Modeling solubility of \( \text{CO}_2-\text{N}_2 \) gas mixtures in aqueous electrolyte systems using artificial intelligence techniques and equations of state

Reza Nakhaei-Kohani\(^1\), Ehsan Taslimi-Renani\(^2\), Fahime Hadavimoghaddam\(^3\)\(^,4\), Mohammad-Reza Mohammadi\(^5\) & Abdolhossein Hemmati-Sarapardeh\(^5\)\(^,6\)*

Determining the solubility of non-hydrocarbon gases such as carbon dioxide (\( \text{CO}_2 \)) and nitrogen (\( \text{N}_2 \)) 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 (\( \text{CO}_2-\text{N}_2 \) 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 \( \text{CO}_2-\text{N}_2 \) 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 \( \text{CO}_2 \) present in gaseous mixtures in aqueous systems, the PC-SAFT model, and for solubility of \( \text{N}_2 \), the VPT EOS had the best results among the EOSs. Also, the sensitivity analysis of input parameters showed that increasing the mole percent of \( \text{CO}_2 \) in gaseous phase, temperature, pressure, and decreasing the ionic strength increase the solubility of \( \text{CO}_2-\text{N}_2 \) 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 \( \text{CO}_2-\text{N}_2 \) 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 \( \text{CO}_2-\text{N}_2 \) gas mixtures in aqueous systems.

Abbreviations

| Abbreviation | Description |
|--------------|-------------|
| 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 |

\(^1\)Department of Chemical & Petroleum Engineering, Shiraz University, Shiraz, Iran. \(^2\)Department of Electrical Engineering, University of Malaya (UM), 50603 Kuala Lumpur, Malaysia. \(^3\)Ufa State Petroleum Technological University, Ufa 450064, Russia. \(^4\)Institute of Unconventional Oil & Gas, Northeast Petroleum University, Heilongjiang, Daqing 163318, China. \(^5\)Department of Petroleum Engineering, Shahid Bahonar University of Kerman, Kerman, Iran. \(^6\)Key Laboratory of Continental Shale Hydrocarbon Accumulation and Efficient Development, Ministry of Education, Northeast Petroleum University, Daqing 163318, China. *Email: hemmati@uk.ac.ir
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. 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. CO₂ as a greenhouse gas has been considered a serious problem in recent decades. Carbon capture and storage (CCS) is a technique that involves capturing CO₂ from major point sources and storing it in formations. 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.

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 recognized. 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. 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 hydrates? 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. 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. 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.

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:

- 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 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. Tomoya et al. measured CO₂ solubility in aqueous solutions and then correlated the experimental data with the Peng-Robinson-Stryjek-Vera EOS. Yiteng et al. 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

- 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 literature, CO₂ solubility in water and aqueous solutions 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. Tomoya et al. measured CO₂ solubility in aqueous solutions and then correlated the experimental data with the Peng-Robinson-Stryjek-Vera EOS. Yiteng et al. 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

In the literature, CO₂ solubility in water and aqueous solutions 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. Tomoya et al. measured CO₂ solubility in aqueous solutions and then correlated the experimental data with the Peng-Robinson-Stryjek-Vera EOS. Yiteng et al. 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
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 condition. Abraham et al. 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.

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. 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 attention. Machine learning techniques have recently found widespread use in forecasting challenges such as hydrate formation, ammonia solubility in liquids, simulating asphaltene behavior, and hydrocarbon-CO2 interfacial tension. They have received much interest as a result of their captivating performance. Samani et al. 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. 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. investigated the solubility of CO2 and N2 in aqueous solutions using Extreme Learning Machine (EML) and LSSVM approaches. Their solubility database was similar to Samani et al.'s work, 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. 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. 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 methods, EOSs, mathematical methods, and intelligent neural networks 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 have investigated the solubility of CO2–N2 mixtures in water and brine. Only two papers 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, which is presented in the Supplementary file. Although two studies 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 is less than 90 data points. The
and
\[ \geq \]
\[ = \]

are regarded positive variables.

here

indicates the transposed matrix,

is the toleration of error,

are regarded positive variables reflecting the lower and higher excessive variations, respectively, and

interprets a positive regularization factor determining the deviance from \( \varepsilon \)

By employing the Lagrange multiplier, Eq. (2) can be converted into a dual optimization problem as follows, which makes it easier to solve

\[
y = \sum_{i=1}^{n} (a_i - a_i^+).K(x_i, x) + b
\]

where \( K(x_i, x) \) is the kernel function, \( a_i \) and \( a_i^+ \) 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)

53,54. 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'
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-fitting. The framework of RF construction is illustrated in Fig. 1.

