Discover Applied Sciences

Research

Modeling CO_2 loading capacity of triethanolamine aqueous solutions using advanced white-box approaches: GMDH, GEP, and GP

Fahimeh Hadavimoghaddam[1,2] · Behnam Amiri-Ramsheh[3] · Saeid Atashrouz[4] · Meftah Ali Abuswer[5] · Qichao Lv[6] · Ahmad Mohaddespour[7] · Mehdi Ostadhassan[8] · Abdolhossein Hemmati-Sarapardeh[3]

Received: 23 July 2023 / Accepted: 11 December 2023
Published online: 29 January 2024

Abstract
The equilibrium solubility of carbon dioxide (CO_2) in the solvents is a key essential characteristic that has to be evaluated for successful absorption-based CO_2 capture procedures. In this study, the CO_2 loading capacity of triethanolamine (TEA) aqueous solutions was estimated using three famous white-box algorithms namely gene expression programming (GEP), genetic programming (GP), and group method of data handling (GMDH). For achieving the aim of this study, 258 data in a wide range of pressure, temperature, and amine concentration were collected from literature. Temperature, partial pressure of CO_2, and amine concentration were used as input parameters. The results demonstrated that GMDH correlation is more accurate than GEP and GP with a determination coefficient (R^2) of 0.9813 and root mean square error of 0.0222. The R^2 values of 0.9713 and 0.9664 for the GEP and GP, respectively, demonstrated that the GEP and GP also showed accurate predictions. In addition, GMDH approach accurately predicted the anticipated trends of the CO_2 loading in response to changes in the partial pressure of CO_2 and temperature. The Pearson and Spearman correlation analyses were also incorporated in this research which showed that temperature and CO_2 partial pressure have almost the same relative effect on CO_2 loading, while amine concentration has the lowest effect on it.

Keywords Triethanolamine (TEA) · Amine aqueous solution · GMDH · CO_2 loading · CO_2 Capture · CCUS · White-box algorithms

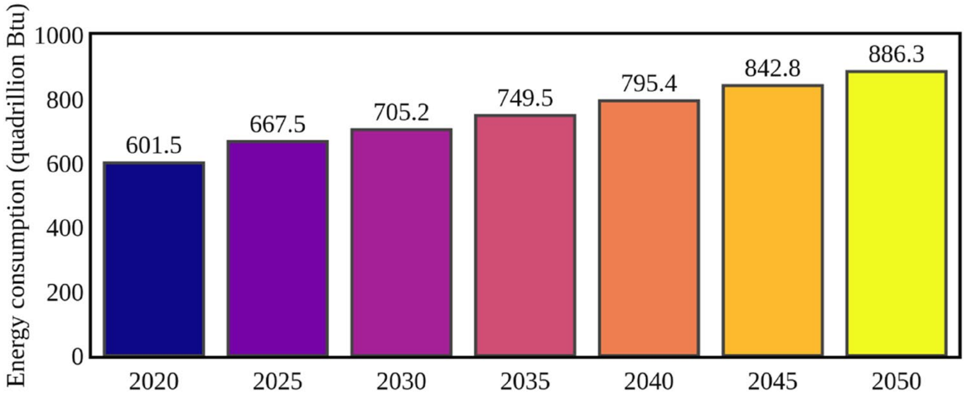
1 Introduction

The global energy consumption would significantly increase within the following decades. Figure 1 depicts the energy consumption from 2020 to 2050 indicating the surge of the indicator analyzed [1, 2].

The first group of main energy sources, that comprises renewable energy, nuclear energy, and fossil energy, all play an important part in providing the need for energy all over the globe. Petroleum, coal, and natural gas (NG) are also the three most significant types of fossil energy. According to various investigations, the primary fuel with the most rapid growth

✉ Ahmad Mohaddespour, ahmad.mohaddespour@mail.mcgill.ca; ✉ Mehdi Ostadhassan, mehdi.ostadhassan@ifg.uni-kiel.de; ✉ Abdolhossein Hemmati-Sarapardeh, hemmati@uk.ac.ir | [1]Institute of Unconventional Oil & Gas, Northeast Petroleum University, Daqing 163318, Heilongjiang, China. [2]Ufa State Petroleum Technological University, Ufa, Russia 450064. [3]Department of Petroleum Engineering, Shahid Bahonar University of Kerman, Kerman, Iran. [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, Egaila 54200, Kuwait. [6]State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum (Beijing), Beijing, China. [7]Department of Chemical Engineering, McGill University, Montreal, QC H3A 0C5, Canada. [8]Institute of Geosciences, Marine and Land Geomechanics and Geotectonics, Christian Albrechts Universität, 24118 Kiel, Germany.

Fig. 1 World energy demand prediction from 2020 to 2050



through 2040 will be natural gas [3–5]. Figure 2 depicts the overall energy usage by type of fuel. Natural gas is widely recognized as the fossil fuel that has the highest levels of safety, cleanliness, and operational effectiveness [5, 6]. This is because the carbon dioxide (CO_2) emission of natural gas is about 41% less than that of other fossil fuels when burned [7].

CO_2 is one of the resources which has been acknowledged as one of the key contributors to the phenomenon of global warming [8, 9]. Consequently, its removal and reuse from the streams and pollutants produced by industrial processes and the search for reliable and cost-effective absorbents have attracted considerable attention [10]. Absorbance in aqueous amine solutions is the method used substantially in industry for eliminating CO_2 from gases. CO_2 is absorbed by aqueous amine solutions via both physical and chemical absorption [11–13]. Figure 3 represents the process schematic of the amine-based CO_2 capture.

Triethanolamine (TEA), a tertiary amine, was recognized as one of the first amines utilized considering this purpose in industrial gas treatment procedures [14]. While it has been replaced by another type of amine solution such as methyldiethanolamine (MDEA) and monoethanolamine (MEA) [15], it is still prescribed for the elimination of acid gas. Recent developments have expanded the repertoire of solvents beyond amine-based solutions, ushering in new possibilities for carbon capture technologies. Notably, amine blends, as discussed in [16], have gained attention for their potential to enhance CO_2 absorption efficiency and reduce energy requirements. Ionic liquids explored comprehensively in literature [17], offer intriguing prospects due to their low volatility and tunable properties, which can be tailored to specific capture scenarios. The utilization of seawater as a solvent, as exemplified by [18], presents an eco-friendly and abundant alternative with unique challenges and advantages. These emerging solvent systems, along with others not mentioned here, constitute a dynamic frontier in CO_2 capture research. Hence, a wide variety of laboratory solubility data (H_2S and CO_2 in aqueous ethanolamine solutions) with a variety range of temperatures, pressures, solvent compositions, and acid gas loadings are now available. This data may be used to better understand the interaction between H_2S and CO_2 [15, 19–22]. At higher power dissipation rates, aqueous TEA solutions can absorb CO_2 under equilibrium conditions in high shear jet absorbers. Both the absorption height and the required flow rate of the solution are decreased as a result of putting solution in a high shear jet absorber. This absorber will be especially effective for the removal of acid gases with low partial pressures, as well as using this in distant fields and offshore activities [23, 24].

Fig. 2 World energy demand by sources

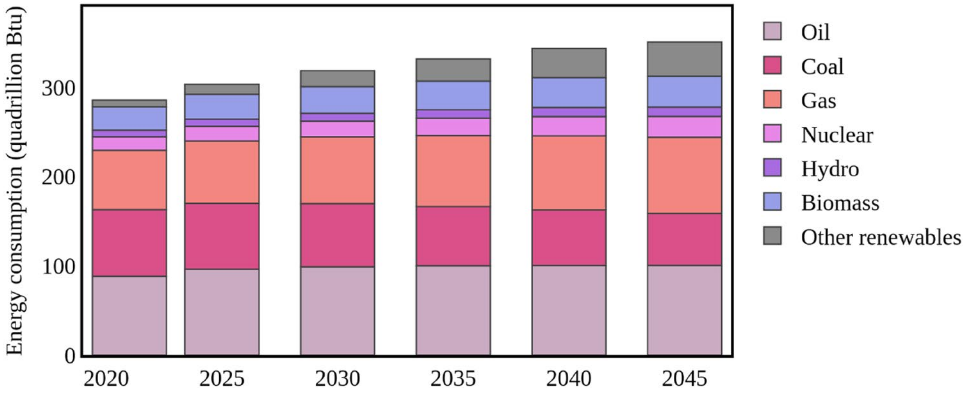
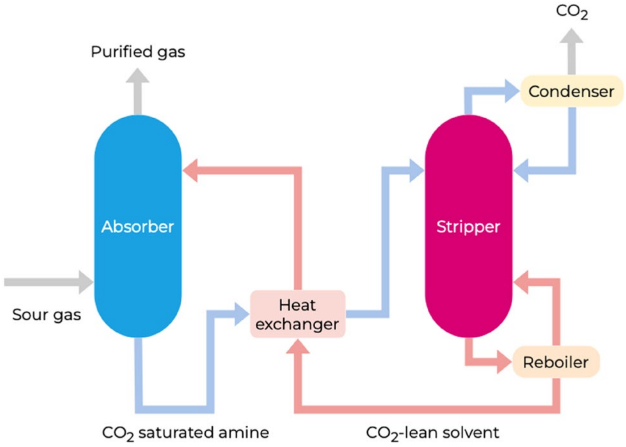


Fig. 3 The process of amine-based CO_2 capture



The solubility of CO_2 in an aqueous alkanolamine solution and the equilibrium CO_2 loading were both calculated using a number of different models [25, 26]. The CO_2 solubility in TEA was investigated by Chung [27] using a modified version of the Kent-Eisenberg model, which represents one of the most precise methods that are currently available. In their research, a modified Kent–Eisenberg model successfully matched the experimental equilibrium loading (solubility)/partial pressure pairs at different temperatures and amine concentration levels. The average absolute relative deviation (AARD) for this model was 18.9%, and it included a maximum of 163 data sets. Fouad et al. [26] also compared experimental data of TEA at 50, 75, and 100 °C. Their findings indicated that the average absolute fitting error ranges between 46.1% and 47.8%. For the purpose of offering an alternate solution method for modeling engineering processes and forecasting the variable of interest, a number of intelligent approaches have been used [13, 28–30]. Yarveicy et al. [31] designed the extra tree (ET) algorithm in order to anticipate the capacity for CO_2 loading. The developed model could predict all the data of TEA (29 data points) with an R^2 of 0.993. The AdaBoost classification and regression trees (AdaBoost-CART) was used by Ghiasi et al. [32] to simulate the CO_2 loading for MEA, Diethanolamine (DEA), and TEA. Their investigation of CO_2 solubility in TEA included 63 data points with an AARD of 1.41%. In addition, the effects of reaction temperature, CO_2 partial pressure, and the concentration of amine on the CO_2 absorption performance of MEA, DEA, and TEA were investigated by using adaptive neuro-fuzzy inference system (ANFIS) by Ghiasi et al. [28]. Unexpectedly, it was discovered that the predominant experimental condition differed for different amins. In particular, the relative effect of inputs on the CO_2 loading was temperature > CO_2 partial pressure > concentration of amines for MEA and TEA, but for DEA, the relative effect was adjusted to CO_2 partial pressure > reaction temperature > concentration of amines. This was the case because DEA reacts more slowly than MEA and TEA. This study provided useful implications on the variety of amine design, but it has the potential to produce error as a result of insufficient amount of experimental data.

