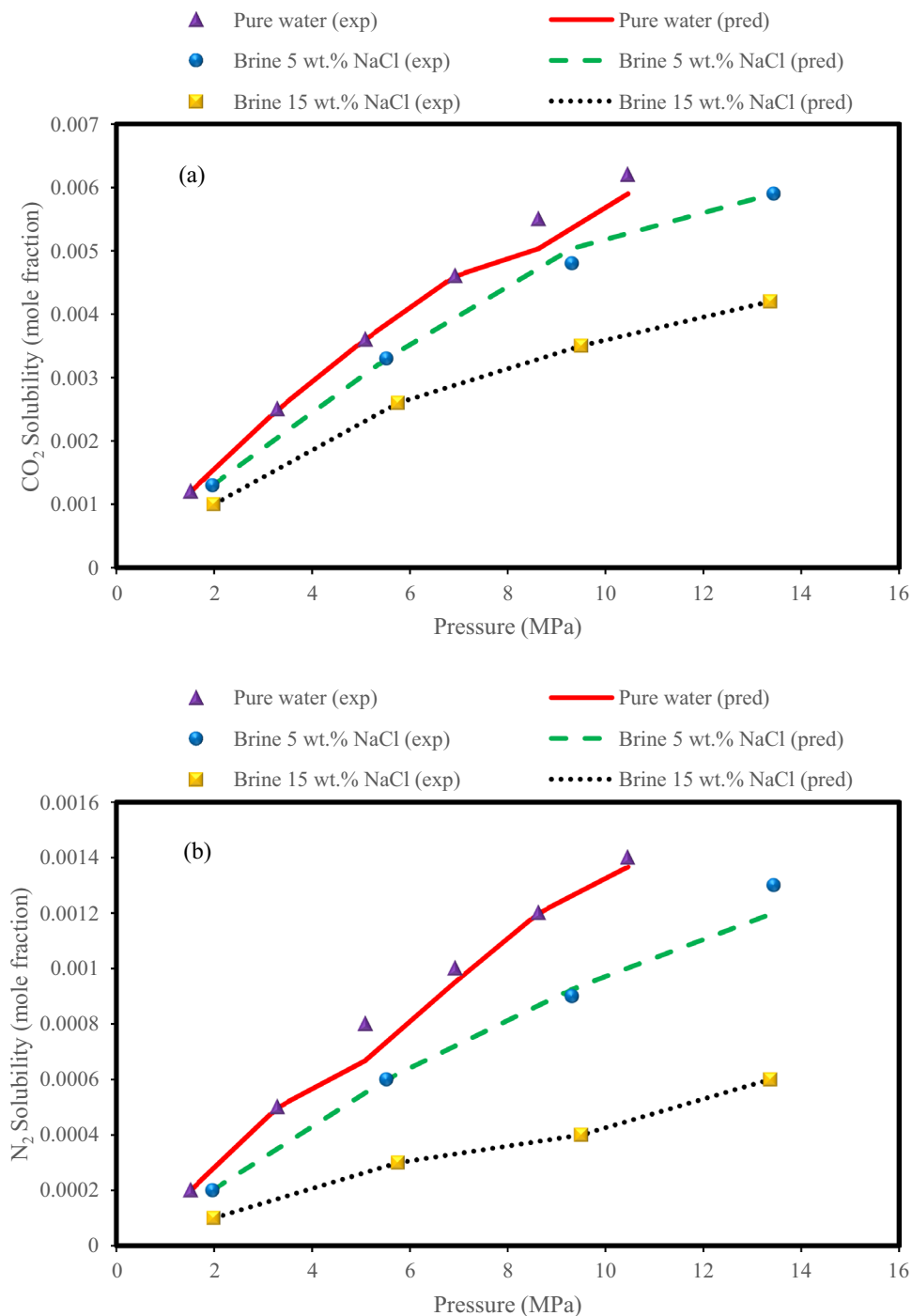


Figure 10. Effect of salinity on (a) CO₂ solubility and (b) N₂ solubility in the N₂ (85.4%) + CO₂ (14.6%) + H₂O (brine) systems at a temperature of 273 K; experimental data¹⁰ with predictions of the Random Forest model.