Suppose $D$ is the training dataset with $n$ number of observations, $D = [(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)]$, 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 follows:

$$H_{\text{OOB}}(x) = \arg\max \sum_{t=1}^{T} I(h_t(x)) = y$$

(4)

The learning error of the OOB can be obtained by:

$$e_{\text{OOB}}(x) = \frac{1}{|D|} \sum_{(x,y) \in D} I(H_{\text{OOB}}(x) \neq y)$$

(5)

The procedure of RF must be random and this feature can be controlled over a parameter formulated as $k$.

The significance of a characteristic of a variable $X_i$ could be obtained as follows:

$$I(X_i) = \frac{1}{B} \sum_{t=1}^{B} \tilde{O}O_{\text{Berr}}_t - OOB_{\text{err}}$$

(6)

where $X_i$ is the $i$th parameter in vector $X$, $B$ indicates the current number of trees in the RF, $\tilde{O}O_{\text{Berr}}_t$ denotes the predicted error of the OOB samples for the feature $X_i$ of tree $t$, and $OOB_{\text{err}}$ is the initial OOB samples including permuted variables.

**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 regression. 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 stopping. The input data is split into branches and decision nodes starting from the root node. The splitting process continues on till a stopping criterion is convinced. The pruning technique implies removing the low-importance branches. A simple architecture of a DT model is illustrated in Fig. 2.
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:

$$g_t(x) = E_y \left[ \frac{\partial \Psi(y, f(x))}{\partial f(x)} \right]_{f(x) = \hat{f}_{i-1}(x)}$$  \hspace{1cm} (7)$$

This converts a potential sophisticated optimization problem into a classic least square minimization:

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

The following are the steps in the GBDT technique:

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

$$\hat{f}_t = \hat{f}_{t-1} + \rho_t h(x_i, \theta)$$  \hspace{1cm} (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 Schapire 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 samples. The following are the general stages of the AdaBoost technique:

- **Figure 2.** Schematic illustration of a typical decision tree.
Weights definition: $w_j = \frac{1}{n}$, $j = 1, 2, \ldots, n$

- Apply the training data to a weak learner $W_l(x)$, weights, and obtain the weighted error for each $i$.

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

(10)

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

(11)

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

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

(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 nonlinear 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 performance.\footnote{1}

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 substantially.\footnote{2} 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 generalizability.\footnote{3} We combined GB and SVR algorithms\footnote{4}. 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.

Figure 3. Schematic illustration of a typical GBDT.
| EOS | PVT relation | Parameters | Reference |
|-----|--------------|------------|-----------|
| SRK | \( p = \frac{RT}{\alpha(T) - \frac{a}{v^2}} \) | \( \alpha(T) = \left[ 1 + m \left( 1 - \sqrt{\frac{RT}{\alpha(T)}} \right) \right]^2 \) | 88 |
| PR  | \( p = \frac{RT}{\alpha(T) - \frac{a}{v^2}} \) | \( \alpha(T) = \left[ 1 + m \left( 1 - \sqrt{\frac{RT}{\alpha(T)}} \right) \right]^2 \) | 88 |
| VPT | \( p = \frac{RT}{\alpha(T) - \frac{a}{v^2}} \) | \( \alpha(T) = a^C + a^A \) | 72–78 |

\[ \tilde{a}^{ic} = m_i \tilde{a}^{ic} + \tilde{a}^{id} = m_i \tilde{a}^{id} - \sum_i x_i (m_i - 1) \ln \tilde{g}_i \]

\[ \tilde{b}^{ic} = \tilde{b}^{ic} + \tilde{b}^{id} = \tilde{b}^{ic} + \tilde{b}^{id} + \tilde{b}^{d} + \tilde{b}^{d} \]

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 volume. These thermodynamic models can characterize the thermal characteristics and volumetric behavior of mixtures and pure materials. 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.
few years. To explain the interactions between the molecules in a system, the SAFT EOSs were constructed using statistical mechanics\(^\text{74,75}\). 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\(^\text{72,83}\):