To the best of the authors' knowledge, there are no published white-box correlations for CO_2 loading capacity in amine-based solutions. Our present study focuses on TEA-based systems and development of interpretable models using advanced white-box approaches, it is paramount to recognize and appreciate the growing diversity of solvent options, each contributing to the overarching goal of mitigating CO_2 emissions. Hence, the purpose of this work is to assess the capacity of robust correlations to predict the equilibrium absorption of CO_2 in TEA aqueous solutions. To this aim, experimental data of equilibrium absorption of CO_2 in TEA aqueous solutions are gathered from the published literatures [19, 27, 33]. To this end, temperature, CO_2 partial pressure, and amine concentration were regarded as input variables and CO_2 loading was the output. Three famous robust correlative algorithms, namely genetic programming (GP), gene expression programming (GEP), and group method of data handling (GMDH) are used to estimate CO_2 loading in an aqueous system containing TEA. In the following section, we will present the summary of our collected databank and the pre-processing of the dataset used. Furthermore, Sect. 3 provides a detailed explanation of the development of intelligent white-box algorithms. In Sect. 4, we represent various error analyses to evaluate the models' performance, statistically. Besides, Sect. 5 provides the equations for predicting CO_2 loading using GP, GEP, and GMDH techniques, and also gives a comprehensive graphical and statistical assessment of these models.

Table 1 Statistical description of the dataset used in this study

Parameters	T (K)	CO_2 partial pressure (kPa)	Amine concentration (mol/L)	CO_2 loading (mol CO_2/mol amine)
Skewness	−0.765	1.299	1.253	0.257
Kurtosis	0.614	2.027	−1.328	−0.856
Mean	316.3	36.45	2.6	0.32
Std	16.45	28.34	0.4	0.163
Min	298	1.43	2	0.034
Max	353.2	153.4	3	0.711

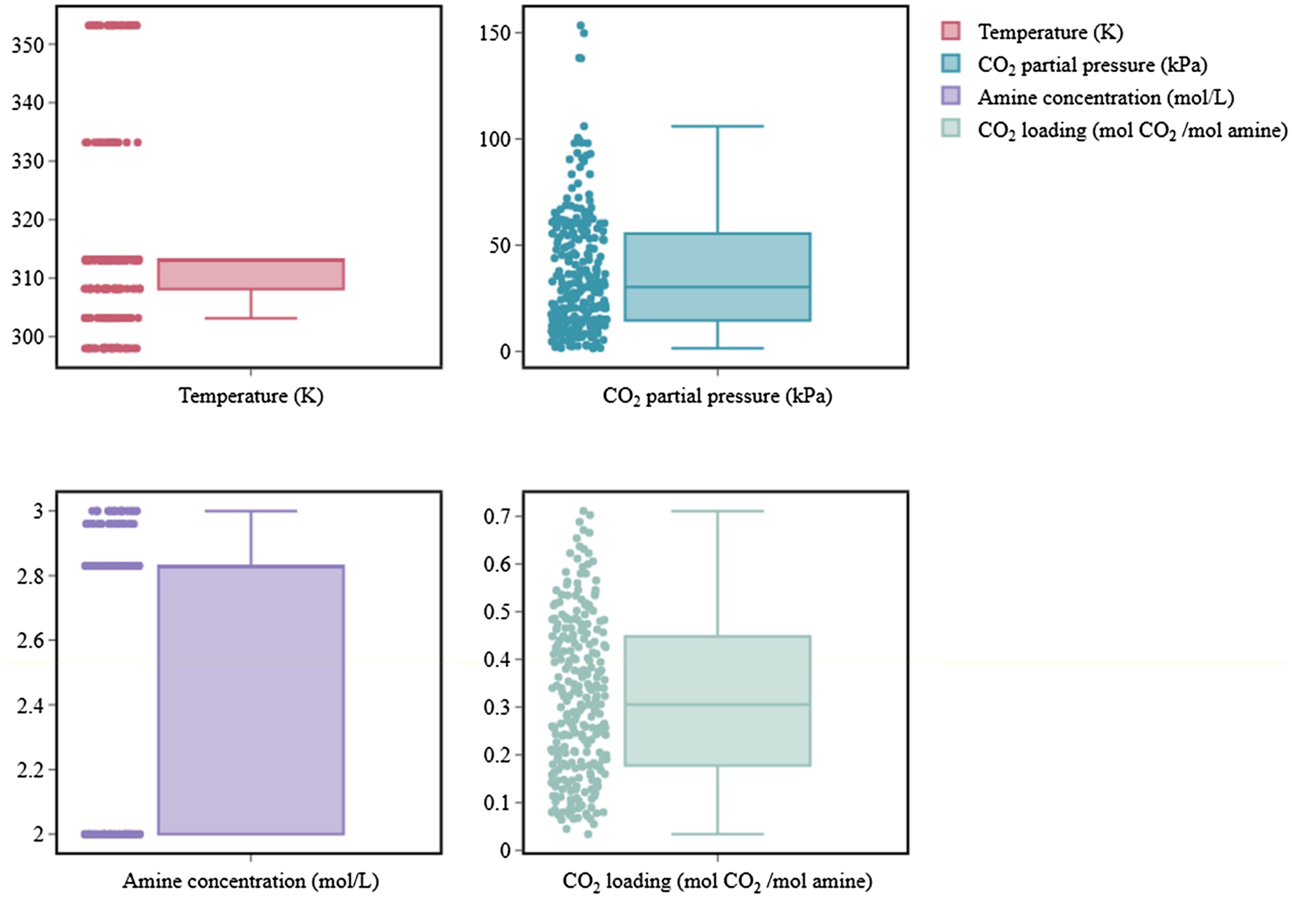


Fig. 4 Box Plot for all parameters in this study

2 Data gathering and preparation

In order to construct comprehensive correlations, a large database was assembled from literature sources. Experimental values for CO_2 absorption in TEA aqueous solutions were gathered from [19, 27, 33]. Table 1 gives specific precise information on the CO_2 partial pressure, temperature, amine concentration, and CO_2 loading capacity of TEA aqueous systems. The table illustrates the ranges of inputs/output and statistical parameters that were used throughout this investigation.

Figure 4 presents the distribution of all parameters in the form of box plots. The forecast distribution is symmetrical when it follows a predictable pattern, sometimes represented as a bell curve. The skewness value is positive when the probability function's left side contains the vast majority of the data, and vice versa. In contrast, kurtosis describes the shape of the distribution in relation to the Gaussian distribution. A positive kurtosis, for instance, demonstrates that the statistical model has a larger peak than the usual range does [34]. Table 1 and Fig. 4 declare that the distribution and fluctuation range of the input variables are sufficiently broad to support the development of a general model for the precise prediction of CO_2 loading.

Figure 5 illustrates the input data part plot. Temperature has the greatest influence on the CO_2 loading. It must be mentioned that the connection is negative, which suggests that as temperature rises, CO_2 loading decreases and vice versa. Another important parameter worth analyzing is pressure. In this case, the relationship is positive.

3 Model development

For achieving the aim of the study, three robust correlation algorithms were developed for estimation of CO_2 loading. The flowchart in Fig. 6 shows the steps of the developed models. Three robust correlations namely genetic programming (GP), gene expression programming (GEP), and group method of data handling (GMDH) were considered in this research. The main goal of this study is to develop a strong correlation considering a white-box algorithm, and the equation developed with this algorithm can be easily used without special software or technical programs, thus, the application of this study is high.

