Estimating Carbon Dioxide Solubility in Brine Using Mixed Effects Random Forest Based on Genetic Algorithm: Implications for Carbon Dioxide Sequestration in Saline Aquifers

Grant Charles Mwakipunda[1], AL-Wesabi Ibrahim[2], Allou Koffi Franck Kouassi[1], Norga Alloyce Komba[3], Edwin Twum Ayimadu[4], Melckzedeck Michael Mgimba[5], Mbega Ramadhani Ngata[1], and Long Yu[1]*

[1]Key Laboratory of Theory and Technology of Petroleum Exploration and Development in Hubei Province, China University of Geosciences, Wuhan 430074, China.
[2]College of Electrical and Information Engineering in Hunan University, Hunan 410083, China
[3]Research Institute of Environmental Law, Wuhan University, No. 299, Luoyu Road, 10 Wuchang District, Wuhan City, Hubei Province, China.
[4]School of Resource and Environmental Science, Wuhan University, No. 299, Luoyu Road, 10 Wuchang District, Wuhan City, Hubei Province, China.
[5]Mbeya University of Science and Technology (MUST), P.O Box 131, Mbeya, Tanzania

Summary

Accurate prediction of carbon dioxide (CO_2) solubility in brine is crucial for the success of carbon capture and storage (CCS) by means of geological formations like aquifers. This study investigates the effectiveness of a novel genetic algorithm-mixed effects random forest (GA-MERF) model for estimating CO_2 solubility in brine. The model's performance is compared with established methods like the group method of data handling (GMDH), backpropagation neural networks (BPNN), and traditional thermodynamic models. The GA-MERF model utilizes experimental data collected from literature, encompassing key factors influencing CO_2 solubility: temperature (T), pressure (P), and salinity. These data are used to train and validate the model's ability to predict CO_2 solubility values. The results demonstrate the superiority of GA-MERF compared to the other models. Notably, GA-MERF achieves a high coefficient of determination (R) of 0.9994 in unseen data, indicating a strong correlation between estimated and actual CO_2 solubility values. Furthermore, the model exhibits exceptionally low error metrics, with a root mean squared error (RMSE) of 2×10^{-8} and a mean absolute error (MAE) of 1.8×10^{-11}, signifying outstanding accuracy in estimating CO_2 solubility in brine. Beyond its high accuracy, GA-MERF offers an additional benefit—reduced computational time compared to the other models investigated, with 65 seconds. This efficiency makes GA-MERF a particularly attractive tool for real-world applications where rapid and reliable CO_2 solubility predictions are critical. In conclusion, this study presents GA-MERF as a powerful and efficient model for predicting CO_2 solubility in brine. Its superior performance compared to existing methods and previous literature highlights its potential as a valuable tool for researchers and engineers working on CCS projects utilizing aquifer storage. The high accuracy, low error rates, and reduced computational time make GA-MERF a promising candidate for advancing the development of effective and efficient CCS technologies.

Introduction

CO_2 is the main greenhouse gas, contributing to 76% of total emitted gas in the atmosphere (Mwakipunda et al. 2024). Atmospheric CO_2 concentrations are rising primarily due to human activity, particularly the burning of fossil fuels for energy. Since the industrial revolution, human CO_2 emissions have increased significantly, from ~280 ppm before the industrial revolution to ~425 ppm in March 2024, continuing to affect global climatic change (Statista 2024). Decarbonization and carbon capture, utilization, and storage (CCUS) are essential strategies in mitigating these CO_2 emissions and combating climate change, with the goal of meeting the Paris Climate Summit agreement to limit global warming to well below 2°C above preindustrial levels and pursuing efforts to further limit it to 1.5°C by 2050. In essence, while decarbonization is crucial for a sustainable future, CCUS complements these efforts by offering solutions for hard-to-decarbonize sectors, reducing carbon intensity, achieving negative emissions, mitigating economic and social impacts, and maintaining energy security.

CCUS is a combination of technologies and processes designed to capture CO_2 emitted from industrial sources, use or convert them for beneficial purposes, and store them to prevent their release into the atmosphere (Mwakipunda et al. 2023a; Nath et al. 2024; Liu and Wu 2024). If the captured CO_2 is not utilized, it can be stored underground permanently in geological formations to prevent its release into the atmosphere. This process is known as CCS (Mwakipunda et al. 2023c; Ngata et al. 2023; Zhang et al. 2024). Suitable storage sites include depleted oil and gas reservoirs, saline aquifers, unmineable coal seams, basalt formations, shale formations, salt caverns, deep ocean storage, etc. There are four mechanisms by which CO_2 can be stored underground, which are structural trapping, residual trapping, solubility trapping, and mineral trapping (Luo et al. 2022). Solubility trapping offers a robust and reliable mechanism for permanent CO_2 storage. Solubility trapping involves the dissolution of CO_2 into formation fluids (brine), for instance, in an aquifer, creating a stable solution. This mechanism offers a long-term and permanent storage solution as the CO_2 is chemically incorporated into the brine, reducing

*Corresponding author; email: yulong36@cug.edu.cn

Original SPE manuscript received for review 9 May 2024. Revised manuscript received for review 31 July 2024. Paper (SPE 223123) peer approved 27 August 2024.

the risk of leakage or migration back to the surface (Ratnakar et al. 2020; Mwakipunda et al. 2023b). Accurate estimation of CO_2 storage in brine formations is a fundamental aspect of planning, implementing, and managing successful CCS projects, contributing to the effective and responsible deployment of these technologies to mitigate CO_2 emissions and combat climate change. It helps to optimize storage capacity, assess project feasibility, ensure safety and environmental integrity, comply with regulations, facilitate monitoring and verification, and engage with the public and stakeholders.

There are several methods for estimating CO_2 solubility in brine: (1) Laboratory experiments, which involve injecting CO_2 into brine at reservoir conditions and measuring the solubility as a function of pressure, temperature, and salinity (Yan et al. 2011; Wang et al. 2014; Mosavat et al. 2014; Mohammadian et al. 2015, 2023; Zhao et al. 2015a; Lu et al. 2023; Wang and Ehlig-Economides 2023; Ji et al. 2024; Mutailipu et al. 2024). However, laboratory experiments face several challenges related to scale, cost and time intensiveness, representativeness, parameter range, potential for contamination, simplifications, scaleup challenges, and safety concerns. (2)Pressure-volume-temperature empirical correlations are derived from laboratory measurements and can provide quick and approximate estimates of CO_2 solubility for a wide range of conditions (Li and Nghiem 1986; Bahadori et al. 2009; Zhao et al. 2015a; Mutailipu et al. 2024). Empirical correlations have several limitations, such as applicability, extrapolation risks, data quality, simplifications, mechanistic understanding, validation requirements, and sensitivity to assumptions. (3) Equation-of-state models, such as the Peng-Robinson (Li and Nghiem 1986; Sodeifian et al. 2023; Hiraga and Ushiki. 2024) or Soave-Redlich-Kwong (Li et al. 2001; Sodeifian et al. 2024; Mehdizade et al. 2024; Costa de Souza et al. 2024) equations, can be used to calculate CO_2 solubility in brine based on thermodynamic principles and phase behavior. However, equations of state require detailed information on the properties of CO_2 and brine, which provides an accurate estimation of CO_2 solubility at various pressures and temperatures. Other researchers, such as Sørensen et al. (2002), Portier and Rochelle (2005), Duan et al. (2006), and Mao et al. (2013), developed some models in CO_2 solubility estimations in brine but these models still have data range applicability limitations.

Recently, a few researchers have applied machine learning (ML) models as alternative techniques in estimating CO_2 solubility in brines because they offer a promising approach for estimating CO_2 solubility in brines due to their ability to handle complex, nonlinear relationships between influencing factors. Traditional methods can be cumbersome and computationally expensive, especially for brines with varying salinity, temperature, and ionic compositions. ML can learn from large data sets of experimental measurements within a short time and is less costly, capturing these intricate dependencies and enabling rapid, accurate CO_2 solubility predictions for diverse brine compositions encountered in CCS applications. For instance, Ratnakar et al. (2023) estimated CO_2 in brine from experimental data, (i.e., pressure, temperature, and salinity) using different ML models, which were random forests (RF), decision trees, and artificial neural networks. It was found that the utilized ML models estimated the CO_2 solubility in brine accurately with a minimum relative error of 2–7% of the experimental data. Decision trees suffer from overfitting and instability, while RFs, though more robust, can be complex and computationally intensive. Artificial neural networks, while powerful, require substantial data and computational resources, and their black-box nature can be problematic for interpretability. Also, Menad et al. (2019) utilized different hybrid ML models in estimating CO_2 solubility in brine from experimental data having pressure, temperature, and salinity as inputs. Among the ML models used, radial basis function neural network (RBFNN)-artificial bee colony (ABC) accurately estimated CO_2 solubility in brine with correlation coefficient (R^2) of 0.9984 and 0.9896 and RMSE of 0.0218 and 0.0572 during training and testing, respectively, followed by RBFNN-particle swarm optimization, multilayer perceptron-Levenberg-Marquardt, and RBFNN-GA. While RBFNN-ABC combines the strengths of both RBFNN and ABC, it also inherits their limitations. These include computational complexity, sensitivity to parameters and initial conditions, slower convergence rates, and potential scalability issues. Careful tuning and validation are required to achieve optimal performance with this combined approach. Further, Zou et al. (2024) estimated the CO_2 solubility of experimental data in brine with pressure, temperature, and salinity as inputs using various ML models. From the used ML models, cascade forward neural network (CFNN)-Levenberg-Marquardt estimated precisely the experimental data with the highest R^2 of 0.9949 and minimum average absolute percent relative error values of 5.37% for the overall data set compared to CFNN-bayesian regularization (BR), cascade forward neural network-Broyden–Fletcher–Goldfarb –Shanno(CNN-BFGS), and (general regression neural network) GRNN. From sensitivity analysis, it was found that pressure has a positive impact on CO_2 solubility, whereas temperature and salinity have negative impacts with CO_2 solubility. While CFNN-Levenberg-Marquardt can be powerful for certain applications, it comes with significant limitations, including high computational and memory requirements, sensitivity to initial conditions, potential for overfitting, and challenges with scalability and interpretability. Careful design, tuning, and validation are necessary to mitigate these limitations and achieve robust model performance.

Hence, in this paper, we apply a novel ML algorithm to estimate CO_2 solubility in brine for CCS applications from experimental data. We combined GA-MERF to form a new hybrid approach that leverages the strengths of both techniques: GA's ability to explore the search space for optimal solutions and MERF's capability to handle complex nonlinear relationships and make accurate estimations. This new hybrid algorithm was established to overcome the limitations of the ML models described in the previous paragraph, such as avoiding overfitting, fast convergence, high accuracy with minimum errors, and handling complex nonlinear relationships. Furthermore, we conducted SHapley Additive exPlanations (SHAP) analysis in this study. To assess its effectiveness, GA-MERF was compared with BPNN, GMDH and other empirical correlations (i.e., the thermodynamic model).

Data Sources and Preprocessing

Data Collection. Data collection from experimental published literature on CO_2 solubility in brine serves as a critical foundation for this paper. The 1,000 data sets utilized in this study to estimate CO_2 solubility in brine were collected from experimental published literature in different types of brines made of several salt types, such as NaCl, KNO_3, $CaCO_3$, $CaSO_4$, $MgSO_4$, $NaHCO_3$, $NaNO_3$, Na_2SO_4, K_2SO_4 KCl, $Mg(NO_3)_2$, $CaCl_2$, and $MgCl_2$, to reflect various types of salts that exist in different aquifers (Markham and Kobe 1941; Nighswander et al. 1989; Rumpf and Maurer 1993; Rumpf et al. 1994; Kiepe et al. 2002; Bando et al. 2003; Koschel et al. 2006; Yan et al. 2011; Zhao et al. 2015b; Mohammadian et al. 2015; Steel et al. 2016; Cruz et al. 2021); 70% of the data was used for training and 30% was used for testing the models. The data inputs include temperatures in the range of 50.5–473.7 K, pressures in the range of 0.0643–986.8 bar, and salinity in the ranges of 0–5.999 mol/kg. The output CO_2 solubility range was from 0 to 0.09388 mole fraction. Temperature plays a pivotal role in governing the thermodynamic behavior of the CO_2-brine system, affecting the solubility and phase behavior of CO_2. Pressure, on the other hand, influences the physical state and density of the system, impacting the solubility and storage capacity of CO_2 in brine. Salinity concentration, representing the dissolved salt content in the brine, influences the chemical interactions, ion activities, and phase equilibrium of the CO_2-brine system.

