Journal of Colloid and Interface Science 647 (2023) 174–187

Contents lists available at ScienceDirect

Journal of Colloid And Interface Science

journal homepage: www.elsevier.com/locate/jcis

Machine learning implemented exploration of the adsorption mechanism of carbon dioxide onto porous carbons

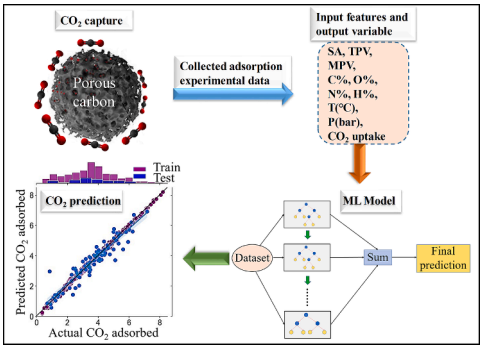
Sarvesh Namdeo [a], Vimal Chandra Srivastava [a,*], Paritosh Mohanty [b]

[a] Department of Chemical Engineering, Indian Institute of Technology Roorkee, Roorkee 247667, Uttarakhand, India
[b] Department of Chemistry, Indian Institute of Technology Roorkee, Roorkee 247667, Uttarakhand, India

HIGHLIGHTS

- CO_2 adsorption onto porous carbons was modelled by machine learning (ML).
- Six ML models used, and feature importance and instance plots explained by SHAP.
- GBDT showed the best prediction, however other models were also performed well.
- CO_2 uptake more governed by adsorption conditions followed by adsorbent properties.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Porous carbons
Adsorption
Machine learning
SHAP explanations
SDG13

ABSTRACT

Adsorption of CO_2 on porous carbons has been identified as one of the promising methods for carbon capture, which is essential for meeting the sustainable developmental goal (SDG) with respect to climate action, i.e., SDG 13. This research implemented six supervised machine learning (ML) models (gradient boosting decision tree (GBDT), extreme gradient boosting (XGB), light boost gradient machine (LBGM), random forest (RF), categorical boosting (Catboost), and adaptive boosting (Adaboost)) to understand and predict the CO_2 adsorption mechanism and adsorption uptake, respectively. The results recommended that the GBDT outperformed the remaining five ML models for CO_2 adsorption. However, XGB, LBGM, RF, and Catboost also represented the prediction in the acceptable range. The GBDT model indicated the accurate prediction of CO_2 uptake onto the porous carbons considering adsorbent properties and adsorption conditions as model input parameters. Next, two-factor partial

Abbreviations: ML, Machine learning; AI, Artificial intelligence; DL, Deep learning; CCS, Carbon capture and storage; MOFs, Metal organic frameworks; MLP, Multilayer perceptron; MLP-LMA, Multilayer perceptron-Levenberg-Marquardt; MLP-BR, Multilayer perceptron-Bayesian Regularization; ELM, Extreme learning machine; GP, Genetic programming; ANN, Artificial neural network; SVM, Support vector machine; KNN, k-nearest neighbour; DNN, Deep neural network; LR, Linear regression; RF, Random Forest; DT, Decision tree; GBT, Gradient boosting tree; GBDT, Gradient boosting decision tree; LBGM, Light boost gradient machine; GBRT, Gradient boost regression tree; MART, Multiple additive regression tree; XGB, Extreme gradient boosting; SA, Surface area; TPV, Total pore volume; MPV, Micropore volume; T, Temperature; P, Pressure; RMSE, Root mean square error; MAE, Mean absolute error; MAPE, Mean absolute percentage error; AARD, Average absolute relative deviation; MSE, Mean squared error; R^2, Coefficient of determination; PCC, Pearson correlation coefficient; DART, Dropouts meet multiple additive regression trees; IQR, Interquartile range; LIME, Local interpretable model-agnostic explanations.

* Corresponding author.
E-mail addresses: snamdeo@ch.iitr.ac.in (S. Namdeo), vimalcsr@yahoo.co.in, vimal.srivastava@ch.iitr.ac.in (V.C. Srivastava), pm@cy.iitr.ac.in (P. Mohanty).

https://doi.org/10.1016/j.jcis.2023.05.052
Received 16 January 2023; Received in revised form 28 April 2023; Accepted 8 May 2023
Available online 18 May 2023

dependence plots revealed a lucid explanation of how the combinations of two input features affect the model prediction. Furthermore, SHapley Additive exPlainations (SHAP), a novel explication approach based on ML models, were employed to understand and elucidate the CO_2 adsorption and model prediction. The SHAP explanations, implemented on the GBDT model, revealed the rigorous relationships among the input features and output variables based on the GBDT prediction. Additionally, SHAP provided clear-cut feature importance analysis and individual feature impact on the prediction. SHAP also explained two instances of CO_2 adsorption. Along with the data-driven insightful explanation of CO_2 adsorption onto porous carbons, this study also provides a promising method to predict the clear-cut performance of porous carbons for CO_2 adsorption without performing any experiments and open new avenues for researchers to implement this study in the field of adsorption because a lot of data is being generated.

1. Introduction

Increasing population and continuous use of fossil fuel lead to an increase in the carbon dioxide (CO_2) concentration (410 ppm to 415 ppm) in the atmosphere [1], which contribute to the major increment in global warming and climate change [2–5]. The major consequences of global warming as depicted by scientists and researchers to the environment, lead to an increment in the seawater level, typhoons, tropical storms, parchedness, and ill health for the people [5–8]. To conquer this issue, the reduction of CO_2 from the environment attracts the attention of almost all nations and researchers over the past decade [3]. The Department of Energy (DOE), United States, has already taken a step back in 2009 to capture 90% of environmental CO_2 with an expenditure of not more than 35% [9]. However, CO_2 reduction from the environment is still a major topic of concern for researchers with green technologies associated with less economy to make the globe greener.

For the reduction of CO_2 concentration in the atmosphere, carbon capture and storage (CCS) technology got the most attention in combating global climate change [10]. However, the CCS technology is considered as expensive procedure because of being CO_2 capture as cost determining step that has part of more than 50% of the total cost of CCS [11]. CO_2 adsorption through porous carbons is a fruitful procedure because adsorbent can be restored and having low cost operations [11–14].

Machine learning (ML), an emerging mathematically advanced tool, is getting more attention to deal with a large amount of data-driven complex problems (i.e., natural language processing, image analysis, language translation) and procured interests of chemical engineers for the past couple of years [15,16]. Using ML models over traditional models (i.e., Prandtl's boundary layer model) for modelling is advantageous because ML models get intelligent over time, trained on real data sets, flexible in nature, robust-designing, accurate, less time-consuming, and easy to use and implement [17,18]. Moreover, ML methods, which are non-parametric, are enriched with a special ability to deal with non-linear data [19]. As a consequence, the application of ML in dealing with environmental and engineering problems is growing rapidly [20–22].