- **Average absolute percent relative error (AAPRE)**
  \[
  AAPRE = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{(S_{\text{EXP}} - S_{\text{PRED}})}{S_{\text{EXP}}} \right) \times 100
  \]
  (13)

- **Standard deviation (SD)**
  \[
  SD = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} \left( \frac{S_{\text{EXP}} - S_{\text{PRED}}}{S_{\text{EXP}}} \right)^2}
  \]
  (14)

- **Coefficient of determination (R\(^2\))**
  \[
  R^2 = 1 - \frac{\sum_{i=1}^{N} (S_{\text{EXP}} - S_{\text{PRED}})^2}{\sum_{i=1}^{N} (S_{\text{EXP}} - \bar{S}_{\text{EXP}})^2}
  \]
  (15)

- **Root mean square error (RMSE)**
  \[
  RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (S_{\text{EXP}} - S_{\text{PRED}})^2}
  \]
  (16)

In the above equations, \(S_{\text{EXP}}, S_{\text{PRED}}, \bar{S}_{\text{EXP}}\) 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 elsewhere\(^\text{72}\).

### Results and discussion

#### Statistical evaluation of models

The models discussed in the previous sections have been developed to predict the solubility of CO\(_2\)–N\(_2\) mixtures in water utilizing 289 laboratory data. In this study, we have employed six algorithms, which were rarely used, to estimate CO\(_2\)–N\(_2\) 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

| Substance | \(M_w\) (g/mol) | \(m\) | \(\sigma\) (Å) | \(\varepsilon/k\) (K) | Reference |
|-----------|----------------|------|---------------|-----------------|-----------|
| N\(_2\)   | 28.013         | 1.2053 | 3.313         | 90.96           | \(^\text{80}\) |
| CO\(_2\)  | 44.01          | 2.0729 | 2.7852        | 169.21          | \(^\text{80}\) |
| H\(_2\)O  | 18.015         | 2     | 2.3533        | 207.84          | \(^\text{80}\) |

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

| Substance | \(P_c\) (MPa) | \(T_c\) (K) | \(Z_c\) | \(\omega\) |
|-----------|---------------|-------------|---------|-----------|
| 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 paper\(^\text{79}\).
statistical parameters for the represented models are summarized in Table 5. In this table, different statistical parameters such as RMSE, $R^2$, 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 automated.

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 CO$_2$–N$_2$ 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 CO$_2$ and N$_2$ in different CO$_2$ + N$_2$ + H$_2$O (brine) systems was calculated using 24 laboratory data points extracted from the literature, 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 CO$_2$ in aqueous solutions, the Random Forest approach out-performs 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 N$_2$ 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%.

