scientific reports

OPEN

Modeling the solubility of light hydrocarbon gases and their mixture in brine with machine learning and equations of state

Mohammad-Reza Mohammadi[1], Fahimeh Hadavimoghaddam[2,3], Saeid Atashrouz[4]✉, Ali Abedi[5], Abdolhossein Hemmati-Sarapardeh[1,6]✉ & Ahmad Mohaddespour[7]✉

Knowledge of the solubilities of hydrocarbon components of natural gas in pure water and aqueous electrolyte solutions is important in terms of engineering designs and environmental aspects. In the current work, six machine-learning algorithms, namely Random Forest, Extra Tree, adaptive boosting support vector regression (AdaBoost-SVR), Decision Tree, group method of data handling (GMDH), and genetic programming (GP) were proposed for estimating the solubility of pure and mixture of methane, ethane, propane, and *n*-butane gases in pure water and aqueous electrolyte systems. To this end, a huge database of hydrocarbon gases solubility (1836 experimental data points) was prepared over extensive ranges of operating temperature (273–637 K) and pressure (0.051–113.27 MPa). Two different approaches including eight and five inputs were adopted for modeling. Moreover, three famous equations of state (EOSs), namely Peng-Robinson (PR), Valderrama modification of the Patel–Teja (VPT), and Soave–Redlich–Kwong (SRK) were used in comparison with machine-learning models. The AdaBoost-SVR models developed with eight and five inputs outperform the other models proposed in this study, EOSs, and available intelligence models in predicting the solubility of mixtures or/and pure hydrocarbon gases in pure water and aqueous electrolyte systems up to high-pressure and high-temperature conditions having average absolute relative error values of 10.65% and 12.02%, respectively, along with determination coefficient of 0.9999. Among the EOSs, VPT, SRK, and PR were ranked in terms of good predictions, respectively. Also, the two mathematical correlations developed with GP and GMDH had satisfactory results and can provide accurate and quick estimates. According to sensitivity analysis, the temperature and pressure had the greatest effect on hydrocarbon gases' solubility. Additionally, increasing the ionic strength of the solution and the pseudo-critical temperature of the gas mixture decreases the solubilities of hydrocarbon gases in aqueous electrolyte systems. Eventually, the Leverage approach has revealed the validity of the hydrocarbon solubility databank and the high credit of the AdaBoost-SVR models in estimating the solubilities of hydrocarbon gases in aqueous solutions.

Abbreviations

AAPRE	Average absolute percent relative error
AdaBoost	Adaptive boosting
AdaBoost-SVR	Adaptive boosting support vector regression
DT	Decision tree
EOS	Equation of state

[1]Department of Petroleum Engineering, Shahid Bahonar University of Kerman, Kerman, Iran. [2]Key Laboratory of Continental Shale Hydrocarbon Accumulation and Efficient Development (Northeast Petroleum University), Ministry of Education, Northeast Petroleum University, Daqing 163318, Heilongjiang, China. [3]Institute of Unconventional Oil and Gas, Northeast Petroleum University, Daqing 163318, China. [4]Department of Chemical Engineering, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran. [5]College of Engineering and Technology, American University of the Middle East, Kuwait City, Kuwait. [6]College of Construction Engineering, Jilin University, Changchun, China. [7]Department of Chemical Engineering, McGill University, Montreal, QC H3A 0C5, Canada. ✉email: s.atashrouz@gmail.com; saeid_atashrouz@aut.ac.ir; hemmati@uk.ac.ir; aut.hemmati@gmail.com; ahmad.mohaddespour@mail.mcgill.ca

ET	Extra tree
exp	Experimental
PR	Peng–Robinson
pred	Predicted
RMSE	Root mean square error
r	Relevancy factor
SD	Standard deviation
SVR	Support vector regression
SRK	Soave–Redlich–Kwong
RF	Random forest
R^2	Coefficient of determination
VPT	Valderrama modification of the Patel–Teja

One of the crucial theoretical and practical challenges in petroleum, chemical, and geochemical engineering is the solubilities of hydrocarbons, such as methane, ethane, propane, *n*-butane, or their mixtures, in pure water and aqueous electrolyte solutions. Achieving optimal conditions for gas and oil transportation, designing thermal separation processes, coal gasification, and hydrate formation require accurate information about the solubilities of hydrocarbon gases in different aqueous phases[1–5]. Natural gases coexist with aqueous solutions in petroleum reservoirs under the circumstances of high temperature and high pressure, which makes the solubilities of gases an important challenge for engineers. The water content of gases can undergo a phase alteration from vapor to gas hydrates, water condensate, and ice in the production and transportation of hydrocarbons. The condensed water phase in the compressor can damage impeller blades. Also, corrosion and pipeline blockage, as two serious flow assurance problems, can be caused by the formation of gas hydrates and/or ice throughout the production and transportation of hydrocarbons[1,6–8]. From an environmental point of view, gases solubility in water is a substantial problem because of the legislation and restrictions on the hydrocarbons contents in the water disposal[9]. In addition, leaking pipelines, underground oil storage tanks, and accidents on oil platforms and ships of the hydrocarbons' transportation are responsible for oil spillage in water[10–12].

Because of complex non-idealities from the strong H-bonding of water molecules, an accurate description of the phase behavior of these systems, utilizing theoretical methods is a challenging issue[13]. Accurate gas solubility data is essential to develop thermodynamic models for giving a qualified evaluation of the water content in the gases phase[9]. Therefore, the objective of thermodynamic calculations is the estimation of the compositions, content, and other equilibrium properties of the phases. Traditional equations of state (EOSs) are mainly applied to estimate thermodynamic and physical properties such as gas solubility. However, accurate estimates of gases solubility in various solvents by EOSs face serious problems such as iterative calculations, limited flexibility, and adjustable parameters at different temperatures and pressures. This makes the application of current conventional approaches, for example EOSs, unreliable and convinces researchers to seek better predictive techniques[14–19].

The petroleum industry needs appropriate and precise knowledge of the correlation between operating conditions (i.e., pressure and temperature), vapor and liquid phases compositions, and the salinity of the aqueous phase for the systems containing aqueous electrolyte solutions and natural gas' components. This knowledge can help design/optimize the operating condition for gas processing units and avoid/diagnose problems accompanying natural gas applications. Literature survey shows that there are many sets of experimental solubility data for various gas – liquid systems. Available experimental sources mainly present the solubility of pure hydrocarbon gases[2,4,20–22], hydrocarbon gas mixtures[1,5,6,9,23–25], and non-hydrocarbon gases (e.g., N_2 and CO_2)[26–30] in water/brine systems. On the other hand, due to the difficulties encountered in measuring the low content of water of gases at low-temperature and high-pressure conditions, experimental data of water content of hydrocarbon and non-hydrocarbon gases are limited and scattered. However, Mohammadi et al.[1] demonstrated that complexities associated with experimental measurement of the water content in natural gas could be eliminated by gas solubilities data, which provides an accurate estimate of water content[1]. Attempts to model the vapor–liquid phase equilibria of non-hydrocarbon and hydrocarbon gases and brine solutions have always been considered by researchers due to the limited number of measurements. The activity coefficient, Henry's constant approach, and EOSs were widely used in thermodynamic models in order to gain information about the equilibrium conditions of non-hydrocarbon and hydrocarbon gases and pure water or aqueous electrolytes solutions[5,9,31–41]. Although Henry's law can appropriately be utilized to estimate the solubilities, this approach has several drawbacks. For instance, this approach is correct for unique compounds at low concentrations under equilibria conditions with no chemical reactions for the aqueous phase. Also, it is appropriate for near-ideal or dilute solutions[42]. Moreover, at low temperatures, there is a limited count of Henry's constants for the systems containing hydrocarbons-aqueous solutions[3]. On the other hand, the advantages such as lower count of parameters, the easiness of implementation, and computational efficiency make the use of EOSs widespread[2,4,9,43]. However, the accuracy of EOSs is highly dependent on the appropriation of empirical adjustments via incorporating the binary interaction parameters. Therefore, reliable sources of experimental data for the vapor–liquid equilibria of binary or even multi-component mixtures are essential to determine these parameters[23,44]. Hence, developing EOS for extensive applications such as calculations of natural gas' solubility faces serious problems, and numerous EOSs developed so far are mostly attributed to limited systems. Due to the above discussions, in recent years, researchers have tried to provide accurate and reliable approaches to predict the solubilities of non-hydrocarbon and hydrocarbon gases in pure water and aqueous electrolyte systems. Literature survey shows that many intelligent models have been proposed to estimate the solubilities of non-hydrocarbon gases, especially CO_2, in water and brine[45–50]. Regarding hydrocarbons solubility in pure water and brine, Safamirzaei et al.[51] utilized a simple artificial neural network (ANN) with overall 101 solubility data points for modeling *n*-alkanes (*n*C1–*n*C6) solubilities in water. They showed that an ANN-based model could be an alternative to other methods such as EOSs with high accuracy[51]. Samani et al.[52] proposed two hybrid models based on least-squares support vector machine and coupled simulated annealing algorithms for estimating the solubility of hydrocarbons (C1–C4) and non-hydrocarbon gases (CO_2 and N_2) in aqueous electrolyte systems. Regarding hydrocarbon gases, their database had 1175 solubility data points, and the average absolute error of their proposed model was 30.6%[52]. Nabipour et al.[53] used a similar database including 1175 data points and an extreme learning machine algorithm to develop a model for predicting hydrocarbon gases (C1–C4) solubility in electrolyte solutions. The mean relative error of their model was 22.05%[53]. Although two relatively comprehensive intelligent models have been developed to predict the solubilities of hydrocarbon gases in aqueous electrolyte systems, the error of these models is slightly high. Also, due to the nature of the data-driven soft computing approaches, incorporating a larger number of data, various operating conditions, and adopting different modeling approaches may propel a comprehensive predictive tool for estimating the solubilities of light hydrocarbon gases and their mixture in water and aqueous electrolyte solutions. Furthermore, the development of easy-to-use mathematical correlations by advanced algorithms can simplify and accelerate the prediction of hydrocarbon gas solubilities in brine.

In this research, a huge database (1836 experimental data points) of hydrocarbon gases solubilities in pure water and aqueous electrolyte systems was accumulated from the literature. Next, for developing predictive tools, six robust machine learning algorithms viz., Random Forest, Extra Tree, adaptive boosting support vector regression (AdaBoost-SVR), Decision Tree, genetic programming (GP), and group method of data handling (GMDH) are implemented in this study by considering two different approaches. Additionally, three famous equations of state (EOSs) viz., Peng–Robinson (PR), Valderrama modification of the Patel–Teja (VPT), and Soave–Redlich–Kwong (SRK) are utilized in comparison with machine learning models. Furthermore, the performance of machine learning-based predictive tools and mathematical correlations is studied by employing various statistical and visual error analyses. Besides, a well-known sensitivity analysis, i.e., the relevancy factor, is identified the relative impact of input variables on hydrocarbon gases solubility in brine. Ultimately, the validity of the solubility databank, along with the application domain of the best-developed predictive tools in the present work, is examined by the Leverage mathematical method.

Data acquisition

In this work, a large databank was collected on the basis of experimental solubility data of light hydrocarbon gases and their mixtures in water and aqueous electrolytes. This databank consists of 1836 data points that are 661 data points more than what is used in Samani et al.[52] and Nabipour et al.[53] works. Table 1 presents the details and references of experimental solubility data for hydrocarbon components of natural gas in pure water and aqueous electrolytes used in this survey. It should be noted that the collected laboratory data for the solubility of gases in pure water and brine is such that most of the solubility values were reported in two-phase conditions (a gaseous phase and an aqueous phase in equilibrium). This means that the temperature and pressure of the system were such that only two phases would exist in equilibrium. This is while there is a possibility of the formation of three phases at conditions of pressure higher than the critical pressure of components or low-temperature conditions. According to the Gibbs phase rule, degrees of freedom are the number of intensive properties that can be altered without varying the number of phases, or the number of components in any phase[54]. Hence, in some studies such as Amirijafari's work[23], for measuring hydrocarbon gas solubility in water under high-pressure conditions, the temperatures were selected such that only two phases (hydrocarbon gas mixture and the liquid water with hydrocarbons dissolved in it) would be present. Adopting this approach makes measuring gas solubilities easier and the obtained data more reliable. Although in some other studies[5,6], in addition to measuring the solubility data in the two-phase state, the solubility values have been measured in the three-phase conditions, i.e. (three-phase equilibrium between the hydrate, the aqueous, and the vapor phase or three-phase equilibrium between water-rich liquid, hydrocarbon-rich liquid, and vapor phase). However, experimental measurements of solubilities in such a condition are challenging and could potentially generate unreliable laboratory data. For example, concentrations of light hydrocarbon gases in water are low, and moreover reaching the equilibrium states near and inside the gas hydrate formation region is a time-consuming process. However, the data collected in this research were all carefully selected from reliable references where considerable time has been spent on conducting experiments and calculated solubility values using specific methods, especially in three-phase conditions. Further explanation of the laboratory process for calculating gas solubility is beyond the scope of this work and interested readers are referred to the literature[6,55,56]. It should be mentioned that what is mentioned as gas solubility in this study is x = mole fraction of hydrocarbon gas in the aqueous liquid phase, which is collected from reliable references reported in Table 1.

Literature survey reveals that the gaseous phase composition, aqueous phase composition, temperature, and pressure highly affect the solubilities of hydrocarbon gases in the aqueous solutions[1,5,6,9,68]. The ionic strength (I) as a single characteristic of aqueous electrolyte solutions was utilized in the modeling process instead of multiple salt concentrations of brine solutions in order to reduce the dimensions of the modeling process. Considering m_i as the molar concentration of each ion and z_i as valance of charged ions in brine solutions, the ionic strength (I) is defined as follows:

$$\mathrm{I} = \frac{1}{2}\sum \mathrm{m_i}|\mathrm{z_i}|^2 \tag{1}$$

In this study, two approaches were considered for modeling. First, hydrocarbon gases solubility (η_h: mole fraction) is assumed to be a function of eight independent parameters: temperature (K), pressure (MPa), ionic strength of the solution (M), the mole percent of each component (C1, C2, C3, and C4) in the gas mixture, and