Data Preprocessing. CO_2 solubility in brine is controlled mostly by three factors, which are pressure, temperature, and salinity concentration of aquifer water. Based on the collected data from the literature, they contain some outliers, as shown in **Fig. 1**, which

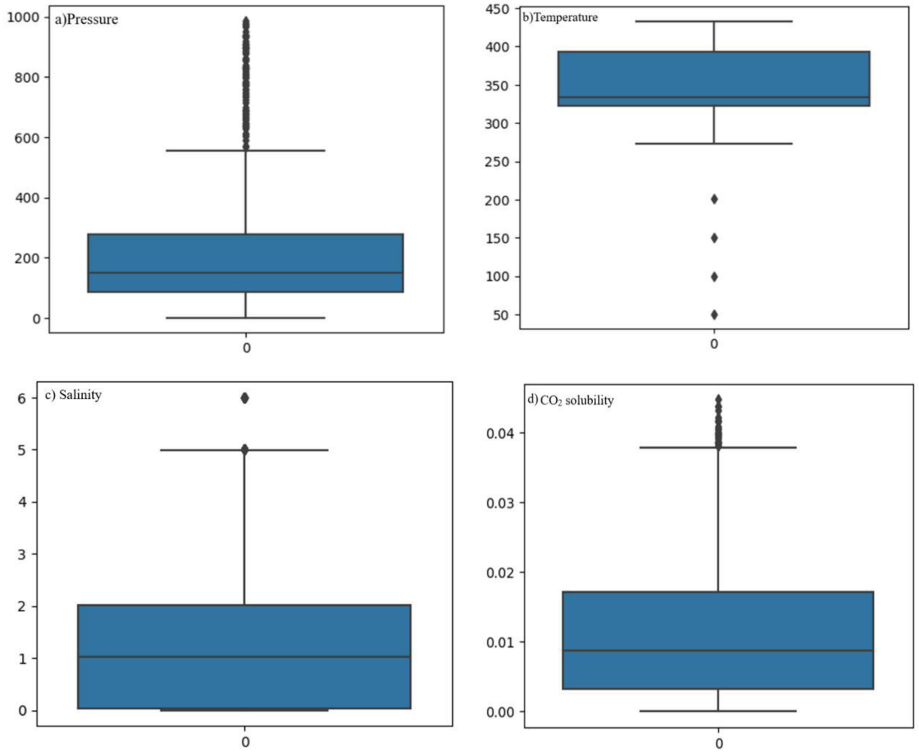

Fig. 1—Experimental data sets with outliers.

usually affect model performance. To enhance the model's performance, outliers were removed using the Z-score technique, as shown in **Fig. 2**. The descriptive statistics of the used data for training and testing the model are shown in **Table 1**. Normalization is a vital preprocessing step in ML, paving the way for better model performance, interpretability, and overall robustness. First, it helps to ensure that features are on a similar scale, which prevents certain features from dominating others during the training process. Second, normalization aids in speeding up the convergence of iterative optimization algorithms, leading to faster training times. Moreover, it can improve the performance of models by making them more robust to outliers and noise in the data. Additionally, normalization facilitates the interpretation of model parameters since the scale of the features no longer affects the magnitude of the weights (Majid et al. 2023; Mkono et al. 2023). In this paper, all the data were normalized in the range of 0 to 1 using Eq. 1.

Experimental Data Sets	Parameters	Max	75%	50%	25%	Min	SD	Mean
Inputs	Pressure (bar)	986.8	275.975	150	84.045	0.84	242.7746	232.2562
	Temperature (K)	433.18	393.1425	333.17	323.1	50.5	45.7701	354.8309
	Salinity (mol/kg)	5.999	2.015	1.0125	0.03	0	1.7465	1.7452
Output	CO_2 solubility in brines (mole fraction)	0.0448	0.01721	0.0088	0.00328	0	0.01147	0.012

Max: Maximum; Min: Minimum; SD: Standard Deviation

Table 1—Descriptive statistics of the used data sets.

$$y_i' = \frac{y_i - y_{\min}}{y_{\max} - y_{\min}}, \tag{1}$$

where y_i', y_i, $y_{\min}$, $y_{\max}$ are the normalized value of y_i, the value to be normalized, the minimum value of y_i, and the maximum value of y_i, respectively.

Methodology

This section discusses different ML algorithms that were used for CO_2 solubility estimation in brine. It includes a novel GA-MERF algorithm compared with GMDH and BPNN, which were utilized to measure its robustness and convergence in predicting CO_2 solubility in brine for CO_2 sequestration implementation.

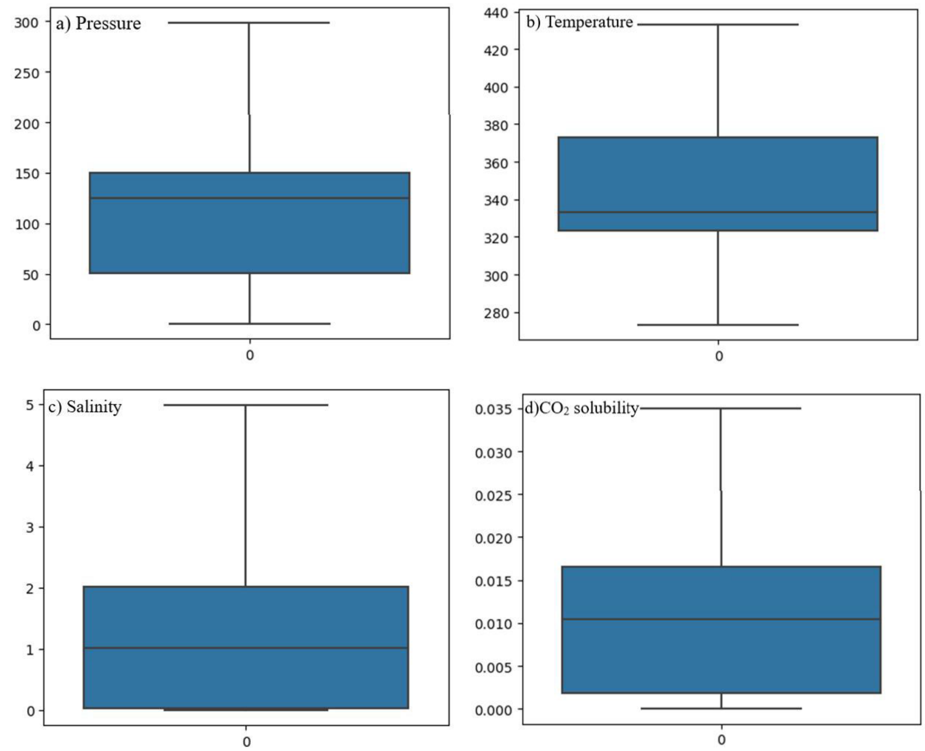

Fig. 2—Experimental data sets without outliers.

BPNN. The BPNN contains three layers, which are input, hidden, and output layers. BPNN architecture was established by various researchers (Yang et al. 2021; Yao et al. 2023; Duan et al. 2023). The configuration of input and output layers depends on the specific problem and study objectives. The minimum requirement for hidden neurons in the hidden layer is two (Buscema 1998). The training process in BPNN involves iteratively adjusting weights to reduce errors and to improve accuracy. In the present study, input variables for CO_2 solubility estimation are pressure, temperature, and salinity. These inputs are transformed through a nonlinear hyperbolic tangent activation function in the hidden layers, producing outputs as defined by Eq. 2 (Majid et al. 2023).

$$f(w) = \tanh(w) = \frac{2}{1 - e^{-2w}} - 1, \tag{2}$$

where w denotes the cumulative weights of inputs. The training function of a BPNN is expressed by the nonlinear equation (Eq. 3),

$$z^* = \operatorname{argmin} V(r) . \tag{3}$$

The goal is to identify the optimal weight connections z^* to minimize the disparity between predicted and actual values. The error function $V(r)$ is expressed in Eq. 4 (Elkatatny and Mahmoud 2018):

$$V(r) = \sum_n V_n(r) . \tag{4}$$

Here, $V_n(r)$ represents output error and n stands for the number of training data, which is defined as:

$$V_n(r) = \frac{1}{2}\sum_j \left(y_{nj} - \hat{y}_{nj}(r)\right)^2 . \tag{5}$$

In this case of the presented model, y_{nj} denotes the predicted values for the nth observations while $\hat{y}_{nj}(r)$ represents the actual values for the jth observations. Substituting Eq. 5 into Eq. 4 yields Eq. 6, which is then optimized to minimize the error between predicted and actual values (Ikiensikimama and Azubuike 2012; Mulashani et al. 2022; Dongare et al. 2024). The training process involves iteratively adjusting the weights of output neurons to modify the inputs until the error converges to the desired value (Hecht-Nielsen 1992; Liu et al. 2023; Al-Bukhaiti et al. 2024).

$$V(r) = \frac{1}{2}\sum_n \sum_j \left(y_{nj} - \hat{y}_{nj}(r)\right)^2 . \tag{6}$$

GMDH. The GMDH relies on a hierarchical network of polynomial equations progressively fitting to the data (Ivakhnenko 1971). The selection of inputs for subsequent layers and nodes in the network is determined through the integration of multilayer techniques. The creation of the GMDH network structure involves combining specific layers and nodes, guided by the performances of earlier layers and nodes (Lv et al. 2023; Mgimba et al. 2023; Rezaie et al. 2023; Zhang and Xue 2024). **Fig. 3** shows the GMDH neural network.

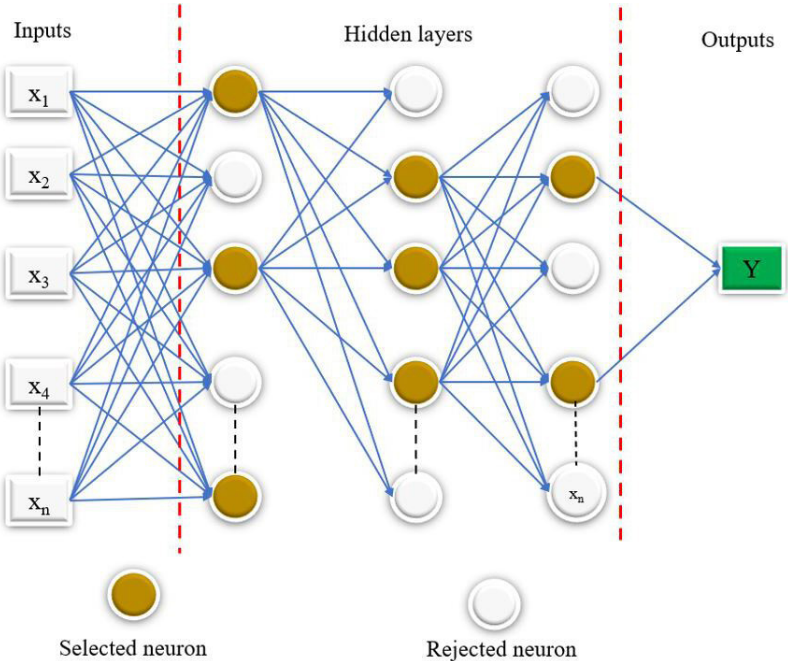

Fig. 3—GMDH network architecture.