ML has been used widely in both fields of adsorption (liquid phase and gas phase) [23]. For instance, Yuan et al. [10] demonstrated the application of ML models (tree-based models-GBDT, LBGM, and XGB) to predict the CO_2 uptake in biomass waste-derived porous carbons. Recently, Amar et al. [3] applied advanced ML models (i.e., MLP-LMA, MLP-BR, ELM, and GP) for the prediction of CO_2 uptake in metal–organic frameworks (MOFs) and MLP-LMA outperformed MLP-BR, ELM, and GP, which gave AARD and R^2 values of 0.9205% and 0.9998, respectively. A study on the identification of adsorption characteristics of supercritical CO_2/CH_4 onto coal is carried out using a machine learning approach where ANN is implemented for the prediction [24]. The adsorbent selection process is conducted for the adsorption of tetracycline and sulfamethoxazole using machine learning models [25]. Where RF and ANN approaches were used for the prediction of the adsorption capacity of carbon-based materials against tetracycline and sulfamethoxazole.

Based on the previous research, this work represents the CO_2 uptake prediction onto porous carbons through the black-box nature of ML interpretation. Six ML models (GBDT, XGB, LBGM, RF, Catboost, and Adaboost) are implemented and compared to predict the adsorption behaviour of porous carbons towards CO_2 adsorption. Thus, this research is particularly aimed at (i) establishing generic ML models to predict the CO_2 adsorption onto porous carbons which are based on the adsorbent properties and CO_2 adsorption conditions, (ii) comparing and finding the best model for the prediction, and (iii) obtain the combined effect of features onto the CO_2 adsorption (Fig. 1). Global model-agnostic method (i.e., two-factor partial dependence plot) and local model-agnostic method (i.e., SHAP) are potential methods to unveil the black-box nature of ML models [26]. The novelty in this work involves the usage of Shapley additive explanations (SHAP) and a two-factor partial dependence plot. These give insight into the CO_2 adsorption mechanism based on SHAP values and the potential correlation between partial dependence of CO_2 and features (i.e., C%, H%) visualized by 3-dimensional graphs based on the ML model. SHAP is considered a unified approach to explaining the output of ML models. SHAP value is useful to understand how a single feature is affecting the output of the ML model (i.e., CO_2 uptake) which can be elucidated by plotting the curve between the SHAP values of the features and the feature value from the dataset. SHAP also explains the feature importance, in simple words, helps in finding the most important feature for the model. Overall, this study refers to a novel data-driven approach to predict and study the CO_2 adsorption onto the porous carbons which promises to help the researchers to study the carbon capture process without conducting any experiments.

2. Method

2.1. Machine learning (ML) algorithms

In the present study, six supervised ML algorithms (i.e., GBDT, XGB, LBGM, RF, Catboost, and Adaboost) were used to predict the targeted output variable i.e., CO_2 uptake. All proposed algorithms were arranged based on classification for categorical values of outputs and regression for real values of outputs respectively. The arrangement of the algorithms also involves ensemble techniques. The cluster of many algorithms within a single model was considered an ensemble algorithm. Therefore, this study represents the comparison among all the algorithms used for the prediction and finding the best algorithm for the prediction of CO_2 uptake based on accuracy. The comparison among all algorithms was considered on the basis of (i) the capability of learning each algorithm; (ii) the potentiality of handling non-linear data; (iii) the quality of handling complicated targeted variables with numerical values [20]. These algorithms are explained in short and their structural schematic diagram can be accessed in Figs. S1-S6 in the supplementary information (SI) [27,28]. Parameters of the ML models used for this study are given in Tables S1-S5 based on the scikit-learn library [29], however, model.get_params() command is used in Python (version 3.10.4) for getting the parameters used by the ML model.

2.1.1. Random forest (RF) algorithm

Breiman proposed an RF model which is a multiple decision tree-

based supervised ML ensemble modelling technique and bootstrap aggregation with nonparametric in nature [30,31]. Three steps were used for the execution of RF and can be understood in Fig. S4. The steps are; (i) the primary data set was divided into multiple subsamples of the dataset with a replacement that is randomly chosen; (ii) these subsamples were sent to decision trees where each decision trees were trained based on the subsample data, each tree is grown to the maximum size based on a bootstrap sample of the observation and each leave node's output represents the all label value's mean present in the node [32]; (iii) the final prediction of the target variable (i.e., the output of RF model) got by the average of the outputs of each decision trees [33]. For getting the accurate model prediction, hyperparameter tuning was required as the number of decision trees (N_{tree}), each node's feature numbers ($N_{feature}$), and input parameters were optimized using the trial-and-error method. For the regression problem, the RF regression predictor is concluded by Eq. (1) (after k trees $T_k(X)$ are grown) [34].

$$f(x) = \frac{\sum_{k=1}^{K} T_k(X)}{K} \tag{1}$$

Where, K = set of decision trees, X = {x_1, x_2, x_3, ….x_β}; β = input dimensional factor to make a forest.

2.1.2. Gradient boosting decision tree (GBDT) algorithm

GBDT is also called by many different names (i.e., GBRT, MART) and it is a supervised ML algorithm [35]. In GBDT, the gradient boosting technique is used for the combination of regression trees. GBDT is not as similar to the traditional boosting methods in which the weightage of positive and negative samples is considered [36]. The series combination of a weak base classifier with a strong base classifier represents the basic idea of GBDT [37]. For getting the global convergence of the algorithm, GBDT follows the negative gradient direction [38].

2.1.3. Extreme gradient boosting (XGB) algorithm

The algorithm, particularly based on the idea of 'boosting', was proposed by Chen and Guestrin, which represents the novel implementation for gradient boosting machines, particularly in the way of K classification and regression trees [39,40]. XGB deals with the overfitting problem and is used in handling many situations (i.e., sparse data, parallel computing, missing values) [41,42].

The learning of XGB is additive learning, which is explained below (Eqs. (2)-(4)). The first learner is fitted based on a complete input data set followed by the fitting of the second model based on residuals to deal with the drawbacks of the weak learners. At step t, the prediction function is represented by Eq. (2).

$$f_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = f_i^{(t-1)} + f_t(x_i) \tag{2}$$

Where, $f_t(x_i)$ is the learner at t step, $f_i^{(t)}$ and $f_i^{(t-1)}$ are the predictions on t and (t-1) steps, respectively, and x_i is the input variable.

$$Obj^{(t)} = \sum_{k=1}^{n} L(\bar{y}_i, y_i) + \sum_{k=1}^{n} \Omega(f_i) \tag{3}$$

Where, L is the loss function, n is the used observations numbers, Ω is the regularization term which is defined by Eq. (4).

$$\Omega(f) = \gamma T + 0.5\lambda||\omega||^2 \tag{4}$$

Where, γ is the minimum loss required for upcoming leaf node partitioning, ω is the scores vector in the leaves, and λ is the regularization parameter. For a detailed explanation of XGB, Chen and Guestrin [40] must be followed.