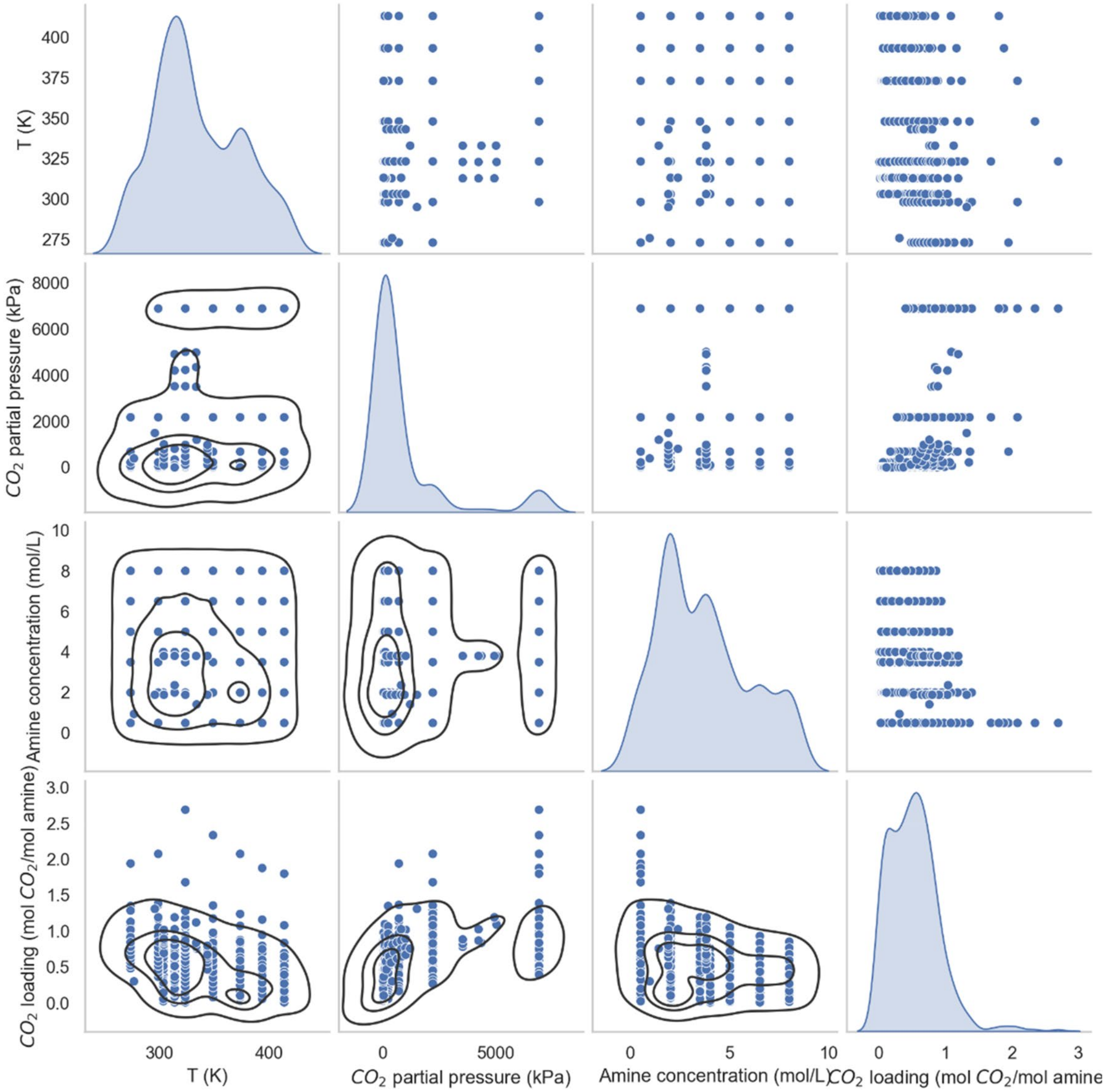


Fig. 5 Correlation matrix of input data in this study

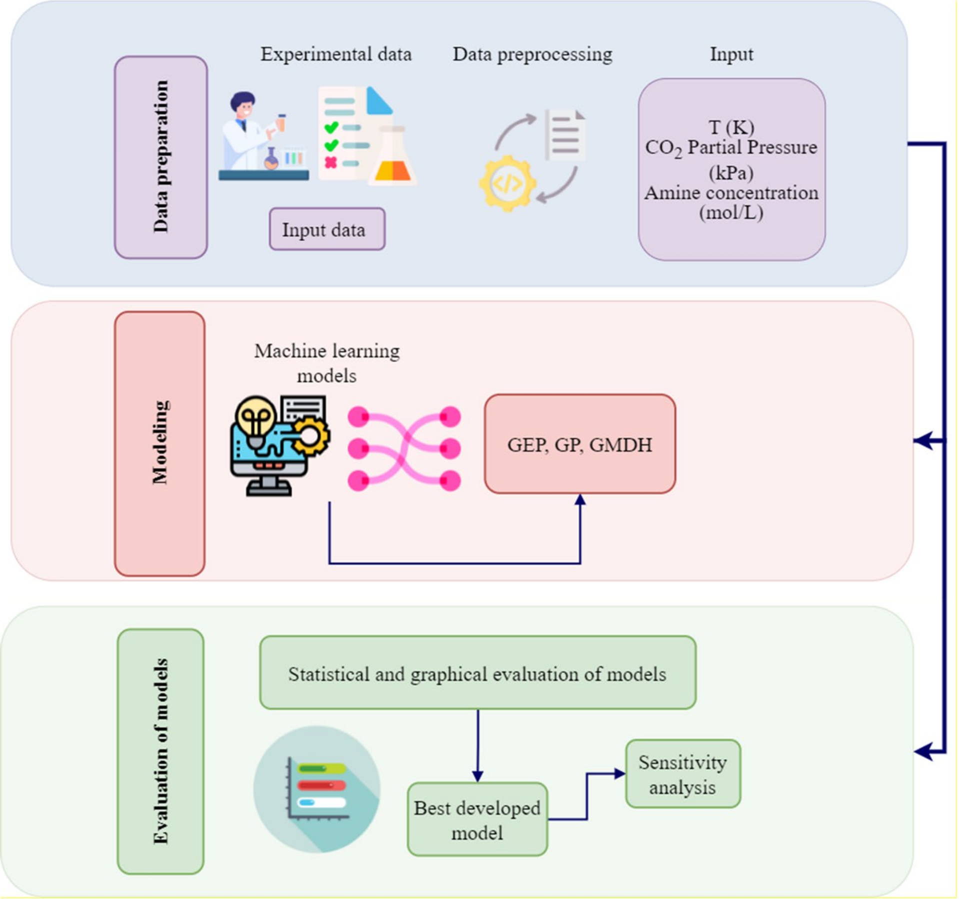


Fig. 6 Flowchart of the developed correlations in this study

3.1 Genetic programming

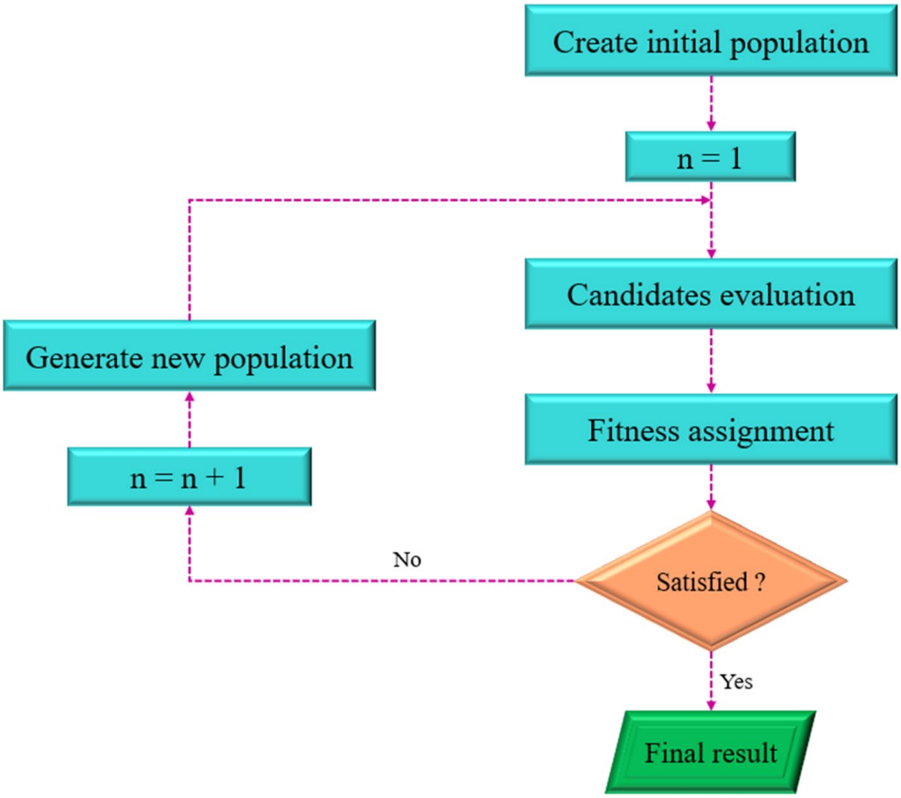
Genetic programming (GP) which was proposed by John Koza in 1994 [35], is a famous robust mathematical paradigm for modeling and optimizing tasks [36, 37]. GP has been generated on the basis of genetic algorithm for the aim of generating precise networks and correlations. GP is capable of recognizing and combining beneficial program sub-expressions to produce a comprehensive network that maximizes the adaptation between inputs and target values [38]. This white-box technique solves problems in extensive ranges of engineering fields, automatically [39]. Besides, it is a machine learning (ML) methodology developing evolutionary computational programs to accomplish issues for solving problems. Due to GP's flexibility, this algorithm can regenerate a mathematical correlation for estimation of various variables in different industries [40, 41]. The notable benefit of the GP method in comparison to other soft computing approaches, is that the GP paradigm prepares white-box techniques which are interpretable by scientists and engineers, readily [42].

In GP structure, to generate chromosome to be operated on a dataset, an initial population of haphazard functions is created [43]. Next, the network's framework is generated simultaneously with tuning the parameters during computation processes. These chromosomes make the next population which takes over for the following generation. These iterations are repeated until a stopping criterion is satisfied [44]. A schematic flowchart of the GP algorithm is shown in Fig. 7.

3.2 Gene expression programming

Gene Expression Programming (GEP) which was proposed by Ferreira in 2001 [45], has appeared as an enhanced artificial intelligence based symbolic regression framework [45, 46]. GEP removes some of the Genetic algorithm and

Fig. 7 Flowchart of the GP technique



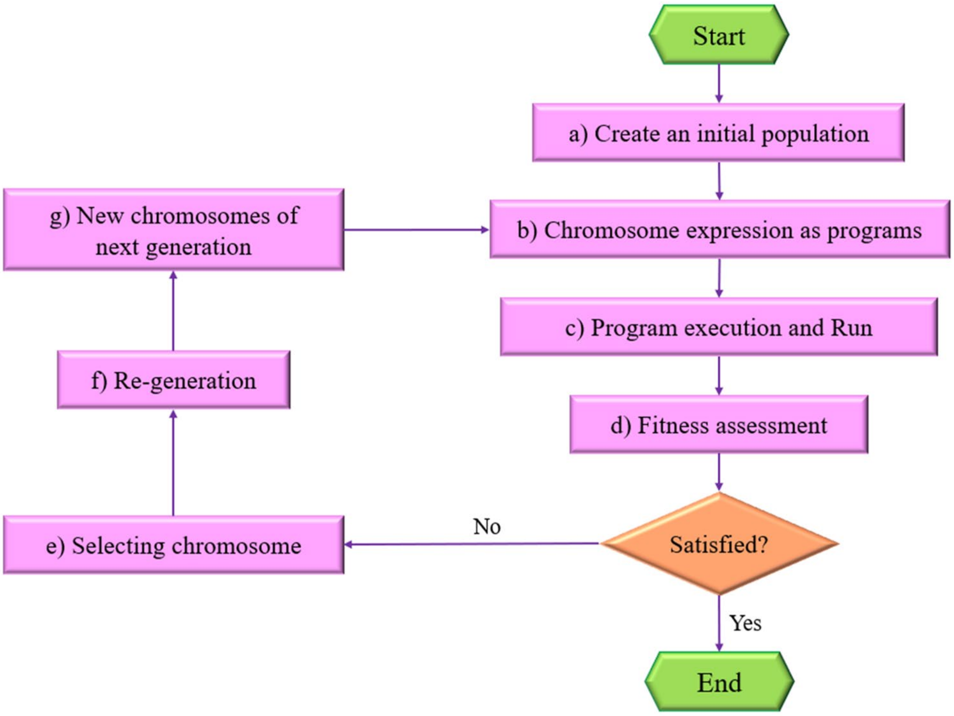
GP technique's restrictions in its procedure, mathematically [47]. This paradigm is a well-known evolutionary algorithm for generation of computer programs, automatically. The two principal parameters in GEP are the expression trees (ETs) and chromosomes. On the other hand, GEP involves linear chromosomes with an established length and expressive parse trees with different shapes and sizes [48].

This fact that no specific functional representation should be detected to find out the optimum estimation for the real measurements, is one of the most noteworthy advantages of the GEP method [49]. In each common GEP, each computer program is encoded by fixed-length gene expression string, usually that is developed through nature-inspired operators like crossover and mutation [50]. Due to the simple rules that detect the platform of the ETs and their interactions, it is possible to conclude the phenotypes given the sequence of the genes, immediately [51]. Each GEP framework has various genes that are created of a head that consists of a terminal and a function. A simple flowchart of the GEP model is presented in Fig. 8. As demonstrated in this figure, steps (b) to (g) will be iterated until a stopping requirement is reached.