The relationship between independent variable Y and dependent input variables x_i is captured by the model as:

$$Y = f\left(x_1, x_2, x_{3,} ....., x_n\right) . \tag{7}$$

Then, the polynomial Kolmogorov-Gabor of Eq. 7 is written as

$$y = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} x_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} a_{ijk} x_i x_j x_k + \dots , \tag{8}$$

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & \dots & x_{1M} \\ x_{21} & x_{22} & \dots & \dots & x_{2M} \\ \dots & \dots & \dots & x_{ij} & x_{iM} \\ \dots & \dots & \dots & \dots & \dots \\ x_{N1} & x_{N2} & \dots & \dots & x_{NM} \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ \dots \\ y_N \end{bmatrix}, \tag{9}$$

where y denotes the output, n signifies the number of inputs, x represents the input systems, and a corresponds to the coefficients. Partial polynomial equations of Eqs. 8 and 9 are expressed as

$$y = G\left(X_i X_j\right), \tag{10}$$

$$\hat{y} = a_0 + a_1 x_i + a_2 x_j + a_3 x_i x_j + a_4 x_i^2 + a_5 x_j^2, \tag{11}$$

$$y_i = A a_i, \tag{12}$$

$$A = \begin{bmatrix} 1 & x_{1p} & x_{1q} & x_{1p}x_{1q} & x_{1p}^2 & x_{1q}^2 \\ 1 & x_{2p} & x_{2q} & x_{2p}x_{2q} & x_{2p}^2 & x_{2q}^2 \\ 1 & \dots & \dots & \dots & \dots & \dots \\ 1 & x_{Np} & x_{Nq} & x_{Np}x_{Nq} & x_{Np}^2 & x_{Nq}^2 \end{bmatrix}. \tag{13}$$

To identify the underlying relationships between input and output data, the least-squares method is used to estimate the coefficients of various quadratic equations in Eq. 14.

$$\mathrm{MSE} = \frac{1}{N}\sum_{i}^{N}\left(y_i - \hat{y}_i\right)^2 \sim \min . \tag{14}$$

Building the initial set of desired outputs is optimally achieved through a least-squares regression based on Eq. 14. This leverages n input data points and incorporates the entire spectrum of potential outputs for two specific values $(y_i, i = 1, 2, 3 \dots, N)$. Hence, $\begin{pmatrix} n \\ 2 \end{pmatrix} = n\left(n-1\right)/2$, based on the observed patterns in the $p : q$ 1;2; 3,... data, the potential for pre-FFN neuronal development level is developed $y_i; x_{pi}, x_{qi}\ \left(i = 1, 2, 3 \dots, N\right)$ (Bueno et al. 2011; Teng et al. 2017). To enrich the data sets and improve model performance, Eq. 15 is utilized to generate N additional data points $y_i; x_{pi}, x_{qi}\ \left(i = 1, 2, 3 \dots, N\right)$ from existing data sets, effectively creating pairs (p, A), where A takes on various values (1, 2, 3, ...) (Teng et al. 2017; Hemmati-Sarapardeh et al. 2020; Mulashani et al. 2022; Zhang and Xue 2024).

$$\begin{bmatrix} x_{ip} & x_{1q} & : & y_1 \\ x_{2p} & x_{2q} & : & y_2 \\ \cdots & \cdots & \cdots & \cdots \\ x_{Np} & x_{Np} & : & y_N \end{bmatrix}. \tag{15}$$

A set of N matrix equations is developed by applying the quadratic subexpression within Eq. 16 to each row, as shown in Eq. 19.

$$A_a = Y, \tag{16}$$

$$a = \left[a_0, a_1, a_2, a_3, a_5\right], \tag{17}$$

$$Y = \left[y_1, y_2, y_3, \ldots, y_N\right]^T, \tag{18}$$

$$A = \begin{bmatrix} 1 & x_{1p} & x_{1q} & x_{1p}x_{1q} & x_{1p}^2 & x_{1q}^2 \\ 1 & x_{2p} & x_{2q} & x_{2p}x_{2q} & x_{2p}^2 & x_{2q}^2 \\ 1 & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{Np} & x_{Nq} & x_{Np}x_{Nq} & x_{Np}^2 & x_{Nq}^2 \end{bmatrix}. \tag{19}$$

For computational convenience, the simplified normal solution can be represented in the familiar form of the least square's solution (Roshani et al. 2020; Sun et al. 2024):

$$a_i = \left(A^T A\right)^{-1} A^T Y. \tag{20}$$

GA-MERF. GA-MERF integrates GA with MERF and aims to optimize the selection of hyperparameters, such as the number of trees, tree depth, and regularization parameters, which can significantly impact the model's performance by reducing overfitting and improving the prediction accuracy to unseen data. GA can efficiently search the hyperparameter space to find optimal or near-optimal configurations, thus enhancing MERF predictive accuracy capability. Moreover, GA assists in selecting the most appropriate model structure, including the choice of fixed and random effects, leading to more interpretable and accurate models. As both GA and MERF continue to evolve, their combined use is expected to yield innovative solutions to complex data analysis problems in various domains.

GA. GA, inspired by the process of natural selection, iteratively improves a population of candidate solutions to achieve optimal or near-optimal results for complex optimization and search problems (Holland 1973; Elyan and Gaber 2017). The core principle of GA hinges on the concept of "survival of the fittest," derived from Darwinian evolution (Holland 1992; Shadkani et al. 2024). Individuals with superior fitness, as measured by a problem-specific objective function, are more likely to be selected for reproduction. This selection process guides the search toward promising regions of the solution space (Katoch et al. 2021). Three fundamental operators orchestrate the evolutionary process in GA (Ray et al. 2023) are as follows:

1. Selection: Selection algorithms determine which individuals from the current population contribute their genetic material to the next generation. Popular selection methods include roulette wheel selection, tournament selection, and elitism, where the fittest individuals are guaranteed to survive.
2. Crossover: Crossover mimics biological reproduction by combining genetic material from two parent chromosomes to generate offspring. Common crossover techniques include single-point crossover and two-point crossover, where sections of the parent chromosomes are swapped to create new combinations.
3. Mutation: Mutation introduces random variations into the offspring's chromosomes, maintaining genetic diversity within the population. This helps prevent premature convergence and explore new regions of the search space. The mutation rate, controlling the frequency of mutations, is a crucial parameter in GA design.

GA operates iteratively. The initial population is typically generated randomly, ensuring a diverse starting point. Subsequently, the selection, crossover, and mutation operators are applied to create a new generation of individuals. This cycle continues for a predefined number of generations or until a termination criterion, such as achieving a desired fitness level, is met (Katoch et al. 2021; Razavi-Termeh et al. 2023; Dhanya and Chitra 2024). Over successive generations, the population gradually evolves toward better solutions, mimicking the process of natural selection.

MERF. MERF is an extension of RF that accommodates both fixed and random effects in the data. It is particularly well-suited for analyzing hierarchical or nested data, such as repeated measures or clustered data. MERF combines the strengths of RF with mixed effects, offering robustness against overfitting and the ability to capture complex nonlinear relationships. This approach can be formulated as (Hajjem et al. 2011, 2017; Rutten 2021; Yang et al. 2022; Krennmair and Schmid 2022; Katreddi et al. 2023):

$$\begin{gathered} y_i = f\left(X_i\right) + Z_i u_i + \varepsilon_i, \\ u_i \sim N\left(0, G\right), \varepsilon_i \sim N\left(0, R_i\right), i = 1, 2, ..., K \end{gathered}, \tag{21}$$

where $y_i = \left(y_{i1}, ..., y_{in}\right)$ stands for vector output for cluster i; X_i and Z_i represent design matrices for fixed effects and RF, respectively; u_i represents unknown vector for random effects; and ε_i stands for residual vector. The constant part $f\left(X_i\right)$ is computed by an RF. In the MERF model, it is assumed that the data from the clusters are independent and u_i as well as ε_i. In addition, a diagonal matrix ($R_i = \sigma^2 I_{in}$) is needed to ensure that the residual structures and sizes are identical across all clusters. Steps for MERF implementation are as follows (Hajjem et al. 2011, 2017; Mayapada et al. 2021; Katreddi et al. 2023):

Step 1: Set the random effects factors to zero at the beginning,$\sigma_R^2 = 1$, and G as the identity matrix ($G = I_m$). k as the iteration number is zero (Hajjem et al. 2014; Rutten 2021; Yang et al. 2022).

Step 2: (i) Escalating the value of k by 1 to $k = k + 1$, followed by subtracting the random component from the output, $y^*_{i(k)} = y_i - Z_i u\_\mathrm{i}\left(k - 1\right)$(Hajjem et al. 2014; Rutten 2021).

(ii) The bagging method is used to train the RF model based on $y^*_{i(k)}$.

(iii) Make predictions for new data points in every observation *j* using trees that were not affected by *j* during the training process (Hajjem et al. 2014; Rutten 2021).

(iv) Subsequently ,update u_i (Hajjem et al. 2014; Rutten 2021):

$$\hat{u}_{i(k)} = \hat{G}_{(k-1)} Z_i^T V_{i(k-1)}^{-1} \left(y_i - \hat{f}(X_i)_{(k)} \right), i = 1, ..., n, \tag{22}$$

where $V_{i(k-1)} = Z_i \hat{G} Z\left(k-1\right)_i^T + \hat{\sigma}_{R(k-1)}^2 I_{ni}$

Step 3: Update the covariance matrix *G* and the estimate σ_R^2 by using the most recent residual values (Hajjem et al. 2014; Rutten 2021).

$$\hat{\sigma}_{R(k)}^2 = \frac{1}{N} \sum_{i=1}^{n} \hat{\varepsilon}_{i(k)}^T \hat{\varepsilon}_{i(k)} + \hat{\sigma}_{R(k-1)}^2 \left(n_i - \hat{\sigma}_{R(k-1)}^2 \cdot \text{trace} \left(V_{i(k-1)} \right) \right), \tag{23}$$

$$\hat{G}_{(k)} = \frac{1}{n} \sum_{i=1}^{n} u_{i(k)}^T u_{i(k)} + \hat{G}_{(k-1)} - \hat{G}_{(k-1)} Z_i^T V_{i(k-1)}^{-1} Z_i \hat{G}_{(k-1)}, \tag{24}$$

where $\varepsilon_{i(k)} = y_i - f(X_i)_{(k)} - Z_i u\_\text{i}(k)$, $i = 1, 2, ..., K$ and is not defined by RF and random effects estimates.

Step 4: Steps 2 and 3 are repeated until the stopping criteria of model fit are met (Hajjem et al. 2014; Rutten 2021; Krennmair and Schmid 2022).

$$GLL\left(f, u|y\right) = \sum_{i=1}^{n} \left(y_i - f\left(X_i\right) - Z_i u_i \right)^T R_i^{-1} \left(y_i - f\left(X_i\right) - Z_i u_i \right) + u_i^T D^{-1} u_i + \log |G| + \log |R_i|). \tag{25}$$

The process is considered to be converged when the general likelihood criteria (GLL_k) is met after the *k*th iteration (Hajjem et al. 2014; Rutten 2021; Krennmair and Schmid 2022; Katreddi et al. 2023):

$$\frac{|GLL_k - GLL_{k-1}|}{GLL_{k-1}} < \delta, \tag{26}$$

for some $\delta > 0$.

In this scenario, relative convergence is chosen above absolute convergence. The GLL absolute value differs substantially per application and sample. Thus, utilizing the actual GLL value to determine convergence is not particularly relevant. The steps for GA-MERF implementation are outlined as follows:

Step 1: After data preprocessing by removing the outliers, the data were divided into training and testing, in which 70% were used for training and 30% for testing the model. A 70/30 split was chosen based on the literature to balance between having enough data for training the model and retaining a sufficient portion for testing to evaluate the model's performance (Majid et al. 2023; Mkono et al. 2023).

Step 2: *Initialization of GA and MERF hyperparameters*: In this paper, GA hyperparameters utilized were population size = 85, mutation rate = 0.2, crossover rate = 0.9, and number of generations = 80. For MERF, the hyperparameters were min_samples_leaf = 11, min_samples_split = 12, max_depth = 8, and number of trees = 200.

Step 3: *Training and evaluation*: For each individual in the population, the encoded hyperparameters were used to build a MERF model, followed by training the MERF model, which was evaluated based on *R*, RMSE, and MAE fitness functions.

Step 4: *Iteration and termination*: Steps 3 and 4 were repeated for a predefined number of generations or until a stopping criterion was met (e.g., reaching a desired fitness level). Over successive generations, the GA should converge toward hyperparameters that result in a better-performing MERF model.

Step 5: *Extract the best model*: The individual in the final population with the highest fitness score was identified, which was used to train the MERF model and test the unseen data based on the utilized fitness function. The simplified GA-MERF flow chart is shown in **Fig. 4**.