5. Increasing the water salinity causes the reduction of CO₂ and N₂ solubility in water.
6. The impact of mole percent of CO₂ in gaseous phase, temperature, and pressure on increasing the solubility of CO₂ and N₂ in water is incremental, and the impact of ionic strength on the solubility of CO₂ and N₂ in water is decreasing.

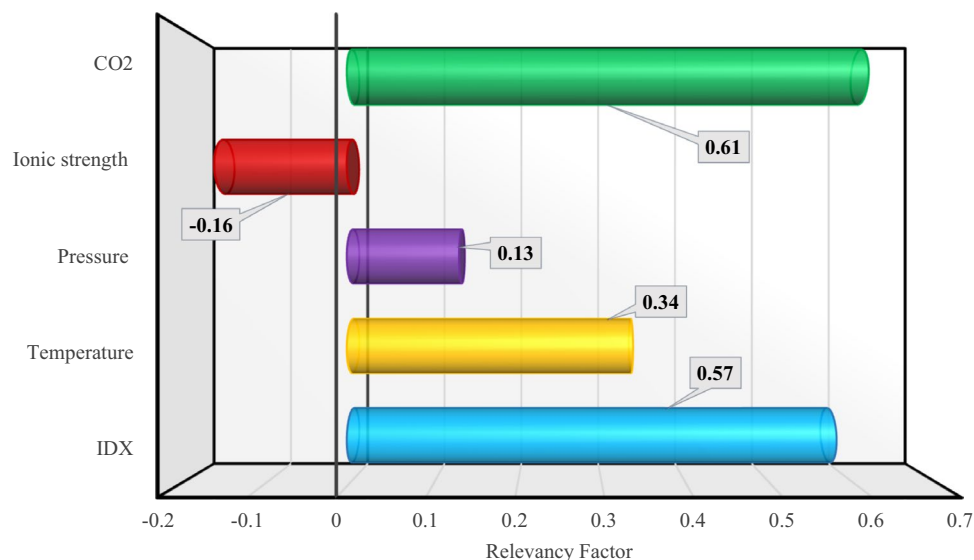


Figure 11. Evaluation of the input parameters' impact.

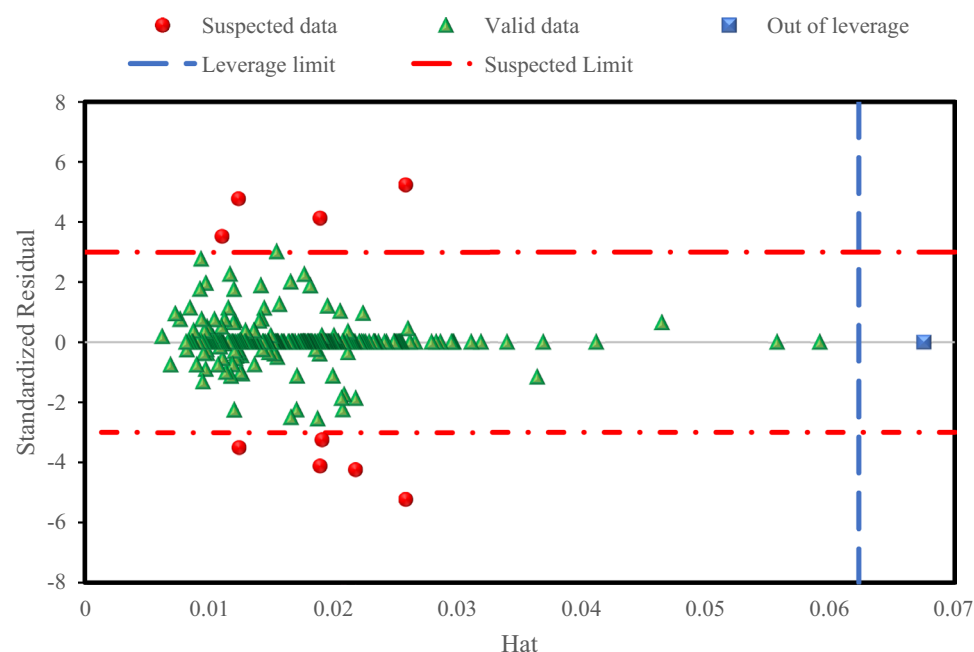


Figure 12. William's plot for the outlier detection using the Random Forest model.