### Table 5. Calculated statistical criteria for the proposed models.

| 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    |

**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) | Exp | DT | GB-DT | AdaBoost-DT | AdaBoost-SVR | GB-SVR | Random Forest | SRK | PR | VPT | PC-SAFT |
|------------------|---------|---------|-----|----|------|-------------|--------------|--------|--------------|-----|----|-----|---------|
| CO₂ (97%) + H₂O, at 303.05 K | 1 | 1.98 | 0.0008 | 0.000791 | 0.000800 | 0.000900 | 0.000900 | 0.000816 | 0.000800 | 0.00100 | 0.00130 | 0.00090 | 0.00080 |
| CO₂ (14.6%) + N₂ (85.4%) + H₂O at 283.15 K | 2 | 2.05 | 0.0003 | 0.000445 | 0.000450 | 0.000450 | 0.000444 | 0.000465 | 0.000450 | 0.000450 | 0.000450 | 0.000445 | 0.000445 |
| CO₂ (61%) + N₂ (39%) + H₂O at 303.05 K | 7 | 1.92 | 0.0002 | 0.000320 | 0.000320 | 0.000320 | 0.000320 | 0.000320 | 0.000320 | 0.000320 | 0.000320 | 0.000320 | 0.000320 |
| CO₂ (14.6%) + N₂ (85.4%) + brine (10 wt.% NaCl), at 273.25 K | 13 | 2.07 | 0.0007 | 0.000810 | 0.000810 | 0.000810 | 0.000810 | 0.000810 | 0.000810 | 0.000810 | 0.000810 | 0.000810 | 0.000810 |
| CO₂ (14.6%) + N₂ (85.4%) + H₂O at 283.15 K | 15 | 2.08 | 0.0008 | 0.000910 | 0.000910 | 0.000910 | 0.000910 | 0.000910 | 0.000910 | 0.000910 | 0.000910 | 0.000910 | 0.000910 |
| CO₂ (61%) + N₂ (39%) + H₂O at 303.05 K | 16 | 2.13 | 0.0009 | 0.001010 | 0.001010 | 0.001010 | 0.001010 | 0.001010 | 0.001010 | 0.001010 | 0.001010 | 0.001010 | 0.001010 |
| CO₂ (14.6%) + N₂ (85.4%) + brine (10 wt.% NaCl), at 273.25 K | 17 | 2.14 | 0.0010 | 0.001110 | 0.001110 | 0.001110 | 0.001110 | 0.001110 | 0.001110 | 0.001110 | 0.001110 | 0.001110 | 0.001110 |
| CO₂ (61%) + N₂ (39%) + H₂O at 303.05 K | 18 | 2.15 | 0.0011 | 0.001210 | 0.001210 | 0.001210 | 0.001210 | 0.001210 | 0.001210 | 0.001210 | 0.001210 | 0.001210 | 0.001210 |
| CO₂ (14.6%) + N₂ (85.4%) + brine (10 wt.% NaCl), at 273.25 K | 19 | 2.16 | 0.0012 | 0.001310 | 0.001310 | 0.001310 | 0.001310 | 0.001310 | 0.001310 | 0.001310 | 0.001310 | 0.001310 | 0.001310 |
| CO₂ (61%) + N₂ (39%) + H₂O at 303.05 K | 20 | 2.17 | 0.0013 | 0.001410 | 0.001410 | 0.001410 | 0.001410 | 0.001410 | 0.001410 | 0.001410 | 0.001410 | 0.001410 | 0.001410 |
| CO₂ (14.6%) + N₂ (85.4%) + brine (10 wt.% NaCl), at 273.25 K | 21 | 2.18 | 0.0014 | 0.001510 | 0.001510 | 0.001510 | 0.001510 | 0.001510 | 0.001510 | 0.001510 | 0.001510 | 0.001510 | 0.001510 |
| CO₂ (61%) + N₂ (39%) + H₂O at 303.05 K | 22 | 2.19 | 0.0015 | 0.001610 | 0.001610 | 0.001610 | 0.001610 | 0.001610 | 0.001610 | 0.001610 | 0.001610 | 0.001610 | 0.001610 |
| CO₂ (14.6%) + N₂ (85.4%) + brine (10 wt.% NaCl), at 273.25 K | 23 | 2.20 | 0.0016 | 0.001710 | 0.001710 | 0.001710 | 0.001710 | 0.001710 | 0.001710 | 0.001710 | 0.001710 | 0.001710 | 0.001710 |
| CO₂ (61%) + N₂ (39%) + H₂O at 303.05 K | 24 | 2.21 | 0.0017 | 0.001810 | 0.001810 | 0.001810 | 0.001810 | 0.001810 | 0.001810 | 0.001810 | 0.001810 | 0.001810 | 0.001810 |

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...
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.

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 water.85.

**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 follows:86:

\[
\text{Relevancy factor} = \frac{\text{Importance of input parameter}}{\text{Total importance of all input parameters}}
\]

\[
\text{Importance of input parameter} = \left( \frac{\text{Change in model output}}{\text{Change in input parameter}} \right) \times \text{Absolute value of input parameter}
\]

**Table 7. Predictions of EOSs and smart models for N₂ solubility in different CO₂ + N₂ + H₂O (brine) systems.**
Here, \( \text{inp}_{i,0} \) and \( \text{inp}_{i,j} \) 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 follows:

\[
    r(inp_{i,j}, O) = \frac{\sum_{j=1}^{n} \left( \text{inp}_{i,j} - \text{inp}_{m,j} \right) (O_j - O_m)}{\sqrt{\sum_{j=1}^{n} \left( \text{inp}_{i,j} - \text{inp}_{m,j} \right)^2 \sum_{j=1}^{n} (O_j - O_m)^2}}
\]

(17)

Figure 4. Cross plots of the developed models in this study.
Figure 5. Error distribution plots of the developed models for training and test sets.
Figure 5. (continued)
(a) 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.

(b) If the relevancy factor = 0, there is no relation between the input parameter and output or this relation is not monotonic.