Solubility system	Pressure (MPa)	Temperature (K)	Solubility (mole fraction)	References
Methane + pure water	0.973–17.998	275.11–313.11	C1: 0.000204–0.002459	9
	2–40.03	283.2–303.2	C1: 0.000563–0.004049	24
	2.5–100	344.25	C1: 0.000127–0.005085	5
	2.53–60.8	293.1–353.1	C1: 0.000361–0.004328	25
	4.13–34.47	310.9–344.2	C1: 0.000602–0.00335	23
	1.327–6.451	297.5–518.3	C1: 0.0002124–0.0010337	20
	9.81–113.27	423.2–633.2	C1: 0.001–0.18	57
	0.101325	273.15–283.15	C1: 0.0000444–0.0000345	58
	0.101325	273.42–353.15	C1: 0.0000188–0.0000445	59
Ethane + pure water	0.5–4	283.2–303.2	C2: 0.000119–0.000864	24
	0.8–69.61	310.92–444.26	C2: 0.0000698–0.0033	21
	2.5–100	344.25	C2: 0.000821–0.001398	5
	0.05074–0.11	275.44–323.15	C2: 0.00002073–0.0000725	60
	0.373–4.952	274.26–343.08	C2: 0.0000854–0.0009696	61
	20–370	473.15–673.15	C2: 0.005–0.34	62
	0.101325	285.5–345.6	C2: 0.000016– 0.0000434	63
Propane + pure water	0.357–3.915	277.62–368.16	C3: 0.0000321–0.0002694	2
	0.0995–3.409	288.7–410.9	C3: 0.0000078–0.000313	64
	0.49–4.269	278.87–422	C3: 0.0000796–0.000366	65
	0.101325	285.45–347.25	C3: 0.0000118–0.0000415	63
n-Butane + pure water	2.5–100	344.25	C4: 0.000021–0.000103	5
	0.12–3.044	310.9–410.9	C4: 0.000016–0.0001771	22
	25.5–83	628.15–637.15	C4: 0.025–0.077	62
	0.101325	277.15–328.15	C4: 0.000011–0.000058	66
Methane/ethane + pure water	1–4	275.2–283.2	C1: 0.000643–0.00115 C2: 0.000098–0.0001475	24
	4.58–54.572	310.9–344.2	C1: 0.00045–0.003336 C2: 0.000232–0.002439	23
Methane/propane + pure water	4.92–55.26	377.59	C1: 0.000862–0.003702 C3: 0.00015–0.001863	23
Ethane/propane + pure water	4.58–55.26	377.59	C2: 0.000208–0.000929 C3: 0.000188–0.000642	23
Methane/ethane/propane + pure water	4.58–34.57	344.26–377.59	C1: 0.000768–0.003276 C2: 0.000119–0.001396 C3: 0.0000019–0.000607	23
Methane/ethane/*n*-butane + pure water	0.987–14.407	278.14–313.12	C1: 0.000218–0.002191 C2: 0.000014–0.000067 C4: 0.00000387–0.0000112	9
Methane + pure water, NaCl	10.13–61.6	324.65–398.15	C1: 0.000805–0.0043	67
Methane + pure water, NaCl, LiCl, NaBr, NaJ, $CaCl_2$	4.09–45.89	298.15–423.15	C1: 0.00017–0.00269	68
Methane + pure water, KCl, LiBr, KBr, LiCl	0.3–10.23	313.1–373.2	C1: 0.00003–0.00154	4
Methane/ethane/propane + pure water, NaCl	6.22–20.1	274.55–299	C1: 0.00099–0.0028 C2: 0.000038–0.00024 C3: 0.000006–0.000042	6

Table 1. The solubility systems of light hydrocarbon gases in pure water and aqueous electrolyte systems.

carbon number (*IDX*: 1, 2, 3, and 4) of the gas component (methane, ethane, propane, and *n*-butane) whose solubility is to be predicted:

$$\eta_h = f(\mathrm{P, T, I, C1, C2, C3, C4, IDX}) \tag{2}$$

The mentioned approach is similar to that utilized in Samani et al.[52] and Nabipour et al.[53] works. The second approach is that hydrocarbon gases solubility (η_h: mole fraction) is assumed to be a function of five input parameters: pressure (MPa), temperature (K), ionic strength of the solution (M), the pseudo-critical temperature of the gas mixture (T_{pc}), and the critical temperature of the gas component (Tc_{gas}) whose solubility is to be predicted:

$$\eta_h = f\left(\mathrm{P, T, I, T_{pc}, Tc_{gas}}\right) \tag{3}$$

Here, if Tc_i is the critical temperature of individual components and y_i is the molar fraction of individual components in the gas mixture of n components, T_{pc} can be calculated as follows[69]:

	IDX	Temperature (K)	Pressure (MPa)	Ionic strength (M)	C1 (mole %)	C2 (mole %)	C3 (mole %)	C4 (mole %)	T_{pc} of gas mixture (K)	Tc_{gas} (k)	Solubility (mole fraction)
Mean	1.829521	341.1801	14.11	3.252	56.65336	20.70442	15.66009	6.98213	258.7715	268.3451	0.002634
SD	0.978137	64.15295	19.78	7.656	45.11362	36.41282	33.02423	24.61655	1.79276	1.96642	0.013492
Minimum	1	273.15	0.051	0	0	0	0	0	190.56	190.56	3.87E−06
Maximum	4	637.15	113.27	37.351	100	100	100	100	425.12	425.12	0.18

Table 2. Statistical description of the solubility databank utilized in the present research.

$$T_{pc} = \sum_{i=1}^{n} y_i Tc_i \tag{4}$$

In the second approach, although the number of parameters has been reduced, by using the parameters of the pseudo-critical temperature of the gas mixture and the critical temperature of gaseous components instead of the mole percent of each component in the gas mixture and the carbon number, the development of the model becomes more general. Table 2 presents the statistical details of the databank (including all inputs utilized in both modeling approaches along with hydrocarbon gases solubility as the models' target) utilized to model the solubility of light hydrocarbon gases and their mixtures in water and aqueous electrolyte solutions.

Table 2 reports that the ionic strength of brine solutions based on molarity is in the range of 0–37.351 M. The mole percent of light hydrocarbon gases (C1-C4) in the gaseous mixture was in the range of 0–100%. The experimental solubility data of light hydrocarbons and their mixtures in water and aqueous electrolyte systems have also been gathered over broad ranges of operating temperatures, 273.15–637.15 (K), and pressures, 0.05–113.27 (MPa). Hence, the variety of input variables is broad enough to provide a general machine learning-based predictive tool for estimating light hydrocarbon gases and their mixtures in water and aqueous electrolyte systems.

Model development

Adaptive boosting (AdaBoost).

The Adaptive boosting (AdaBoost) technique established by Freund and Schapire[70] seeks to develop a powerful classifier by integrating weak classifiers and benefiting from their failures. In other words, it repeatedly chooses the training inputs in order to complement several classifiers and apply the proper weight for every classifier depending on its performance, with larger weights allocated to miscategorized data sets. The following are the common parts of the AdaBoost procedure[71]:

Step 1: Weights determination: $w_j = \frac{1}{n}.j = 1.2.\ldots.n$

Step 2: Providing the training data to a weak learner $Wl_i(x)$, assigning weights, and calculating the weighted error for each i.

$$Err_i = \frac{\sum_{j=1}^{n} w_j I(t_j \neq wl_i(x))}{\sum_{j=1}^{n} w_j}, I(x) = \begin{cases} 0 if x = false \\ 1 if x = true \end{cases}$$

Step 3: The weights should be calculated for each i for estimators: $\beta_i = log\left(\frac{(1-Err_i)}{Err_i}\right)$

Step 4: Changing the weights of the data for each i to N (N refers to the count of the learner).

Step 5: Setting a weak learner to the data test (x) as a response.

Support vector regressors are utilized as weak learners in the AdaBoost algorithm in this research.

Support vector machine for regression (SVR).

Although support vector machine is a collection of controlled machine learning techniques that may be applied for regression and classification[72], support vector regression (SVR) is routinely used for soft calculation since it has a well-defined mathematical model. Because of its consistency in simulating numerous complicated structures, SVR has recently piqued researchers' curiosity. Since the main theory of SVR has been published[73], it is just shortly presented in this work for the sake of brevity. The SVR objective is to catch a regressor *f(x)* for such a sample data $[(x_1.y_1)\ldots..(x_n.y_n)]$, having $x \in R_d$ as the d-dimensional input dataset and $y \in R$ as the output variable (which relies on the inputs), in order to calculate the output:

$$f(x) = w.\phi(x_i) + b \tag{5}$$

Here w denotes weight, b indicates bias vectors, and $\phi(x)$ represents the kernel function. To get the proper aforementioned parameters, Vapnik et al.[74] developed the following minimizing method:

$$minimize \frac{1}{2} w^T w + C \sum_{j=1}^{N} \left(\zeta_j^- + \zeta_j^+\right)$$

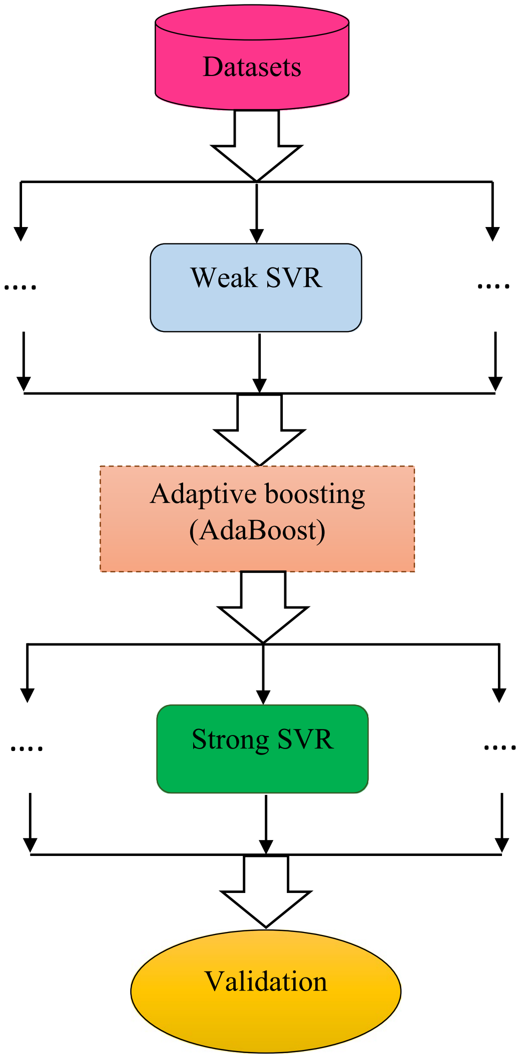


Figure 1. Schematic illustration of the proposed AdaBoost-SVR.

$$\begin{cases} (w.\varnothing(x_i) + b) - y_i \leq \varepsilon + \zeta_j^- \\ y_i - (w.\varnothing(x_i) + b) \leq \varepsilon + \zeta_j^+ \\ \zeta_j^+.\zeta_j^- \geq 0.i = 1.2\ldots.m \end{cases} \tag{6}$$

where transposed matrix of w is represented by w^T, error connivance by ε, positive factors expressing the lower and higher extra variances by ζ_j^+ and ζ_j^-, and positive regularization parameter indicating the variation from ε by C.

The abovementioned constraints optimization issue is transformed into a dual function utilizing Lagrange multipliers, yielding the subsequent solution:

$$f(x) = \sum_{j=1}^{n} (a_k - a_k^*)K(x_k, x_l) + b \tag{7}$$

where a_k^* and a_k indicate the Lagrange multipliers, while $K(x_k.x_l)$ is the kernel function. Figure 1 presents a schematic image of the proposed AdaBoost-SVR in this study.

Decision tree (DT). This method[75] is derived from natural sources and may be used to tackle both regression and classification problems. Root nodes, leaf nodes, internal nodes, and branches make up this system. The inputs are carried by the root node, which is the initial portion of the proposed technique. The last section of the diagram, known as the leaf nodes or final nodes, represents the model's output. Between the root and leaf nodes are internal nodes. The nodes are linked together by branches. Pruning, dividing, and halting are the three major activities used to build a decision tree[76]. The data dividing stage begins from the root node just before

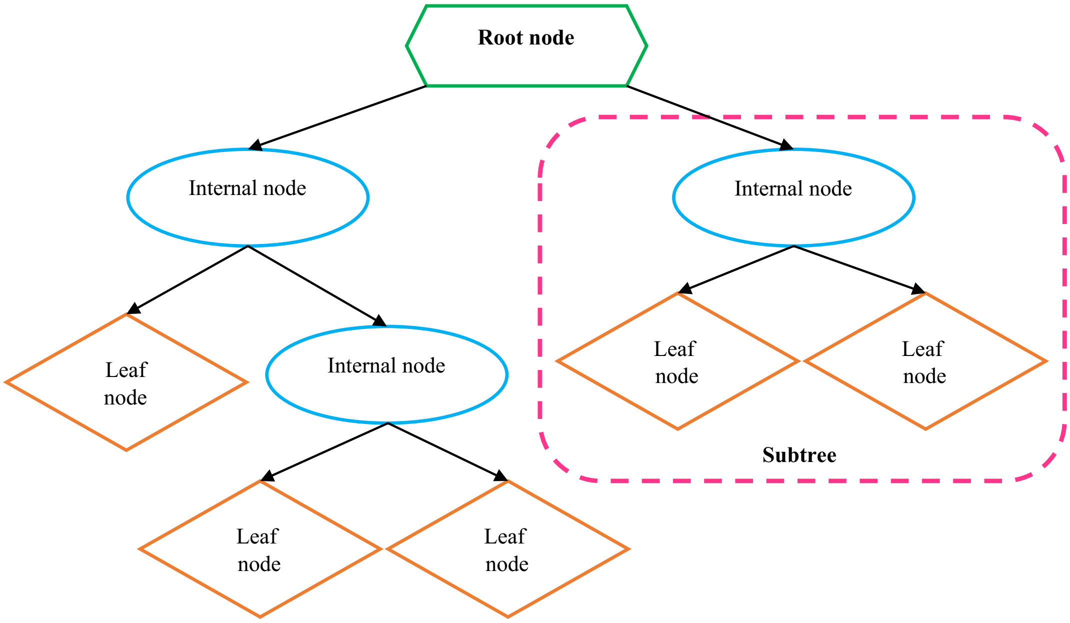


Figure 2. Schematic illustration of a typical decision tree.

data is presented to the system. This process of separating proceeds until a stopping condition is met[77]. Figure 2 depicts the basic DT.

Random forest (RF).

The decision tree is an effective machine learning technique; however, it has two flaws. First, while the estimation error of the decision tree is typically low in training data, the forecasting deviation is sometimes high because it is susceptible to small disturbances in the training samples; second, while the separating law in each node is desirable, according to the previous section, this greedy strategy cannot assure that the overall decision tree is the best. By simultaneously training many trees and transforming several weak learners into powerful learners, ensemble techniques can address these two problems. A random forest is made up of a set of different decision trees that are all being learned at the same time. The system determines the superiority and significance of each decision tree[78]. Furthermore, a constructed attribute of the Classification model that is used to choose different attributes allows the RF to govern various inputs characteristics without the requirement to remove a set of variables for dimension decrement [79]. The RF approach uses a process called Bagging throughout the simulation to increase the variety of trees in the forest. Usually, the system provides the number of trees as an input, and the algorithm divides datasets into distinct groupings as a result. Bagging is a sort of sample selection approach that uses only a third of the datasets in the learning phase of the subtree creation procedure, with the other inputs being known as the out-of-bag data (OOB). Moreover, verification of outputs is not necessary for the RF during model building since the correctness of the model may be assessed utilizing OOB's errors[80]. The RF technique is shown in Fig. 3. If the system is provided with a training dataset as a prerequisite, the training procedure will be completed. If you have a training sample in the form of $D = [(x_1.y_1).(x_2.y_2)\ldots.(x_n.y_n)]$, D_t is the described training data for tree h_t, and the final estimation of the out-of-bag dataset of sample x is H^{oob}, as shown:

$$H^{oob}(x) = argmax\sum_{t=1}^{T} I(h_t(x)) = y \tag{8}$$

The error of the OOB data is extended as following for modeling purposes:

$$\varepsilon^{oob}(x) = \frac{1}{|D|}\sum_{(x.y)\epsilon D} I(H^{oob}(x) \neq y) \tag{9}$$

The functioning of the RF must be randomized, and this characteristic is regulated by the variable $k = log_2 d$ [80]. The following equation may be used to determine the importance of a feature of a parameter X_i:

$$I(X_i) = \frac{1}{B}\sum_{t}^{B} \widetilde{OOBerr}_{t^i} - OOBerr_t \tag{10}$$

Correspondingly, the ith component is characterized by X_i in the X vector, B represents the number of trees in the existing RF, the original OOB datasets are offered as the $\utilde{OOBerr}_t$, which involves the replaced parameters, and the estimated error of the OOB samples is described by $\widetilde{OOBerr}_{t^i}$, which refers to the attribute X_i of tree t.

Extra tree (ET).