Received: 26 September 2021; Accepted: 9 February 2022

Published online: 07 March 2022

References

- Nabipour, N., Qasem, S. N., Salwana, E. & Baghban, A. Evolving LSSVM and ELM models to predict solubility of non-hydrocarbon gases in aqueous electrolyte systems. *Measurement* **164**, 107999 (2020).
- Wu, H., Zheng, K., Wang, G., Yang, Y. & Li, Y. Modeling of gas solubility in hydrocarbons using the perturbed-chain statistical associating fluid theory equation of state. *Ind. Eng. Chem. Res.* **58**, 12347–12360 (2019).
- Marshall, B. D. A PC-SAFT model for hydrocarbons IV: Water-hydrocarbon phase behavior including petroleum pseudo-components. *Fluid Phase Equilib.* **497**, 79–86 (2019).
- Zaidin, M. F., Kantaatmadja, B. P., Chapoy, A., Ahmadi, P. & Burgass, R. in *SPE Middle East Oil and Gas Show and Conference*. (OnePetro).
- Zhao, Y., Gani, R., Afzal, R. M., Zhang, X. & Zhang, S. Ionic liquids for absorption and separation of gases: An extensive database and a systematic screening method. *AIChE J.* **63**, 1353–1367 (2017).

6. Kang, X. *et al.* Prediction of Henry's law constant of CO₂ in ionic liquids based on SEP and $\Sigma\sigma$ -profile molecular descriptors. *J. Mol. Liq.* **262**, 139–147 (2018).
7. Pan, M., Zhao, Y., Zeng, X. & Zou, J. Efficient absorption of CO₂ by introduction of intramolecular hydrogen bonding in chiral amino acid ionic liquids. *Energy Fuels* **32**, 6130–6135 (2018).
8. Keith, D. W. Why capture CO₂ from the atmosphere?. *Science* **325**, 1654–1655 (2009).
9. Haszeldine, R. S. Carbon capture and storage: How green can black be?. *Science* **325**, 1647–1652 (2009).
10. Hassanpouryouzband, A. *et al.* Solubility of flue gas or carbon dioxide-nitrogen gas mixtures in water and aqueous solutions of salts: Experimental measurement and thermodynamic modeling. *Ind. Eng. Chem. Res.* **58**, 3377–3394 (2019).
11. Foltran, S. *et al.* Understanding the solubility of water in carbon capture and storage mixtures: An FTIR spectroscopic study of H₂O+ CO₂+ N₂ ternary mixtures. *Int. J. Greenhouse Gas Control* **35**, 131–137 (2015).
12. Hendriks, C., Blok, K. & Turkenburg, W. *Climate and Energy: The Feasibility of Controlling CO₂ Emissions 125–142* (Springer, 1989).
13. Pires, J., Martins, F., Alvim-Ferraz, M. & Simões, M. Recent developments on carbon capture and storage: An overview. *Chem. Eng. Res. Des.* **89**, 1446–1460 (2011).
14. Boot-Handford, M. E. *et al.* Carbon capture and storage update. *Energy Environ. Sci.* **7**, 130–189 (2014).
15. Hassanpouryouzband, A. *et al.* CO₂ capture by injection of flue gas or CO₂-N₂ mixtures into hydrate reservoirs: Dependence of CO₂ capture efficiency on gas hydrate reservoir conditions. *Environ. Sci. Technol.* **52**, 4324–4330 (2018).
16. Chambwera, M. *et al.* *Economics of Adaptation* (Springer, 2014).
17. Baghban, A., Ahmadi, M. A. & Shahraki, B. H. Prediction carbon dioxide solubility in presence of various ionic liquids using computational intelligence approaches. *J. Supercrit. Fluids* **98**, 50–64 (2015).