Results and Discussion

Models' Performance Indicators. Performance indicators, or evaluation metrics, play a crucial task in evaluating the competence of ML models. The selection of performance metrics is contingent upon the problem being addressed (classification, regression, clustering, etc.) and the particular objectives of the analysis (Naser and Alavi 2021; Arshad et al. 2023). Three performance indicators were used in this paper, which are correlation coefficient (*R*), RMSE, and MAE, as shown in Eqs. 27 through 29, respectively (Willmott and Matsuura 2005; Mgimba et al. 2023; Nadege et al. 2024). In this paper, Python 3.12.3 was used to train and test the models. According to research findings, the model's performance is considered excellent when it achieves an *R*-value near unity, and RMSE and MAE approach zero during training and testing, which signifies a strong alignment between the estimated values by the model and the experimental data (Chai and Draxler 2014; Mkono et al. 2023; Arshad et al. 2023).

$$R = \frac{\sum_{i=1}^{N} \left(y_{\text{act}} - \overline{y_{\text{act}}} \right) \left(Y_{prd} - \overline{Y_{prd}} \right)}{\left(\sqrt{\sum_{i=1}^{N} \left(y_{\text{act}} - \overline{y_{\text{act}}} \right)^2} \right) \left(\sqrt{\sum_{i=1}^{N} \left(Y_{prd} - \overline{Y_{prd}} \right)^2} \right)}, \tag{27}$$

$$\text{RMSE} = \sqrt{\left(\frac{1}{N} \sum_{i=1}^{N} \left(y_{\text{act}} - Y_{prd} \right)^2 \right)}, \tag{28}$$

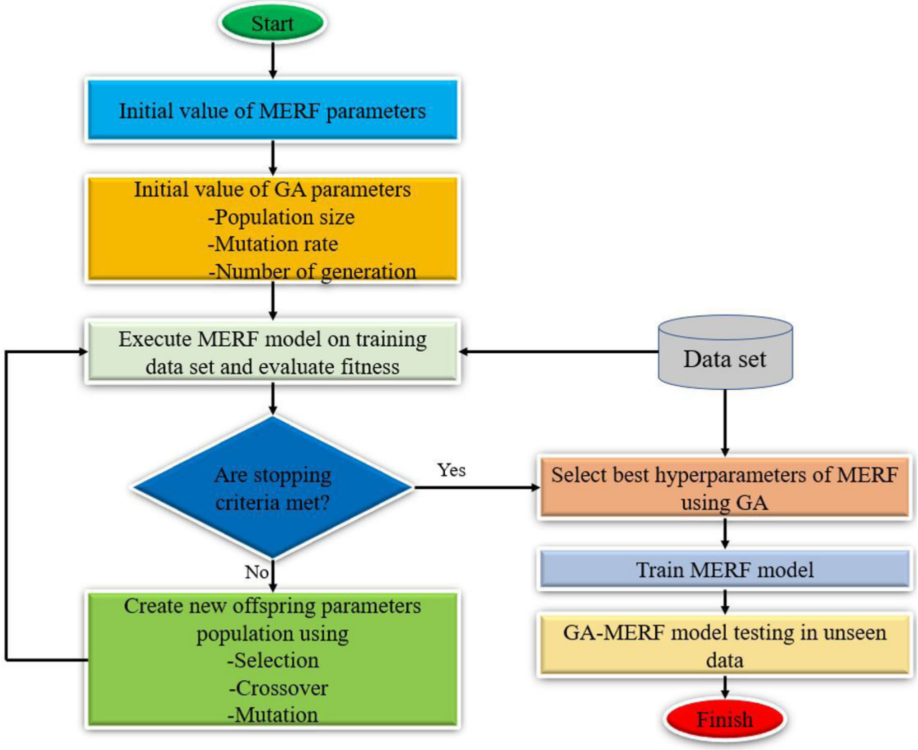
Start
Initial value of MERF parameters
Initial value of GA parameters
-Population size
-Mutation rate
-Number of generation
Execute MERF model on training data set and evaluate fitness
Data set
Are stopping criteria met?
Yes
No
Select best hyperparameters of MERF using GA
Train MERF model
GA-MERF model testing in unseen data
Create new offspring parameters population using
-Selection
-Crossover
-Mutation
Finish

Fig. 4—Flow chart for GA-MERF model.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{n} |y_{\mathrm{act}} - Y_{prd}|. \tag{29}$$

Models' Statistical Analysis. The GA-MERF model demonstrates exceptional *R* values of 0.9999 in training, indicating an excellent fit to the training data. However, its *R* drops to 0.9994 in testing, suggesting the potential ability to estimate new data, as shown in **Table 2**. Similarly, GMDH maintains strong *R* values of 0.9979 during training and 0.9592 during testing, intensifying its incapability to capture the variance in CO_2 solubility estimation to unseen data sets. Further, BPNN models exhibit excellent *R*-value (0.9961) in training but face challenges in maintaining predictive power in the testing phase (0.8846), as shown in **Table 2**, reflecting potential overfitting to the training data and limitations in estimating new data sets. For the RMSE, the GA-MERF model exhibits outstanding performance during the training phase with an RMSE of 1.8×10^{-8}, suggesting a close fit to the training data. However, during testing, the RMSE increases slightly to 2×10^{-8}, as shown in **Fig. 5**, highlighting the potential ability of the model to estimate new data. GMDH performed well in the training phase with low RMSE values of 2.8×10^{-8} and increased to 2×10^{-7} during testing, as shown in **Fig. 5**, showing its robustness in accurately estimating CO_2 solubility in brine. The BPNN model, while effective in training with an RMSE of 3.9×10^{-8}, experienced higher RMSE values of 9.9×10^{-7} during testing, as shown in **Fig. 5**, indicating potential difficulties in maintaining accuracy with unseen data. Further, for MAE analysis, the GA-MERF model demonstrates superior performance in the training phase with an MAE of 1.1×10^{-11} **(Fig. 6)**. However, the MAE increases slightly to 1.8×10^{-11} during testing, as shown in **Fig. 6**, suggesting its ability to adapt to new data. Also, GMDH stands out as consistently reliable in training and testing, with low MAE values of 2.4×10^{-11} during training and 1.4×10^{-10} during testing, as shown in **Fig. 6**, showing its ability to provide accurate CO_2 solubility estimation. In contrast, BPNN models experience an increase in MAE during testing by 8.89×10^{-10}, signaling potential limitations in maintaining precision with unseen data. On the other hand, GA-MERF used less computational time (65 seconds) compared to GMDH and BPNN, as summarized in **Table 3**, showing its fast convergence during model developments.

	R		RMSE		MAE		
Model	Training	Testing	Training	Testing	Training	Testing	Computational Time (seconds)
GA-MERF	0.9999	0.9994	1.8×10^{-8}	2×10^{-8}	1.1×10^{-11}	1.8×10^{-11}	65
GMDH	0.9979	0.9592	2.8×10^{-8}	2.1×10^{-7}	2.4×10^{-11}	1.4×10^{-10}	103
BPNN	0.9961	0.8846	3.9×10^{-8}	9.9×10^{-7}	3.1×10^{-11}	9.2×10^{-10}	188

Table 2—Training and testing results of the utilized models.

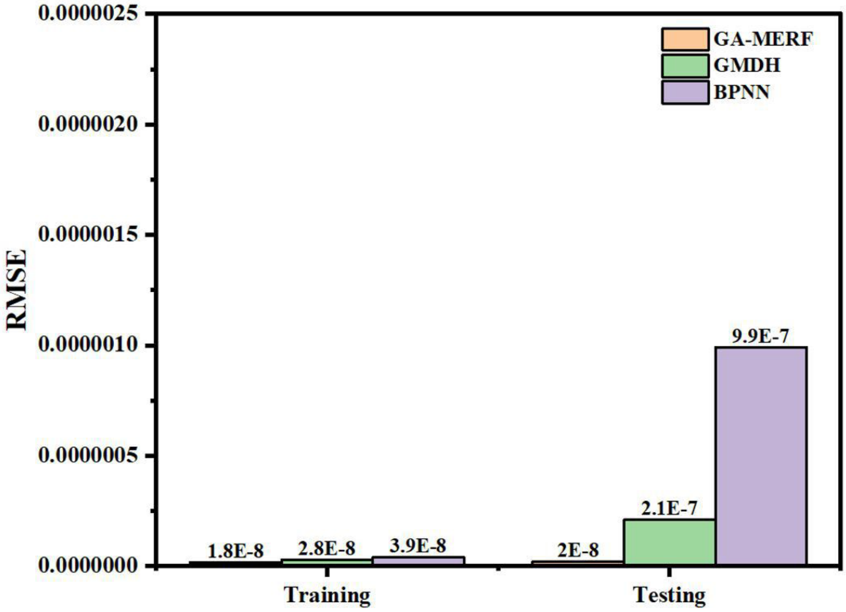

Fig. 5—RMSE comparisons for used models.

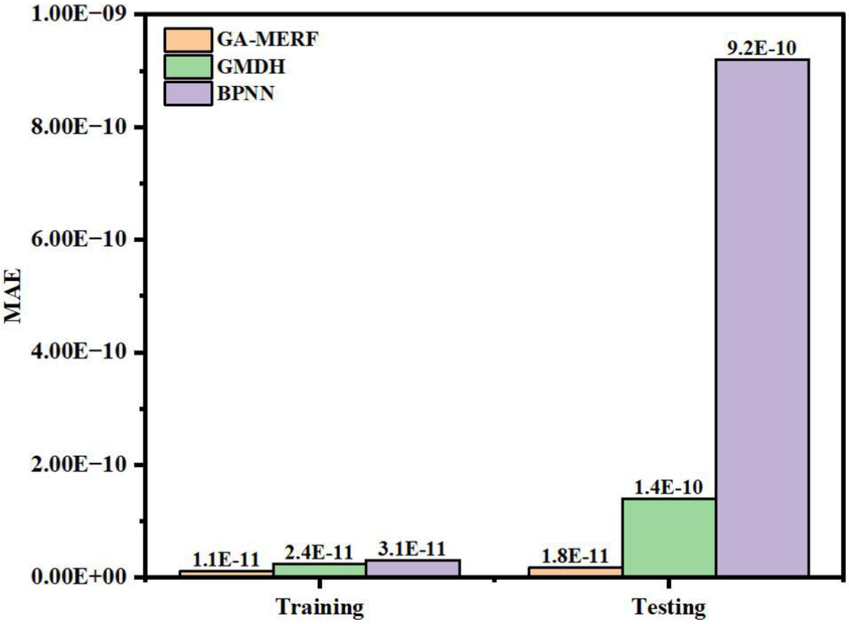

Fig. 6—MAE comparisons for used models.

Models	RMSE	MAE
Thermodynamic model (Sadeghi et al. 2015)	9.25×10^{-6}	4.79×10^{-8}
GA-MERF model	1.72×10^{-8}	1.36×10^{-11}

Table 3—Comparison between GA-MERF and traditional model.

In summarizing this section, GA-MERF outperformed both GMDH and BPNN in CO_2 solubility estimation with high *R* and lower RMSE and MAE during training and testing, respectively, as detailed in **Table 2**. The performance rank is GA-MERF > GMDH > BPNN. This is because GA-MERF addressed the limitations of the other methods, such as overfitting in GMDH and the need for manual tuning in BPNN. The RF component of GA-MERF helps reduce overfitting and capture complex relationships. At the same time, the GA automates hyperparameter optimization, leading to a more robust and accurate model.

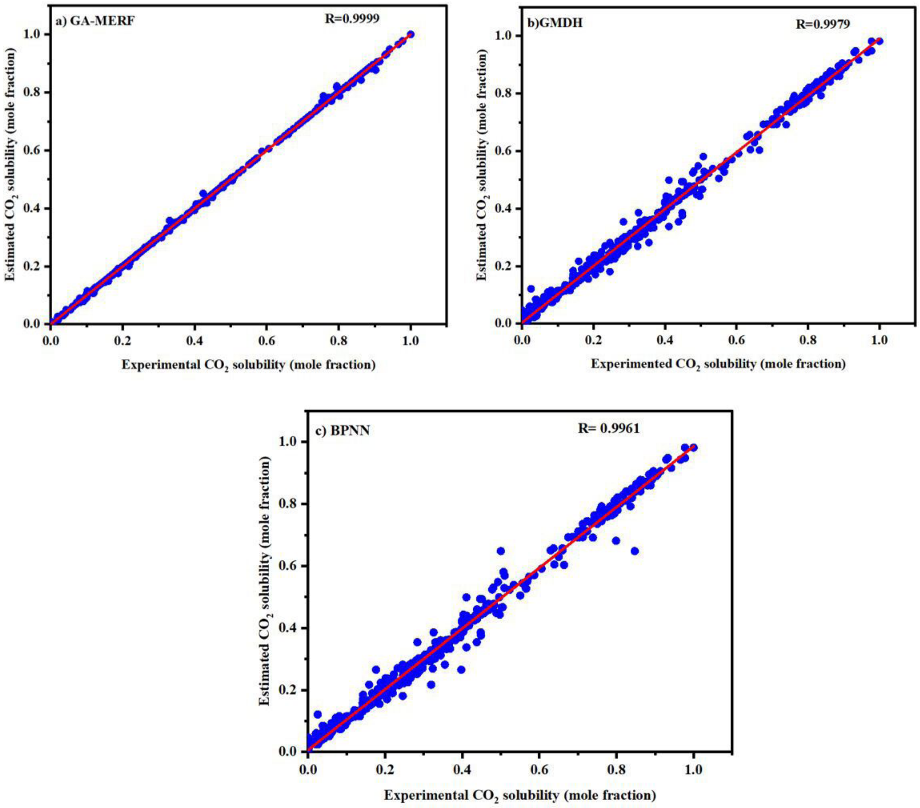

Fig. 7—Crossplots for the used models during training.