The Extra trees [81] are a novel machine learning approach that was created as an improvement of the random forest model and is less prone to over-fit a database[81]. Extra tree (ET) randomly selects a set

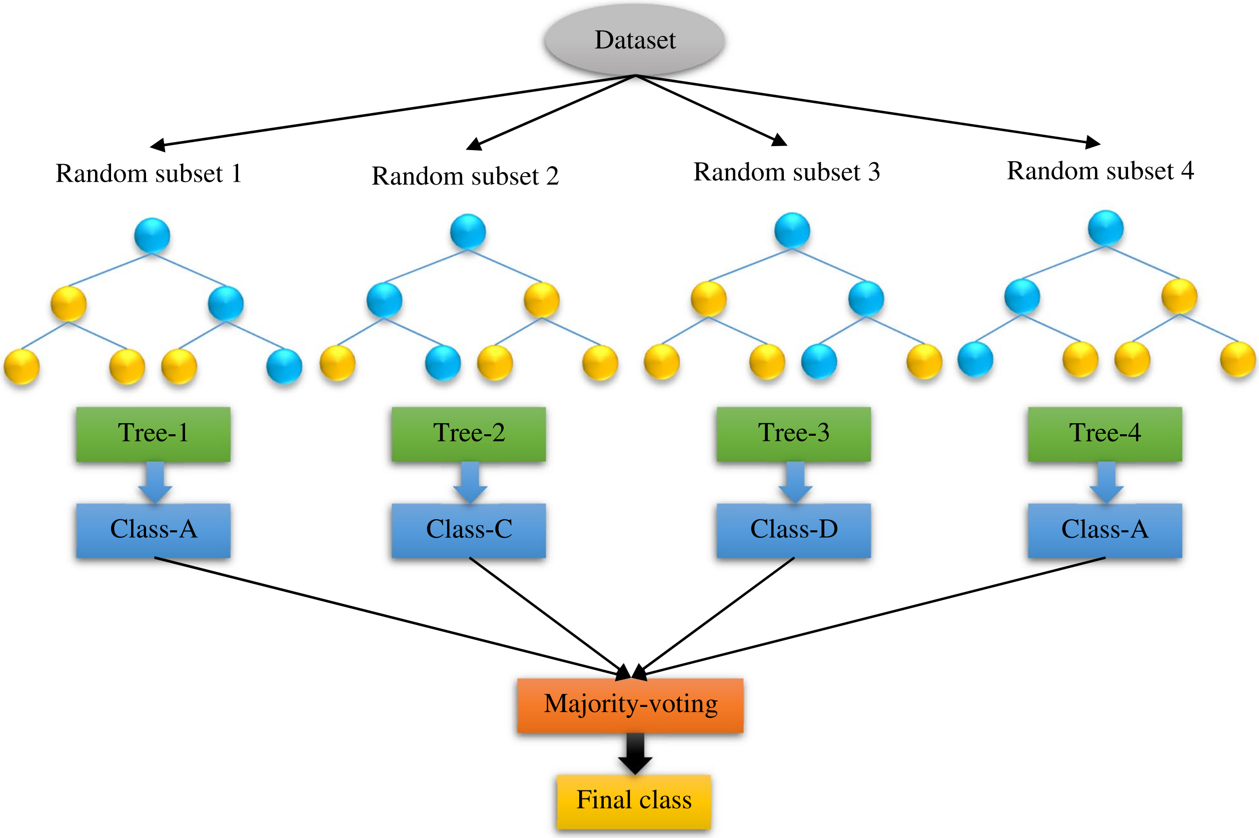


Figure 3. A schematic of the random forest model.

of attributes to train a basic predictor[82], using the same idea as random forest. For dividing the node, it chooses the best characteristic and the matching value at random[82]. For every regression tree, ET utilizes all training data. In contrast, RF's model is trained using a bootstrap replica.

Genetic programming (GP). GP is an organized method for getting machines to automatically solve a problem beginning with a high-level statement of what ought to be accomplished. GP is a systematic approach that is independent of a problem domain, that genetically reproduces a population of programs to solve a problem[83,84]. Programs are 'bred' through the continuous progress of an initially random population of programs. Actually, in this iterative improvement approach, at each new step of the algorithm, it selects only the fittest of the descendant to pass and regenerate in the subsequent production, which is occasionally referred to as a fitness function[85]. More explanations related to the application of this algorithm in the implementation of symbolic regression can be found elsewhere in the literature[86–88].

Group method of data handling (GMDH). GMDH[89] features fully automatic structural and parametric optimization of models and is a kind of inductive algorithm for computer-based mathematical modeling of multi-parametric datasets. In the inner levels of the GMDH method[90], there are multiple independent neurons. All neurons per layer are attached in couples via a quadratic polynomial and form individual neurons in the structure of polynomials in the subsequent layer[91]. Each GMDH neuron's generated value is determined by employing a quadratic polynomial representative that comprises the preceding neuron[92,93]. The quadratic polynomial procedures merging the neurons in the earlier levels will create the neurons in the subsequent layers[94]. To amend the limitations of the primary GMDH method[89], the hybrid GMDH is usually utilized which has more than two independent variables that can be combined concurrently and it permits the intersection of nodal within diverse layers. The succeeding formula shows the final form of the hybrid GMDH[95]:

$$\mathrm{Y_i} = \mathrm{a} + \sum_{i=1}^{M}\sum_{j=1}^{M}\cdots\sum_{k=1}^{M} b_{ij\ldots k}x_i^n x_j^n \ldots x_k^n \quad n = 1, 2, \ldots, 2^l \tag{11}$$

Here, M is the count of inputs, l stands for the count of layers, x_i, x_j, …, x_k are the inputs, a, $b_{ij\ldots k}$ denote the polynomial coefficients, and Y indicates the model output.

Equations of state (EOSs). An EOS is utilized to relate pressure, volume, and temperature (PVT) for both systems of a pure substance and for multi-component mixtures. There are many EOSs in the thermodynamic literature that is used to describe vapor–liquid-equilibria, solubility estimation, thermal features, and volumetric properties of a substance or multi-component mixtures[71]. In this work, three famous EOSs, namely SRK, VPT, and PR, have been utilized to estimate the solubility of light hydrocarbon gases in water with the purpose of comparing them with machine learning algorithms. Tables S1 in the Supplementary file presents the PVT relationships of these EOSs. Also, the parameters of considered EOSs are presented in Table S2. Besides, acentric factors and critical properties of the light hydrocarbon gases and water are represented in Table S3 used in EOSs.

Assessment of models

The following statistical factors viz., determination coefficient (R^2), average absolute percent relative error (AAPRE), root mean square error (RMSE), and standard deviation (SD) were employed to assess the accuracy of the machine learning models. The mathematical formula of these statistical criteria is defined below[96,97]:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\eta_{i,\exp} - \eta_{i,pred}\right)^2} \tag{12}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\eta_{i,\exp} - \eta_{i,pred})^2}{\sum_{i=1}^{N}(\eta_{i,\exp} - \overline{\eta_{\exp}})^2} \tag{13}$$

$$AAPRE = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{\eta_{i,\exp} - \eta_{i,pred}}{\eta_{i,\exp}}\right| \tag{14}$$

$$SD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{\eta_{i,\exp} - \eta_{i,pred}}{\eta_{i,\exp}}\right)^2} \tag{15}$$

where N refers to the count of data, $\eta_{i,exp}$ shows the experimental hydrocarbon gases solubility, and $\eta_{i,pred}$ is predicted hydrocarbon gases solubility in the liquid phase by presented models.

In the present research, the subsequent graphical analyses are utilized simultaneously to assess the performance of machine learning-based models and correlations:

Histogram plot: in this graph, the discrepancy between the experiments data and prediction of the model can be seen statistically, which helps to evaluate the model's performance.

Cross-plot: the cross-plot graph illustrates the correlation between experimental solubilities and predicted values by models with the fact that the higher the concentration of data nearby the unit-slope line, the better the model's prediction.

Error distribution plot: the scatter of error (exp-pred) around the zero-error line is evaluated to check for possible error trends.

Trend plot: the experiments data and prediction of the model are plotted versus a special property to assess the model's validation by checking the coverage of these data. High data coverage shows the high validity of the model.

Cumulative frequency graph: it is a statistical plot for quantifying the precision of the models, which is shown by drawing the cumulative frequency of data against absolute error (exp-pred).

Results and discussion

Correlations' development. As mentioned earlier, this work employed white-box modeling approaches to create precise predictive correlations for the solubility of light hydrocarbon gases and their mixture in brine. The correlations utilize the second modeling approach having five inputs (P, T, I, T_{pc} of gas mixture, Tc_{gas}) to calculate hydrocarbon gases solubility. The reason for choosing five parameters for the development of mathematical correlations was that, firstly, a simpler mathematical expression was obtained and solubility calculations become easier, and secondly, the correlation become more general by using the pseudo-critical of the gas mixture instead of using the percentage of gas (C1–C4) composition. The proposed correlations by GMDH and GP methods are presented below:

GMDH correlation:

$$Solubility = -0.000257478 + N_6 * 0.104357 + N_1 * 0.995504$$

$$N_1 = -0.000402032 + P * 3.34159e - 05 + N_2 * 0.976721$$

$$N_2 = 0.000417773 + N_5 * 0.163256 + N_3 * 0.277835 + {N_3}^2 * 6.25097$$

Statistical criteria		RMSE	SD	R^2	AARPE, %
Random forest (8 inputs)	Train	0.001099	0.47198	0.9928	15.092
	Test	0.001628	0.47280	0.9886	16.089
	Total	0.001223	0.47217	0.9917	15.292
Decision tree (8 inputs)	Train	0.000154	0.27784	0.9998	17.019
	Test	0.000383	0.63358	0.9991	20.762
	Total	0.000220	0.37761	0.9997	17.769
AdaBoost-SVR (8 inputs)	Train	0.000099	0.20911	0.9999	10.433
	Test	0.000101	0.25008	0.9999	11.497
	Total	0.000099	0.21807	0.9999	10.647
Extra tree (8 inputs)	Train	0.000218	0.23459	0.9997	11.979
	Test	0.002642	0.69527	0.9585	25.802
	Total	0.001199	0.37821	0.9921	14.750
Random forest (5 inputs)	Train	0.001099	0.61834	0.9928	15.365
	Test	0.001803	0.37921	0.9860	14.314
	Total	0.001272	0.57841	0.9911	15.154
Decision tree (5 inputs)	Train	0.000170	0.43871	0.9998	18.313
	Test	0.000391	0.85103	0.9991	21.875
	Total	0.000231	0.54727	0.9997	19.027
AdaBoost-SVR (5 inputs)	Train	0.000102	0.25916	0.9999	11.613
	Test	0.000109	0.44120	0.9999	13.643
	Total	0.000104	0.30470	0.9999	12.020
Extra tree (8 inputs)	Train	0.000331	0.22614	0.9994	11.413
	Test	0.002457	1.06098	0.9642	31.982
	Total	0.001138	0.52128	0.9928	15.536
GMDH correlation (5 inputs)	Train	0.001973	1.06744	0.9769	17.470
	Test	0.006485	0.88234	0.8190	34.834
	Total	0.003397	1.03387	0.9365	20.951
GP correlation (5 inputs)	Train	0.002456	0.57392	0.9643	13.640
	Test	0.006386	0.53905	0.8245	27.615
	Total	0.003605	0.56727	0.9286	16.441

Table 3. Statistical error analysis for the developed models and correlations.

$$N_3 = 0.000769644 + N_4 * N_5 * 81.1485 - {N_4}^2 * 31.6265 - {N_5}^2 * 30.9349$$

$$N_4 = 0.0113595 - T^2 * 1.51522e - 07 + \mathrm{T} * P * 3.24299e - 09 + \mathrm{T}^{\,4} * 4.06799e - 13 - P * 0.000290132 - \mathrm{P}^2 * 1.23427e - 06$$

$$\begin{aligned} N5 = 0.00995312 + \mathrm{Tc},^2 * 4.48223e - 08 - \mathrm{Tc}^2 * T_{pc}^2 * 5.36312e - 13 \\ + (\mathrm{Tc})^4 * 3.23202e - 14 - T_{pc}^2 * 1.85458e - 07 + T_{pc}^4 * 9.26622e - 13 \end{aligned} \tag{16}$$

$$N6 = 0.0128381 - \mathrm{Tc}^2 * 2.05784e - 07 + \mathrm{Tc}^2 * I * 5.76622e - 09 + (\mathrm{Tc})^4 * 8.16174e - 13 - I * 0.00081115 + \mathrm{I}^2 * 1.35367e - 05$$

GP correlation:

$$Solubility = \left(\left(\frac{\log(\log(c_0P + c_1))}{\frac{c_2Tc}{\exp(\frac{(c_3T)}{c_4I})}} - (\exp(c_5)\exp(c_6T) - \left(c_7T_{pc} + \log\left(\log\left(\log((c_8P + c_9))\right)\right)\right)\right)c_{10} + c_{11}\right) \tag{17}$$

$$c_0 = 0.909;\ c_1 = -19.076;\ c_2 = 0.45799;\ c_3 = 0.6495;\ c_4 = 15.867;\ c_5 = 4.777;\ c_6 = 0.026667;$$
$$c_7 = 0.87809;\ c_8 = 0.909;\ c_9 = -19.194;\ c_{10} = 9.7169E - 12;\ c_{11} = 0.0018755$$

Evaluation of the models. In the current study, R^2, AAPRE, SD, and RMSE were utilized to appraise the models' estimates. The results of these statistical criteria for all predictive tools are presented in Table 3. As can be observed in this table, for both modeling approaches, AdaBoost-SVR, Extra Tree, Random Forest, and DT models can be classified in terms of high exactness for predicting the whole dataset, respectively. However, for the test subset, AdaBoost-SVR, Random Forest, DT, and Extra Tree models, respectively, had the best estimates,