(c) 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 CO2 in gaseous phase on the solubility of CO2–N2 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 CO2 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 CO2 in gaseous phase, the solubility of CO2–N2 mixture in water and brine solutions increases, and with increasing ionic strength, the solubility decreases.

Implementation of Leverage method. The Leverage method88–90 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:

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

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 domain47.

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
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.
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 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 values⁹ 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.
5. Increasing the water salinity causes the reduction of CO$_2$ and N$_2$ solubility in water.
6. The impact of mole percent of CO$_2$ in gaseous phase, temperature, and pressure on increasing the solubility of CO$_2$ and N$_2$ in water is incremental, and the impact of ionic strength on the solubility of CO$_2$ and N$_2$ in water is decreasing.

Figure 10. Effect of salinity on (a) CO$_2$ solubility and (b) N$_2$ solubility in the N$_2$ (85.4%) + CO$_2$ (14.6%) + H$_2$O (brine) systems at a temperature of 273 K; experimental data$^{10}$ with predictions of the Random Forest model.
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
1. 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).
2. 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).
3. 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).
4. Zaidin, M. F., Kantaatmadja, B. P., Chapoy, A., Ahmadi, P. & Burgass, R. in *SPE Middle East Oil and Gas Show and Conference* (OnePetro).
5. 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 Sø-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. & Shahraei, 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. Gillfillan, 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. Garagheizi, F., Eslamimanesh, A., Mohammadi, A. H. & Richon, D. Representation/prediction of solubilities of pure compounds in water using artificial neural network-group contribution method. Ind. Chem. Eng. Data 56, 720–726 (2011).

23. Peng, D.-Y. & Robinson, D. B. A new two-constant equation of state. J. Chem. Phys. 15, 59–64 (1976).

24. Bamberger, A., Sieder, G. & Maurer, G. High-pressure (vapor+ liquid) equilibrium in binary mixtures of (carbon dioxide + water) at temperatures up to 350 K. Ind. Eng. Chem. Res. 17, 97–108 (1980).

25. Spycher, N., Pruess, K. & Ennis-King, J. CO2–H2O mixtures in the geologic formation 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. Chapy, A., Mohammadi, A., Charetton, 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. Zabalyol, M., Mabe, G., Bottini, S. & Brignole, E. Vapour 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. Buji, 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 amines based on the mobile order theory. Pharm. Res. 11, 201–205 (1994).

35. Austegard, A., Solbøe, E., De Koeijer, G. & Mohvik, 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., Lernraski, 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 behaviour modelling of asphaltene precipitation utilizing MLP-ANN approach. Pet. Sci. Technol. 35, 2009–2015 (2017).

40. Sulaymani, M. & Remani, 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, 1026–1033 (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, 256–263 (2017).

45. Zhang, J., Lee, S. & Lee, J. W. Solubility of CO2, N2, and CO2+ N2 gas mixtures in isooctane. J. Chem. Eng. Data 53, 1321–1324 (2008).

46. Esalamimanesh, A., Garagheizi, 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 1-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. Vagnnik, 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 method for introduction to Random Forest for genetic association studies and signal processing. 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. Pref. 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 outdoor 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. Mach. Learn. 89, 295–335 (2012).

61. Shaikhina, T. et al. 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).

62. Ye, J. & Yang, L. in 2018 5th International Conference on Systems and Informatics (ICSAI), 139–143 (IEEE).

63. Hu, D., Zhang, C., Cao, W., Lv, X. & Xie, S. Grain yield predict based on GRA-AdaBoost-SVR model. Procedia Comput. Sci. 1, 193–201 (2010).

64. Mohammad, M.-R. et al. Modeling hydrogen solubility in hydrocarbons using extreme gradient boosting and equations of state. Sci. Rep. 11, 1–20 (2021).

65. Madani, S. A. et al. Comparision of experimental and geochemical modeling methods. Int. J. Hydrogen Energy 44, 23975–23983 (2019).

66. Chandran, J. M. & Sreekumar, A. C. S. Prediction of hydrogen solubility in hydrocarbon fuels. Fuel 231, 1–8 (2018).

67. Valderrama, I. O. A generalized Patel-Teja equation of state for polar and nonpolar fluids and their mixtures. J. Chem. Eng. Data. 33, 176–186 (1988).

68. Valderrama, I. O. A generalized Patel-Teja equation of state for polar and nonpolar fluids and their mixtures. J. Chem. Eng. Data. 33, 176–186 (1988).
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