Models' Comparisons. In this paper, we use three distinct models for CO_2 solubility estimation, as detailed in the methodology section. To assess their effectiveness and robustness, we compared the models based on their predictive accuracy. This involved utilizing crossplots (also known as scatter plots) to analyze the relationship between actual and estimated CO_2 solubility values. In a crossplot, the closeness of points to the line $y = x$ reveals a model's predictive accuracy. Tightly clustered points along this line indicate strong agreement between predicted and actual values, reflected in a high correlation coefficient (approaching unity) and signifying good model performance. As points stray from the line, the model's accuracy weakens, hinting at higher errors and lower-quality predictions. Ideally, perfect predictions would yield all data points on the $y = x$ line, resulting in a flawless $R = 1$ correlation. This simple visual analysis in a crossplot provides a powerful tool for quickly assessing a model's capability to match actual data. The crossplots for GA-MERF, GMDH, and BPNN during training are presented in **Fig. 7a, 7b, and 7c,** respectively. For testing, the crossplots for GA-MERF, GMDH and BPNN are shown in **Fig. 8a, 8b, and 8c,** respectively. The analysis of **Figs. 7 and 8** indicates that the CO_2 solubility values estimated by GA-MERF outperformed those estimated by GMDH and BPNN during training and testing, in which R values are close to unity compared to other models, followed by GMDH and BPNN. GA-MERF surpasses both GMDH and BPNN in its flexibility and optimization power. Compared to GMDH with limited pairwise connections, GA-MERF thrives on complex data due to its RF structure. By combining multiple decision trees, each focusing on different aspects of the data, GA-MERF can capture the intricate relationships between factors like pressure, temperature, and salinity that influence CO_2 solubility in brine. Additionally, averaging predictions from these trees helps prevent overfitting, a common issue with complex data, leading to a more robust model for this specific task. Its GA optimization escapes local traps, finding superior model structures unlike BPNN gradient descent, making it more robust for diverse problems. This synergy of adaptable architecture and powerful optimization unlocks superior accuracy, particularly in complex, nonlinear domains, where GMDH struggles and BPNN can get stuck in suboptimal solutions.

Moreover, Taylor's diagram assessed the model's performances in fracture permeability prediction. A Taylor diagram is a valuable tool in meteorology and other fields to visually compare the performance of different models or data sets against a reference one. It allows you to assess three critical aspects of this comparison: (1) Correlation: How well does the model reproduce the overall data pattern of the overall pattern? Points lying closer to the reference (actual) data indicate a higher correlation. (2) RMSE: How much does the model differ from the reference data regarding magnitude? The distance between the points on the diagram represents this. Points closer to the center indicate lower error. (3) Standard deviation: How well does the model capture the variability of the reference data? This is represented by the length of the radial line from the origin to the point representing the model. Points on the same circle as the reference indicate a similar standard deviation (Taylor 2001; Xu and Han 2020). **Fig. 9** reveals that all ML models predicted CO_2 solubility in brine successfully. However, the GA-MERF model surpassed GMDH and BPNN in the accuracy estimation of CO_2 solubility in brine. Its standard deviation and R values demonstrate the closest alignment with reference data, showcasing its superior performance. Notably, the performance rank of the utilized models is GA-MERF > GMDH > BPNN, which matches the other results sections in this paper.

Moreover, a spider (radar) plot was used to compare ML model's performance on CO_2 solubility in brine. Based on **Fig. 10**, it appears that the GA-MERF performed better overall than GMDH and BPNN. This is because it has the lowest error values, particularly MAE and RMSE, during both training and testing. Also, GA-MERF has the highest R in both training and testing compared to other models. This is because GA-MERF is a more flexible model compared to GMDH and BPNN. This flexibility allows it to capture the complex nonlinear relationships between the input parameters and CO_2 solubility in brine. However, it is important to note that the differences between the models are relatively small for all models, especially during training, because the model usually memorizes specific patterns in the

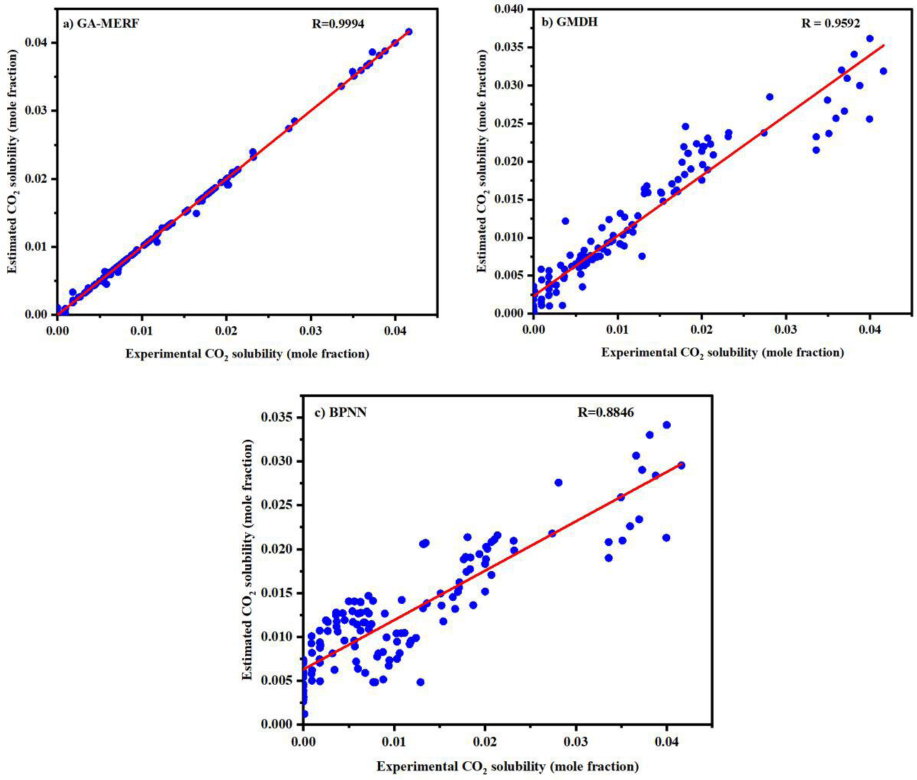

Fig. 8—Crossplots for the used models during testing.

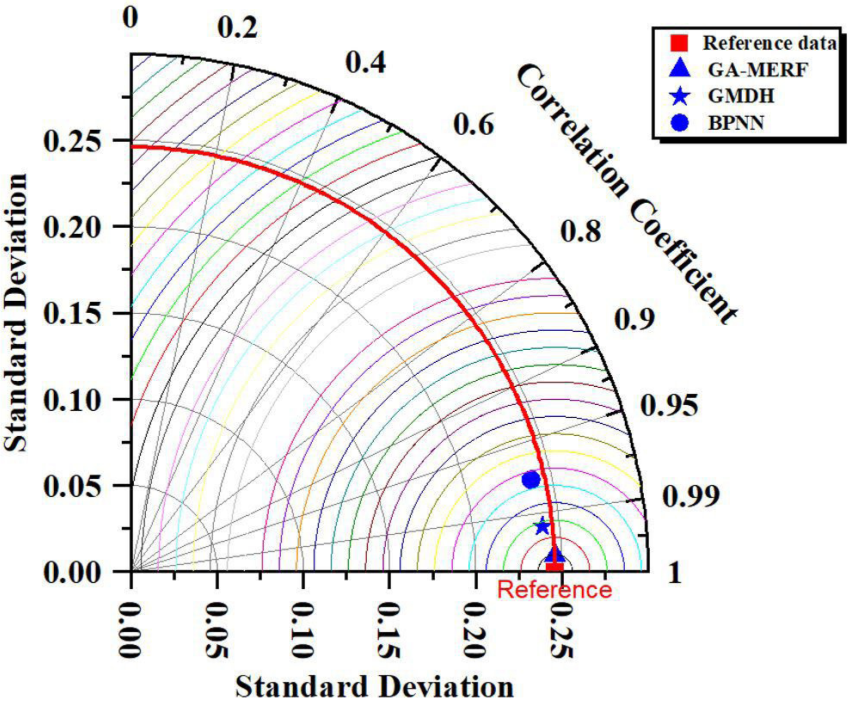

Fig. 9—Model's comparison using Taylor's diagram.

training data rather than learning the underlying relationships before being used in unseen data. On the other hand, GMDH performed better than BPNN with the highest *R*, and lower MAE and RMSE. This result confirms the superiority and applicability of GA-MERF in CO_2 solubility in brine estimation. The ranks in model's performance are GA-GMDH > GMDH > BPNN, which agrees with other results in this paper.

In addition, in this paper we compare the performance of a traditional thermodynamic model with a GA-MERF, which is the best ML model compared to others used. The thermodynamic model was developed by Sadeghi et al. (2015) as the combination of the

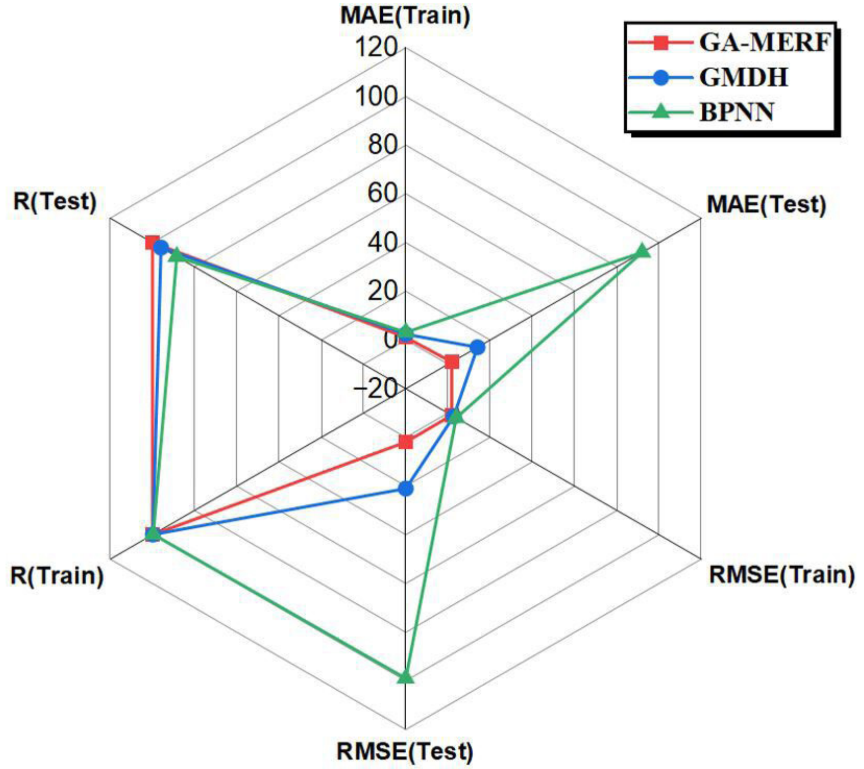

Fig. 10—Radar plot in models comparisons.

Redlich-Kwong equation of state and Pitzer expansion. The combination of Redlich-Kwong and Pitzer expansion generally offers a good balance between accuracy, versatility, and potentially reduced complexity compared to individual traditional models. We found that GA-MERF achieved lower errors (RMSE and MAE) compared to the thermodynamic model in estimating CO_2 solubility in brines, as summarized in **Table 3**. Unlike other ML models, GA-MERF incorporates a GA for optimization, potentially leading to superior performance. This enhanced capability of GA-MERF led to lower errors (RMSE and MAE) compared to the traditional thermodynamic model. This comparison approves the potential of GA-MERF for improved CO_2 solubility estimation in brines.

SHAP Analysis. SHAP analysis is a technique used in ML to explain the inner workings of a model, particularly how different features influence the final prediction (Parsa et al. 2020). It is not directly concerned with evaluating overall model performance but rather with providing insights into how each input contributes to the model's decisions. SHAP assigns a value to each feature, indicating how much it influenced the prediction. Positive values mean the feature pushed the prediction in a certain direction, while negative values indicate the opposite effect. The magnitude of the value reflects how strong the influence is. Unlike some methods that provide a single importance score for each feature, SHAP explains feature influence for each prediction. This is helpful because a feature's impact can vary depending on the values of other features (Abdulalim Alabdullah et al. 2022; Cakiroglu et al. 2024). **Fig. 11** shows that the increase in salinity results in a decrease in the model output. This indicates that the model obeys salting-out effects where salts in brine (like NaCl) tend to crowd out CO_2 molecules, making it harder for them to dissolve. So, as salinity increases, CO_2 solubility decreases, which implies that aquifers with high salinity are not effective for CO_2 sequestration. Also, the increase in temperature results in the decrease of the model output because increased temperature increases the movement of both gas and liquid molecules, making it less favorable for the gas to stay dissolved. So, as temperature increases, CO_2 solubility decreases. Because as depth increases, the temperatures increase; this implies that a deep saline aquifer is not as efficient as a medium or shallow aquifer in CO_2 sequestration. Further, the increase in pressure results in an increase in the model output because as pressure increases, CO_2 molecules get squeezed together and become more likely to dissolve in the brine. This happens because it is harder for them to escape the liquid, and the higher pressure also makes the interactions between CO_2 and water molecules more favorable for dissolving, hence more effective CO_2 sequestration. Also, **Fig. 12** shows that salinity has a great influence on the model output, followed by pressure, while temperature has a lower influence on the model output compared to others.