Solubility system	Data No.	P (MPa)	Gas solubility, mole fraction													
			Exp	DT (8 inputs)	Extra tree (8 inputs)	AdaBoost-SVR (8 inputs)	Random forest (8 inputs)	DT (5 inputs)	Extra tree (5 inputs)	AdaBoost-SVR (5 inputs)	Random forest (5 inputs)	GMDH correlation (5 inputs)	GP correlation (5 inputs)	PR	SRK	VPT
Methane + water, at 275 K[9]	1	0.973	0.000399	0.000263	0.000349	0.000393	0.000324	0.000263	0.000351	0.000379	0.000363	0.000331	0.000364	0.000298	0.000302	0.000351
	2	1.565	0.000631	0.000668	0.000524	0.000666	0.000784	0.000667	0.000581	0.000666	0.000737	0.000559	0.000593	0.000401	0.000703	0.000553
	3	2.323	0.000901	0.000668	0.000636	0.000863	0.000901	0.000668	0.000694	0.000866	0.000787	0.000764	0.000966	0.000608	0.001005	0.000805
	4	2.82	0.001061	0.000668	0.000624	0.000939	0.000919	0.000669	0.000698	0.000944	0.000902	0.000766	0.001046	0.000802	0.001204	0.000947
Ethane + water, at 303 K[61]	5	0.373	0.000134	0.000192	0.000150	0.000138	0.000170	0.000192	0.000156	0.000141	0.000134	0.000157	0.000171	0.000103	0.000102	0.000101
	6	0.719	0.000240	0.000192	0.000225	0.000245	0.000210	0.000192	0.000247	0.000246	0.000197	0.000240	0.000222	0.000193	0.000201	0.000205
	7	1.093	0.000346	0.000412	0.000353	0.000346	0.000388	0.000415	0.000356	0.000347	0.000355	0.000328	0.000275	0.000284	0.000296	0.000311
	8	1.598	0.000472	0.000675	0.000492	0.000491	0.000511	0.000675	0.000462	0.000487	0.000522	0.000452	0.000471	0.000396	0.000414	0.000414
	9	2.299	0.000630	0.000675	0.000598	0.000611	0.000616	0.000676	0.000570	0.000620	0.000606	0.000623	0.000629	0.000487	0.000508	0.000539
	10	2.932	0.000742	0.000675	0.000694	0.000734	0.000728	0.000677	0.000654	0.000740	0.000722	0.000727	0.000741	0.000584	0.000610	0.000638
	11	3.977	0.000883	0.000685	0.000755	0.000844	0.000800	0.000679	0.000731	0.000844	0.000767	0.000812	0.000882	0.000680	0.000702	0.000802
Propane + water, at 368 K[2]	12	0.41	0.000032	0.000053	0.000047	0.000046	0.000052	0.000053	0.000047	0.000042	0.000045	0.000041	0.000060	0.000027	0.000028	0.000027
	13	1.028	0.000089	0.000114	0.000105	0.000096	0.000110	0.000114	0.000105	0.000093	0.000114	0.000095	0.000089	0.000073	0.000075	0.000073
	14	1.433	0.000120	0.000114	0.000134	0.000122	0.000135	0.000115	0.000131	0.000121	0.000137	0.000123	0.000121	0.000100	0.000102	0.000102
	15	1.94	0.000159	0.000150	0.000169	0.000167	0.000177	0.000150	0.000170	0.000168	0.000173	0.000161	0.000158	0.000129	0.000132	0.000139
	16	2.495	0.000199	0.000202	0.000202	0.000205	0.000204	0.000202	0.000209	0.000203	0.000204	0.000205	0.000181	0.000156	0.000160	0.000170
	17	2.997	0.000224	0.000259	0.000226	0.000228	0.000232	0.000258	0.000230	0.000225	0.000235	0.000230	0.000230	0.000176	0.000181	0.000193
	18	3.503	0.000248	0.000271	0.000249	0.000249	0.000257	0.000233	0.000246	0.000288	0.000263	0.000311	0.000254	0.000214	0.000210	0.000212
	19	3.915	0.000260	0.000271	0.000253	0.000257	0.000259	0.000234	0.000256	0.000257	0.000274	0.000261	0.000266	0.000230	0.000223	0.000221
n-Butane + water, at 410 K[22]	20	0.2792	0.000022	0.000013	0.000052	0.000027	0.000020	0.000013	0.000041	0.000025	0.000019	0.000033	0.000018	0.000033	0.000032	0.000027
	21	1.003	0.000076	0.000114	0.000087	0.000074	0.000093	0.000114	0.000082	0.000073	0.000081	0.000072	0.000063	0.000059	0.000058	0.000056
	22	1.486	0.000110	0.000114	0.000117	0.000107	0.000110	0.000114	0.000112	0.000106	0.000116	0.000105	0.000086	0.000096	0.000093	0.000088
	23	1.727	0.000123	0.000150	0.000124	0.000122	0.000121	0.000151	0.000124	0.000124	0.000127	0.000120	0.000116	0.000111	0.000109	0.000103
	24	2.43	0.000157	0.000163	0.000156	0.000157	0.000163	0.000155	0.000158	0.000158	0.000159	0.000151	0.000161	0.000150	0.000142	0.000133
	25	3.044	0.000177	0.000163	0.000171	0.000173	0.000177	0.000177	0.000175	0.000176	0.000177	0.000166	0.000177	0.000173	0.000164	0.000158
AAPRE, %				21.20	16.04	5.45	11.46	20.91	13.19	5.13	9.79	10.06	10.02	20.05	17.07	15.02

Table 4. Estimates of EOSs, mathematical correlations, and machine-learning models for the solubilities of light hydrocarbon gases in pure water.

which is the most important part of the assessment of models. AAPRE values of 10.64% for the total collection, 11.49% for the test collection, and 10.43% for the train collection, as well as a total R^2 value of 0.9999, indicating that the AdaBoost-SVR model developed with 8 inputs had the most precise predictions of hydrocarbon gases solubilities in aqueous electrolyte solutions. After that, in terms of accuracy, the AdaBoost-SVR model developed with 5 inputs with an AAPRE of 12.02% for the total collection and a total R^2 value of 0.9999 ranks second among all models. AdaBoost-SVR models have the least overall values of RMSE, SD, and AAPRE along with the highest overall R^2 value among the other machine learning models leading us to conclude that this model is the most accurate model for predicting light hydrocarbon gases and their mixtures in water and aqueous electrolyte solutions. Moreover, despite the expected poorer performance than machine learning models, the mathematical correlations yielded by GP and GMDH methods show satisfying results with AAPRE values of 16.44% and 20.95%, respectively.

In the next step, the performance of the machine learning algorithms was compared with SRK, PR, and VPT EOSs. To this end, the solubilities data of light hydrocarbon gases in pure water at different operating conditions, acquired from the literature[2,9,22,61], was predicted by the developed machine-learning models, mathematical correlations, and three EOSs. Table 4 reports the predictions of these predictive tools and EOSs as well as calculated AAPRE. Aa represented in Table 4, AdaBoost-SVR models are superior to all machine learning-based predictive tools and EOSs showing AAPRE values of 5.13% (AdaBoost-SVR model with 5 inputs) and 5.45% (AdaBoost-SVR model with 8 inputs), which is the least among these predictive tools. Among the EOSs, VPT, SRK, and PR are ranked in terms of good predictions, respectively. Moreover, the mathematical correlations generated by the GMDH and GP techniques demonstrate satisfactory results with an AAPRE of approximately 10%.

To gain a better vision of the validity of the machine learning models in the training and testing stages, graphical error analyses were conducted along with statistical analyses. First, cross plots of all models are compared in Fig. 4. As pointed out earlier, the nearer the data to the X = Y line, the greater precision of the model in prognosticating hydrocarbon gases and their mixtures in water and aqueous electrolyte systems. As can be observed in Fig. 4, the AdaBoost-SVR models (developed with 8 and 5 inputs) have the high closest data around the X = Y line compared to the other suggested models and correlations, which exhibits the great robustness and validness of these models for the prediction of hydrocarbon gases solubility in aqueous electrolyte systems. However, other models have also performed well. Next, the error distribution graphs of all developed predictive tools based on temperature and pressure are illustrated in Fig. S1 in the supplementary file. These plots help to distinguish the performance of the models at different pressures and temperatures. Fig. S1(a) shows the low scatter of errors around the zero-error line for all models at different pressures, especially AdaBoost-SVR and DT models. Fig. S1(b) demonstrates that the AdaBoost-SVR models have the least scattering of errors around the zero-error line compared to other models and correlations at different temperatures. In relation to Random Forest, Extra Tree, and GMDH models, it seems that although the predictions of these models show a low error

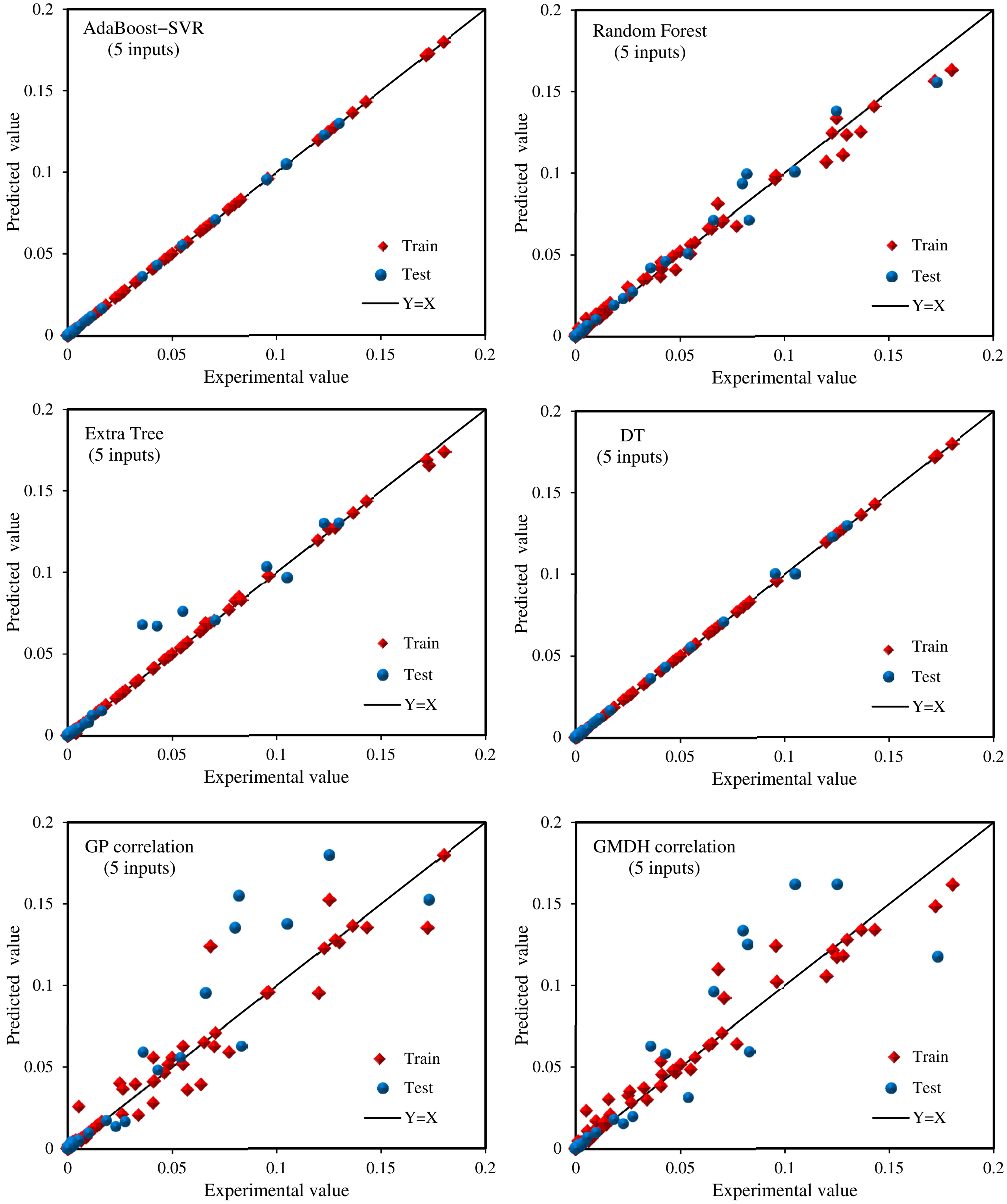


Figure 4. Cross-plots of the developed machine learning models and mathematical correlations.

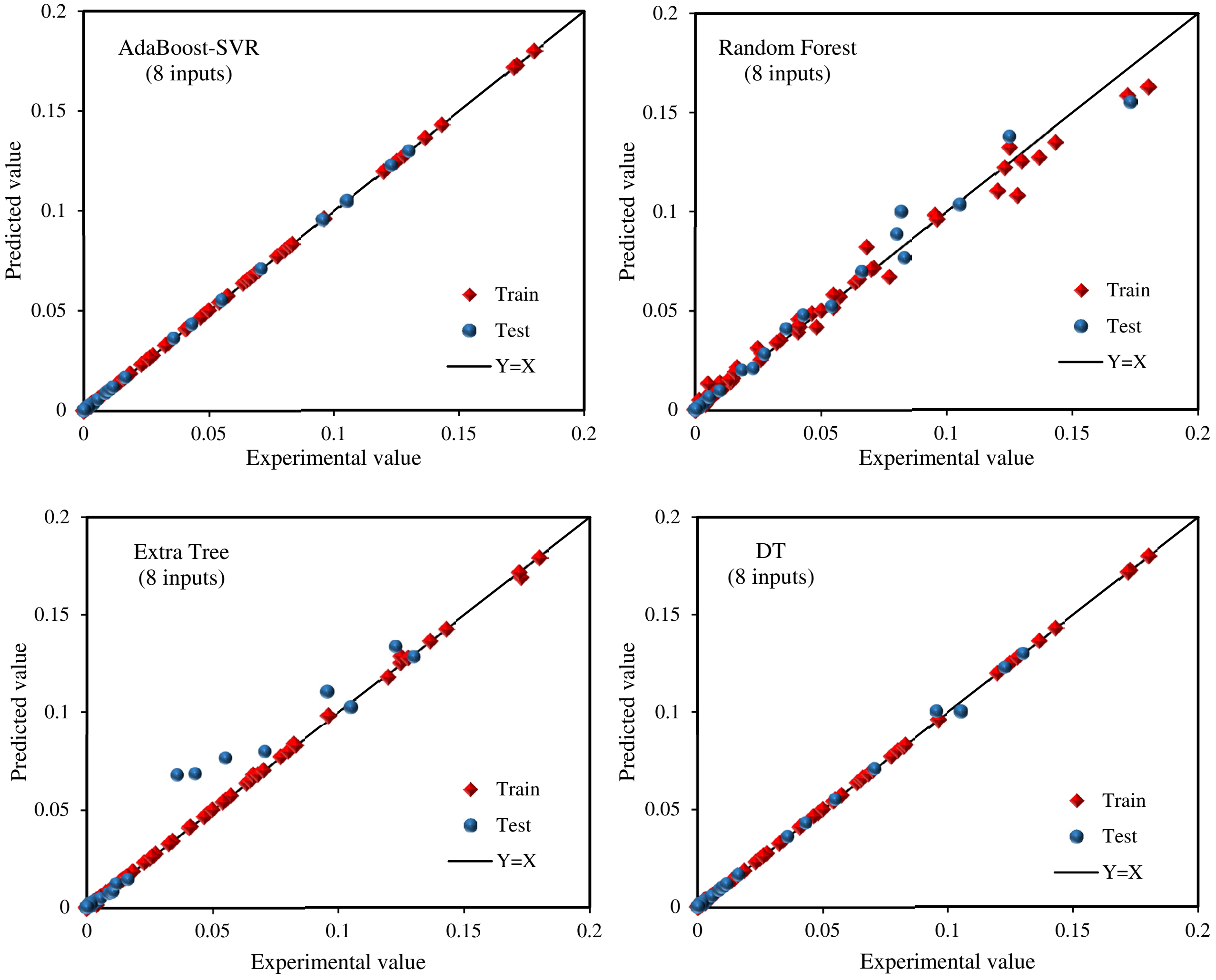


Figure 4. (continued)

at low temperatures, at high temperatures, the scattering of error is high. Overall, the AdaBoost-SVR models are superior to other machine learning models in different temperature and pressure ranges.

In the next step, the histograms of errors between experimental solubilities and prognosticated values associated with all models are illustrated in Fig. 5. The computed error values for all models are located in a narrow scope from −0.001 to 0.001. This figure shows that the histograms of all machine learning models benefit from normal distributions. However, despite the excellent performance in the training phase, the histogram of the Extra Tree model seems to be a bit skewed in the testing phase. As can be observed in Fig. 5, all histogram plots benefit from the bursts of growing at zero-error value, which indicates the excellent match between the estimated solubility data and experimental values. However, again AdaBoost-SVR and DT models display less error for more data during both testing and training stages in both modeling approaches.

The next step of graphical error analysis is a helpful statistical plot for quantifying the precision of the models and correlations, named cumulative frequency plot. As shown in Fig. 6, the cumulative frequency curves of the AdaBoost-SVR models are very close to the vertical axis, which indicates the high accuracy of these models. Besides, more than 70% of predicted gas solubility data by the AdaBoost-SVR models have an absolute error of less than 0.00004, and more than 90% of the predicted data have an error of less than 0.00013. Meanwhile, other models and correlations including Extra Tree, DT, Random Forest, GP, and GMDH represent absolute errors of 0.00015–0.0003 for 90% of the data, respectively. Therefore, this conclusion can be drawn that the AdaBoost-SVR models are superior to other models and correlations in estimating the solubility of hydrocarbon gases and their mixtures in water and aqueous electrolytes.