18. Liu, Y. *et al.* Phase equilibria of CO₂+ N₂+ H₂O and N₂+ CO₂+ H₂O+ NaCl+ KCl+ CaCl₂ systems at different temperatures and pressures. *J. Chem. Eng. Data* **57**, 1928–1932 (2012).
19. Garapati, N., McGuire, P. & Anderson, B. J. in *Unconventional Resources Technology Conference. 1942–1951* (Society of Exploration Geophysicists, American Association of Petroleum ...).
20. Gilfillan, S. M. *et al.* Solubility trapping in formation water as dominant CO₂ sink in natural gas fields. *Nature* **458**, 614–618 (2009).
21. Rosenqvist, J., Kilpatrick, A. D. & Yardley, B. W. Solubility of carbon dioxide in aqueous fluids and mineral suspensions at 294 K and subcritical pressures. *Appl. Geochem.* **27**, 1610–1614 (2012).
22. Gharagheizi, F., Eslamimanesh, A., Mohammadi, A. H. & Richon, D. Representation/prediction of solubilities of pure compounds in water using artificial neural network-group contribution method. *J. Chem. Eng. Data* **56**, 720–726 (2011).
23. Peng, D.-Y. & Robinson, D. B. A new two-constant equation of state. *Ind. Eng. Chem. Fundam.* **15**, 59–64 (1976).
24. Bamberger, A., Sieder, G. & Maurer, G. High-pressure (vapor+ liquid) equilibrium in binary mixtures of (carbon dioxide + water or acetic acid) at temperatures from 313 to 353 K. *J. Supercrit. Fluids* **17**, 97–110 (2000).
25. Spycher, N., Pruess, K. & Ennis-King, J. CO₂-H₂O mixtures in the geological sequestration of CO₂. I. Assessment and calculation of mutual solubilities from 12 to 100 C and up to 600 bar. *Geochim. Cosmochim. Acta* **67**, 3015–3031 (2003).
26. Wang, Z. *et al.* Near-infrared spectroscopic investigation of water in supercritical CO₂ and the effect of CaCl₂. *Fluid Phase Equilib.* **338**, 155–163 (2013).
27. Chapoy, A., Mohammadi, A., Chareton, A., Tohidi, B. & Richon, D. Measurement and modeling of gas solubility and literature review of the properties for the carbon dioxide-water system. *Ind. Eng. Chem. Res.* **43**, 1794–1802 (2004).
28. Zabaloy, M., Mabe, G., Bottini, S. & Brignole, E. Vapor liquid equilibria in ternary mixtures of water-alcohol-non polar gases. *Fluid Phase Equilib.* **83**, 159–166 (1993).
29. Ferreira, O., Brignole, E. A. & Macedo, E. A. Modelling of phase equilibria for associating mixtures using an equation of state. *J. Chem. Thermodyn.* **36**, 1105–1117 (2004).
30. Tsuji, T. *et al.* CO₂ solubility in water containing monosaccharides, and the prediction of pH using Peng–Robinson equation of state. *Fluid Phase Equilib.* **441**, 9–16 (2017).
31. Li, Y., Qiao, Z., Sun, S. & Zhang, T. Thermodynamic modeling of CO₂ solubility in saline water using NVT flash with the cubic-Plus-association equation of state. *Fluid Phase Equilib.* **520**, 112657 (2020).
32. Abraham, M. H., Chadha, H. S. & Mitchell, R. C. Hydrogen bonding 33. Factors that influence the distribution of solutes between blood and brain. *J. Pharm. Sci.* **83**, 1257–1268 (1994).