Models' Validation. In this study, we compared three different models (GA-MERF, GMDH, and BPNN) for their ability to predict CO_2 solubility in brine. It was found that GA-MERF performed significantly better than the other two models. To confirm this finding, the GA-MERF model was tested on a completely new data set from the literature (Liu et al. 2011; Ratnakar et al. 2020). This new data contained the same key factors (pressure, temperature, and salinity) that influence CO_2 solubility, but the actual CO_2 solubility values were assumed missing. Essentially, the model was given only the input information (pressure, temperature, and salinity) and tasked with predicting the missing output (CO_2 solubility). As illustrated in **Fig. 13**, GA-MERF estimated CO_2 solubility values with exceptional accuracy and minimal errors. The accuracy was measured to be 99.08% with a relative error of 1.12%, which strongly suggests that GA-MERF can be reliably used to estimate CO_2 solubility in new data, even when the actual CO_2 solubility values are not available. This makes GA-MERF a valuable tool for researchers and engineers working with CO_2 sequestration in brine formations.

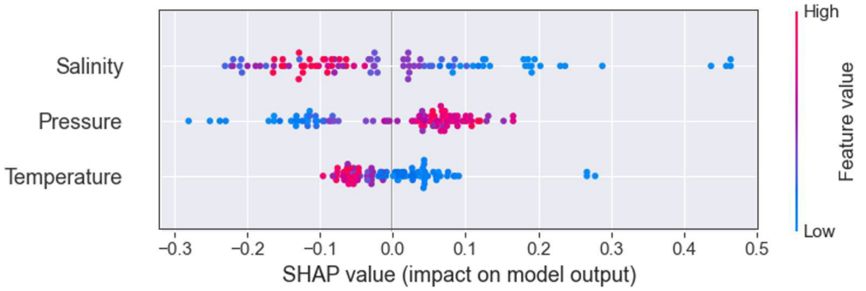

Fig. 11—Effects of each input to the model (GA-MERF) output.

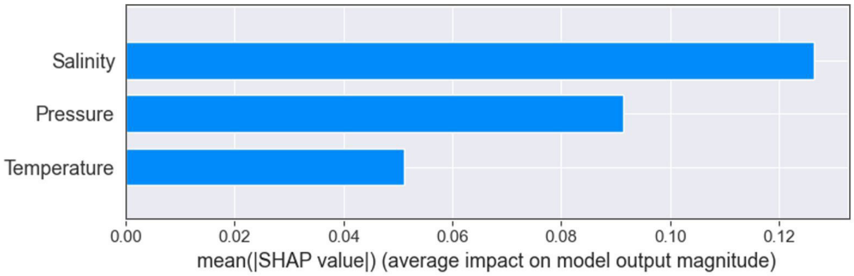

Fig. 12—Influence of the inputs to the model (GA-MERF) output.

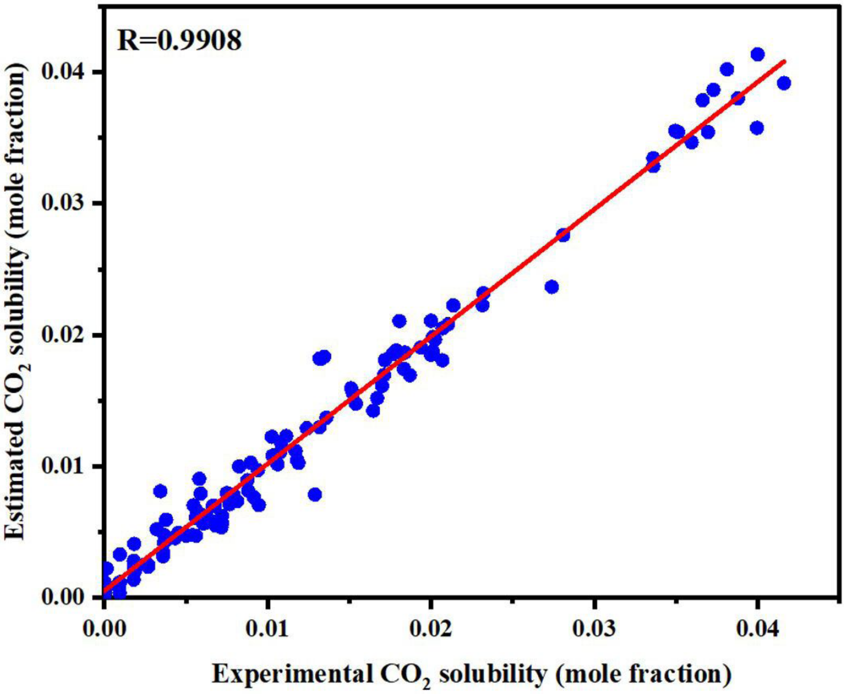

Fig. 13—GA-MERF model validation.

Comparisons of GA-MERF with Previous ML Models. Table 4 compares the performances of different models [GA-MERF, RBFNN-ABC, least-squares support vector machine (LSSVM), and fusion model] in predicting CO_2 solubility in brine. The metrics used for comparison are *R*, RMSE, and MAE for both training and testing data sets. The GA-MERF model outperforms the other models (RBFNN-ABC, LSSVM, and fusion model) in predicting CO_2 solubility in brine, as evidenced by its superior metrics—the highest *R* of 0.9999 in training and 0.9994 in testing and the lowest error values (RMSE and MAE), as shown in **Table 4**. While the fusion model also shows high performance with *R* values of 0.9998 in training and testing, GA-MERF significantly excels with notably lower RMSE and MAE values. The RBFNN-ABC and LSSVM models, though performing well, lag in both error metrics and *R* values, making GA-MERF the most accurate and reliable model for this task. The reason behind GA-MERF outperforming other models is because of its ability to leverage the optimization capabilities of GAs and the robust predictive power of MERFs. The GA optimizes feature selection and parameter tuning, enhancing the model's accuracy. MERF handles complex nonlinear relationships and accounts for both fixed and random effects, capturing underlying patterns and variability in the data. This combination results in a model with superior error minimization and consistent performance across training and testing data sets, making GA-MERF the best model for predicting CO_2 solubility in brine. The overall rank model performance is GA-MERF > fusion model > LSSVM > RBFNN-ABC.

	Training			Testing			Data Sets	
Model	*R*	RMSE	MAE	*R*	RMSE	MAE		References
GA-MERF	0.9999	1.8×10^{-8}	1.1×10^{-11}	0.9994	2×10^{-8}	1.8×10^{-11}	1,000	This study
RBFNN-ABC	0.9992	0.0218	–	0.9948	0.0572	–	–	Menad et al. (2019)
LSSVM	0.9997	0.0184	–	0.9952	0.0170	–	570	Ali Ahmadi (2016)
Fusion model	0.9998	0.0271	0.0169	0.9998	0.0272	0.0172	2,784	Wei et al. (2024)

Table 4—Comparison between GA-MERF with previous ML studies.

Conclusion and Recommendations

This study investigated the efficacy of a novel ML approach, the GA-MERF, for predicting CO_2 solubility in brine. Accurate prediction of CO_2 solubility is crucial for the development and implementation of CCS strategies, particularly those utilizing saline aquifers. The following points have been noted in this study:

1. GA-MERF demonstrated superior performance to GMDH and BPNN in estimating CO_2 solubility in brines by achieving high *R* values and minimum errors in the training and testing phases. The *R* values for GA-MERF were 0.9999 and 0.9994 throughout the training and testing stages. The RMSE and MAE values were 1.8×10^{-8} and 1.1×10^{-11} during the training phase. The RMSE and MAE values were 2×10^{-8} and 1.8×10^{-11} during the testing stage. The model's performance ranking was GA-MERF > GMDH > BPNN. Based on these results, GA-MERF emerges as a promising alternative method for CO_2 solubility estimation tasks. An additional advantage of GA-MERF lies in its computational efficiency. The model demonstrably requires less computational time of 65 seconds compared to the alternative models evaluated. This characteristic is particularly advantageous for large-scale CCS applications, where efficient data processing and analysis are paramount.
2. From SHAP analysis, it has been discovered that salinity has a great influence on the model output, followed by pressure. In contrast, temperature has the lowest impact on the model output compared to other output. Also, the increase in salinity and temperature results in a decrease in the model output. This indicates that the model obeys salting-out effects where salts in brine (like NaCl) tend to crowd out CO_2 molecules, making it harder for them to dissolve. So, as salinity increases, CO_2 solubility decreases, which implies that aquifers with high salinity are not effective for CO_2 sequestration. Further, because as depth increases, the temperatures increase, this means that deep saline aquifer is not as efficient as medium or shallow aquifer in CO_2 sequestration. In contrast, the increase in pressure results in an increase in the model output because as pressure increases, CO_2 molecules get squeezed together and become more likely to dissolve in the brine. This happens because it is harder for them to escape the liquid, and the higher pressure also makes the interactions between CO_2 and water molecules more favorable for dissolving, hence more effective CO_2 sequestration.
3. Upon applying the proposed GA-MERF model to estimate CO_2 solubility for new experimental data, which assumed that there was no CO_2 solubility data to validate the model, it was discovered that the model estimated CO_2 solubility data with 99.08% accuracy in new experimental data, which is important for CO_2 sequestration in aquifer.

In conclusion, this study presents compelling evidence that GA-MERF offers a powerful and efficient approach for predicting CO_2 solubility in brine. Its exceptional accuracy, minimal error rates, and computational efficiency make it a valuable tool for researchers and engineers working on CCS projects utilizing saline aquifers. Further investigations could explore the generalizability of GA-MERF to other CCS scenarios and its potential for real-time CO_2 sequestration monitoring and optimization.

Acknowledgments

The authors acknowledge the support from the Chinese Scholarship Council (2022GXZ005733).

References

Abdulalim Alabdullah, A., Iqbal, M., Zahid, M. et al. 2022. Prediction of Rapid Chloride Penetration Resistance of Metakaolin Based High Strength Concrete Using Light GBM and XGBoost Models by Incorporating SHAP Analysis. *Constr Build Mater* **345**: 128296. https://doi.org/10.1016/j.conbuildmat.2022.128296.

Al-Bukhaiti, K., Yanhui, L., Shichun, Z. et al. 2024. Based on BP Neural Network: Prediction of Interface Bond Strength between CFRP Layers and Reinforced Concrete. *Pract Period Struct Des Constr* **29** (2): 04023067. https://doi.org/10.1061/PPSCFX.SCENG-1421.

Ali Ahmadi, M. 2016. Applying a Sophisticated Approach to Predict CO2 Solubility in Brines: Application to CO2 Sequestration. *Int J Low-Carbon Technol* **11** (3): 325–332.

Arshad, S., Kazmi, J. H., Javed, M. G. et al. 2023. Applicability of Machine Learning Techniques in Predicting Wheat Yield Based on Remote Sensing and Climate Data in Pakistan, South Asia. *Eur J Agron* **147**: 126837. https://doi.org/10.1016/j.eja.2023.126837.

Bahadori, A., Vuthaluru, H. B., and Mokhatab, S. 2009. New Correlations Predict Aqueous Solubility and Density of Carbon Dioxide. *Int J Greenh Gas Control* **3** (4): 474–480. https://doi.org/10.1016/j.ijggc.2009.01.003.

Bando, S., Takemura, F., Nishio, M. et al. 2003. Solubility of CO2 in Aqueous Solutions of NaCl at (30 to 60) C and (10 to 20) MPa. *J Chem Eng Data* **48** (3): 576–579. https://doi.org/10.1021/je0255832.

Bueno, E. I., Pereira, I. M., and Teixeira e Silva, A. 2011. GMDH and Neural Networks Applied in Monitoring and Fault Detection in Sensors in Nuclear Power Plants. Paper presented at the International Nuclear Atlantic Conference—INAC 2011, Belo Horizonte, MG, Brazil, 24–28 October.

Buscema, M. 1998. Back Propagation Neural Networks. *Substance Use & Misuse* **33** (2): 233–270. https://doi.org/10.3109/10826089809115863.

Cakiroglu, C., Demir, S., Hakan Ozdemir, M. et al. 2024. Data-Driven Interpretable Ensemble Learning Methods for the Prediction of Wind Turbine Power Incorporating SHAP Analysis. *Expert Syst Appl* **237**. https://doi.org/10.1016/j.eswa.2023.121464.

Chai, T. and Draxler, R. R. 2014. Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? – Arguments against Avoiding RMSE in the Literature. *Geosci Model Dev* **7** (3): 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014.

Costa de Souza, A., Emmanuel da Silva Calixto, E., Pellegrini Pessoa, F. L. et al. 2024. Modelling a CO2 Meter for a Petroleum Multiphase Mixture at Subsea Conditions. *Flow Meas Instrum* **95**: 102489. https://doi.org/10.1016/j.flowmeasinst.2023.102489.

Cruz, J. L., Neyrolles, E., Contamine, F. et al. 2021. Experimental Study of Carbon Dioxide Solubility in Sodium Chloride and Calcium Chloride Brines at 333.15 and 453.15 K for Pressures up to 40 MPa. *J Chem Eng Data* **66** (1): 249–261. https://doi.org/10.1021/acs.jced.0c00592.