According to the results of statistical and graphical analyses of machine learning models, it can be concluded that the AdaBoost-SVR models (developed with 8 and 5 inputs) are more precise in estimating the solubility of hydrocarbons in water and brine solutions than other models suggested in this work. To assess the accuracy of the proposed AdaBoost-SVR models against the available predictive models in the literature for estimating the solubility of hydrocarbon gases, the AdaBoost-SVR results were compared with two machine learning models, including Samani et al.[52] and Nabipour et al.[53], which are shown in Table 5. As depicted in Table 5, the AdaBoost-SVR models proposed in this study have the lowest AAPRE values plus the highest R^2 value, indicating that the

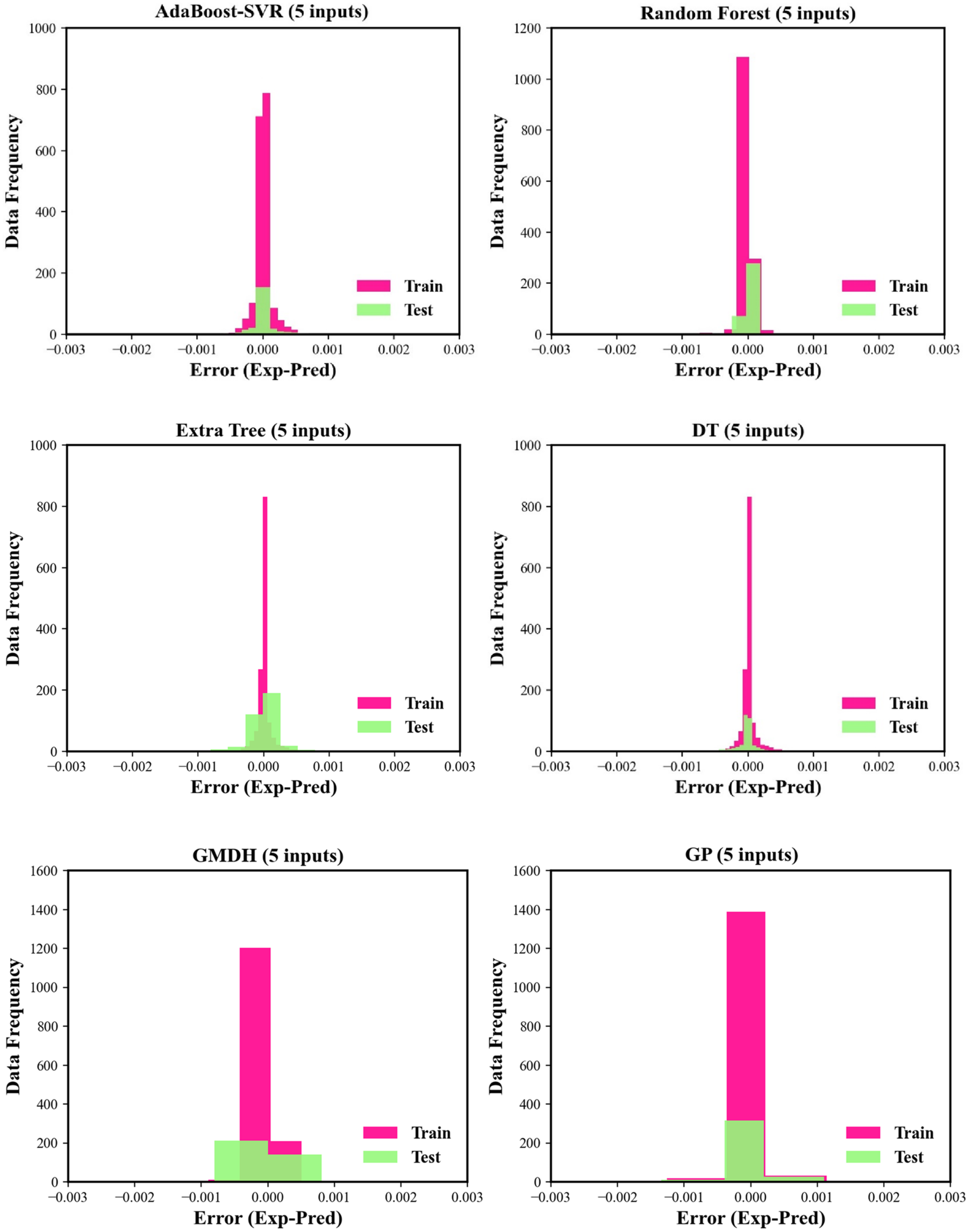


Figure 5. Histograms of residuals for the machine learning models and correlations.

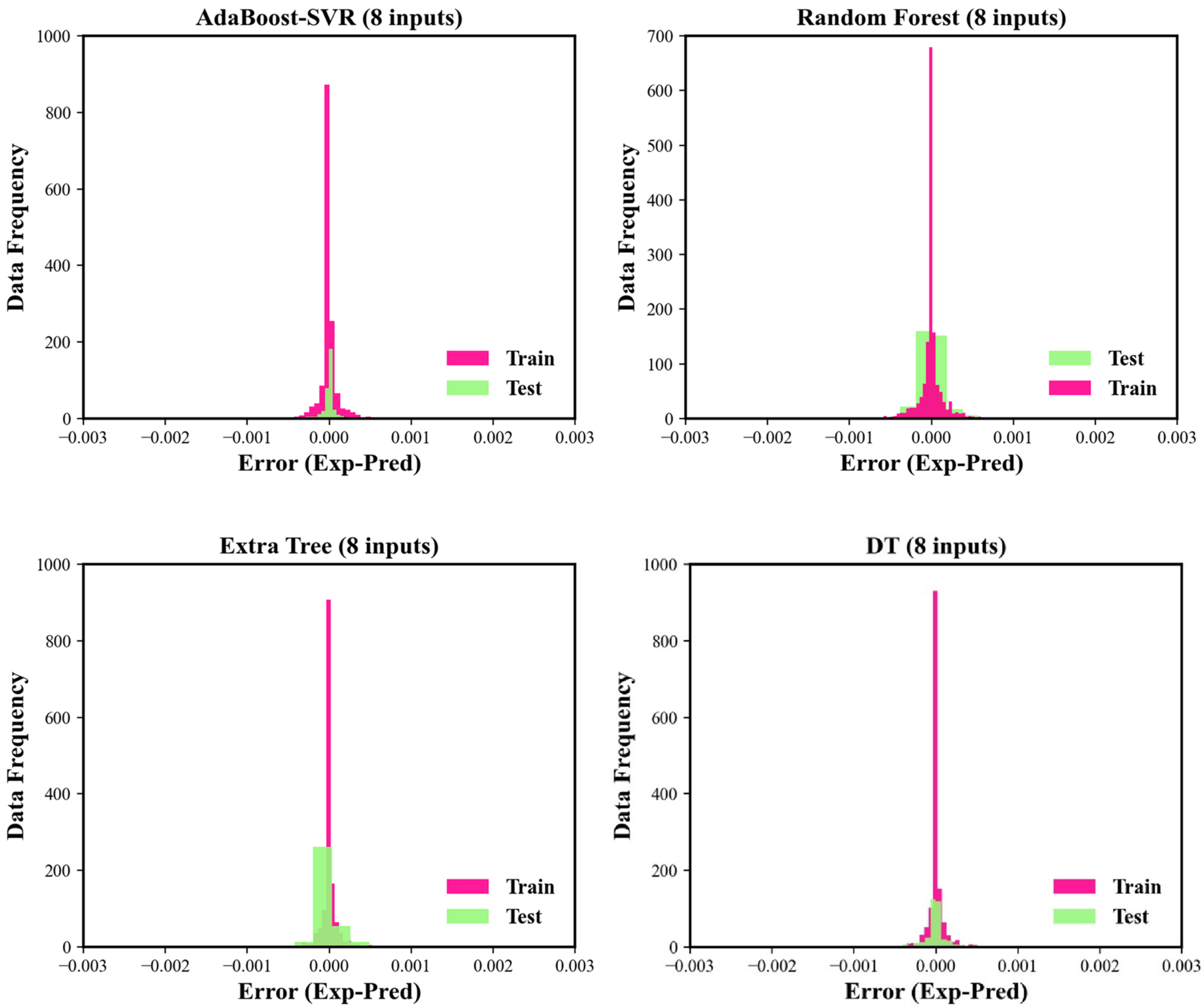


Figure 5. (continued)

AdaBoost-SVR models are more precise than other artificial intelligence models presented in the literature for estimating the solubility of hydrocarbon gases.

Trend analysis. As mentioned earlier, the AdaBoost-SVR models are more accurate in predicting the solubility of light hydrocarbon gases in aqueous solutions than other models. Hence, the solubilities of hydrocarbon gases in several solubility systems have been investigated to evaluate the ability of the AdaBoost-SVR models in estimating the true physical trend of gases solubility in the liquid phase. In the beginning, the solubilities of methane, ethane, and *n*-butane in a gas mixture + pure water system at a temperature of 283 K[9] were estimated utilizing the AdaBoost-SVR models and three EOSs, and the outcomes are depicted in Fig. 7. As demonstrated in Fig. 7, EOSs overestimated or underestimated the solubilities of hydrocarbon gases in water at low-temperature conditions. However, VPT EOS again is superior to SRK and PR EOSs and provides better estimations. Nevertheless, both AdaBoost-SVR models (developed with 8 and 5 inputs) offer an exceptional ability to track solubility data of hydrocarbon gases with increasing pressure at low-temperature conditions compared to EOSs. Although the accuracy of EOSs has been lower than machine learning models, this does not mean questioning the capabilities of these thermodynamic equations. EOSs predict solubility data based on the thermodynamic variables within an analytical framework and they are valuable tools in the modeling of a wide range of industrial processes. Here, only a comparison between predictions of developed models and EOSs was made to clarify the high predictability of these models. Hence, machine learning models can be considered as an alternative to achieve accurate and fast predictions of the solubility of gases in brine in order to cover the disadvantages of EOSs mentioned earlier.

Next, the solubilities of methane and propane mixtures in pure water, which has been experimentally investigated by Amirijafari[23] at a temperature of 377.59 K under high-pressure conditions, was predicted by the AdaBoost-SVR models, as demonstrated in Fig. 8. As depicted in the figure, both AdaBoost-SVR models correctly predicted the solubilities of methane and propane in pure water by increasing the pressure as an important parameter affecting solubility.

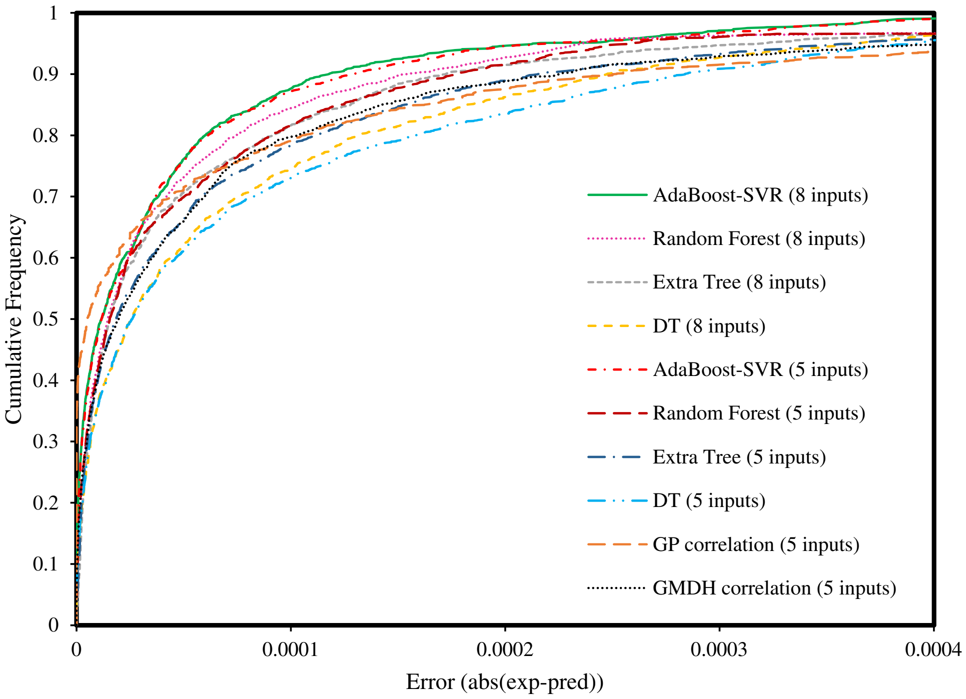


Figure 6. Cumulative frequency plot of the proposed predictive tools for estimating the solubility of hydrocarbon gases.

Models		RMSE	R^2	AARPE, %
Samani et al.[52]	Train	0.00013	0.9893	28.78
	Test	0.00017	0.9834	37.84
	Total	0.00014	0.9880	30.60
Nabipour et al.[53]	Train	0.0001	0.9850	22.049
	Test	0.0001	0.9870	22.054
	Total	0.0001	0.9850	22.050
AdaBoost-SVR (8 inputs)	Train	0.000099	0.9999	10.433
	Test	0.000101	0.9999	11.497
	Total	0.000099	0.9999	10.647
AdaBoost-SVR (5 inputs)	Train	0.000102	0.9999	11.613
	Test	0.000109	0.9999	13.643
	Total	0.000104	0.9999	12.020

Table 5. Statistical factors for the available hydrocarbon gases solubility predictive models and the proposed AdaBoost-SVR models.

In the next step, the solubility of methane in water versus pressure at different temperatures was predicted by the AdaBoost-SVR models, which has been examined in the literature[9]. The solubilities of methane, as the basic constituent of natural gas, in pure water and aqueous electrolyte systems at different pressure and temperature is crucial for the petroleum industry. As shown in Fig. 9, the solubility of methane in water at various pressure and temperature conditions is accurately predicted by the AdaBoost-SVR models. As can be seen, the temperature has a decreasing impact on the methane' solubility in water at the studied pressures, which is correctly estimated by the AdaBoost-SVR models.

Eventually, the solubilities of methane in pure water and in aqueous NaCl solutions with different salt concentrations at a temperature of 324.65 K, which has been studied experimentally in the literature[67], was predicted by the AdaBoost-SVR models. As can be observed in Fig. 10, the solubility of methane has an appreciable decrease with an increase in salt concentration or ionic strength of the solution. Again, both AdaBoost-SVR models provide accurate predictions for the systems of methane + water and methane + aqueous salt solution with different concentrations at different pressures with very little deviation from the experimental data.