3.3 Group method of data handling

The first version of Group Method of Data Handling or GMDH algorithm was introduced by Ivakhnenko in the 1960s [52]. GMDH tries to solve different problems, mathematically using a set of spectrums of polynomial procedures. This data-driven algorithm can overcome the complexity and non-linearity of the networks as it permits producing precise and explicit correlations between inputs and output variables [53]. GMDH also known as polynomial neural network (PNN) consists of a group of inductive paradigms and can be used in various fields such as optimization, data mining, pattern recognition, modeling, and prediction [54]. By applying this heuristic technique, a system can be presented as a group of neurons in which different neuron couples in every layer are linked through a quadratic polynomial, and thus generate new neurons in the next layer [55]. These layers and relevant neurons provide the linking of input variables to the desired output. Figure 9 depicts a schematic flowchart of the GMDH paradigm, and Fig. 10 shows the scheme of the GMDH applied in this paper. Possessing a self-organizing nature and smooth accessibility for the users are two remarkable benefits of the GMDH method [56]. The output value concluded by the primary GMDH method is calculated as [57]:

Fig. 8 Flowchart of the GEP algorithm


Start
a) Create an initial population
b) Chromosome expression as programs
c) Program execution and Run
d) Fitness assessment
Satisfied?
No
Yes
End
e) Selecting chromosome
f) Re-generation
g) New chromosomes of next generation

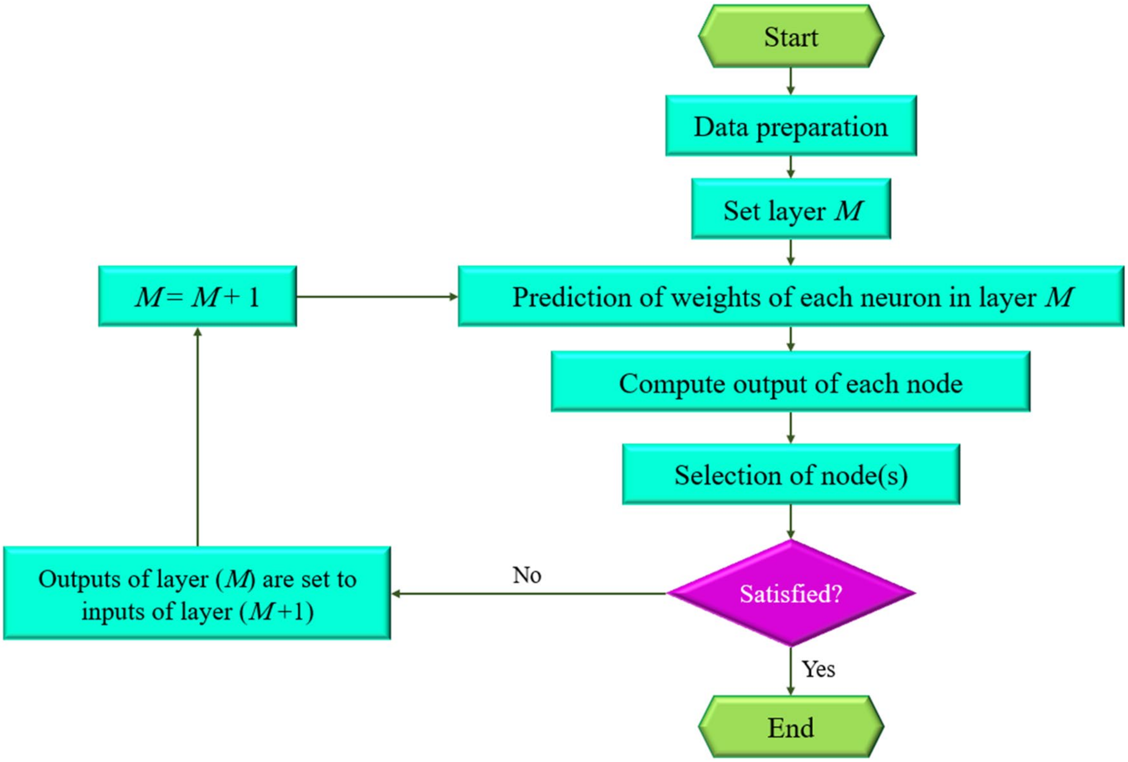

Start
Data preparation
Set layer M
Prediction of weights of each neuron in layer M
Compute output of each node
Selection of node(s)
Satisfied?
No
Yes
End
Outputs of layer (M) are set to inputs of layer (M+1)
M= M+ 1

Fig. 9 Flowchart of the GMDH technique

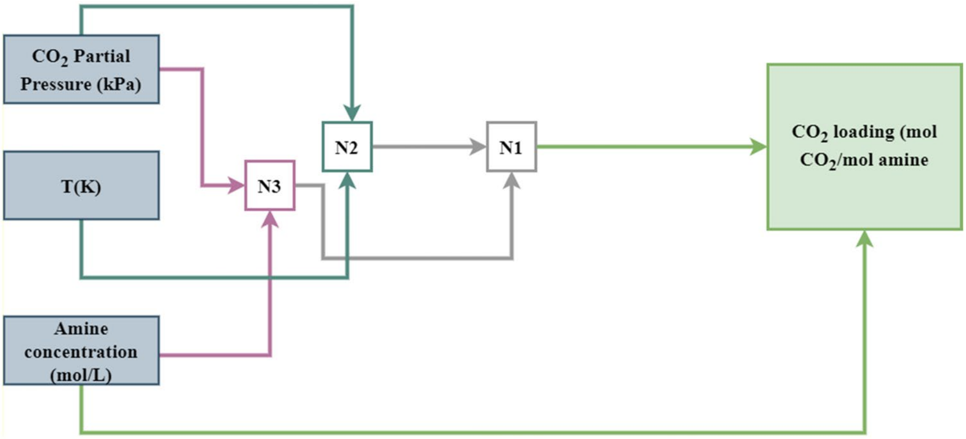

CO$_2$ Partial Pressure (kPa)
T(K)
Amine concentration (mol/L)
N3
N2
N1
CO$_2$ loading (mol CO$_2$/mol amine

Fig. 10 Flowchart of the developed GMDH in this study

$$y = a_0 + \sum_{i=1}^{N} a_i x_i + \sum_{i=1}^{N}\sum_{j=1}^{N} a_{i,j} x_i x_j + \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} a_{i,j,k} x_i x_j x_k + \cdots \tag{1}$$

where, $x_{i,j,k,\ldots}$ show the input vectors, $a_{0,i,j,k,\ldots}$ are the polynomial coefficients, and N denotes the number of input variables. Therefore, the quadratic polynomial functions are performed for mixing the neurons in the previous layer in order to generate new variables using the following equation:

$$P_i^{GMDH} = a_0 + a_1 x_i + a_2 x_j + a_3 x_i x_j + a_4 x_i^2 + a_5 x_j^2 \tag{2}$$

Eventually, the best combination of the two independent variables is recognized according to Eq. 3.

$$\delta_j^2 = \sum_{i=N_t+1}^{N} \left(y_i - P_i^{GMDH}\right)^2 < \varepsilon, \quad j = 1, 2, \ldots, \binom{N}{2} \tag{3}$$

In the above formula, N_t stands for the number of training data. Hence, the subsequent independent variable will be saved if the prementioned stopping condition is reached [58].

4 Evaluation of models

Using multiple statistical indicators, the precision of the suggested models was evaluated. These are the descriptions of the measures listed [59]:

Root mean square error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{4}$$

Standard deviation (SD):

$$\mathrm{SD} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2} \tag{5}$$

Mean absolute percentage error (MAPE%):

$$E_r = \frac{1}{n}\sum_{i=1}^{n} abs\left(\left[\frac{y_i - \hat{y}_i}{y_i}\right]\right) \times 100 \tag{6}$$

Mean absolute value (MAE):

This prognosis is equivalent to the value that was anticipated for the absolute error loss or the l1-norm loss, both of which serve as measures of risk. If $\widetilde{yi}$ is the predicted value of the i-th sample, and yi is the actual value, then the following formula may be used to determine the mean absolute error (MAE) over n "samples."

$$\text{MAE}(y, \hat{y}) = \frac{1}{n}\sum_{i=0}^{n_{\text{samples}}-1} \left|y_i - \hat{y}_i\right|. \tag{7}$$

Mean Bias Error (MBE):

$$MBE = \frac{1}{n}\sum_{i=1}^{n} \left(\tilde{y}_i - y_i\right) \tag{8}$$

The Coefficient of determination (R^2):

$$R^2 = 1 - \frac{\sum_{t=1}^{T}\left(y_i - \tilde{y}_i\right)^2}{\sum_{t=1}^{T}\left(y_i - \overline{y}_i\right)^2} \tag{9}$$

where $\tilde{y}_i$ is the predicted value of the i-th sample, yi is the actual value, and $\overline{y}_i$ is the mean of experimental data.

Moreover, graphical analysis was used to validate the models' correctness.

5 Results and discussion

5.1 Development of the correlations

In this study, three robust correlations of GP, GEP, GMDH were developed for prediction of CO_2 loading in TEA aqueous solutions. The modeling details, namely hyperparameters for proposed models, are depicted in Table 2. One of the primary

Table 2 Optimal features for implemented models

Model	Hyperparameter	Value
GMDH	K-fold validation	5
	Neuron input	2
	Max number of Layer	12
	Initial layer width	4
GP	The population size (P)	3000
	The number of generations	50
	The set of function symbol	+, −, *, /, sin, cos, exp, ln
	The mutation factor	Rand(0,1)
	The crossover factor	Rand(0,1)
GEP	The population size (P)	5000
	The number of generations	150
	The set of function symbol	+, −, *, /, sin, cos, exp, ln
	The mutation factor	Rand(0,1)
	The crossover factor	Rand(0,1)
	The crossover rate	0.3
	The number of iterations	1000

advantages of these white-box approaches utilized is that it is rather possible and simple to review and apply their anticipation power employing the comprehensible equations. That is why, this study is dedicated to the presentation of the developed formulas allowing to estimate CO_2 loading with three input parameters including CO_2 partial pressure (kPa), amine concentration ($\frac{mol}{L}$), and temperature (K).