33. Ruelle, P., Sarraf, E. & Kesselring, U. W. Prediction of carbazole solubility and its dependence upon the solvent nature. *Int. J. Pharm.* **104**, 125–133 (1994).
34. Ruelle, P. & Kesselring, U. W. Solubility predictions for solid nitriles and tertiary amides based on the mobile order theory. *Pharm. Res.* **11**, 201–205 (1994).
35. Austegard, A., Solbraa, E., De Koeijer, G. & Mølnvik, M. Thermodynamic models for calculating mutual solubilities in H₂O–CO₂–CH₄ mixtures. *Chem. Eng. Res. Des.* **84**, 781–794 (2006).
36. Song, K. Y. & Kobayashi, R. The water content of a carbon dioxide-rich gas mixture containing 5.31 Mol% methane along the three-phase and supercritical conditions. *J. Chem. Eng. Data* **35**, 320–322 (1990).
37. Baghban, A. *et al.* Phase equilibrium modelling of natural gas hydrate formation conditions using LSSVM approach. *Pet. Sci. Technol.* **34**, 1431–1438 (2016).
38. Baghban, A., Bahadori, M., Lemraski, A. S. & Bahadori, A. Prediction of solubility of ammonia in liquid electrolytes using least square support vector machines. *Ain Shams Eng. J.* **9**, 1303–1312 (2018).
39. Zarei, F. & Baghban, A. Phase behavior modelling of asphaltene precipitation utilizing MLP-ANN approach. *Pet. Sci. Technol.* **35**, 2009–2015 (2017).
40. Suleymani, M. & Bemani, A. Prediction of the interfacial tension between hydrocarbons and carbon dioxide. *Pet. Sci. Technol.* **36**, 227–231 (2018).
41. Choubin, B. *et al.* Spatial hazard assessment of the PM₁₀ using machine learning models in Barcelona, Spain. *Sci. Total Environ.* **701**, 134474 (2020).
42. Samani, N. N. *et al.* Solubility of hydrocarbon and non-hydrocarbon gases in aqueous electrolyte solutions: A reliable computational strategy. *Fuel* **241**, 1026–1035 (2019).
43. Hemmati-Sarapardeh, A., Amar, M. N., Soltanian, M. R., Dai, Z. & Zhang, X. Modeling CO₂ solubility in water at high pressure and temperature conditions. *Energy Fuels* **34**, 4761–4776 (2020).
44. Saghafi, H. & Arabloo, M. Modeling of CO₂ solubility in MEA, DEA, TEA, and MDEA aqueous solutions using AdaBoost-Decision Tree and Artificial Neural Network. *Int. J. Greenhouse Gas Control* **58**, 256–265 (2017).
45. Zhang, J., Lee, S. & Lee, J. W. Solubility of CO₂, N₂, and CO₂+ N₂ gas mixtures in isoctane. *J. Chem. Eng. Data* **53**, 1321–1324 (2008).
46. Eslamimanesh, A., Gharagheizi, F., Mohammadi, A. H. & Richon, D. Artificial neural network modeling of solubility of supercritical carbon dioxide in 24 commonly used ionic liquids. *Chem. Eng. Sci.* **66**, 3039–3044 (2011).
47. Garg, S. *et al.* Experimental data, thermodynamic and neural network modeling of CO₂ solubility in aqueous sodium salt of l-phenylalanine. *J. CO₂ Util.* **19**, 146–156 (2017).
48. Smola, A. J. & Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **14**, 199–222 (2004).
49. Schölkopf, B., Smola, A. J., Williamson, R. C. & Bartlett, P. L. New support vector algorithms. *Neural Comput.* **12**, 1207–1245 (2000).