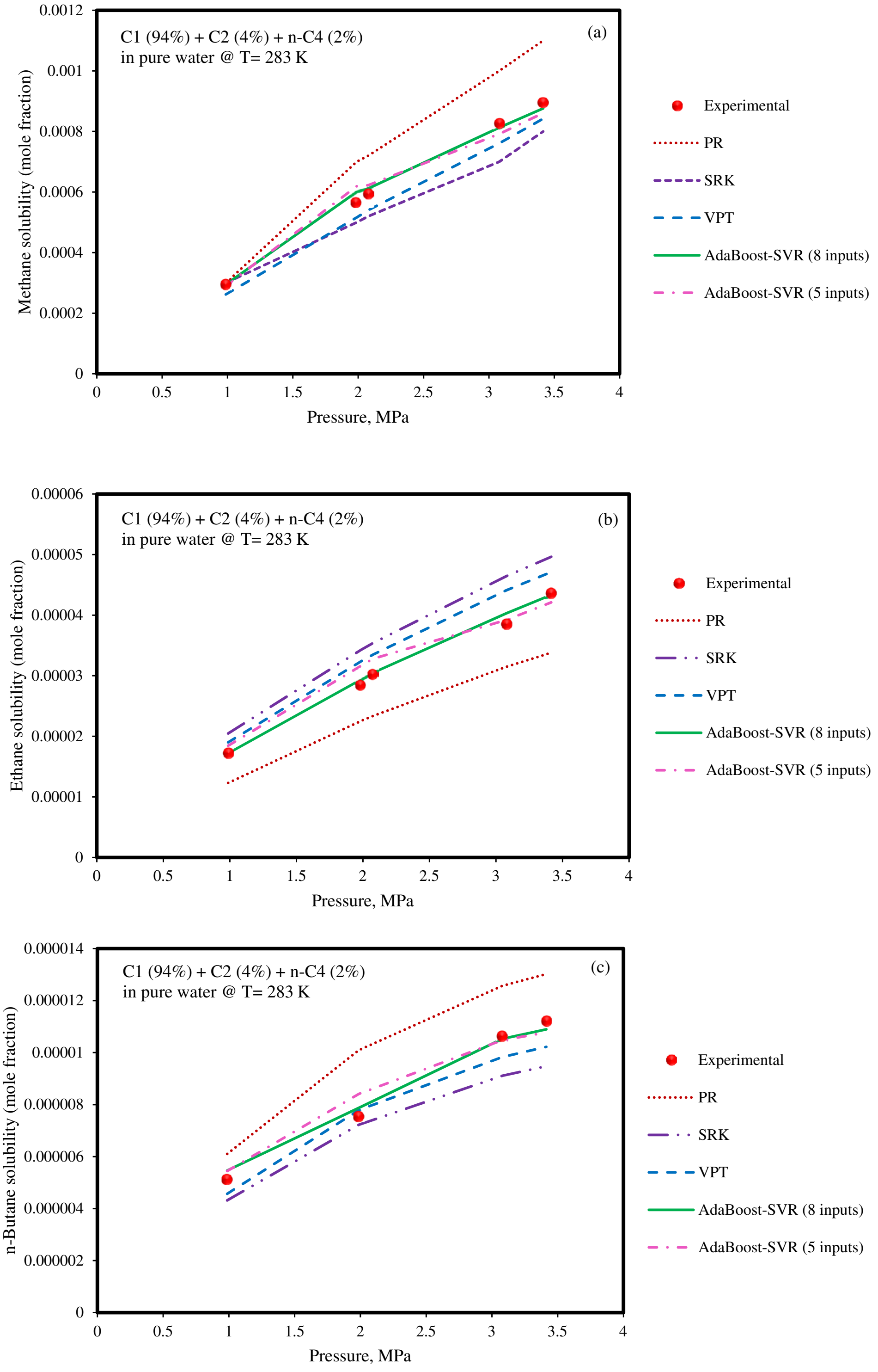


Figure 7. Experimental values and estimations of the solubilities of (**a**) methane, (**b**) ethane, and (**c**) *n*-butane in the aqueous phase of the gas mixture + water system by EOSs and AdaBoost-SVR models.

Sensitivity analysis.

In parametric studies, identifying the impacts of all inputs on the output can be valuable. As stated earlier, two modeling approaches with 8 and 5 inputs were adopted in this work. The first approach was that there were 8 inputs including the temperature, pressure, ionic strength of the solution, the mole percent of each component (C1, C2, C3, and C4) in the gas mixture, and carbon number (IDX) of the gas component whose solubility is to be predicted. On the other hand, the second approach considered 5 inputs containing the temperature, pressure, ionic strength of the solution, the pseudo-critical temperature of the gas mixture, and the critical temperature of the gas component whose solubility is to be predicted. To check the relative effects

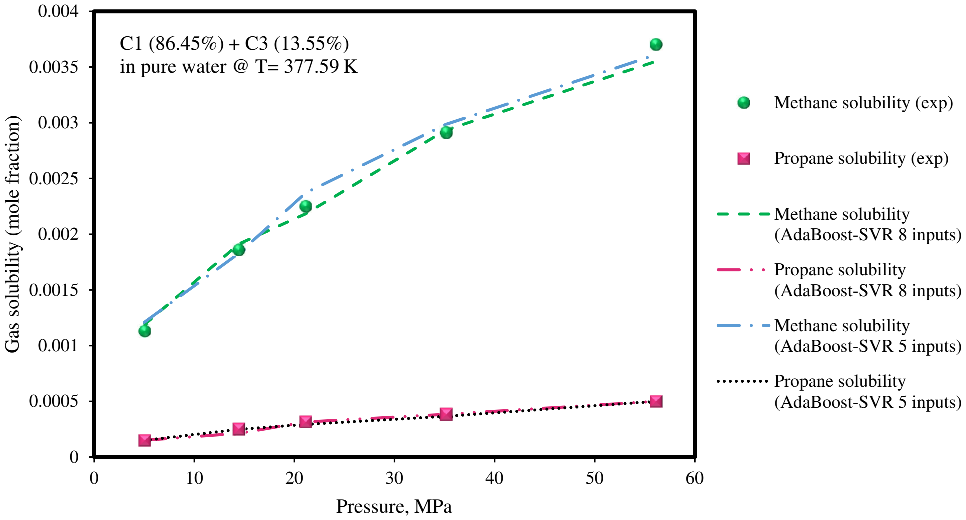


Figure 8. Experimental solubility data of methane and propane mixture in water at different operating pressures along with AdaBoost-SVR models predictions.

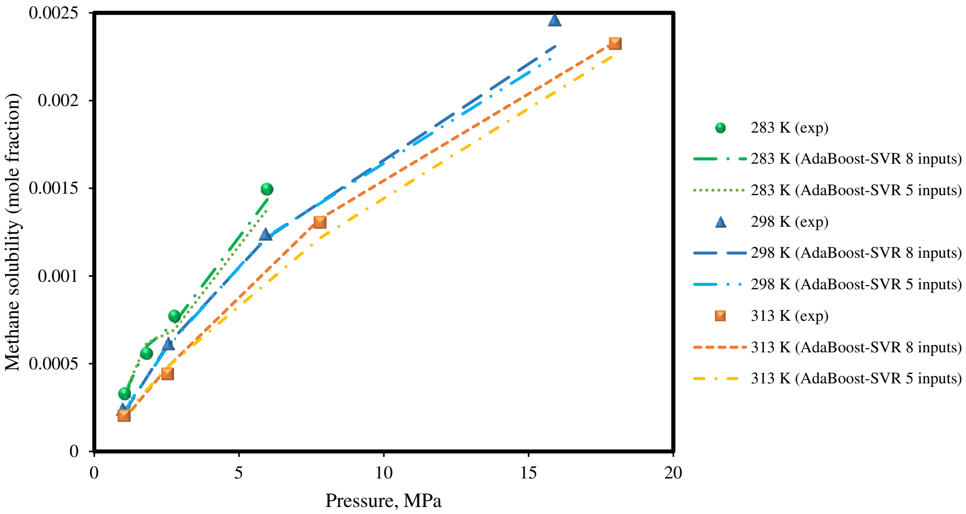


Figure 9. Experimental methane solubility data and AdaBoost-SVR models predictions for the methane + pure water system at different temperatures.

of these input variables on the solubilities of hydrocarbon gases in water and aqueous electrolyte systems, the relevancy factor (r)[98] was employed in this research. It should be mentioned that the outcomes of all developed models and correlations developed in this work along with experimental data have been utilized for sensitivity analysis to make a comparison between the results of all models in both modeling approaches. Positive or negative values of r for an input parameter indicate a direct or inverse relationship between that parameter and the output, respectively. The higher value of r between an input variable and output, the greater the impact of that input on the solubilities of hydrocarbon gases in water and aqueous electrolyte systems. The subsequent equation is utilized for calculating the r-values for the input parameters[99]:

$$r(inp_i, \eta) = \frac{\sum_{j=1}^{n} \left(inp_{i,j} - inp_{m,i}\right)\left(\eta_j - \eta_m\right)}{\left(\sum_{j=1}^{n} \left(inp_{i,j} - inp_{m,i}\right)^2 \sum_{j=1}^{n} \left(\eta_j - \eta_m\right)^2\right)^{0.5}} \tag{18}$$

where i could be any of the input parameters considered for modeling; $inp_{m,i}$ and $inp_{i,j}$ respectively indicate the mean and jth value of the ith input parameter. η_m stands for the mean of predicted solubility of hydrocarbon gases in water and aqueous electrolyte systems and η_j is the jth value of predicted solubilities of hydrocarbon gases. Figure 11 illustrates the relative impacts of considered input variables on the solubilities of hydrocarbon

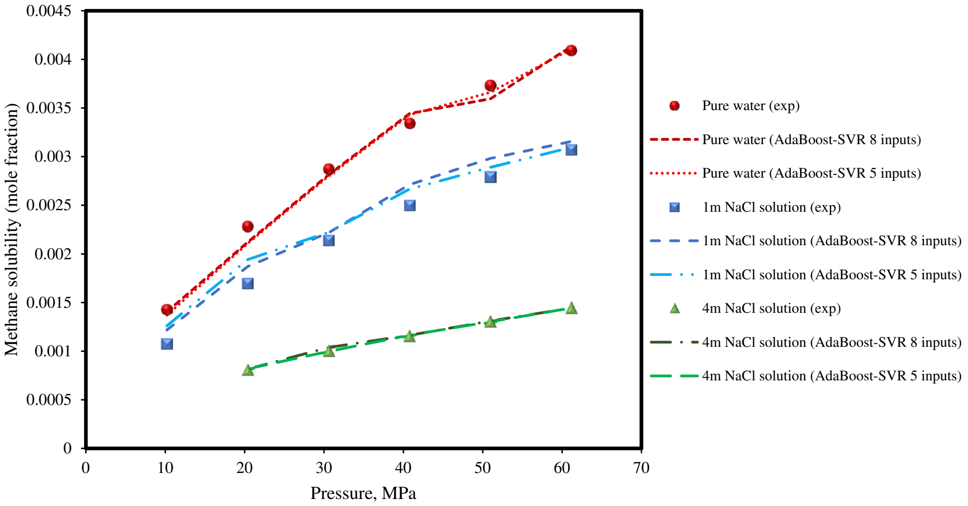


Figure 10. Experimental methane solubilities in water and aqueous NaCl solutions at a temperature of 324.65 K along with AdaBoost-SVR models predictions.

gases in water and brine solutions. As seen in Fig. 11a, in the first modeling approach, the temperature, pressure, and methane (mole %) in the gas mixture had the greatest effects on hydrocarbon gases solubility. Also, the mole percent of the *n*-butane in the gas mixture was the least effective parameter for estimating the solubilities of hydrocarbon gases. Based on results, the temperature, pressure, and mole percent of methane and *n*-butane in the gas mixture have direct effects, and mole percent of ethane and propane in the gas mixture, IDX, and ionic strength of the solutions have reverse effects on the solubility of investigated hydrocarbon gas. An increase in the ionic strength of the solution decreases the solubilities of hydrocarbon gases in aqueous electrolyte systems. In the second modeling approach, as shown in Fig. 11b, the results of sensitivity analysis for temperature, pressure, and ionic strength variables have been obtained quite similarly to the previous case. Moreover, the pseudo-critical temperature of the gas mixture and the critical temperature of the gas components have negative effects on the solubility of light hydrocarbon gases and their mixture in brine, which exhibits that the solubility decreases with the rise of these parameters. As inferred from the results of the sensitivity analysis of both modeling approaches, the feature-solubility correlations are completely independent of machine learning frameworks and the impact of each specific input variable applied for modeling in each model or correlation developed in this work are the same and similar to the laboratory results.

Implementation of Leverage method. Finally, the degree of precision of utilized data along with the application scope of the AdaBoost-SVR models was examined using the Leverage approach[100–102], which can assess the validity of these model and solubility databank. The subsequent equation was utilized to calculate the variations of the prognosticated solubility values by the model from the real data, which is named standardized residuals (R)[103]:

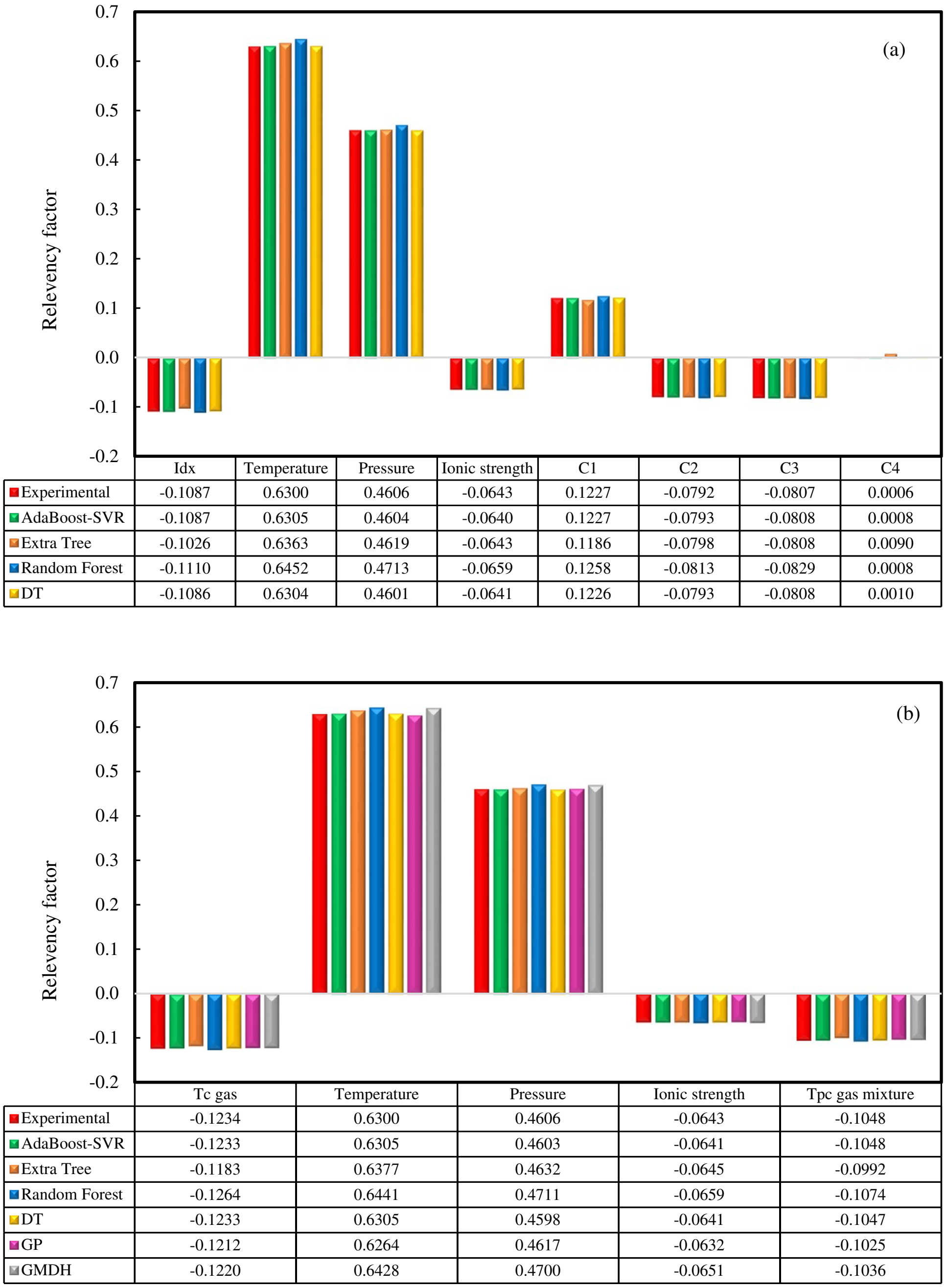
$$R_z = \frac{e_z}{(MSE(1 - H_{zz}))^{0.5}} \tag{19}$$

in which, the mean square error of the predictive tool is shown by MSE; H_{zz} shows Leverage of the zth data; and e_z denotes the variation of the estimations from the experiments of the zth data. Afterward, the following formula is utilized to calculate the values of Hat matrix Leverage[104]:

$$\mathrm{H} = \mathrm{K}\,(\mathrm{K}^{\mathrm{T}}\mathrm{K})^{-1}\mathrm{K}^{\mathrm{T}} \tag{20}$$

where K^T shows the transpose of the matrix K, which is $(g \times c)$ matrix; g and c indicate the number of databank points and the number of input variables, respectively. Besides, the critical Leverage limit (H^*) is achieved using $3(c + 1)/g$.