For GP algorithm, the following correlation was developed:

$$\begin{aligned}
&\text{GP}\\
&CO_2\,\text{loading}\left(\frac{molCO_2}{molamine}\right)=\left(\frac{\left(a_0\times \mathrm{T}+\frac{(a_1\times \mathrm{T}+a_2\times \mathrm{P})}{(a_3\times \mathrm{T}+a_4)}\right)\times\left(a_5\times \mathrm{C}+a_6\right)\times a_7}{\mathrm{T}\times\left(a_8\times \mathrm{T}+a_9\right)\cdot\left(\frac{a_{10}\times \mathrm{T}}{a_{11}\times \mathrm{P}}+a_{12}\right)\times a_{13}}+a_{14}\right)\\
&a_0=0.097705\\
&a_1=1.5222\\
&a_2=0.65395\\
&a_3=-0.08609\\
&a_4=15.551\\
&a_5=0.14348\\
&a_6=-1.4122\\
&a_7=-2224.4\\
&a_8=-0.082256\\
&a_9=14.859\\
&a_{10}=1.0087\\
&a_{11}=1.2999\\
&a_{12}=12.532\\
&a_{13}=1.4848\\
&a_{14}=0.045162
\end{aligned}\tag{10}$$

For GEP algorithm, the following correlation was developed:

$$\begin{aligned}
&\text{GEP}\\
&CO_2\,\text{loading}\left(\frac{molCO_2}{molamine}\right)=\left(\exp\left(\exp\left(\frac{\mathrm{P}\times\exp(a_0\times \mathrm{T})\times \mathrm{T}\times(a_1\times \mathrm{T}+a_2)\cdot a_3\cdot a_4}{\ln\left(a_5\times \mathrm{P}\right)}\right)\cdot a_6+a_7\right)\right.\\
&c_0=-0.051245\\
&c_1=-1.4244\\
&c_2=2.2228\\
&c_3=4.6629\\
&c_4=0.51884\\
&c_5=3.2356\\
&c_6=-0.52106\\
&c_7=1.4637
\end{aligned}\tag{11}$$

For GMDH algorithm, the following correlation was developed:

$$\begin{aligned}
&\mathrm{GMDH}:\\
&CO_2loading\left(\frac{molCO_2}{molamine}\right) = -1.15938 + \mathrm{C}\times 0.95809 - \mathrm{C}\times \mathrm{N}_1 \times 0.175229 - \mathrm{C}^2\times 0.189917 + \mathrm{N}_1\times 1.36004 + \mathrm{N}_1^2\times 0.143343\\
&N_1 = 0.0182987 + N_2\times 0.617315 + N_2\times N_3\times 2.06751 - N_2^2\times 0.542134 + N_3\times 0.257519 - N_3^2 * 1.31129\\
&\mathrm{N}_2 = 346.54 - \ln(\mathrm{T}) * 120.329 - \ln(\mathrm{T}) * \ln(\mathrm{P})\times 0.851806 + (\ln(\mathrm{T}))^2\times 10.4474 + \ln(\mathrm{P}\times 4.88665 + (\ln(\mathrm{P})^2\times 0.0244435\\
&\mathrm{N}_3 = -7.38024 - \ln(\mathrm{P}\times 0.199505 + \ln(\mathrm{P}\times \mathrm{C}\times 0.0662905 + (\ln(\mathrm{P})^2\times 0.023719 + \mathrm{C}\times 6.41163 - (\mathrm{C}))^2\times 1.32279
\end{aligned} \tag{12}$$

where C is amine concentration mol/L, P is CO_2 partial pressure kPa, and T is temperature (K).

5.2 Statistical evaluation of the models

From Table 3, it is observed that all the models utilized are of high reliability. RMSE values are all between 0.02 and 0.036 which is an extremely low indicators, and SD represents that predictions of all datasets are close to their corresponding experimental values, as all of the SD values are below 0.12. Moreover, R^2 is equal to or more than 0.94 in all models, and MAPE% values are not more than 9.1%. In addition, MBE and MAE in all cases are far less than even 0.1. Concerning all said above, GEP, GP, and GMDH are very strong and robust techniques for the forecasting of CO_2 loading in TEA aqueous solutions.

It should be mentioned that despite the effectiveness of all the algorithms applied, GMDH is the most precise and credible one. It has the highest R^2 values and lowest RMSE, SD, MAPE, and MAE figures for "Train", "Test" and "All" groups.

5.3 Graphical evaluation of the models

The most evident representation of all algorithms' performance is the visual materials. In this paper, five diverse types of graphs were utilized including cross plots, data index graphs, residual error plots, error distribution, and cumulative frequency graphs.

Cross plot is the method to compare both real and anticipated data. As it is seen in Fig. 11, GEP is the most unprecise algorithm among the three developed having the greatest number of substantial outliers. Speaking about GP, errors are quite a few, and data points are generally located within the ± 10% error line, however, observations are slightly remote from 0% error line which reduces the efficiency of the approach. GMDH, on the other hand, has the "thinnest" line of observations located right at 0% line.

Data Index plot is another great way to visualize how good a correlation is in making predictions. This graph depicts the comparison between what each data point really is and what the robust correlation predicted it to be. From Fig. 12, it is understood that all models utilized cope with the forecast problem really well. The visual imprecision is rather small for GP, GEP, and GMDH.

Residual error plot depicts the difference between experimental and anticipated estimates as a function of experimental CO_2 loading data. As evident from Fig. 13, GMDH has the lowest spread range between −0.07 and 0.14. Nevertheless, the majority of the points lie within −0.05 and 0.05 intervals. That makes the GMDH the approach with the smallest number of outliers. GP and GEP are practically similar in this case having the scope of around −0.07 and 0.16.

Table 3 Statistical assessment of the developed white-box correlations

Algorithms	Dataset	RMSE	SD	R^2	MAPE%	MBE	MAE
GEP	Test	0.03540	0.11643	0.94802	8.29930	0.00191	0.02379
	Train	0.02512	0.11587	0.97659	7.35200	−0.00064	0.01738
	All	0.02750	0.11617	0.97138	7.54290	−0.00013	0.01867
GP	Test	0.03274	0.11838	0.95555	9.05040	−0.00142	0.02383
	Train	0.02903	0.11950	0.96873	8.53590	−0.00399	0.02109
	All	0.02982	0.11943	0.96637	8.63960	−0.00347	0.02164
GMDH	Test	0.02770	0.09465	0.96818	5.91090	−0.00274	0.01648
	Train	0.02062	0.10154	0.98422	5.72520	0.00134	0.01335
	All	0.02223	0.10020	0.98130	5.76260	0.00052	0.01398

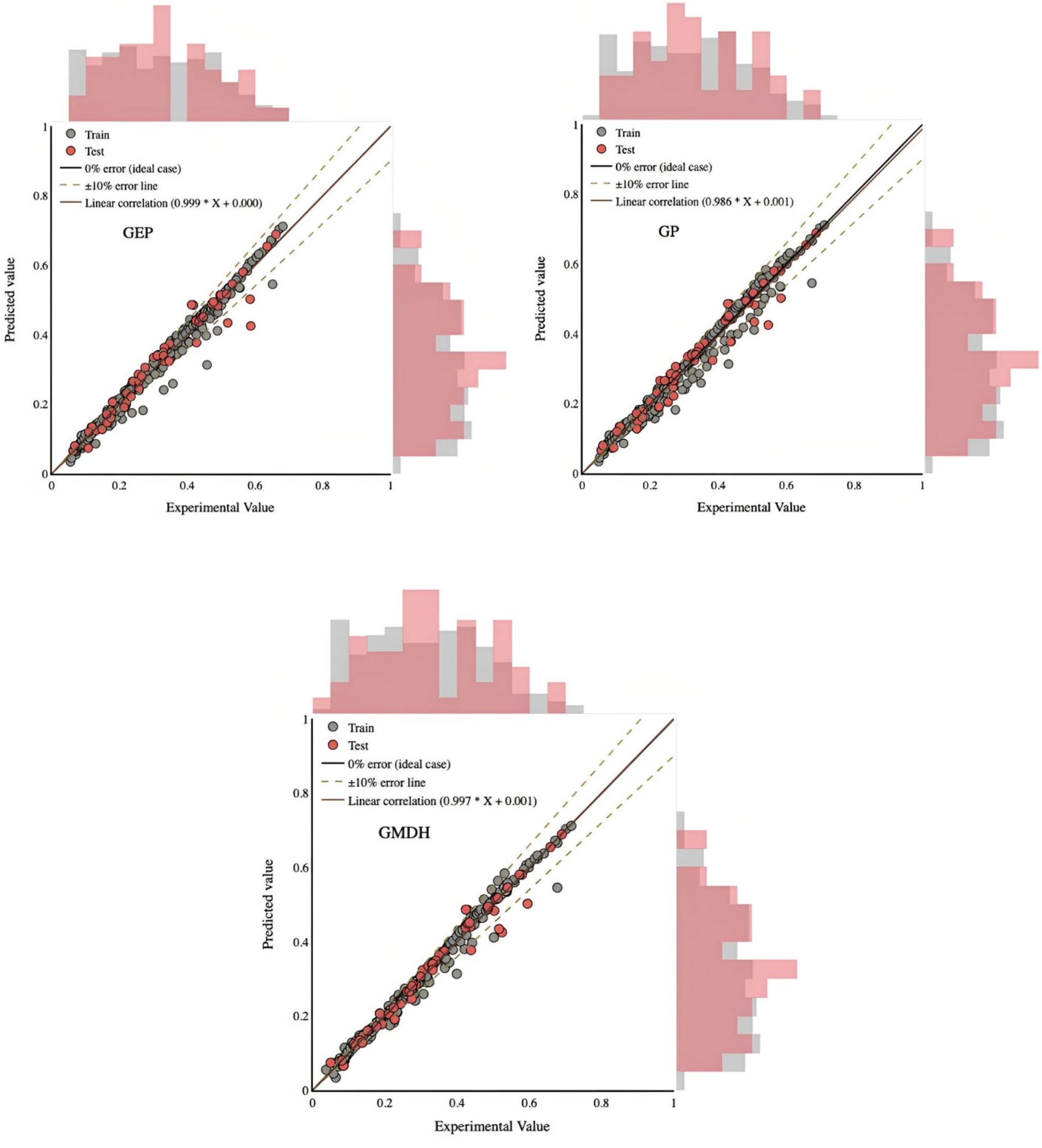
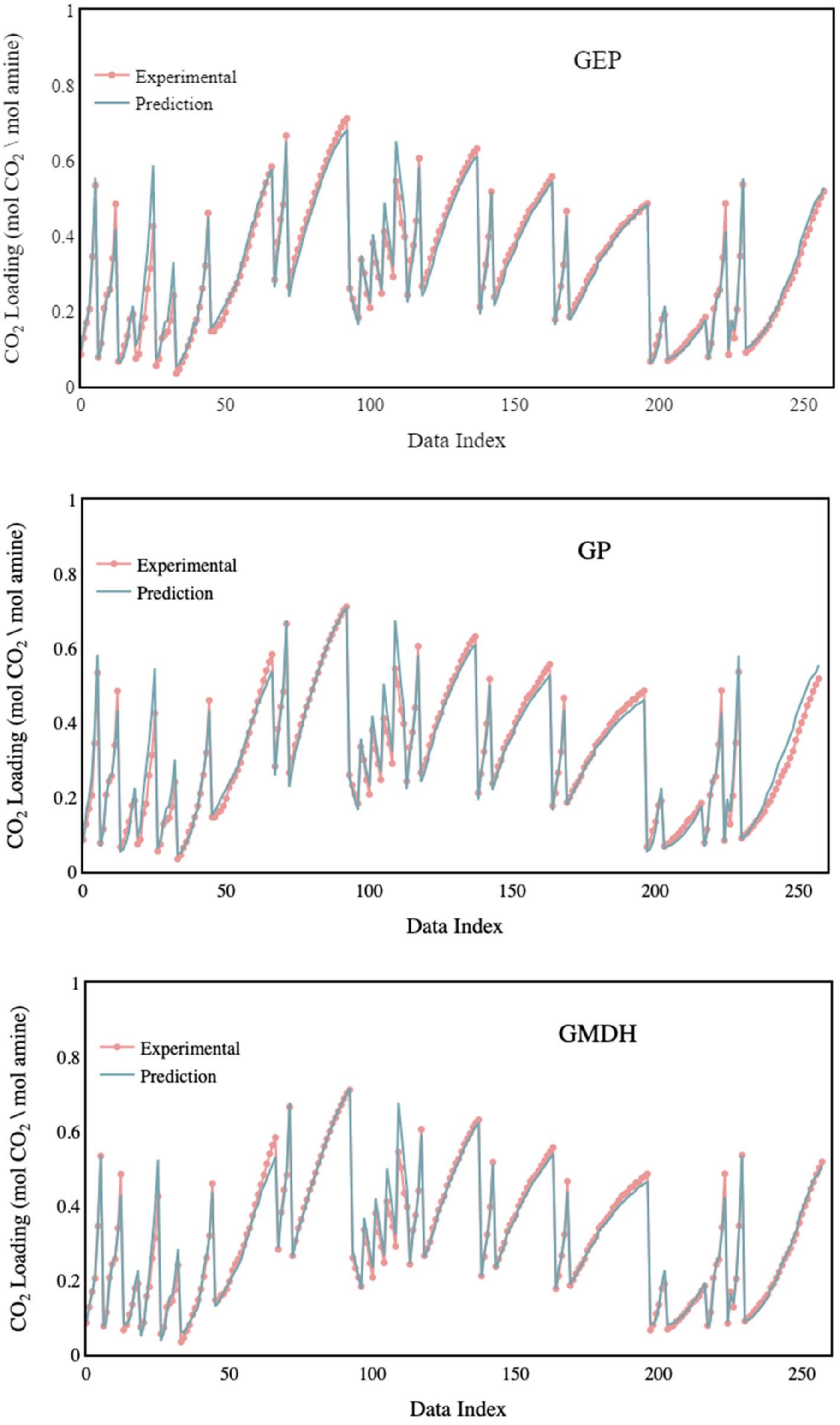