50. Vapnik, V., Golowich, S. E. & Smola, A. Support vector method for function approximation, regression estimation, and signal processing. *Adv. Neural Inf. Process. Syst.* **1**, 281–287 (1997).
51. Briec, M. S., Waters, C. D., Drinan, D. P. & Naish, K. A. A practical introduction to Random Forest for genetic association studies in ecology and evolution. *Mol. Ecol. Resour.* **18**, 755–766 (2018).
52. Granitto, P. M., Gasperi, F., Biasioli, F., Trainotti, E. & Furlanello, C. Modern data mining tools in descriptive sensory analysis: A case study with a Random forest approach. *Food Qual. Prefer.* **18**, 681–689 (2007).
53. Wu, Y. & Misra, S. Intelligent image segmentation for organic-rich shales using random forest, wavelet transform, and hessian matrix. *IEEE Geosci. Remote Sens. Lett.* **17**, 1144–1147 (2019).
54. Shaikhina, T. et al. Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. *Biomed. Signal Process. Control* **52**, 456–462 (2019).
55. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
56. Chen, J. et al. A parallel random forest algorithm for big data in a spark cloud computing environment. *IEEE Trans. Parallel Distrib. Syst.* **28**, 919–933 (2016).
57. Amar, M. N., Shateri, M., Hemmati-Sarapardeh, A. & Alamatsaz, A. Modeling oil-brine interfacial tension at high pressure and high salinity conditions. *J. Pet. Sci. Eng.* **183**, 106413 (2019).
58. Song, Y.-Y. & Ying, L. Decision tree methods: Applications for classification and prediction. *Shanghai Arch. Psychiatry* **27**, 130 (2015).
59. Patel, N. & Upadhyay, S. Study of various decision tree pruning methods with their empirical comparison in WEKA. *Int. J. Comput. Appl.* **60**, 20–25 (2012).
60. Natekin, A. & Knoll, A. Gradient boosting machines, a tutorial. *Front. Neurobot.* **7**, 21 (2013).
61. Chen, Y., Jia, Z., Mercola, D. & Xie, X. A gradient boosting algorithm for survival analysis via direct optimization of concordance index. *Comput. Math. Methods Med.* **2013**, 1–8 (2013).
62. Sun, R., Wang, G., Zhang, W., Hsu, L.-T. & Ochieng, W. Y. A gradient boosting decision tree based GPS signal reception classification algorithm. *Appl. Soft Comput.* **86**, 105942 (2020).
63. Zhou, K., Zhang, J., Ren, Y., Huang, Z. & Zhao, L. A gradient boosting decision tree algorithm combining synthetic minority oversampling technique for lithology identification. *Geophysics* **85**, 147–158 (2020).
64. 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).
65. Dargahi-Zarandi, A., Hemmati-Sarapardeh, A., Shateri, M., Menad, N. A. & Ahmadi, M. Modeling minimum miscibility pressure of pure/impure CO₂-crude oil systems using adaptive boosting support vector regression: Application to gas injection processes. *J. Pet. Sci. Eng.* **184**, 106499 (2020).
66. Margineantu, D. D. & Dietterich, T. G. in *ICML*. 211–218 (Citeseer).
67. Zerrouki, N., Harrou, F., Sun, Y. & Houacine, A. Vision-based human action classification using adaptive boosting algorithm. *IEEE Sens. J.* **18**, 5115–5121 (2018).
68. Mishra, S., Mishra, D. & Santra, G. H. Adaptive boosting of weak regressors for forecasting of crop production considering climatic variability: An empirical assessment. *J. King Saud Univ. Comput. Inf. Sci.* **32**, 949–964 (2020).
69. Ye, J. & Yang, L. in *2018 5th International Conference on Systems and Informatics (ICSAI)*. 139–143 (IEEE).
70. Hu, D., Zhang, C., Cao, W., Lv, X. & Xie, S. Grain yield predict based on GRA-AdaBoost-SVR model. *J. Big Data* **3**, 65 (2021).
71. Budka, M. & Gabrys, B. Ridge regression ensemble for toxicity prediction. *Procedia Comput. Sci.* **1**, 193–201 (2010).
72. Mohammadi, M.-R. et al. Modeling hydrogen solubility in hydrocarbons using extreme gradient boosting and equations of state. *Sci. Rep.* **11**, 1–20 (2021).