The reliable zone is considered to be the cut-off area of R-values (−3 and 3) and $H_{zz} \le H^*$, as shown in William's plot in Fig. 12. This figure exhibits that the bulk of data, called valid data, rested in the reliable zone that proves the high reliability of the hydrocarbon solubility databank and high validation of the AdaBoost-SVR models. For the AdaBoost-SVR model developed with 8 inputs, as depicted in Fig. 12a, quantitative identification of the outliers of the used databank shows that only 54 data points (2.94% of the whole data) have an R-value outside the range of −3 to 3, which is considered suspected data. In addition, only 35 data points (1.91% of the whole data) have $H_{zz} > 0.0147$, which is regarded as out of Leverage data, while other data have acceptable Leverage $(H_{zz} \le 0.0147)$. For the AdaBoost-SVR model developed with five inputs, due to the reduction of the number of input variables, the critical Leverage limit value is reduced to $H^* = 0.0098$, and the application scope of the model becomes more limited. However, there is no specific change in the number of suspected data points (54



	Idx	Temperature	Pressure	Ionic strength	C1	C2	C3	C4
🟥 Experimental	-0.1087	0.6300	0.4606	-0.0643	0.1227	-0.0792	-0.0807	0.0006
🟩 AdaBoost-SVR	-0.1087	0.6305	0.4604	-0.0640	0.1227	-0.0793	-0.0808	0.0008
🟧 Extra Tree	-0.1026	0.6363	0.4619	-0.0643	0.1186	-0.0798	-0.0808	0.0090
🟦 Random Forest	-0.1110	0.6452	0.4713	-0.0659	0.1258	-0.0813	-0.0829	0.0008
🟨 DT	-0.1086	0.6304	0.4601	-0.0641	0.1226	-0.0793	-0.0808	0.0010

	Tc gas	Temperature	Pressure	Ionic strength	Tpc gas mixture
🟥 Experimental	-0.1234	0.6300	0.4606	-0.0643	-0.1048
🟩 AdaBoost-SVR	-0.1233	0.6305	0.4603	-0.0641	-0.1048
🟧 Extra Tree	-0.1183	0.6377	0.4632	-0.0645	-0.0992
🟦 Random Forest	-0.1264	0.6441	0.4711	-0.0659	-0.1074
🟨 DT	-0.1233	0.6305	0.4598	-0.0641	-0.1047
🟪 GP	-0.1212	0.6264	0.4617	-0.0632	-0.1025
⬜ GMDH	-0.1220	0.6428	0.4700	-0.0651	-0.1036

Figure 11. The impact of input variables on hydrocarbon gases solubility in water and aqueous electrolyte systems in the (**a**) first and (**b**) second modeling approaches.

data points means 2.94% of the whole data), and only the out of Leverage data has increased to 70 (3.81% of the whole data). As shown in Fig. 12b, these points are also predicted by the model with a very low error, and they are just statistically beyond the critical Leverage limit. Hence, it cannot be considered a negative point for the model. The results of the Leverage mathematical method reveal the validity of the hydrocarbon solubility databank and the high credit of both AdaBoost-SVR models in estimating the solubility of hydrocarbon gases in water and brine solution systems.

Conclusions

In the present study, the solubilities of the principal hydrocarbon components of natural gas in water and aqueous electrolyte solutions were modeled utilizing six machine learning algorithms. A large databank (1836 experimental data points) of hydrocarbon gases solubility was gathered from numerous sources of literature to cover a wide range of temperature and pressure conditions. Two different approaches including eight and five inputs were adopted for modeling. Also, three famous EOSs, including PR, VPT, and SRK were used in comparison

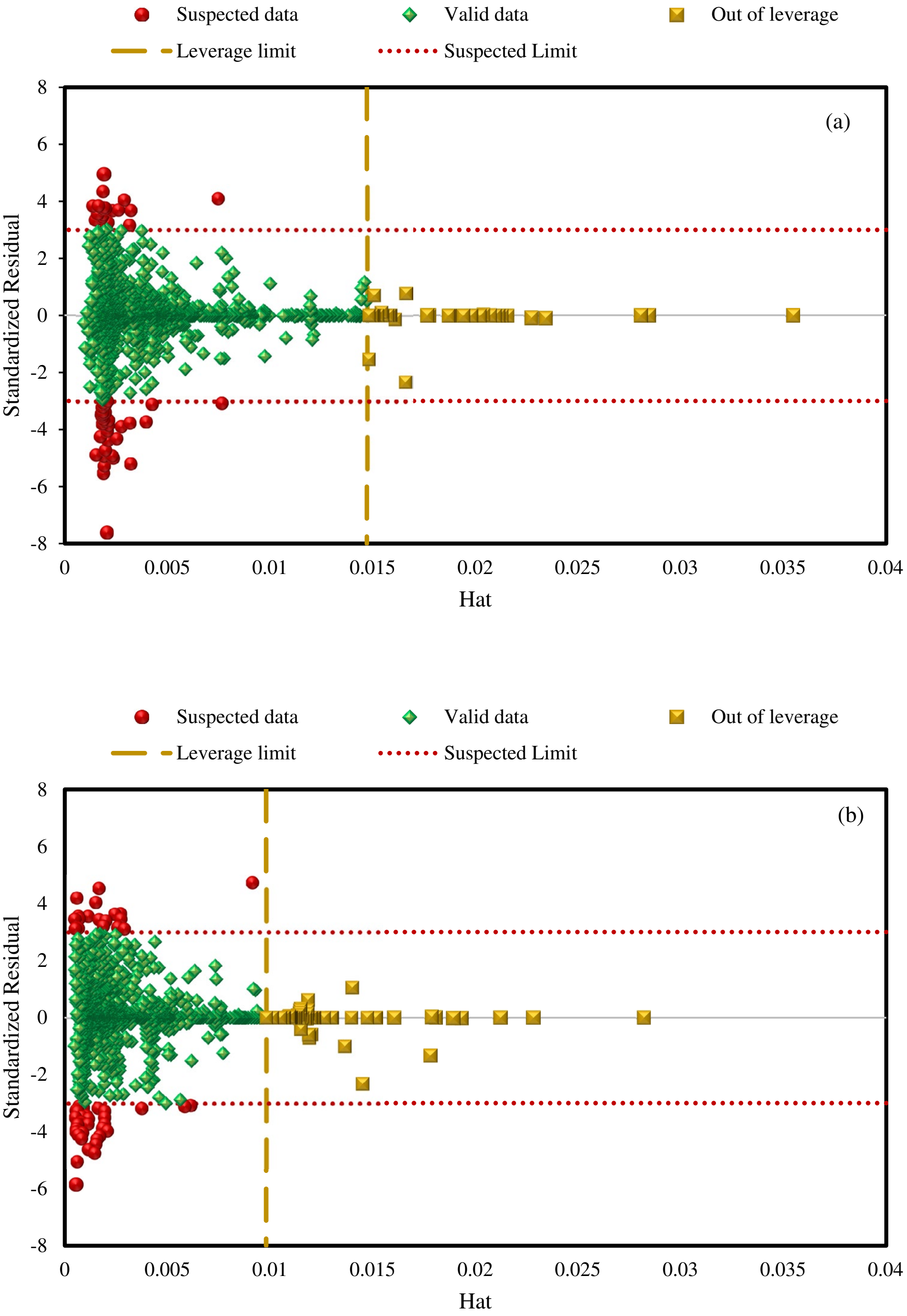


Figure 12. Detection of applicability area, suspected data, and outliers of AdaBoost-SVR models developed with (**a**) 8 inputs and (**b**) 5 inputs.

with machine learning models. Based on graphical and statistical analyses, the best-developed models in this work, namely AdaBoost-SVR developed with eight and five inputs, are able to predict the solubility of hydrocarbon gases and their mixture with an overall AAPRE of 10.65% and 12.02%, respectively, and R^2 of 0.9999. The AdaBoost-SVR models outperform other models developed in this work, EOSs, and intelligence models proposed in the literature. Also, the Random Forest, DT, and Extra Tree models are positioned subsequent to the AdaBoost-SVR model in terms of high precision in predicting test collection in both modeling approaches. Despite higher errors than machine learning models, two mathematical correlations generated by the GMDH and GP techniques had satisfactory outcomes. Among the EOSs, VPT, SRK, and PR are ranked in terms of good predictions, respectively. Based on sensitivity analysis, the temperature and pressure had the greatest effect on hydrocarbon gases solubility in both modeling approaches. Regarding the gas mixture composition (C1–C4), the percentage of methane and *n*-butane in the gas mixture was the most and least effective parameter for predicting the solubility of hydrocarbon gases in brine, respectively. Additionally, an increase in the ionic strength of the solution and the pseudo-critical temperature of the gas mixture decreases the solubilities of hydrocarbon gases in aqueous electrolyte systems. Moreover, the influence of input variables on light hydrocarbon gases solubility is completely independent of machine learning frameworks. Eventually, the investigation of the Leverage technique proved the high validity of the hydrocarbon solubility databank and the high credit of the AdaBoost-SVR models in predicting hydrocarbon gases solubility in water and aqueous electrolyte systems.

Data availability

All the data have been collected from literature. We cited all the references of the data in the manuscript. However, the data will be available from the corresponding author on reasonable request.

Received: 31 December 2021; Accepted: 23 August 2022
Published online: 02 September 2022

References

1. Mohammadi, A. H., Chapoy, A., Tohidi, B. & Richon, D. Gas solubility: A key to estimating the water content of natural gases. *Ind. Eng. Chem. Res.* **45**, 4825–4829 (2006).
2. Chapoy, A. *et al.* Solubility measurement and modeling for the system propane–water from 277.62 to 368.16 K. *Fluid Phase Equilib.* **226**, 213–220 (2004).
3. Chapoy, A., Haghighi, H. & Tohidi, B. Development of a Henry's constant correlation and solubility measurements of *n*-pentane, *i*-pentane, cyclopentane, *n*-hexane, and toluene in water. *J. Chem. Thermodyn.* **40**, 1030–1037 (2008).
4. Kiepe, J., Horstmann, S., Fischer, K. & Gmehling, J. Experimental determination and prediction of gas solubility data for methane+ water solutions containing different monovalent electrolytes. *Ind. Eng. Chem. Res.* **42**, 5392–5398 (2003).
5. Dhima, A., de Hemptinne, J.-C. & Moracchini, G. Solubility of light hydrocarbons and their mixtures in pure water under high pressure. *Fluid Phase Equilib.* **145**, 129–150 (1998).
6. Marinakis, D. & Varotsis, N. Solubility measurements of (methane+ ethane+ propane) mixtures in the aqueous phase with gas hydrates under vapour unsaturated conditions. *J. Chem. Thermodyn.* **65**, 100–105 (2013).
7. Kondori, J., Zendehboudi, S. & Hossain, M. E. A review on simulation of methane production from gas hydrate reservoirs: Molecular dynamics prospective. *J. Petrol. Sci. Eng.* **159**, 754–772 (2017).
8. Kondori, J., Zendehboudi, S. & James, L. Evaluation of gas hydrate formation temperature for gas/water/salt/alcohol systems: Utilization of extended UNIQUAC model and PC-SAFT equation of state. *Ind. Eng. Chem. Res.* **57**, 13833–13855 (2018).
9. Chapoy, A., Mohammadi, A. H., Richon, D. & Tohidi, B. Gas solubility measurement and modeling for methane–water and methane–ethane–*n*-butane–water systems at low temperature conditions. *Fluid Phase Equilib.* **220**, 113–121 (2004).
10. Abha, S. & Singh, C. S. Hydrocarbon pollution: Effects on living organisms, remediation of contaminated environments, and effects of heavy metals co-contamination on bioremediation. in *Introduction to Enhanced Oil Recovery (EOR) Processes and Bioremediation of Oil-Contaminated Sites*. 185–206 (2012).
11. Latimer, J. S., Hoffman, E. J., Hoffman, G., Fasching, J. L. & Quinn, J. G. Sources of petroleum hydrocarbons in urban runoff. *Water Air Soil Pollut.* **52**, 1–21 (1990).
12. Husaini, A., Roslan, H., Hii, K. & Ang, C. Biodegradation of aliphatic hydrocarbon by indigenous fungi isolated from used motor oil contaminated sites. *World J. Microbiol. Biotechnol.* **24**, 2789–2797 (2008).
13. Li, Z. & Firoozabadi, A. Cubic-plus-association equation of state for water-containing mixtures: Is "cross association" necessary?. *AIChE J.* **55**, 1803–1813 (2009).
14. Alvarez, E., Riverol, C., Correa, J. & Navaza, J. Design of a combined mixing rule for the prediction of vapor−liquid equilibria using neural networks. *Ind. Eng. Chem. Res.* **38**, 1706–1711 (1999).
15. Urata, S., Takada, A., Murata, J., Hiaki, T. & Sekiya, A. Prediction of vapor–liquid equilibrium for binary systems containing HFEs by using artificial neural network. *Fluid Phase Equilib.* **199**, 63–78 (2002).
16. Mohanty, S. Estimation of vapour liquid equilibria of binary systems, carbon dioxide–ethyl caproate, ethyl caprylate and ethyl caprate using artificial neural networks. *Fluid Phase Equilib.* **235**, 92–98 (2005).
17. Torrecilla, J. S., Palomar, J., García, J., Rojo, E. & Rodríguez, F. Modelling of carbon dioxide solubility in ionic liquids at sub and supercritical conditions by neural networks and mathematical regressions. *Chemom. Intell. Lab. Syst.* **93**, 149–159 (2008).
18. Safamirzaei, M. & Modarress, H. Hydrogen solubility in heavy *n*-alkanes; Modeling and prediction by artificial neural network. *Fluid Phase Equilib.* **310**, 150–155 (2011).
19. Moosanezhad-Kermani, H., Rezaei, F., Hemmati-Sarapardeh, A., Band, S. S. & Mosavi, A. Modeling of carbon dioxide solubility in ionic liquids based on group method of data handling. *Eng. Appl. Comput. Fluid Mech.* **15**, 23–42 (2021).
20. Crovetto, R., Fernández-Prini, R. & Japas, M. L. Solubilities of inert gases and methane in H2O and in D2O in the temperature range of 300 to 600 K. *J. Chem. Phys.* **76**, 1077–1086 (1982).
21. Culberson, O. & McKetta, J. Phase equilibria in hydrocarbon-water systems II—The solubility of ethane in water at pressures to 10,000 psi. *J. Petrol. Technol.* **2**, 319–322 (1950).
22. Le Breton, J. & McKetta, J. Jr. Low-pressure solubility of *n*-butane in water. *Hydrocarb. Proc. Petr. Ref.* **43**, 136–138 (1964).
23. Amirijafari, B. *Solubility of Light Hydrocarbons in Water Under High Pressures* (The University of Oklahoma, 1969).
24. Wang, L.-K., Chen, G.-J., Han, G.-H., Guo, X.-Q. & Guo, T.-M. Experimental study on the solubility of natural gas components in water with or without hydrate inhibitor. *Fluid Phase Equilib.* **207**, 143–154 (2003).
25. Vul'fson, A. & Borodin, O. A thermodynamic analysis of the solubility of gases in water at high pressures and supercritical temperatures. *Russ. J. Phys. Chem. A* **81**, 510–514 (2007).
26. Tong, D., Trusler, J. M. & Vega-Maza, D. Solubility of CO2 in aqueous solutions of CaCl2 or MgCl2 and in a synthetic formation brine at temperatures up to 423 K and pressures up to 40 MPa. *J. Chem. Eng. Data* **58**, 2116–2124 (2013).
27. Teng, H. & Yamasaki, A. Solubility of liquid CO2 in synthetic sea water at temperatures from 278 K to 293 K and pressures from 6.44 MPa to 29.49 MPa, and densities of the corresponding aqueous solutions. *J. Chem. Eng. Data* **43**, 2–5 (1998).
28. Chapoy, A., Mohammadi, A. H., Tohidi, B. & Richon, D. Gas solubility measurement and modeling for the nitrogen+ water system from 274.18 K to 363.02 K. *J. Chem. Eng. Data* **49**, 1110–1115 (2004).
29. Smith, N. O., Kelemen, S. & Nagy, B. Solubility of natural gases in aqueous salt solutions—II: Nitrogen in aqueous NaCl, CaCl2, Na2SO4 and MgSO4 at room temperatures and at pressures below 1000 psia. *Geochim. Cosmochim. Acta* **26**, 921–926 (1962).
30. Bando, S., Takemura, F., Nishio, M., Hihara, E. & Akai, M. Solubility of CO2 in aqueous solutions of NaCl at (30 to 60) C and (10 to 20) MPa. *J. Chem. Eng. Data* **48**, 576–579 (2003).
31. Dhima, A., de Hemptinne, J.-C. & Jose, J. Solubility of hydrocarbons and CO2 mixtures in water under high pressure. *Ind. Eng. Chem. Res.* **38**, 3144–3161 (1999).