Fig. 11 Cross plots of the developed models

Figure 14 shows the error distribution for testing and training data. Error distribution is a kind of the visual representation which shows the spread of residual error along x-axis. In Fig. 14, GEP and GP are almost identical having the major data portion lying at the center and the spread range from – 0.2 and 0.6. On the other hand, the GMDH situation is different as it has the interval of roughly – 0.35 and 0.35. The spread is slightly lower compared to the first two algorithms,

Fig. 12 Comparison of experimental and predicted values by data index



and the tails are much more centered when GP and GEP have disproportional data allocation which directly influences their accuracy.

Figure 15 is the cumulative frequency which shows absolute relative error versus data frequency. In accordance with it, GMDH is the most accurate one and can predict 80% of the data with less than 0.015 absolute residual error. The corresponding values for GEP and GP robust correlations were 0.03 and 0.06.

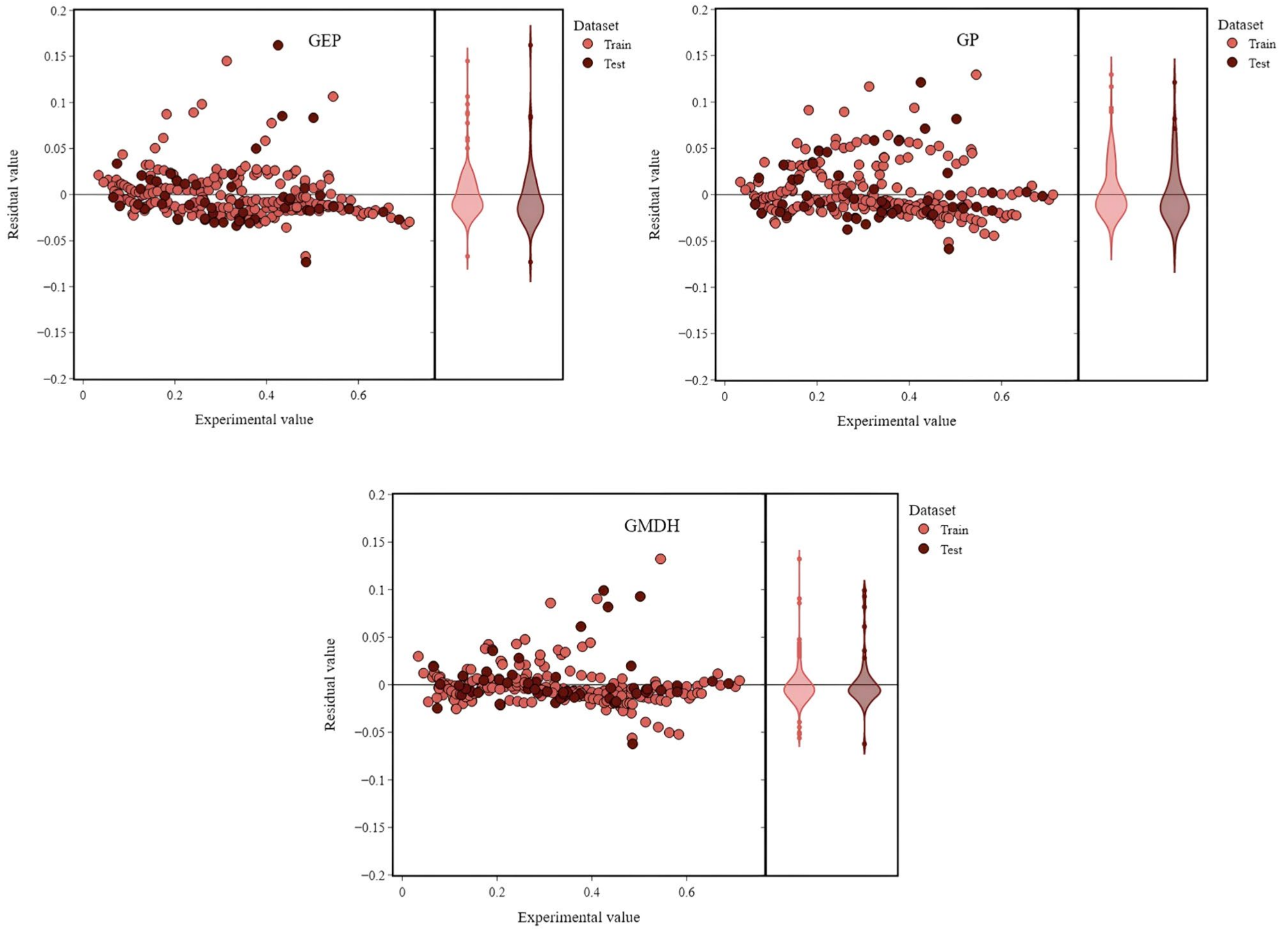


Fig. 13 Residual error plots for the developed models

5.4 Trend analysis

As shown, the GMDH algorithm can estimate CO_2 loading based on temperature, CO_2 partial pressure (kPa), and amine content (mol/L). It needs to be noted that the samples used in this research were acquired from validated experimental findings. As seen in Figs. 16 and 17, the GMDH model well predicts the experimental trend of the different temperatures. Figure 16 shows the expected and experimental outcomes for the samples at 298 K and C = 2.83 mol/L.

The anticipated and experimental outcomes of CO_2 loading are also shown in Fig. 17. The created GMDH model could precisely forecast the behavior of samples at 353.2 K and C = 2 mol/L.

5.5 Sensitivity analysis

The relevancy coefficient (r) (also known as the Pearson coefficient) and the output of the GMDH model are considered to estimate the relative importance of the input coefficients for CO_2 loading. This formula is also known as the Pearson correlation coefficient. The r value for each input parameter is calculated using the following procedure [60].

$$r(inp,x) = \frac{\sum_{j=1}^{n}(inp_{i,j} - inp_{m,i})(x_j - x_m)}{\left(\sum_{j=1}^{n}\left(inp_{i,j} - inp_{m,i}\right)^2 \sum_{j=1}^{n}\left(x_j - x_m\right)^2\right)^{0.5}} \tag{13}$$

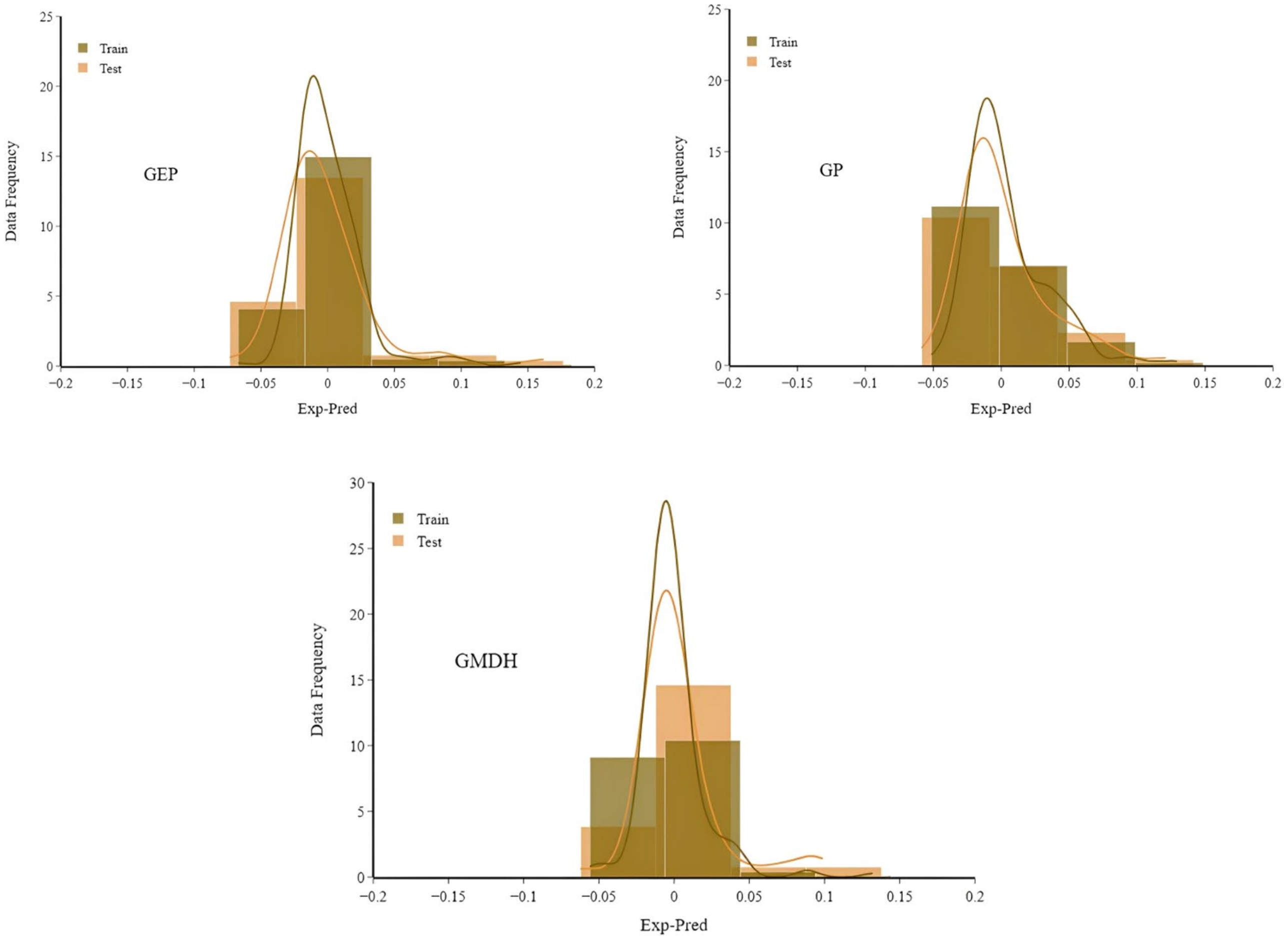
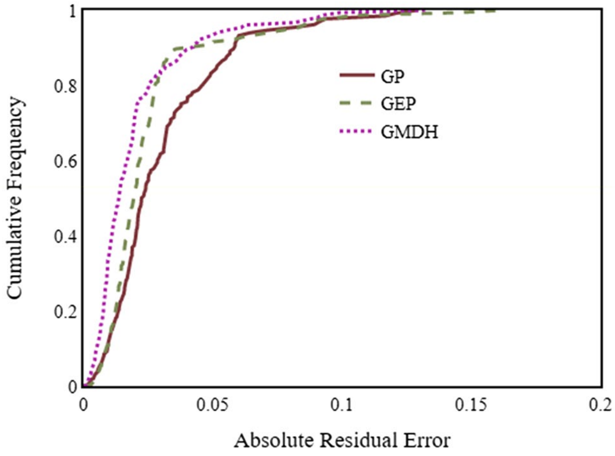