2.1.4. Light boost gradient machine (LBGM) algorithm

LBGM is considered a GBDT-based novel algorithm proposed by Ke and colleagues, recently, [43]. LBGM is a boosting implemented decision tree (moderately weaker model) based on a gradient training structure, therefore, segmentation precision is not considered an important ingredient [44,45]. Compared to XGB; LBGM is filled with some unique features (i.e., employing histogram-based algorithms). Therefore, the utilization of memory is minimized while accelerating the

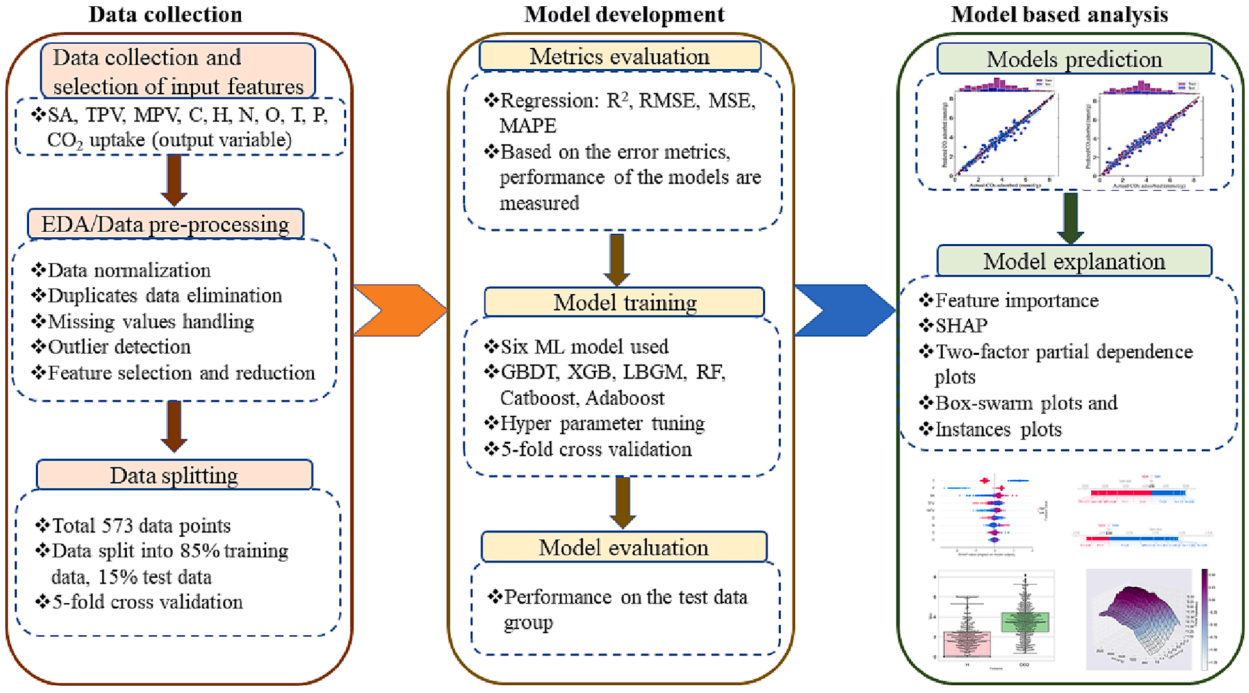


Fig. 1. Schematic diagram of workflow for this work. CO_2 adsorption data on porous carbons is collected from literature. Based on the data collected, selection of features is done. After data collection, an exploratory data analysis and data pre-processing are conducted for effective implementation of ML algorithms on the processed data. After finding the best algorithm (GBDT) for the prediction of CO_2 uptake, ML model interpretable analysis is conducted (i.e., two-factor partial dependence plots, SHAP) which gives the insight into the CO_2 adsorption mechanism onto porous carbons, and feature importance.

training sequence and the leaf-wise growth process with deep constraint is employed [46]. To deal with the over-fitting problem and for better model accuracy, the DART booster is coupled with LBGM while training the model [47]. LBGM flourishes the tree in a vertical direction compared to other tree-based algorithms (i.e., XGB, GBDT) which grow the tree horizontally. Therefore, LBGM is considered an effective algorithm while processing large-scale data sets and features [48]. The objective function of LBGM is represented by Eq. (5). However, Sun et al. [48] represented a detailed explanation of the LBGM algorithm.

$$G = \left(\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda} \right) \tag{5}$$

Where, I_L and I_R are sample sets of the left and right branches, respectively. g_i and h_i are first- and second-order gradient statistics of the loss function, respectively.

2.1.5. Catboost algorithm

Prokhorenkova et al. [49] proposed the Catboost algorithm, which is an advanced form of gradient boosting algorithm. Catboost focuses on solving the gradient bias and prediction shift, which comes up with its enhanced accuracy and generalization ability. The training process in Catboost is coupled with the process of classification features which makes it to prevent dependence on data sorting [50]. Target-based statistics, used by the traditional GBDT are described by Eq. (6).

$$\widehat{x}_k^i = \frac{\sum_{j=1}^{n} I_{\{x_j^i = x_k^i\}} \times y_j}{\sum_{j=1}^{n} I_{\{x_j^i = x_k^i\}}} \tag{6}$$

Where, I is the indicator function, and x_j^i is the i^{th} subtype feature of the k^{th} training sample.

To minimize the impact of noise and low-frequency category data, Catboost introduces the priority factor and priority weight coefficients on the data distribution, explained by Eq. (7).

$$\widehat{x}_k^i = \frac{\sum_{j=1}^{n} I_{\{x_j^i = x_k^i\}} \times y_j + \beta p}{\sum_{j=1}^{n} I_{\{x_j^i = x_k^i\}} + \beta} \tag{7}$$

Where, p is the prior value, β is the prior value weight. To tackle the overfitting problem, Catboost uses a greedy approach, where combinations are considered.

2.1.6. Adaboost algorithm

The adaboost algorithm, also known as the adaptive boosting algorithm, can be applied to the modelling of generic regression problems [45]. This algorithm got also a good hand in being a recognition algorithm. In Adaboost, multiple weak predictors are inherently integrated for the establishment of strong predictors [51]. While in the training process of Adaboost, the distribution of sample weights is done according to the quantity of error. They are high and low if the error is large and small, respectively [52]. Afterwards, output prediction is refined based on the training of samples according to the distribution of new weight [53]. The generic working principle of Adaboost is explained in the following steps: (i) initialization of the weight distribution for the first iteration, (ii) quantification of output prediction and error based on the weight distribution, (iii) calculation of proportional error, (iv) connection weight computation, and (v) updating step of weight distribution followed by the final prediction, which is acquired after P iterations by the Eq. (8) [51,54,55].

$$F(x) = \sum_{p=1}^{p} w_p f_p(x) \tag{8}$$

Where, p = 1, 2, 3, …., P represents the weight distribution at p^{th} iteration, w_p = connection weight.