32. Zheng, K. *et al.* A comparative study of the perturbed-chain statistical associating fluid theory equation of state and activity coefficient models in phase equilibria calculations for mixtures containing associating and polar components. *Ind. Eng. Chem. Res.* **57**, 3014–3030 (2018).
33. Ahmed, S. *et al.* A new PC-SAFT model for pure water, water–hydrocarbons, and water–oxygenates systems and subsequent modeling of VLE, VLLE, and LLE. *J. Chem. Eng. Data* **61**, 4178–4190 (2016).
34. Lee, M.-T. & Lin, S.-T. Prediction of mixture vapor–liquid equilibrium from the combined use of Peng–Robinson equation of state and COSMO-SAC activity coefficient model through the Wong-Sandler mixing rule. *Fluid Phase Equilib.* **254**, 28–34 (2007).
35. Yan, Y. & Chen, C.-C. Thermodynamic modeling of CO2 solubility in aqueous solutions of NaCl and Na2SO4. *J. Supercrit. Fluids* **55**, 623–634 (2010).
36. Tabasinejad, F. *et al.* Water solubility in supercritical methane, nitrogen, and carbon dioxide: measurement and modeling from 422 to 483 K and pressures from 3.6 to 134 MPa. *Ind. Eng. Chem. Res.* **50**, 4029–4041 (2011).
37. Shabani, B. & Vilcáez, J. Prediction of CO2–CH4–H2S–N2 gas mixtures solubility in brine using a non-iterative fugacity-activity model relevant to CO2-MEOR. *J. Petrol. Sci. Eng.* **150**, 162–179 (2017).
38. Liu, G. *et al.* Investigation of gas solubility and its effects on natural gas reserve and production in tight formations. *Fuel* **295**, 120507 (2021).
39. Avaji, S., Amani, M. J. & Ghaedi, M. Modeling the equilibrium of two and three-phase systems including water, alcohol, and hydrocarbons with CPA EOS and its improvement for electrolytic systems by Debye-Huckel equation. *J. Nat. Gas Sci. Eng.* **90**, 103905 (2021).
40. Sun, L. & Liang, J. Solubility calculations of methane and ethane in aqueous electrolyte solutions. *J. Solut. Chem.* **50**, 1–21 (2021).
41. He, H., Sun, B., Wang, Z., Liu, Y. & Sun, X. A constitutive model for predicting the solubility of gases in water at high temperature and pressure. *J. Petrol. Sci. Eng.* **192**, 107337 (2020).
42. Battino, R. & Clever, H. L. The solubility of gases in liquids. *Chem. Rev.* **66**, 395–463 (1966).
43. Oliveira, M., Coutinho, J. & Queimada, A. Mutual solubilities of hydrocarbons and water with the CPA EoS. *Fluid Phase Equilib.* **258**, 58–66 (2007).
44. 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).
45. 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).
46. Sayahi, T., Tatar, A., Rostami, A., Anbaz, M. A. & Shahbazi, K. Determining solubility of CO_2 in aqueous brine systems via hybrid smart strategies. *Int. J. Comput. Appl. Technol.* **65**, 1–13 (2021).
47. Jeon, P. R. & Lee, C.-H. Artificial neural network modelling for solubility of carbon dioxide in various aqueous solutions from pure water to brine. *J. CO2 Util.* **47**, 101500 (2021).
48. Hemmati-Sarapardeh, A., Amar, M. N., Soltanian, M. R., Dai, Z. & Zhang, X. Modeling CO_2 solubility in water at high pressure and temperature conditions. *Energy Fuels* **34**, 4761–4776 (2020).
49. Menad, N. A., Hemmati-Sarapardeh, A., Varamesh, A. & Shamshirband, S. Predicting solubility of CO_2 in brine by advanced machine learning systems: Application to carbon capture and sequestration. *J. CO2 Util.* **33**, 83–95 (2019).
50. Ali Ahmadi, M. & Ahmadi, A. Applying a sophisticated approach to predict CO_2 solubility in brines: Application to CO_2 sequestration. *Int. J. Low-Carbon Technol.* **11**, 325–332 (2016).
51. Safamirzaei, M. & Modarress, H. Modeling and predicting solubility of *n*-alkanes in water. *Fluid Phase Equilib.* **309**, 53–61 (2011).
52. 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).
53. Nabipour, N., Mosavi, A., Baghban, A., Shamshirband, S. & Felde, I. Extreme learning machine-based model for solubility estimation of hydrocarbon gases in electrolyte solutions. *Processes* **8**, 92 (2020).
54. Ott, J. B. & Boerio-Goates, J. *Chemical Thermodynamics: Advanced Applications: Advanced Applications* (Elsevier, 2000).
55. McKetta, J. J. & Katz, D. L. Methane–*n*-butane–water system in two-and three-phase regions. *Ind. Eng. Chem.* **40**, 853–863 (1948).
56. Eslamimanesh, A., Mohammadi, A. H. & Richon, D. Thermodynamic consistency test for experimental solubility data in carbon dioxide/methane+ water system inside and outside gas hydrate formation region. *J. Chem. Eng. Data* **56**, 1573–1586 (2011).
57. Sultanov, R., Skripka, V. & Namiot, A. Y. Solubility of methane in water at high temperatures and pressures. *Gazova Promyshlennost* **17**, 6–7 (1972).
58. Namiot, A. Y. Solubility of nonpolar gases in water. *J. Struct. Chem.* **2**, 381–389 (1961).
59. Winkler, L. Solubility of gas in water. *Ber. Deut. Chem. Ges* **34**, 1408–1422 (1901).
60. Rettich, T. R., Handa, Y. P., Battino, R. & Wilhelm, E. Solubility of gases in liquids. 13. High-precision determination of Henry's constants for methane and ethane in liquid water at 275 to 328 K. *J. Phys. Chem.* **85**, 3230–3237 (1981).
61. Mohammadi, A. H., Chapoy, A., Tohidi, B. & Richon, D. Measurements and thermodynamic modeling of vapor–liquid equilibria in ethane–water systems from 274.26 to 343.08 K. *Ind. Eng. Chem. Res.* **43**, 5418–5424 (2004).
62. Danneil, A., Tödheide, K. & Franck, E. Verdampfungsgleichgewichte und kritische Kurven in den Systemen Äthan/Wasser und *n*-Butan/Wasser bei hohen Drücken. *Chem. Ing. Tec.* **39**, 816–822 (1967).
63. Morrison, T. & Billett, F. The salting-out of non-electrolytes. Part II. The effect of variation in non-electrolyte. *J. Chem. Soc. (Resumed)* **730**, 3819–3822 (1952).
64. Azarnoosh, A. & McKetta, J. The solubility of propane in water. *Petrol. Refiner* **37**, 275–278 (1958).
65. Kobayashi, R. & Katz, D. Vapor-liquid equilibria for binary hydrocarbon-water systems. *Ind. Eng. Chem.* **45**, 440–446 (1953).
66. Kresheck, G. C., Schneider, H. & Scheraga, H. A. The effect of D2O on the thermal stability of proteins. Thermodynamic parameters for the transfer of model compounds from H_2O to $D_2O_{1,2}$. *J. Phys. Chem.* **69**, 3132–3144 (1965).
67. O'Sullivan, T. D. & Smith, N. O. Solubility and partial molar volume of nitrogen and methane in water and in aqueous sodium chloride from 50 to 125 deg. and 100 to 600 atm. *J. Phys. Chem.* **74**, 1460–1466 (1970).
68. Michels, A., Gerver, J. & Bijl, A. The influence of pressure on the solubility of gases. *Physica* **3**, 797–808 (1936).
69. Danesh, A. *PVT and Phase Behaviour of Petroleum Reservoir Fluids* (Elsevier, 1998).
70. 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).
71. Mohammadi, M.-R. *et al.* Modeling hydrogen solubility in hydrocarbons using extreme gradient boosting and equations of state. *Sci. Rep.* **11**, 1–20 (2021).
72. Smola, A. J. & Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **14**, 199–222 (2004).
73. Schölkopf, B., Smola, A. J., Williamson, R. C. & Bartlett, P. L. New support vector algorithms. *Neural Comput.* **12**, 1207–1245 (2000).
74. Vapnik, V., Golowich, S. E. & Smola, A. Support vector method for function approximation, regression estimation, and signal processing. in *Advances in Neural Information Processing Systems*. 281–287 (1997).
75. Amar, M. N., Shateri, M., Hemmati-Sarapardeh, A. & Alamatsaz, A. Modeling oil-brine interfacial tension at high pressure and high salinity conditions. *J. Petrol. Sci. Eng.* **183**, 106413 (2019).

76. Song, Y.-Y. & Ying, L. Decision tree methods: Applications for classification and prediction. *Shanghai Arch. Psychiatry* **27**, 130 (2015).
77. Patel, N. & Upadhyay, S. Study of various decision tree pruning methods with their empirical comparison in WEKA. *Int. J. Comput. Appl.* **60**, 12 (2012).
78. 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).
79. 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).
80. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
81. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **63**, 3–42 (2006).
82. John, V., Liu, Z., Guo, C., Mita, S. & Kidono, K. *Image and Video Technology.* 721–733 (Springer, 2021).
83. Koza, J. R. & Poli, R. *Search Methodologies*. 127–164 (Springer, 2005).
84. Whigham, P. A. *Proceedings of the Workshop on Genetic Programming: From Theory to Real-World Applications.* 33–41 (Citeseer, 2021).
85. Angeline, P. J. & Spector, L. *Advances in Genetic Programming* Vol. 1 (MIT Press, 1994).
86. Augusto, D. A. & Barbosa, H. J. *Proceedings*. Vol. 1. *Sixth Brazilian Symposium on Neural Networks.* 173–178 (IEEE, 2021).
87. Haeri, M. A., Ebadzadeh, M. M. & Folino, G. Statistical genetic programming for symbolic regression. *Appl. Soft Comput.* **60**, 447–469 (2017).
88. Mohammadi, M.-R. *et al.* Toward predicting SO_2 solubility in ionic liquids utilizing soft computing approaches and equations of state. *J. Taiwan Inst. Chem. Eng.* **133**, 104220 (2022).
89. Ivakhnenko, A. G. Polynomial theory of complex systems. *IEEE Trans. Syst. Man Cybern.* **4**, 364–378 (1971).
90. Rostami, A., Hemmati-Sarapardeh, A. & Mohammadi, A. H. Estimating *n*-tetradecane/bitumen mixture viscosity in solvent-assisted oil recovery process using GEP and GMDH modeling approaches. *Pet. Sci. Technol.* **37**, 1640–1647 (2019).
91. Huang, W. *et al.* Application of modified GMDH network for CO_2-oil minimum miscibility pressure prediction. *Energy Sour. Part A Recov. Util. Environ. Effects* **42**, 2049–2062 (2020).
92. Menad, N. A. *et al.* Modeling temperature dependency of oil-water relative permeability in thermal enhanced oil recovery processes using group method of data handling and gene expression programming. *Eng. Appl. Comput. Fluid Mech.* **13**, 724–743 (2019).
93. Rostami, A. *et al.* Modeling heat capacity of ionic liquids using group method of data handling: A hybrid and structure-based approach. *Int. J. Heat Mass Transf.* **129**, 7–17 (2019).
94. Mahdaviara, M., Menad, N. A., Ghazanfari, M. H. & Hemmati-Sarapardeh, A. Modeling relative permeability of gas condensate reservoirs: Advanced computational frameworks. *J. Petrol. Sci. Eng.* **189**, 106929 (2020).
95. Mohammadi, M.-R., Hemmati-Sarapardeh, A., Schaffie, M., Husein, M. M. & Ranjbar, M. Application of cascade forward neural network and group method of data handling to modeling crude oil pyrolysis during thermal enhanced oil recovery. *J. Petrol. Sci. Eng.* **205**, 108836 (2021).
96. Nakhaei-Kohani, R., Taslimi-Renani, E., Hadavimoghaddam, F., Mohammadi, M.-R. & Hemmati-Sarapardeh, A. Modeling solubility of CO_2–N_2 gas mixtures in aqueous electrolyte systems using artificial intelligence techniques and equations of state. *Sci. Rep.* **12**, 1–23 (2022).
97. 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 (2022).
98. Chen, G. *et al.* The genetic algorithm based back propagation neural network for MMP prediction in CO_2-EOR process. *Fuel* **126**, 202–212 (2014).
99. 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).
100. Leroy, A. M. & Rousseeuw, P. J. Robust regression and outlier detection. *rrod* (1987).
101. Goodall, C. R. *13 Computation Using the QR Decomposition*. (1993).
102. Gramatica, P. Principles of QSAR models validation: Internal and external. *QSAR Comb. Sci.* **26**, 694–701 (2007).
103. Mohammadi, M.-R. *et al.* Modeling hydrogen solubility in alcohols using machine learning models and equations of state. *J. Mol. Liq.* **346**, 117807 (2021).
104. Mohammadi, M.-R. *et al.* Modeling of nitrogen solubility in unsaturated, cyclic, and aromatic hydrocarbons: Deep learning methods and SAFT equation of state. *J. Taiwan Inst. Chem. Eng.* **131**, 104123 (2021).

Author contributions

Mohammad-Reza Mohammadi: Investigation, Data curation, Visualization, Writing-Original Draft, Fahime Hadavimoghaddam: Conceptualization, Validation, Modeling, Saeid Atashrouz: Writing-Review & Editing, Validation, Ali Abedi: Writing-Review & Editing, Validation, Abdolhossein Hemmati-Sarapardeh: Methodology, Validation, Supervision, Writing-Review & Editing, Ahmad Mohaddespour: Writing-Review & Editing, Validation.

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/s41598-022-18983-2.

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

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.