Fig. 14 Error distribution of predicted value for all developed algorithms

Fig. 15 Cumulative frequency for all developed models



where $np_{i,j}$, $inp_{m,i}$ represent the *j*th and average values of the *i*th input value and $inp_{i,j}$ are T(K), CO_2 partial pressure (kPa), and amine concentration. x_m represents the average of the predicted CO_2 loading and x_j is the jth value of the predicted CO_2 loading. The input parameters for the sensitivity analysis are shown in Fig. 18 and are temperature (in Kelvin), CO_2 partial pressure (in kPa), and amine concentration (in moles per liter). The data visualization suggests that temperature is the most important factor in determining CO_2 loading. Then, CO_2 partial pressure and amine concentration are important, respectively.

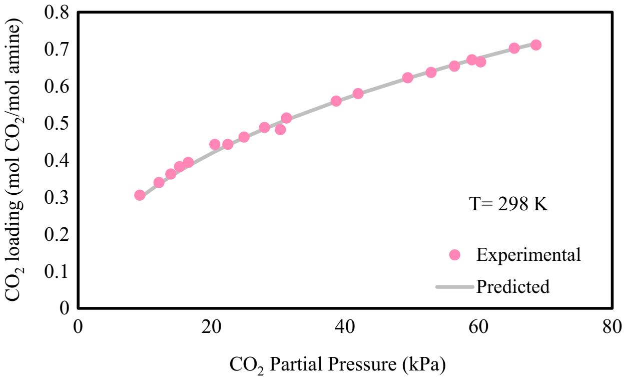


Fig. 16 Trend analysis of the developed model, relationship between CO_2 loading versus partial pressure at T = 298 K and C = 2.83 mol/L

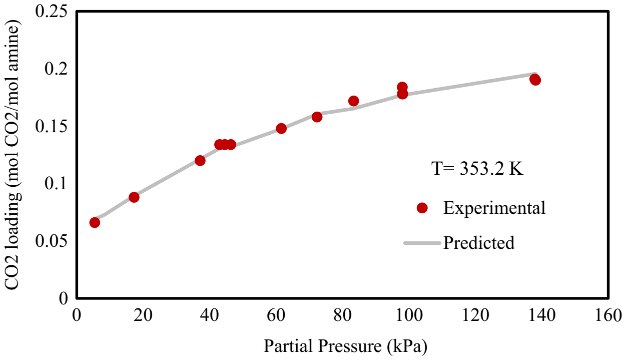


Fig. 17 Trend analysis of the developed model, relationship between CO_2 loading versus partial pressure at T = 353.2 K and C = 2 mol/L

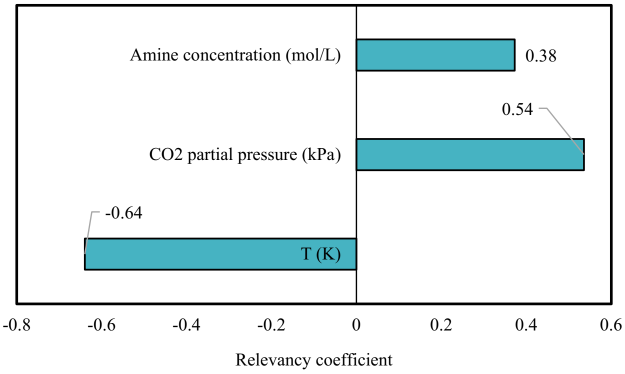


Fig. 18 Sensitivity analysis using the GMDH model

The nonparametric equivalent of the Pearson correlation coefficient, the Spearman correlation measures the strength of the relationship between two variables based on their rankings. One of the following formulae may be used to get the Spearman correlation coefficient according to whether there are ties in the sorting (the same rank being given to two or more observations) or not.

In the absence of ties, the following formula will work:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{14}$$

where the difference between two rankings is called di. The total number of observations is n.

The whole Spearman correlation formula, which is a slightly modified version of Pearson's r, must be employed to handle tied ranks.

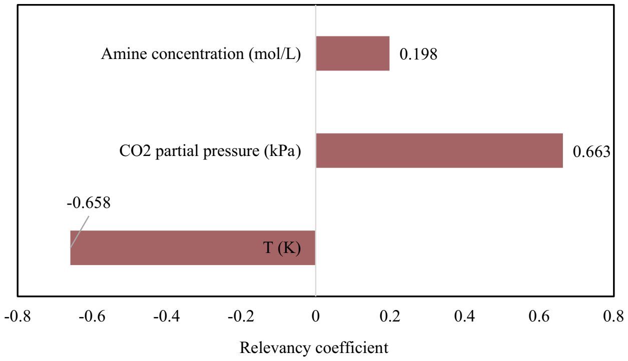


Fig. 19 Spearman analysis using the GMDH model

$$\rho = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(R\left(x_i\right)-\bar{R}(x)\right).\left(R\left(y_i\right)-\bar{R}(y)\right)}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}\left(R\left(x_i\right)-\bar{R}(x)\right)^{2}\right).\frac{1}{n}\sum_{i=1}^{n}\left(R\left(y_i\right)-\bar{R}(y)\right)^{2}}} \tag{15}$$

where, x and y variables' rankings are R(x) and R(y), respectively. The mean rankings are $\bar{R}(x)$ and $\bar{R}(y)$. Figure 19 shows Spearman correlation analysis. As can be seen in this figure, temperature has the highest negative impact, while CO_2 partial pressure has the highest positive effect on CO_2 loading.

The main difference between the Pearson and Spearman coefficients that is the Pearson coefficient works only with a linear relationship between variables, while the Spearman coefficient works with non-linear relationship. It should be also mentioned that Spearman works with rank-ordered variables, while Pearson works with raw data values. The Spearman coefficient is higher than Pearson which means that the data used in this study have correlation which is monotonic but not linear. Pearson coefficient has a lower coefficient for all parameters as pressure and temperature may have a nonlinear relationship with CO_2 loading. Overall, it could be concluded that temperature and CO_2 partial pressure have almost the same absolute relative effect on CO_2 loading.

6 Conclusions

In this study, three advanced white-box algorithms were developed for correlating CO_2 loading capacity of triethanolamine (TEA) aqueous solutions using GMDH, GP, and GEP approaches. Temperature of the system, partial pressure of CO_2, and amine concentration in the aqueous phase were considered as input parameters. Sensitive analysis (Pearson and Spearman) was used to investigate the impact of input parameters on target value (CO_2 loading). The following main conclusions are found in this research:

- According to statistical and graphical analyses, the GMDH robust correlation showed the highest accuracy compared with GEP and GP. The statistical parameters of R^2, RMSE, and MAPE are obtained 0.9813, 0.0222, and 5.76% for GMDH; 0.9713, 0.0275, and 7.54% for GEP and 0.9664, 0.0298, and 8.63% for GP. It can be concluded that the accuracy order of the model for CO_2 loading prediction in TEA is GMDH > GEP > GP.
- The trend analysis of CO_2 loading versus CO_2 partial pressure at constant temperature was investigated. The trend analysis findings demonstrated that the developed GMDH correlation successfully predicted the variation of CO_2 loading with pressure. In order to provide precise predictions, the suggested GMDH model may be used instead of complicated thermodynamic models.
- In addition, two approaches of sensitivity analysis (Spearman coefficient and Pearson coefficient) were used to examine the effect of input parameters on CO_2 loading. The Pearson coefficient showed that temperature is the most important factor in determining the CO_2 loading, whereas CO_2 partial pressure and amine concentration play smaller roles. The Spearman coefficient has different and higher coefficient than Pearson which shows that the dataset has a nonlinear relationship between variables. This coefficient showed that the pressure and temperature have almost the same impact on CO_2 loading. Overall, it could be concluded that temperature and CO_2 partial pressure have almost the same absolute relative effect on CO_2 loading, while amine concentration has the lowest effect on it.

Author contributions FH: Writing—Original draft, Visualization, Data curation, Software, Methodology, BA: Writing—Original draft, Software, Validation, Methodology, SA: Software, Data curation, Validation, Investigation, Visualization, MA: Conceptualization, Methodology, Visualization, QL: Conceptualization, Methodology, Data curation, AM: Supervision, Conceptualization, Methodology, Reviewing and Editing, MO: Supervision, Conceptualization, Reviewing and Editing, AH: Supervision, Conceptualization, Reviewing and Editing, Investigation.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Competing interests The authors have not disclosed any competing interests.