2.2. Data collection and formatting

For this study, a literature survey of CO_2 adsorption has been done on different scientific databases (i.e., scopus, google scholar) for the last decade (2013–2023) and some specific research papers (i.e., CO_2 adsorption on to porous carbons) from reputed journals were selected. These publications represented the work in two ways; (i) synthesis and characterization of adsorbent, (ii) CO_2 capture application of adsorbent by adsorption-based instruments at different adsorption conditions (i.e., T = 0℃, 25℃, P = 1 bar, 0.15 bar). The data is collected directly from the tables given in the publications and where ever needed, data is also scratched from the graphs by the software WebPlotDigitizer [56]. Based on the collected data of porous carbons used in this study for the prediction of CO_2 adsorption, the porous carbons are found mainly in waste feed derived (i.e., bee-collected pollen [57], sawdust [58]). From the perspective of pore size, they are found microporous and hierarchical. However, many of them are enriched with micropores along with some mesopores. Therefore, they are classified into microporous and hierarchical sample types. Collected data of CO_2 adsorption onto porous carbons are referred to in Table S6 in the SI file. While data collection, the following points were kept in mind:

(i) All the data were scratched efficiently to avoid repetition and duplicate entries of data.
(ii) Where ever data of some features were not available, those points were kept empty without any previous judgment.
(iii) All the data was collected for the features which are affecting the CO_2 uptake along with the targeted output variable (CO_2 uptake in this study). Textural properties and elemental composition of adsorbent (i.e., pore surface area, SA (m^2/g), total pore volume, TPV (cm^3/g), micro-pore volume, MPV (cm^3/g), and C%, H%, N %, O%, S%), respectively, were considered as the features for data collection along with the different adsorption conditions of CO_2 uptake (i.e., T (℃), P (bar)).
(iv) Textural properties, the elemental composition of adsorbent, and adsorption conditions are considered input variables to the ML models.
(v) CO_2 uptake is considered the output variable of the ML models.

2.3. Pre-processing of data

After data collection and feature selection, data of particular features were converted into the same units for unit consistency (i.e., all CO_2 uptake data has mmol/g unit). This data was then arranged into 573 rows and 11 columns including the output variable (CO_2 uptake). Some missing data was found for TPV and MPV features. Some authors only represented SA with TPV where MPV was missing, and some represented SA with MPV where TPV was missing. Therefore, for prediction and imputation of the missing values through ML algorithms (i.e., RF, multilinear regression) were implemented.

Pearson's correlation coefficient (PCC) is used to find the linear dependency between any two features (i.e., TPV, MPV, SA) or between any feature and targeted output variable (i.e., CO_2 uptake). PCC is given by the Eq. (9):

$$\eta_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\sum_{i=1}^{n}(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{9}$$

Where, η_{xy} is the PCC value between any two features or between feature and target variable; $\bar{x}$ is the mean of input feature x; and $\bar{y}$ is the mean of target output y.

In this work, PCC is used to find the collinearity between any two features which can be understood by checking its value. PCC value lies in the range of −1 to 1, where 0 represents no linear correlation found while tending PCC value toward +1 and −1 represents high or positive

correlation and negative correlation, respectively.

Before training the ML models (i.e., GBDT, XGB, LBGM, RF, Catboost, Adaboost), the data of input features are normalized according to the Z-score standardization technique given in Eq. (10).

$$x_i^* = \frac{(x_i - \mu)}{\sigma} \tag{10}$$

Where x_i* is the normalized value of the input features; x_i is the original value of input features, μ is the mean value of x_i; and σ is the standard deviation of x_i.

2.4. Error metrics

For checking the accuracy of six applied ML algorithms, statistical and graphical analyses were carried out to find out the performance of applied algorithms. R^2, mean squared error (MSE), root mean square error (RMSE), and mean absolute percentage error (MAPE) were tabulated for applied algorithms. The higher the R^2 value and the lower the MAPE, RMSE and MSE values, the better would be the model prediction for the CO_2 uptake onto the porous carbon adsorbent. R^2, MSE, RMSE, and MAPE are explained by Eqs. (11)-(14), respectively.

$$R^2 = 1 - \frac{\sum_{n=1}^{N}\left(y_{pred} - y_{act}\right)^2}{\sum_{n=1}^{N}\left(y_{pred} - y_{mean}\right)^2} \tag{11}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_{act} - y_{pred}\right)^2 \tag{12}$$

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}\left(y_{pred} - y_{act}\right)^2}{N}} \tag{13}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_{act} - y_{pred}}{y_{act}}\right| \times 100 \tag{14}$$

Where, the data point is represented by n for any given instance, and the total number of data points is represented by N, y_{pred}, y_{act}, and y_{mean} point out the predicted, actual and mean value of the target variable (i.e., CO_2 uptake), respectively.

2.5. Shapley additive exPlainations (SHAP)

SHAP represents clear explanations of ML models' predictions through game theory [59], where input features are considered as players, and output (ML prediction) is referred to as the payout. SHAP explains the importance of the contribution way of each and every player. Since the different categories of ML models are available, therefore, SHAP has also different versions (i.e., Kernel SHAP, Linear SHAP, Deep SHAP, and Tree SHAP) [59]. Tree SHAP is used for this work to elucidate the ML models' prediction. Eq. (15) represents the Shapley values that are required to estimate the initial prediction of the model.

$$h(Z) = \Phi_0 + \sum_{i=1}^{N}\Phi_i Z_i \tag{15}$$

Where, h is the model explanation, Z is the basic feature, N is the collation size (maximum), and Φ is the feature attribution. Eqs. (16) and (17) were used to identify the contribution of each feature [45].

$$\Phi_i = \sum_{K \subseteq M\{i\}} \frac{|K|!(N - |K| - 1)!}{N!}\left[g_X(K \cup \{i\}) - g_X(K)\right] \tag{16}$$

$$g_X(K) = E[g(X)|X_K] \tag{17}$$

Where, K is the subset of the input feature, M = all input set, E[g(X) | X_K] is the function's expected value at the K subset.

2.6. Hyperparameter tuning

ML algorithms are self-learning algorithms; therefore, they are focused to adjust their internal parameter (also called model parameters or parameters in short) based on their self-learning ability. Irrespective of internal parameters, there are other parameters also attached to the ML algorithms which are not tuned automatically. Such parameters need to be adjusted before the training of any ML algorithm starts for better prediction accuracy of the ML models. These parameters are called hyperparameters [60]. There are many methods (i.e., grid search, manual search, random search, particle swarm optimization, genetic algorithm, and Bayesian optimization) available for hyperparameter tuning in the literature [10,60,61]. The grid search method is chosen for the tuning of the hyperparameters for this work because of its user-friendly implementation and a small set of features (nine features) selected for the prediction of CO_2 uptake. Open source Python (version 3.10.4) language is used for the implementation of ML algorithms with the use of open source scikit-learn library [29].

3. Result and discussion

3.1. Pearson correlation matrix

Fig. 2 represents Pearson's correlation matrix between all features. As can be seen in Fig. 2, SA, TPV, and MPV (textural properties) are strongly and positively correlated. The PCC between SA and TPV is + 0.92 which means that if SA increases then TPV increases, the value 0.92 represents the magnitude of correlation. There is also a negative correlation found between C% and O% with a PCC value of −0.93. However, there is only a weak correlation found between the remaining features, which explains these input features are affecting the output variable (i.e., CO_2 uptake) individually. Therefore, these input features help in building the ML model by contributing individually. Regarding the problem of missing values in the data (TPV and MPV), such points are not neglected or discarded. The missing data points were imputed by the RF model. The features which have a strong correlation between them, are used as the input and output features for the model which is used for the missing data imputation.