73. Madani, S. A. et al. Modeling of nitrogen solubility in normal alkanes using machine learning methods compared with cubic and PC-SAFT equations of state. *Sci. Rep.* **11**, 1–20 (2021).
74. Chapman, W. G., Gubbins, K. E., Jackson, G. & Radosz, M. New reference equation of state for associating liquids. *Ind. Eng. Chem. Res.* **29**, 1709–1721 (1990).
75. Chapman, W. G., Gubbins, K. E., Jackson, G. & Radosz, M. SAFT: Equation-of-state solution model for associating fluids. *Fluid Phase Equilib.* **52**, 31–38 (1989).
76. Nasrifar, K., Bolland, O. & Moshfeghian, M. Predicting natural gas dew points from 15 equations of state. *Energy Fuels* **19**, 561–572 (2005).
77. Valderrama, J. O. A generalized Patel-Teja equation of state for polar and nonpolar fluids and their mixtures. *J. Chem. Eng. Jpn.* **23**, 87–91 (1990).
78. Avlonitis, D., Danesh, A. & Todd, A. Prediction of VL and VLL equilibria of mixtures containing petroleum reservoir fluids and methanol with a cubic EoS. *Fluid Phase Equilib.* **94**, 181–216 (1994).
79. Danesh, A. *PVT and Phase Behaviour of Petroleum Reservoir Fluids* (Elsevier, 1998).
80. Gross, J. & Sadowski, G. Perturbed-chain SAFT: An equation of state based on a perturbation theory for chain molecules. *Ind. Eng. Chem. Res.* **40**, 1244–1260 (2001).
81. Chen, Y., Mutelet, F. & Jaubert, J.-N.L. Modeling the solubility of carbon dioxide in imidazolium-based ionic liquids with the PC-SAFT equation of state. *J. Phys. Chem. B* **116**, 14375–14388 (2012).
82. Grenner, A., Schmelzer, J., von Solms, N. & Kontogeorgis, G. M. Comparison of two association models (Elliott–Suresh–Donohue and simplified PC-SAFT) for complex phase equilibria of hydrocarbon–water and amine-containing mixtures. *Ind. Eng. Chem. Res.* **45**, 8170–8179 (2006).
83. Menad, N. A., Noureddine, Z., Hemmati-Sarapardeh, A. & Shamshirband, S. Modeling temperature-based oil-water relative permeability by integrating advanced intelligent models with grey wolf optimization: application to thermal enhanced oil recovery processes. *Fuel* **242**, 649–663 (2019).
84. Probst, P., Wright, M. N. & Boulesteix, A. L. Hyperparameters and tuning strategies for random forest. *Wiley Interdiscipl. Rev.* **9**, e1301 (2019).
85. Steel, L., Liu, Q., Mackay, E. & Maroto-Valer, M. M. CO₂ solubility measurements in brine under reservoir conditions: A comparison of experimental and geochemical modeling methods. *Greenhouse Gases Sci. Technol.* **6**, 197–217 (2016).
86. Mohammadi, M.-R. et al. Application of robust machine learning methods to modeling hydrogen solubility in hydrocarbon fuels. *Int. J. Hydrogen Energy* **47**, 320–338 (2021).
87. 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. Pet. Sci. Eng.* **205**, 108836 (2021).
88. Leroy, A. M. & Rousseeuw, P. J. Robust regression and outlier detection. *RROD* (1987).
89. Goodall, C. R. *13 Computation Using the QR Decomposition* (Springer, 1993).
90. Gramatica, P. Principles of QSAR models validation: Internal and external. *QSAR Comb. Sci.* **26**, 694–701 (2007).
91. 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).

Acknowledgements

Fahimeh Hadavimoghaddam would like to acknowledge the support of the Ministry of Science and Higher Education of the Russian Federation under agreement No. 075-15-2020-900 within the framework of the development program for a world-class Research Center.

Authors contribution

R.N.-K.: Writing-Original Draft, Data curation; Formal analysis, E.T.-R.: Writing-Original Draft, Validation, F.H.: Writing-Review & Editing, Validation, Methodology, M.-R.M.: Writing-Review & Editing, Validation, Data curation, Formal analysis, A.H.-S.: Writing-Review & Editing, Methodology, Validation, Supervision.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-022-07393-z>.

Correspondence and requests for materials should be addressed to 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.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022