Dhanya, L. and Chitra, R. 2024. A Novel Autoencoder Based Feature Independent GA Optimised XGBoost Classifier for IoMT Malware Detection. *Expert Syst Appl* **237**. https://doi.org/10.1016/j.eswa.2023.121618.

Dongare, Y., Shende, A., Dhumane, A. et al. 2024. Enhanced Rainfall Prediction with Weighted Linear Units Using Advanced Recurrent Neural Network. *Int J Intell Syst Appl Eng* **12** (1s): 549–556.

Duan, Z., Sun, R., Zhu, C. et al. 2006. An Improved Model for the Calculation of CO2 Solubility in Aqueous Solutions Containing Na+, K+, Ca2+, Mg2+, Cl−, and SO42−. *Mar Chem* **98** (2–4): 131–139. https://doi.org/10.1016/j.marchem.2005.09.001.

Duan, H., Yin, X., Kou, H. et al. 2023. Regression Prediction of Hydrogen Enriched Compressed Natural Gas (HCNG) Engine Performance Based on Improved Particle Swarm Optimization Back Propagation Neural Network Method (IMPSO-BPNN). *Fuel* **331**: 125872. https://doi.org/10.1016/j.fuel.2022.125872.

Elkatatny, S. and Mahmoud, M. 2018. Development of New Correlations for the Oil Formation Volume Factor in Oil Reservoirs Using Artificial Intelligent White Box Technique. *Petrol* **4** (2): 178–186. https://doi.org/10.1016/j.petlm.2017.09.009.

Elyan, E. and Gaber, M. M. 2017. A Genetic Algorithm Approach to Optimising Random Forests Applied to Class Engineered Data. *Inf Sci (Ny)* **384**: 220–234. https://doi.org/10.1016/j.ins.2016.08.007.

Hajjem, A., Bellavance, F., and Larocque, D. 2011. Mixed Effects Regression Trees for Clustered Data. *Stat Probab Lett* **81** (4): 451–459. https://doi.org/10.1016/j.spl.2010.12.003.

Hajjem, A., Bellavance, F., and Larocque, D. 2014. Mixed-Effects Random Forest for Clustered Data. *J Stat Comput Simul* **84** (6): 1313–1328. https://doi.org/10.1080/00949655.2012.741599.

Hajjem, A., Larocque, D., and Bellavance, F. 2017. Generalized Mixed Effects Regression Trees. *Stat Probab Lett* **126**: 114–118. https://doi.org/10.1016/j.spl.2017.02.033.

Hecht-Nielsen, R. 1992. III.3 - Theory of the Backpropagation Neural Network. In *Neural Networks for Perception: Computation, Learning, and Architectures*, 65–93. San Diego, California, USA: Academic Press, Inc. https://doi.org/10.1016/B978-0-12-741252-8.50010-8.

Hemmati-Sarapardeh, A., Hajirezaie, S., Soltanian, M. R. et al. 2020. Modeling Natural Gas Compressibility Factor Using a Hybrid Group Method of Data Handling. *Eng Appl Comput Fluid Mech* **14** (1): 27–37. https://doi.org/10.1080/19942060.2019.1679668.

Hiraga, Y. and Ushiki., I. 2024. Prediction of Ionic Liquid Solubilities in Supercritical CO2 + Co-Solvent Systems Using Peng–Robinson Equation of State with Accurate Critical Temperature. *J Mol Liq* **398**. https://doi.org/10.1016/j.molliq.2024.124324.

Holland, J. H. 1973. Genetic Algorithms and the Optimal Allocation of Trials. *SIAM J Comput* **2** (2): 88–105. https://doi.org/10.1137/0202009.

Holland, J. H. 1992. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. Cambridge, Massachusetts, USA: The MIT press.

Ikiensikimama, S. S. and Azubuike, I. I. 2012. Modeling Approach for Niger-Delta Oil Formation Volume Factor Prediction Using Artificial Neural Network. Paper presented at the Nigeria Annual International Conference and Exhibition, Lagos, Nigeria, 6−8 August. https://doi.org/10.2118/162987-MS.

Ivakhnenko, A. G. 1971. Polynomial Theory of Complex Systems. *IEEE Trans Syst, Man, Cybern* **1** (4): 364–378. https://doi.org/10.1109/TSMC.1971.4308320.

Ji, Z., Wang, H., Wang, M. et al. 2024. Experimental and Modeling Study of CO2 Solubility in Formation Brines at In-Situ Conditions. *J Clean Prod* **438**. https://doi.org/10.1016/j.jclepro.2024.140840.

Katoch, S., Chauhan, S. S., and Kumar, V. 2021. A Review on Genetic Algorithm: Past, Present, and Future. *Multimed Tools Appl* **80** (5): 8091–8126. https://doi.org/10.1007/s11042-020-10139-6.

Katreddi, S., Thiruvengadam, A., Thompson, G. J. et al. 2023. Mixed Effects Random Forest Model for Maintenance Cost Estimation in Heavy-Duty Vehicles Using Diesel and Alternative Fuels. *IEEE Access* **11**: 67168–67179. https://doi.org/10.1109/ACCESS.2023.3290994.

Kiepe, J., Horstmann, S., Fischer, K. et al. 2002. Experimental Determination and Prediction of Gas Solubility Data for $CO_2 + H_2O$ Mixtures Containing NaCl or KCl at Temperatures between 313 and 393 K and Pressures up to 10 MPa . *Ind Eng Chem Res* **41** (17): 4393–4398. https://doi.org/10.1021/ie020154i.

Koschel, D., Coxam, J.-Y., Rodier, L. et al. 2006. Enthalpy and Solubility Data of CO2 in Water and NaCl(Aq) at Conditions of Interest for Geological Sequestration. *Fluid Phase Equilibria* **247** (1–2): 107–120. https://doi.org/10.1016/j.fluid.2006.06.006.

Krennmair, P. and Schmid, T. 2022. Flexible Domain Prediction Using Mixed Effects Random Forests. *J R Stat Soc Ser C Appl Stat* **71** (5): 1865–1894. https://doi.org/10.1111/rssc.12600.

Li, Y. K. and Nghiem, L. X. 1986. Phase Equilibria of Oil, Gas and Water/Brine Mixtures from a Cubic Equation of State and Henry's Law. *Can J Chem Eng* **64** (3): 486–496. https://doi.org/10.1002/cjce.5450640319.

Li, J., Topphoff, M., Fischer, K. et al. 2001. Prediction of Gas Solubilities in Aqueous Electrolyte Systems Using the Predictive Soave−Redlich−Kwong Model. *Ind Eng Chem Res* **40** (16): 3703–3710. https://doi.org/10.1021/ie0100535.

Liu, Y., Hou, M., Yang, G. et al. 2011. Solubility of CO2 in Aqueous Solutions of NaCl, KCl, CaCl2 and Their Mixed Salts at Different Temperatures and Pressures. *J Supercrit Fluids* **56** (2): 125–129. https://doi.org/10.1016/j.supflu.2010.12.003.

Liu, Y., Jiang, C., Lu, C. et al. 2023. Increasing the Accuracy of Soil Nutrient Prediction by Improving Genetic Algorithm Backpropagation Neural Networks. *Symmetry* **15** (1): 151. https://doi.org/10.3390/sym15010151.

Liu, P. and Wu, J. 2024. Study on the Diffusion of CCUS Technology under Carbon Trading Mechanism: Based on the Perspective of Tripartite Evolutionary Game among Thermal Power Enterprises, Government and Public. *J Clean Prod* **438**. https://doi.org/10.1016/j.jclepro.2024.140730.

Lu, T., Li, Z., and Du, L. 2023. Enhanced CO2 Geological Sequestration Using Silica Aerogel Nanofluid: Experimental and Molecular Dynamics Insights. *Chem Eng J* **474**: 145566. https://doi.org/10.1016/j.cej.2023.145566.

Luo, A., Li, Y., Chen, X. et al. 2022. Review of CO2 Sequestration Mechanism in Saline Aquifers. *Nat Gas Ind B* **9** (4): 383–393. https://doi.org/10.1016/j.ngib.2022.07.002.

Lv, Q., Zhou, T., Zheng, R. et al. 2023. Application of Group Method of Data Handling and Gene Expression Programming for Predicting Solubility of CO2-N2 Gas Mixture in Brine. *Fuel* **332**: 126025. https://doi.org/10.1016/j.fuel.2022.126025.

Majid, A., Mwakipunda, G. C., and Guo, C. 2023. Solution Gas/Oil Ratio Prediction from Pressure/Volume/Temperature Data Using Machine Learning Algorithms. *SPE J.* **29** (2): 1–16. https://doi.org/10.2118/217979-PA.

Mao, S., Zhang, D., Li, Y. et al. 2013. An Improved Model for Calculating CO2 Solubility in Aqueous NaCl Solutions and the Application to CO2–H2O–NaCl Fluid Inclusions. *Chem Geol* **347**: 43–58. https://doi.org/10.1016/j.chemgeo.2013.03.010.

Markham, A. E. and Kobe, K. A. 1941. The Solubility of Carbon Dioxide and Nitrous Oxide in Aqueous Salt Solutions. *J Am Chem Soc* **63** (2): 449–454. https://doi.org/10.1021/ja01847a027.

Mayapada, R., Susetyo, B., and Sartono, B. 2021. A Comparison between Random Forest and Mixed Effects Random Forest to Predict Students' Math Performance in Indonesia. *Int J Sci Basic Appl Res* **57**: 1–8.

Mehdizade, N., Bonyadi, M., Darvishi, P. et al. 2024. Modeling H2S Solubility in Aqueous MDEA, MEA and DEA Solutions by the Electrolyte SRK-CPA EOS. *J Mol Liq* **400**: 124441. https://doi.org/10.1016/j.molliq.2024.124441.

Menad, N. A., Hemmati-Sarapardeh, A., Varamesh, A. et al. 2019. Predicting Solubility of CO2 in Brine by Advanced Machine Learning Systems: Application to Carbon Capture and Sequestration. *J CO2 Util* **33**: 83–95. https://doi.org/10.1016/j.jcou.2019.05.009.

Mgimba, M. M., Jiang, S., Nyakilla, E. E. et al. 2023. Application of GMDH to Predict Pore Pressure from Well Logs Data: A Case Study from Southeast Sichuan Basin, China. *Nat Resour Res* **32** (4): 1711–1731. https://doi.org/10.1007/s11053-023-10207-2.

Mkono, C. N., Chuanbo, S., Mulashani, A. K. et al. 2023. Deep Learning Integrated Approach for Hydrocarbon Source Rock Evaluation and Geochemical Indicators Prediction in the Jurassic - Paleogene of the Mandawa Basin, SE Tanzania. *Energy* **284**: 129232. https://doi.org/10.1016/j.energy.2023.129232.

Mohammadian, E., Hadavimoghaddam, F., Kheirollahi, M. et al. 2023. Probing Solubility and pH of CO2 in Aqueous Solutions: Implications for CO2 Injection into Oceans. *J CO2 Util* **71**: 102463. https://doi.org/10.1016/j.jcou.2023.102463.

Mohammadian, E., Hamidi, H., Asadullah, M. et al. 2015. Measurement of CO_2 Solubility in NaCl Brine Solutions at Different Temperatures and Pressures Using the Potentiometric Titration Method. *J Chem Eng Data* **60** (7): 2042–2049. https://doi.org/10.1021/je501172d.

Mosavat, N., Abedini, A., and Torabi, F. 2014. Phase Behaviour of CO2–Brine and CO2–Oil Systems for CO2 Storage and Enhanced Oil Recovery: Experimental Studies. *Energy Procedia* **63**: 5631–5645. https://doi.org/10.1016/j.egypro.2014.11.596.

Mulashani, A. K., Shen, C., Nkurlu, B. M. et al. 2022. Enhanced Group Method of Data Handling (GMDH) for Permeability Prediction Based on the Modified Levenberg Marquardt Technique from Well Log Data. *Energy* **239**: 121915. https://doi.org/10.1016/j.energy.2021.121915.

Mutailipu, M., Song, Y., Yao, Q. et al. 2024. Solubility and Interfacial Tension Models for CO2–Brine Systems under CO2 Geological Storage Conditions. *Fuel* **357**: 129712. https://doi.org/10.1016/j.fuel.2023.129712.

Mwakipunda, G. C., Abelly, E. N., Mgimba, M. M. et al. 2023a. Critical Review on Carbon Dioxide Sequestration Potentiality in Methane Hydrate Reservoirs via CO_2 –CH_4 Exchange: Experiments, Simulations, and Pilot Test Applications . *Energy Fuels* **37** (15): 10843–10868. https://doi.org/10.1021/acs.energyfuels.3c01510.