3.2. Box-swarm plots for descriptive statistics

Fig. 3 represents the box-swarm plots, which are a combination of box-plot and swarm-plot, used for the data distribution visualization (i.e., all input features including output variable (CO_2 uptake) 11 × 573 metrics) in this study. Box plot (Fig. S7) is a standardized way of understanding the distribution of data which is based on a summary of five numbers. The numbers are 'minimum (Q1–1.5 × IQR)', 'first quartile (Q1/25th percentile)', 'median (Q2/50th percentile)', 'third quartile (Q3/75th percentile)', and 'maximum (Q3 + 1.5 × IQR)', respectively. The skewness of data, data symmetry, and outliers can be easily detected using boxplots. Whereas, a swarm plot is used for the better distribution of the values. It gives the exact number of values. Plotting swarm-plot with box-plot is a good complement, where all observations along with some representation of underlying distribution can be shown.

Fig. 3 represents the visualization of data distribution of CO_2 adsorption onto porous carbons, which is explained in the following manner, in (mmol/g) unit; mean value is 3.52, the minimum value is 0.36, first quartile/25th percentile is 2.50, median/50th percentile is 3.51, third quartile/75th percentile is 4.41, and the maximum value is 8.20. The Christmas tree-like structure of black points in Fig. 3 represents the swarm plot in the box plot, the number of points represents the exact number of values distributed. The distribution of SA of porous carbons is explained as follows (m^2/g unit): the range is (2.5 – 3337.0) with the mean value 1515.4, where 2.5 is minimum SA and 3337.0 is maximum SA; first quartile/25th percentile, median/50th percentile and, third quartile/75th percentile are 982, 1446, and 2053, respectively.

Fig. 2. Pearson's correlation matrix for all the features (including the output variable; CO_2 uptake) which are affecting the output variable, CO_2 uptake.

Data distribution of TPV and MPV is expressed in the following way (cm^3/g unit); minimum and maximum are 0.02, 3.09, and 0.04, 1.32 with the mean value of 0.78 and 0.53, respectively; first quartile/25th percentile, median/50th percentile, and third quartile/75th percentile are 0.45, 0.68, 1.03, and 0.31, 0.48, 0.71, respectively. For the remaining input features (i.e., C%, H%, N%, O%, T (℃), and P (bar), all the data distribution is explained in Table 1 and the data visualization of C%, O%, and H% is expressed in Fig. 3.

3.3. ML model prediction

Six ML models (GBDT, XGB, LBGM, RF, Catboost, and Adaboost) are used for the prediction of CO_2 uptake on the porous carbons for this study. These models used physicochemical properties (SA, TPV, MPV, C %, O%, N%, H%) of the adsorbent (porous carbon) and adsorption conditions (T℃, P bar) as the input parameters. In this study, data points are divided into two parts: (i) training data which is 85% of the total data, and (ii) test data which is 15% of the total data. These models were trained based on the training data (85% of total data) of the features and the model evaluation is done on the basis of test data (15% of total data). Since the training data of ML models were obtained from the adsorption isotherm of CO_2 onto porous carbons, some references [57,58] are referred to for a better understanding of the adsorption isotherm data. While training the model based on the training data, hyperparameter tuning is conducted by implementing a grid search technique coupled with five-fold cross-validation (Fig. S8). K-fold cross-validation is the technique that is used to prevent the overfitting problem in the ML model's training process. This technique also helps ML models to give better and more accurate predictions [62]. In the hyperparameter tuning, the model's prediction accuracy increases because model parameters are set on the optimal values for the prediction. After hyperparameter tuning (optimized model), model prediction accuracy is evaluated on the basis of the test data set. Fig. 4 represents the prediction of all models (i.e., GBDT, XGB, LBGM, RF, Catboost, and Adaboost). Typically, Fig. 4 presents the scatter plot of actual values (experimental) versus all models' predicted values of CO_2 adsorption on the porous carbon. As can be seen in Fig. 4, a classic comparison of all models is explained in the form of training and testing data R^2. Here, all the proposed models explained the close fit of the data around the line of equality ($y = x$), represented by black colour in Fig. 4, except the Adaboost model. Purple points and blue points are representing the train and test data points in each model prediction, respectively. The blue shade is representing the regression line of 95% confidence intervals based on the test data points. Among all the six models, the GBDT model gives the best prediction based on the value of R^2 (goodness of fit) on the test data set (test $R^2 = 0.88$) compared to all other models. All other models are not having the test R^2 value as high as the GBDT model. The training and testing R^2 values for the model GBDT, XGB, LBGM, RF, Catboost, and Adaboost are 0.99, 0.99, 0.94, 0.97, 0.98, and 0.74; and 0.89, 0.86, 0.84, 0.83, 0.87, and 0.71, respectively. Another justification for the GBDT model to be the best one among all models is, it has the lowest MAPE and RMSE values which are 12% and 0.48, respectively. The remaining values of RMSE, MSE, and MAPE for training, testing, and the full dataset are given in Table 2 for all six models. These ML models are directly trained on the basis of real-life experimental data from the literature. There are no assumptions deployed in these ML models like