References

1. IEA. World Energy Outlook 2022. IEA, Paris, France; 2022.
2. Outlook STE. US Energy Information Administration. 2023.
3. Kidnay AJ, Parrish WR, McCartney DG. Fundamentals of natural gas processing. Cambridge: CRC Press; 2019.
4. Nord LO, Anantharaman R, Bolland O. Design and off-design analyses of a pre-combustion CO2 capture process in a natural gas combined cycle power plant. Int J Greenhouse Gas Control. 2009;3(4):385–92.
5. Speight JG. Natural gas: a basic handbook. Houston: Gulf Professional Publishing; 2018.
6. Schoots K, Rivera-Tinoco R, Verbong G, Van der Zwaan B. Historical variation in the capital costs of natural gas, carbon dioxide and hydrogen pipelines and implications for future infrastructure. Int J Greenhouse Gas Control. 2011;5(6):1614–23.
7. Armaroli N, Balzani V. Energy for a sustainable world, from the oil age to a sun-powered future Copyright© 2011 WILEY. Weinheim: VCH Verlag GmbH & Co. KGaA; 2011.
8. Kumar S, Cho JH, Moon I. Ionic liquid-amine blends and CO2BOLs: Prospective solvents for natural gas sweetening and CO2 capture technology—a review. Int J Greenhouse Gas Control. 2014;20:87–116.
9. Mudhasakul S, Ku H-M, Douglas PL. A simulation model of a CO2 absorption process with methyldiethanolamine solvent and piperazine as an activator. Int J Greenhouse Gas Control. 2013;15:134–41.
10. Shahid MZ, Kim J-K. Design and economic evaluation of a novel amine-based CO2 capture process for SMR-based hydrogen production plants. J Clean Prod. 2023;402:136704.
11. Bhide B, Voskericyan A, Stern S. Hybrid processes for the removal of acid gases from natural gas. J Membr Sci. 1998;140(1):27–49.
12. Dortmundt D, Doshi K. Recent developments in CO2 removal membrane technology. UOP LLC 1999;1.
13. Ghiasi MM, Mohammadi AH. Rigorous modeling of CO2 equilibrium absorption in MEA, DEA, and TEA aqueous solutions. J Nat Gas Sci Eng. 2014;18:39–46.
14. Campbell JM, Maddox RN, Lilly LL, Hubbard RA. Gas conditioning and processing. Campbell Petroleum Series Norman, Oklahoma; 1984.
15. Li Y-G, Mather AE. Correlation and prediction of the solubility of CO2 and H2S in aqueous solutions of triethanolamine. Ind Eng Chem Res. 1996;35(12):4804–9.
16. Aghel B, Janati S, Wongwises S, Shadloo MS. Review on CO2 capture by blended amine solutions. Int J Greenhouse Gas Control. 2022;119:103715.
17. Hasib-ur-Rahman M, Siaj M, Larachi F. Ionic liquids for CO2 capture—Development and progress. Chem Eng Process. 2010;49(4):313–22.
18. Aghel B, Gouran A, Behaien S, Vaferi B. Experimental and modeling analyzing the biogas upgrading in the microchannel: Carbon dioxide capture by seawater enriched with low-cost waste materials. Environ Technol Innov. 2022;27:102770.
19. Mason JW, Dodge BF. Equilibrium absorption of carbon dioxide by solutions of the ethanolamines. Verlag nicht ermittelbar; 1936.
20. Jou FY, Otto F, Mather A. Equilibria of H2S and CO2 in triethanolamine solutions. Can J Chem Eng. 1985;63(1):122–5.
21. Jou F-Y, Otto FD, Mather AE. Solubility of mixtures of hydrogen sulfide and carbon dioxide in aqueous solutions of triethanolamine. J Chem Eng Data. 1996;41(5):1181–3.
22. Lyudkovskaya M. Solubility of carbon dioxide in solutions of ethanolamines under pressure. 1963.
23. Chakma A, Lemonier J, Chornet E, Overend R. Absorption of CO2 by aqueous triethanolamine (TEA) solutions in a high shear jet absorber. Gas Sep Purif. 1989;3(2):65–70.
24. Nakhjiri AT, Heydarinasab A, Bakhtiari O, Mohammadi T. Experimental investigation and mathematical modeling of CO2 sequestration from CO2/CH4 gaseous mixture using MEA and TEA aqueous absorbents through polypropylene hollow fiber membrane contactor. J Membr Sci. 2018;565:1–13.
25. Horng S-Y, Li M-H. Kinetics of absorption of carbon dioxide into aqueous solutions of monoethanolamine+ triethanolamine. Ind Eng Chem Res. 2002;41(2):257–66.
26. Fouad WA, Berrouk AS. Prediction of H2S and CO2 solubilities in aqueous triethanolamine solutions using a simple model of Kent-Eisenberg type. Ind Eng Chem Res. 2012;51(18):6591–7.
27. Chung P-Y, Soriano AN, Leron RB, Li M-H. Equilibrium solubility of carbon dioxide in the amine solvent system of (triethanolamine+ piperazine+ water). J Chem Thermodyn. 2010;42(6):802–7.

28. Ghiasi MM, Arabloo M, Mohammadi AH, Barghi T. Application of ANFIS soft computing technique in modeling the CO2 capture with MEA, DEA, and TEA aqueous solutions. Int J Greenhouse Gas Control. 2016;49:47–54.
29. Mores P, Scenna N, Mussati S. CO2 capture using monoethanolamine (MEA) aqueous solution: modeling and optimization of the solvent regeneration and CO2 desorption process. Energy. 2012;45(1):1042–58.
30. Saghafi H, Arabloo M. Modeling of CO2 solubility in MEA, DEA, TEA, and MDEA aqueous solutions using AdaBoost-Decision Tree and Artificial Neural Network. Int J Greenhouse Gas Control. 2017;58:256–65.
31. Yarveicy H, Saghafi H, Ghiasi MM, Mohammadi AH. Decision tree-based modeling of CO2 equilibrium absorption in different aqueous solutions of absorbents. Environ Prog Sustain Energy. 2019;38(s1):S441–8.
32. Ghiasi MM, Abedi-Farizhendi S, Mohammadi AH. Modeling equilibrium systems of amine-based CO2 capture by implementing machine learning approaches. Environ Prog Sustain Energy. 2019;38(5):13160.
33. Xiao M, Liu H, Gao H, Liang Z. CO2 absorption with aqueous tertiary amine solutions: equilibrium solubility and thermodynamic modeling. J Chem Thermodyn. 2018;122:170–82.
34. Hemmati-Sarapardeh A, Varamesh A, Husein MM, Karan K. On the evaluation of the viscosity of nanofluid systems: modeling and data assessment. Renew Sustain Energy Rev. 2018;81:313–29.
35. Koza JR. Genetic programming as a means for programming computers by natural selection. Stat Comput. 1994;4:87–112.
36. Augusto DA, Barbosa HJ. Symbolic regression via genetic programming. Proceedings. Vol. 1. Sixth Brazilian Symposium on Neural Networks. IEEE; 2000:173–8.
37. Koza JR, Keane MA, Streeter MJ, Mydlowec W, Yu J, Lanza G. Genetic programming IV: Routine human-competitive machine intelligence. Berlin: Springer; 2005.
38. Arnaldo I, Krawiec K, O'Reilly U-M. Multiple regression genetic programming. In: Proceedings of the 2014 annual conference on genetic and evolutionary computation. 2014:879–86.
39. Kaydani H, Najafzadeh M, Mohebbi A. Wellhead choke performance in oil well pipeline systems based on genetic programming. J Pipeline Syst Eng Pract. 2014;5(3):06014001.
40. Parhizgar H, Dehghani MR, Eftekhari A. Modeling of vaporization enthalpies of petroleum fractions and pure hydrocarbons using genetic programming. J Petrol Sci Eng. 2013;112:97–104.
41. Luchian H, Băutu A, Băutu E. Genetic programming techniques with applications in the oil and gas industry. Artif Intell Approaches Petrol Geosci 2015:101–26.
42. Rostami A, Ebadi H, Arabloo M, Meybodi MK, Bahadori A. Toward genetic programming (GP) approach for estimation of hydrocarbon/water interfacial tension. J Mol Liq. 2017;230:175–89.
43. Mahmoodpour S, Kamari E, Esfahani MR, Mehr AK. Prediction of cementation factor for low-permeability Iranian carbonate reservoirs using particle swarm optimization-artificial neural network model and genetic programming algorithm. J Petrol Sci Eng. 2021;197:108102.
44. Fathinasab M, Ayatollahi S. On the determination of CO2–crude oil minimum miscibility pressure using genetic programming combined with constrained multivariable search methods. Fuel. 2016;173:180–8.
45. Ferreira C. Gene expression programming: a new adaptive algorithm for solving problems. arXiv preprint cs/0102027 2001.
46. Kakati D, Roy S, Banerjee R. Development and validation of an artificial intelligence platform for characterization of the exergy-emission-stability profiles of the PPCI-RCCI regimes in a diesel-methanol operation under varying injection phasing strategies: a Gene Expression Programming approach. Fuel. 2021;299:120864.
47. Rostami A, Arabloo M, Kamari A, Mohammadi AH. Modeling of CO2 solubility in crude oil during carbon dioxide enhanced oil recovery using gene expression programming. Fuel. 2017;210:768–82.
48. Hong T, Jeong K, Koo C. An optimized gene expression programming model for forecasting the national CO2 emissions in 2030 using the metaheuristic algorithms. Appl Energy. 2018;228:808–20.
49. Amiri-Ramsheh B, Nait Amar M, Shateri M, Hemmati-Sarapardeh A. On the evaluation of the carbon dioxide solubility in polymers using gene expression programming. Sci Rep. 2023;13(1):12505.
50. Zhong J, Feng L, Ong Y-S. Gene expression programming: a survey. IEEE Comput Intell Mag. 2017;12(3):54–72.
51. Ferreira C. Gene expression programming in problem solving. Soft computing and industry: recent applications 2002:635–53.
52. Ivakhnenko AG. The group method of data handling A rival of stochastic approximation. Soviet Automatic Control. 1968;13:43–55.
53. Ivakhnenko A, Ivakhnenko G. The review of problems solvable by algorithms of the group method of data handling (GMDH). Pattern recognition and image analysis c/c of raspoznavaniye obrazov i analiz izobrazhenii. 1995;5:527–35.
54. Behvandi R, Mirzaie M. A novel correlation for modeling interfacial tension in binary mixtures of CH_4, CO_2, and N_2+ normal alkanes systems: Application to gas injection EOR process. Fuel. 2022;325:124622.
55. Khosravi A, Machado L, Nunes R. Estimation of density and compressibility factor of natural gas using artificial intelligence approach. J Petrol Sci Eng. 2018;168:201–16.
56. Mahdaviara M, Rostami A, Shahbazi K. State-of-the-art modeling permeability of the heterogeneous carbonate oil reservoirs using robust computational approaches. Fuel. 2020;268:117389.
57. Armaghani DJ, Momeni E, Asteris PG. Application of group method of data handling technique in assessing deformation of rock mass. 1 2020;1(1):001.
58. Ayoub MA, Elhadi A, Fatherlhman D, Saleh M, Alakbari FS, Mohyaldinn ME. A new correlation for accurate prediction of oil formation volume factor at the bubble point pressure using Group Method of Data Handling approach. J Petrol Sci Eng. 2022;208:109410.
59. Hemmati-Sarapardeh A, Larestani A, Nait Amar M, Hajirezaie S. Chapter 1 - Introduction. In: Hemmati-Sarapardeh A, Larestani A, Nait Amar M, Hajirezaie S, editors. Applications of artificial intelligence techniques in the petroleum industry. Houston: Gulf Professional Publishing; 2020. p. 1–22.
60. Rashidi-Khaniabadi A, Rashidi-Khaniabadi E, Amiri-Ramsheh B, Mohammadi M-R, Hemmati-Sarapardeh A. Modeling interfacial tension of surfactant–hydrocarbon systems using robust tree-based machine learning algorithms. Sci Rep. 2023;13(1):10836.