Mwakipunda, G. C., Mgimba, M. M., Ngata, M. R. et al. 2024. Recent Advances on Carbon Dioxide Sequestration Potentiality in Salt Caverns: A Review. *Int J Greenh Gas Control* **133**: 104109. https://doi.org/10.1016/j.ijggc.2024.104109.

Mwakipunda, G. C., Ngata, M. R., Mgimba, M. M. et al. 2023b. Carbon Dioxide Sequestration in Low Porosity and Permeability Deep Saline Aquifer: Numerical Simulation Method. *J Energy Resour Technol* **145** (7): 073401. https://doi.org/10.1115/1.4056612.

Mwakipunda, G. C., Wang, Y., Mgimba, M. M. et al. 2023c. Recent Advances in Carbon Dioxide Sequestration in Deep Unmineable Coal Seams Using CO_2 -ECBM Technology: Experimental Studies, Simulation, and Field Applications . *Energy Fuels* **37** (22): 17161–17186. https://doi.org/10.1021/acs.energyfuels.3c03004.

Nadege, M. N., Jiang, S., Mwakipunda, G. C. et al. 2024. Brittleness Index Prediction Using Modified Random Forest Based on Particle Swarm Optimization of Upper Ordovician Wufeng to Lower Silurian Longmaxi Shale Gas Reservoir in the Weiyuan Shale Gas Field, Sichuan Basin, China. *Geoenergy Sci Eng* **233**: 212518. https://doi.org/10.1016/j.geoen.2023.212518.

Naser, M. and Alavi, A. H. 2021. Error Metrics and Performance Fitness Indicators for Artificial Intelligence and Machine Learning in Engineering and Sciences. *Arch Struct Constr*: 1–19.

Nath, F., Mahmood, M. N., and Yousuf, N. 2024. Recent Advances in CCUS: A Critical Review on Technologies, Regulatory Aspects and Economics. *Geoenergy Sci Eng* **238**: 212726. https://doi.org/10.1016/j.geoen.2024.212726.

Ngata, M. R., Yang, B., Khalid, W. et al. 2023. Review on Experimental Investigation into Formation Damage during Geologic Carbon Sequestration: Advances and Outlook. *Eng Fuels* **37** (9): 6382–6400. https://doi.org/10.1021/acs.energyfuels.3c00427.

Nighswander, J. A., Kalogerakis, N., and Mehrotra, A. K. 1989. Solubilities of Carbon Dioxide in Water and 1 Wt. % Sodium Chloride Solution at Pressures up to 10 MPa and Temperatures from 80 to 200.Degree.C. *J Chem Eng Data* **34** (3): 355–360. https://doi.org/10.1021/je00057a027.

Parsa, A. B., Movahedi, A., Taghipour, H. et al. 2020. Toward Safer Highways, Application of XGBoost and SHAP for Real-Time Accident Detection and Feature Analysis. *Accid Anal Prev* **136**: 105405. https://doi.org/10.1016/j.aap.2019.105405.

Portier, S. and Rochelle, C. 2005. Modelling CO2 Solubility in Pure Water and NaCl-Type Waters from 0 to 300 C and from 1 to 300 Bar: Application to the Utsira Formation at Sleipner. *Chem Geol* **217** (3–4): 187–199. https://doi.org/10.1016/j.chemgeo.2004.12.007.

Ratnakar, R. R., Chaubey, V., and Dindoruk, B. 2023. A Novel Computational Strategy to Estimate CO2 Solubility in Brine Solutions for CCUS Applications. *Appl Energy* **342**. https://doi.org/10.1016/j.apenergy.2023.121134.

Ratnakar, R. R., Venkatraman, A., Kalra, A. et al. 2020. On the Prediction of Gas Solubility in Brine Solutions with Single or Mixed Salts: Applications to Gas Injection and CO2 Capture/Sequestration. *J Nat Gas Sci Eng* **81**: 103450. https://doi.org/10.1016/j.jngse.2020.103450.

Ray, R., Choudhary, S. S., Roy, L. B. et al. 2023. Reliability Analysis of Reinforced Soil Slope Stability Using GA-ANFIS, RFC, and GMDH Soft Computing Techniques. *Case Studies Constr Mater* **18**. https://doi.org/10.1016/j.cscm.2023.e01898.

Razavi-Termeh, S. V., Sadeghi-Niaraki, A., Seo, M. et al. 2023. Application of Genetic Algorithm in Optimization Parallel Ensemble-Based Machine Learning Algorithms to Flood Susceptibility Mapping Using Radar Satellite Imagery. *Sci Total Environ* **873**: 162285. https://doi.org/10.1016/j.scitotenv.2023.162285.

Rezaie, F., Panahi, M., Bateni, S. M. et al. 2023. Spatial Modeling of Geogenic Indoor Radon Distribution in Chungcheongnam-Do, South Korea Using Enhanced Machine Learning Algorithms. *Environ Int* **171**: 107724. https://doi.org/10.1016/j.envint.2022.107724.

Roshani, M., Sattari, M. A., Muhammad Ali, P. J. et al. 2020. Application of GMDH Neural Network Technique to Improve Measuring Precision of a Simplified Photon Attenuation Based Two-Phase Flowmeter. *Flow Meas Instrum* **75**: 101804. https://doi.org/10.1016/j.flowmeasinst.2020.101804.

Rumpf, B. and Maurer, G. 1993. An Experimental and Theoretical Investigation on the Solubility of Carbon Dioxide in Aqueous Solutions of Strong Electrolytes. *Ber Bunsenges Phys Chem* **97** (1): 85–97. https://doi.org/10.1002/bbpc.19930970116.

Rumpf, B., Nicolaisen, H., cal, C. et al. 1994. Solubility of Carbon Dioxide in Aqueous Solutions of Sodium Chloride: Experimental Results and Correlation. *J Solution Chem* **23** (3): 431–448. https://doi.org/10.1007/BF00973113.

Rutten, T. 2021. *Mixed-Effects Random Forest Model for Quantifying Relations in Clustered Data*. Graduation Thesis, Eindhoven University of Technology, Eindhoven, The Netherlands.

Sadeghi, M., Salami, H., Taghikhani, V. et al. 2015. A Comprehensive Study on CO2 Solubility in Brine: Thermodynamic-Based and Neural Network Modeling. *Fluid Phase Equilib* **403**: 153–159. https://doi.org/10.1016/j.fluid.2015.06.021.

Shadkani, S., Hashemi, S., Pak, A. et al. 2024. Random Forest and Multilayer Perceptron Hybrid Models Integrated with the Genetic Algorithm for Predicting Pan Evaporation of Target Site Using a Limited Set of Neighboring Reference Station Data. *Earth Sci Inform* **17** (2): 1261–1280. https://doi.org/10.1007/s12145-024-01237-2.

Sodeifian, G., Bagheri, H., Razmimanesh, F. et al. 2024. Supercritical CO2 Utilization for Solubility Measurement of Tramadol Hydrochloride Drug: Assessment of Cubic and Non-Cubic EoSs. *J Supercrit Fluids* **206**. https://doi.org/10.1016/j.supflu.2024.106185.

Sodeifian, G., Hsieh, C.-M., Tabibzadeh, A. et al. 2023. Solubility of Palbociclib in Supercritical Carbon Dioxide from Experimental Measurement and Peng-Robinson Equation of State. *Sci Rep* **13** (1). https://doi.org/10.1038/s41598-023-29228-1.

Sørensen, H., Pedersen, K. S., and Christensen, P. L. 2002. Modeling of Gas Solubility in Brine. *Org Geochem* **33** (6): 635–642. https://doi.org/10.1016/S0146-6380(02)00022-0.

Statista. 2024. Average Monthly Carbon Dioxide (CO_2) Levels in the Atmosphere Worldwide From 1990 to 2024. https://www.statista.com/statistics/1091999/atmospheric-concentration-of-co2-historic/.

Steel, L., Liu, Q., Mackay, E. et al. 2016. CO_2 Solubility Measurements in Brine under Reservoir Conditions: A Comparison of Experimental and Geochemical Modeling Methods . *Greenhouse Gas Sci Technol* **6** (2): 197–217. https://doi.org/10.1002/ghg.1590.

Sun, C., Fares, M. N., Sajadi, S. M. et al. 2024. Numerical Examination of Exergy Performance of a Hybrid Solar System Equipped with a Sheet-and-Sinusoidal Tube Collector: Developing a Predictive Function Using Artificial Neural Network. *Case Studies in Thermal Eng* **53**: 103828. https://doi.org/10.1016/j.csite.2023.103828.

Taylor, K. E. 2001. Summarizing Multiple Aspects of Model Performance in a Single Diagram. *J Geophys Res* **106** (D7): 7183–7192. https://doi.org/10.1029/2000JD900719.

Teng, G., Xiao, J., He, Y. et al. 2017. Use of Group Method of Data Handling for Transport Energy Demand Modeling. *Energy Sci Eng* **5** (5): 302–317. https://doi.org/10.1002/ese3.176.

Wang, J. and Ehlig-Economides, C. 2023. Salinity Effect on CO2 Solubility in Live Formation Water Under Reservoir Conditions. Paper presented at the SPWLA 64th Annual Logging Symposium, Lake Conroe, Texas, USA, 10−14 June.

Wang, L., Shen, Z., Hu, L. et al. 2014. Modeling and Measurement of CO2 Solubility in Salty Aqueous Solutions and Application in the Erdos Basin. *Fluid Phase Equilib* **377**: 45–55. https://doi.org/10.1016/j.fluid.2014.06.016.

Wei, W., Lu, P., Zhu, C. et al. 2024. Advanced Machine Learning Models for CO_2 and H_2S Solubility in Water and NaCl Brine: Implications for Geoenergy Extraction and Carbon Storage . *Energy Fuels* **38** (12): 11119–11136. https://doi.org/10.1021/acs.energyfuels.4c01423.

Willmott, C. J. and Matsuura, K. 2005. Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance. *Clim Res* **30** (1): 79–82. https://doi.org/10.3354/cr030079.

Xu, Z. and Han, Y. 2020. Short Communication Comments on ‘DISO: A Rethink of Taylor Diagram. *Intl J Climatol* **40** (4): 2506–2510. https://doi.org/10.1002/joc.6359.

Yang, S.-I., Brandeis, T. J., Helmer, E. H. et al. 2022. Characterizing Height-Diameter Relationships for Caribbean Trees Using Mixed-Effects Random Forest Algorithm. *For Ecol Manage* **524**. https://doi.org/10.1016/j.foreco.2022.120507.

Yan, W., Huang, S., and Stenby, E. H. 2011. Measurement and Modeling of CO2 Solubility in NaCl Brine and CO2–Saturated NaCl Brine Density. *Int J Greenh Gas Control* **5** (6): 1460–1477. https://doi.org/10.1016/j.ijggc.2011.08.004.

Yang, H., Jin, J., Hou, F. et al. 2021. An ANN-Based Method for Predicting Zhundong and Other Chinese Coal Slagging Potential. *Fuel* **293**: 120271. https://doi.org/10.1016/j.fuel.2021.120271.

Yao, P., Yu, Z., Zhang, Y. et al. 2023. Application of Machine Learning in Carbon Capture and Storage: An in-Depth Insight from the Perspective of Geoscience. *Fuel* **333**: 126296. https://doi.org/10.1016/j.fuel.2022.126296.

Zhang, Q., Liu, J., Wang, G. et al. 2024. A New Optimization Model for Carbon Capture Utilization and Storage (CCUS) Layout Based on High-Resolution Geological Variability. *Appl Energy* **363**. https://doi.org/10.1016/j.apenergy.2024.123065.

Zhang, B. and Xue, X. 2024. Ultimate Axial Strength Prediction of Concrete-Filled Double-Skin Steel Tube Columns Using Soft Computing Methods. *Eng Appl Artif Intell* **129**. https://doi.org/10.1016/j.engappai.2023.107676.

Zhao, H., Dilmore, R., Allen, D. E. et al. 2015a. Measurement and Modeling of CO2 Solubility in Natural and Synthetic Formation Brines for CO2 Sequestration. *Environ Sci Technol* **49** (3): 1972–1980. https://doi.org/10.1021/es505550a.

Zhao, H., Dilmore, R. M., and Lvov, S. N. 2015b. Experimental Studies and Modeling of CO2 Solubility in High Temperature Aqueous CaCl2, MgCl2, Na2SO4, and KCl Solutions. *AIChE J* **61** (7): 2286–2297. https://doi.org/10.1002/aic.14825.

Zou, X., Zhu, Y., Lv, J. et al. 2024. Toward Estimating CO_2 Solubility in Pure Water and Brine Using Cascade Forward Neural Network and Generalized Regression Neural Network: Application to CO_2 Dissolution Trapping in Saline Aquifers. *ACS Omega* **9** (4): 4705–4720. https://doi.org/10.1021/acsomega.3c07962.