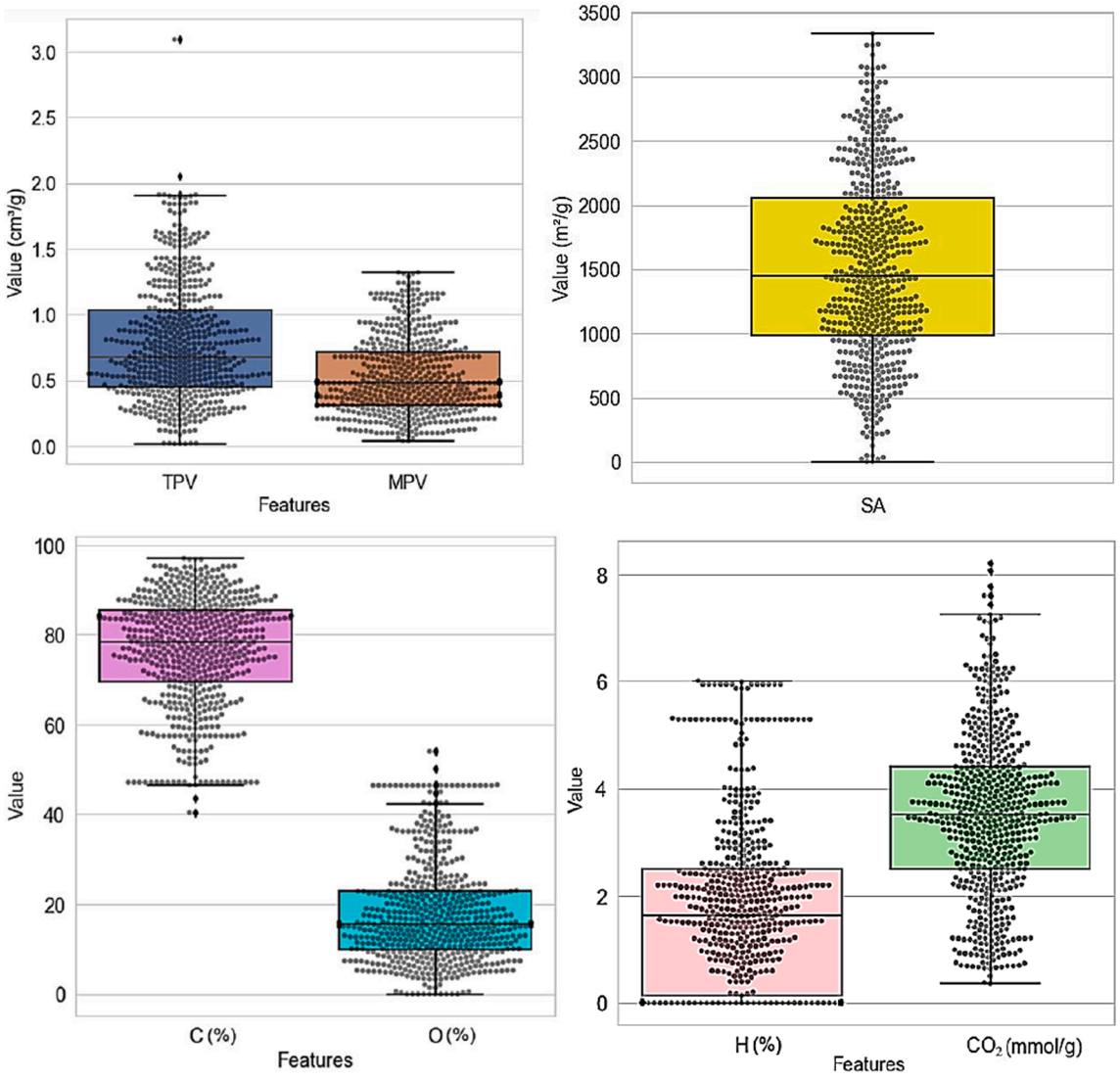


Fig. 3. Box-swarm plots of input features along with CO_2 uptake. These plots represent the descriptive statistics of distribution of data graphically. Here, swarm plot is combined with the box-plot for the better understating of the data distribution (i.e., SA in the range of 2.5–3337 m^2/g), however, swarm plot represents the number of points distribution. Fig. S7 is referred for the structural schematic diagram of box-plot.

Table 1
Summary of porous carbon's property (i.e., SA (m^2/g), TPV (cm^3/g), MPV (cm^3/g), C%, H%, N%, O%), adsorption conditions (T°C, P bar) and CO_2 uptake (mmol/g) based on the data collected from literature.

Variables	SA (m^2/g)	TPV (cm^3/g)	MPV (cm^3/g)	C (%)	H (%)	N (%)	O (%)	T (°C)	P (bar)	CO_2 uptake (mmol/g)
count	573	573	573	573	573	573	573	573	573	573
mean	1515.4	0.78	0.53	76.28	1.79	3.29	18.08	17.98	0.88	3.52
std	728.5	0.46	0.29	12.42	1.59	3.70	11.48	11.25	0.30	1.55
min	2.5	0.02	0.04	40.40	0.00	0.00	0.00	0.00	0.10	0.36
25%	982.0	0.45	0.31	69.51	0.14	0.75	9.92	0.00	1.00	2.50
50%	1446.0	0.68	0.48	78.36	1.63	2.20	15.65	25.00	1.00	3.51
75%	2053.0	1.03	0.71	85.36	2.50	4.45	22.93	25.00	1.00	4.41
max	3337.0	3.09	1.32	96.96	6.01	38.15	54.00	30.00	1.01	8.20

the traditional adsorption isotherm models (i.e., Langmuir and Freundlich isotherms). Therefore, the generalization ability and predictions of ML models can be trusted more by researchers and could be used in place of those traditional models. However, the goodness of ML models depends on the quantity and quality of data of input and output features. Thus, the deployment of ML models to understand the adsorption mechanism of CO_2 uptake onto porous carbons reduces the experimental time consumption and cost of designing and application of the adsorbent. At last, this study can refer to the implementation of the ML model to understand the adsorbate-adsorbent mechanism process.

3.4. *Two-factor partial dependence plots*

Two-factor partial dependence plots were employed in this study the demonstration how combined two factors are affecting CO_2 uptake. Fig. 5 represents the dependence of CO_2 uptake on the different pairs of

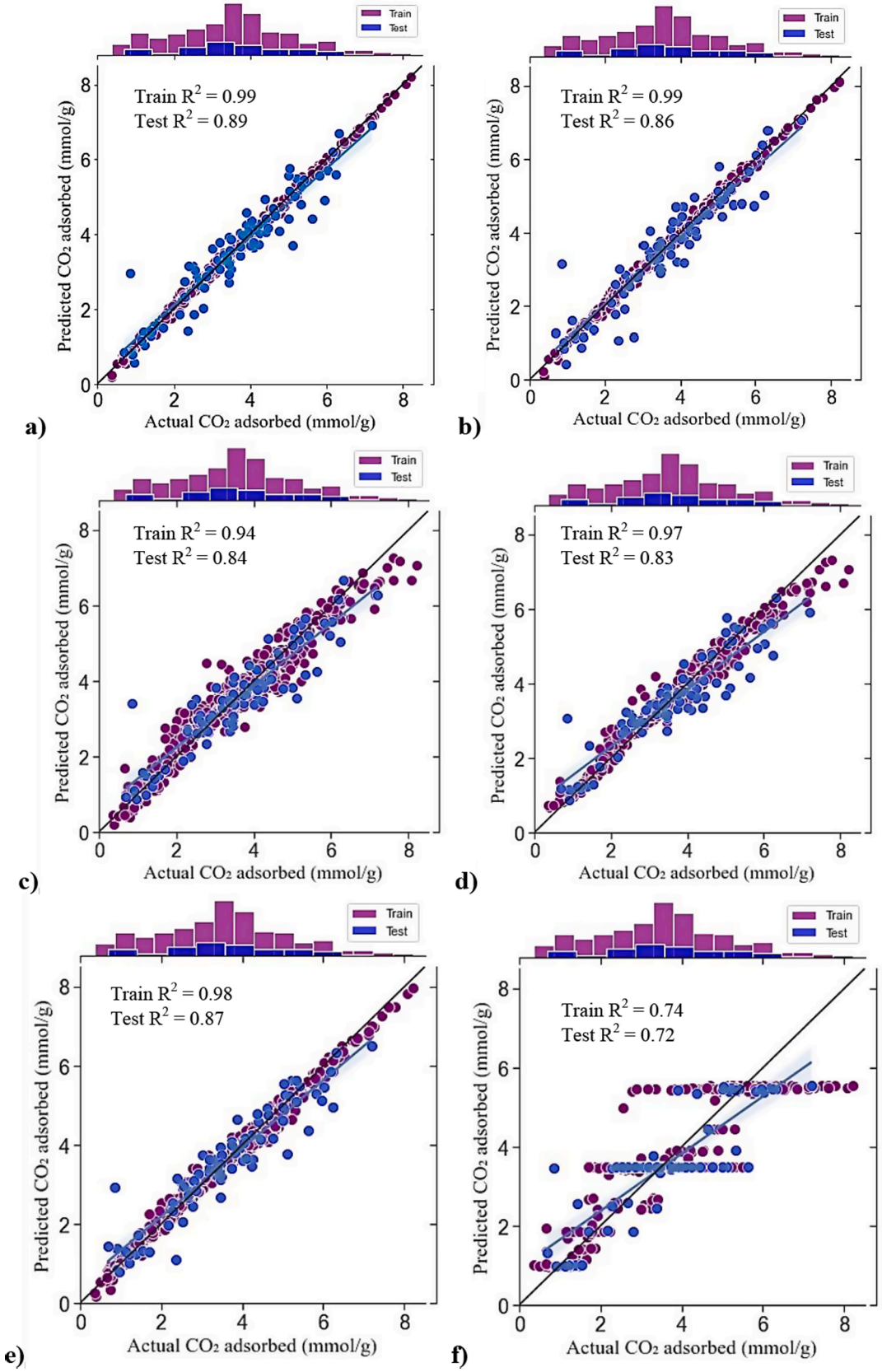


Fig. 4. Predicted CO_2 adsorbed onto porous carbon by machine learning models (a) GBDT, (b) XGB, (c) LBGM, (d) RF, (e) Catboost and, (f) Adaboost. The comparative study of model can be evaluated by comparing the prediction on the train and test data sets. Five-fold cross validation is implemented onto all models. Based on the R^2 value GBDT model which obtained train $R^2 = 0.99$ and test $R^2 = 0.89$, gave the best prediction result.

Table 2
Assessment of prediction accuracy for CO_2 uptake by ML algorithms (GBDT, XGB, LBGM, RF, Catboost, and Adaboost) used in this study, compared for training data (85% of all data), testing data (15% of all data) sets. Based on the value of R^2, MSE, RMSE and MAPE, GBDT model is found best model, having good accuracy (training $R^2 = 0.99$, test $R^2 = 0.89$) and lowest MAPE (12%).

Model type	Dataset	R^2	MSE	RMSE	MAPE
GBDT	Training	0.99	0.006	0.075	12%
	Test	0.89	0.235	0.480	
XGB	Training	0.99	0.006	0.075	15%
	Test	0.86	0.291	0.540	
LBGM	Training	0.94	0.147	0.384	14%
	Test	0.84	0.331	0.570	
RF	Training	0.97	0.071	0.267	15%
	Test	0.83	0.337	0.580	
Catboost	Training	0.98	0.032	0.178	14%
	Test	0.87	0.256	0.501	
Adaboost	Training	0.74	0.640	0.800	20%
	Test	0.72	0.583	0.760	

combined features (i.e., SA-TPV, SA-MPV). As shown in Fig. 5a, the combined effect of SA and TPV on CO_2 adsorption is described as GBDT model accesses. Fig. 5a clearly states that as soon as SA increases the partial dependence increases. A high slope of the curve is observed for the SA range of (500–1400 m^2/g) which explains for this regime partial dependence increases quickly. For the SA greater than 1400 m^2/g, a high slope of the curve is not observed, however, little variation in the curve is monitored but partial dependence still increases. Maximum partial dependence is reported for SA in the range of 2000–2300 m^2/g and TPV for 0.3–0.8 cm^3/g, which is clearly examined by the dark colour side of the curve. The higher the dark colour of the curve, the higher the partial dependence would be. The above explanation states that the porous carbon materials having SA in the range of 2000–2300 m^2/g and TPV in the range of 0.3–0.8 cm^3/g, are more viable for CO_2 adsorption. Further, Fig. 5b elucidates the combined effect of SA and MPV on CO_2 uptake. According to Fig. 5b, high partial dependence is witnessed when the SA ranged from 2000-2300 m^2/g and MPV from 0.4 to 0.8 cm^3/g because of the combined effect. Next, Fig. 5c represents the combined effect of SA with nitrogen percentage in the adsorbent. For which the partial dependence was found higher for SA in the range of 2000–2300 m^2/g and N% in 6–8. The partial dependence of CO_2 uptake on SA and O% is shown in Fig. 5d, where it is observed high for SA in the range of 2000–2300 m^2/g and for O% in the range of 10–15. Temperature is considered the most critical factor for adsorptive behaviour as it is involved in the mass transfer kinetics and adsorbate-adsorbent interactions [63]. And it is also the most affecting feature among all features of the CO_2 uptake according to the SHAP global explanation (Fig. 6). The combined effect of SA and the adsorption temperature is shown in Fig. 5e, where the high partial dependence is found for SA (2000–2300 m^2/g) and low values of T (0-5℃). It clearly shows that the higher values of temperature are not in favour of significant CO_2 adsorption. There is a wide spectrum found for the MPV (0.6–1.1 cm^3/g) combined with TPV (0.3–0.6 cm^3/g) feature for the higher partial dependence (Fig. 5f). There are a lot of spikes observed in Fig. 5g which are dense in the colour, and clearly stating the higher partial dependence at those points. These spikes are found for C% in the range of 60–70 and 80–90, and H% in 0.50–30 (Fig. 5g). Partial dependence on N% and H% is shown in Fig. 5h, it is observed higher for N% and H% in the range of 6–10 and 0.5–2, respectively.

3.5. SHAP explanations

Consequently, there are no proper ML interpretability methods used for the explanation of the ML model predictions in the previous studies of CO_2 adsorption. A post-hoc explanation is intended for the models when they become complex. Therefore, the SHAP explanation is used for the inherent processing of the models, affected outcome due to the interaction of each variable, and a given instance explanation. For being the best model among all the models, GBDT is used for the SHAP explanation. Fig. 6 represents the SHAP summary plot, which gives the bird's eye view of the feature's importance and explains its driving strategy. Colour coding is used in Fig. 6 for the explanation of the feature values, where pink colour depicts the higher feature values and blue colour depicts the lower feature values. Fig. 6 is the combination of many dots of different colours; where each dot contains three characteristics: (i) vertical location (y-axis of Fig. 6) of the dot represents the feature which it is depicting, (ii) colour of the dot represents the intensity of the feature whether the feature is high or low for that particular row (blue and red colour represent low and high intensity respectively), (iii) horizontal location of the dot (x-axis) represents the SHAP value of that particular feature which elucidate how much it is causing the prediction (low or high). According to Fig. 6, the temperature feature (T) is considered to be a much-influencing factor on CO_2 uptake prediction as the GBDT algorithm observed. As higher and lower values of T can be precisely observed in the negative section and positive section regimes of Fig. 6, respectively. Therefore, higher values are reducing the prediction by accompanying the particular SHAP values (-1–0) and lower values help in increasing the prediction by SHAP values in the range of 0.8–2. Further, the pressure feature implied the most effect on the CO_2 uptake. However, the pressure feature indicated a completely reverse effect on CO_2 uptake compared to the temperature feature. Lower values of pressure lie in the negative section of Fig. 6 which clearly shows that lower pressure help in decreasing prediction, and higher pressure helps in increasing the prediction. This explanation can also be assured by the experimental observation of CO_2 adsorption [64] where higher CO_2 uptake is found at higher optimal pressure values. Next, the surface area appears to be the dominant feature that is affecting the output variable CO_2 uptake. According to Fig. 6, a higher value of the surface area, shown in red colour, leads to a positive considerable impact on the prediction. The lump of higher SA values in Fig. 6 represents that most of them are in between the same patch. Lower SA values are not in favour of significant CO_2 adsorption. The above explanation of the effect of the SA feature proved a good match with the experimental explanation, where higher SA of materials helps in significant CO_2 adsorption [65].

This proves that ML model predictions and model explanations are excellent for this study. Here the increase in prediction represents that the particular feature dominantly helps in the CO_2 adsorption (output variable) on the porous material. Lower TPV content has a notable effect on CO_2 uptake, as TPV content increases it tends to have a negative effect on the prediction. Further, higher MPV content in the adsorbent helps in the prediction while lower values reduce prediction. The effect of compositional features (O%, N%, C%, H%) of porous carbons on CO_2 uptake is comparable with each other. As is clearly mentioned in Fig. 6, oxygen is the most affecting feature of CO_2 uptake than nitrogen, carbon, and hydrogen, respectively. Higher oxygen content leads to a considerable negative impact on CO_2 uptake and vice-versa. However, nitrogen content had a completely reversed influence on prediction compared to oxygen content. Carbon content is not affecting the CO_2 uptake as much as oxygen and nitrogen do. A slightly pronounced variation was detected from the hydrogen content on the CO_2 uptake. However, the hydrogen content is found the least affecting feature among all the compositional features. The range of hydrogen in the porous carbon adsorbent used in this study for CO_2 adsorption is 0–6% Table 1. That is of course not a wide range. According to Fig. 6, SHAP considered the maximum value of hydrogen (6%) as a low value even if it could be considered a significant value. Thus, the feature range is also considered an important factor in the model explanation. Besides SHAP global explanation or summary plot Fig. 6, the local explanation also known as a particular prediction for selected two instances which is obtained from the SHAP force plot is shown in Fig. 7.

However, for the interpretation of individual feature importance, SHAP values are especially useful. For the interpretation to be more

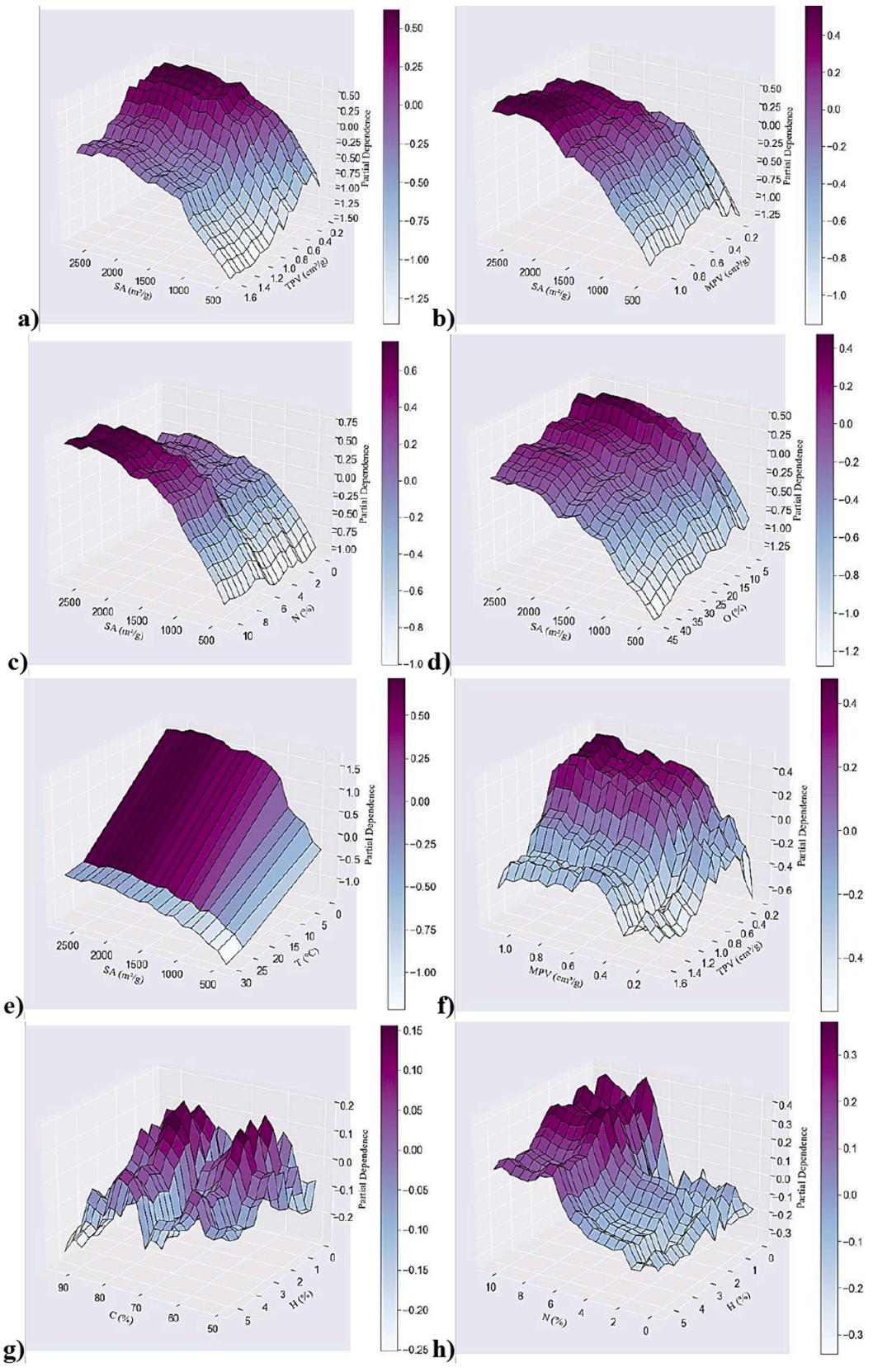


Fig. 5. Combined factor partial dependence of CO_2 uptake on different sets of influential variables (input parameters). Figures of combined factor partial dependence are given in order: **(a)** SA (m^2/g) – TPV (cm^3/g), **(b)** SA(m^2/g) – MPV (cm^3/g), **(c)** SA (m^2/g) **–** N (%), **(d)** SA(m^2/g) – O (%), **(e)** SA(m^2/g) – Temperature (℃), **(f)** MPV (cm^3/g) – TPV (cm^3/g), **(g)** C (%) – H (%), **(h)** N (%) – H (%)) based on GBDT algorithm. These plots revealed the relationship between input parameters (i.e., N%, H %, and CO_2 uptake (mmol/g)) graphically and quantified the range of parameters for the maximum CO_2 adsorption.

intuitive, the abstraction from the general model should be lowered to a single prediction. This is explained by the force plot in Fig. 7. Where SHAP values of the final prediction are represented on the horizontal axis. And each feature's contribution is represented by the blocks shown in pink and blue colours. The feature pushing the prediction higher is shown in pink color and the feature pushing the prediction lower is shown in blue color. The distance of the base value (3.516) to the output value or function value shown in a dark colour (2.64 and 4.16 for Fig. 7a

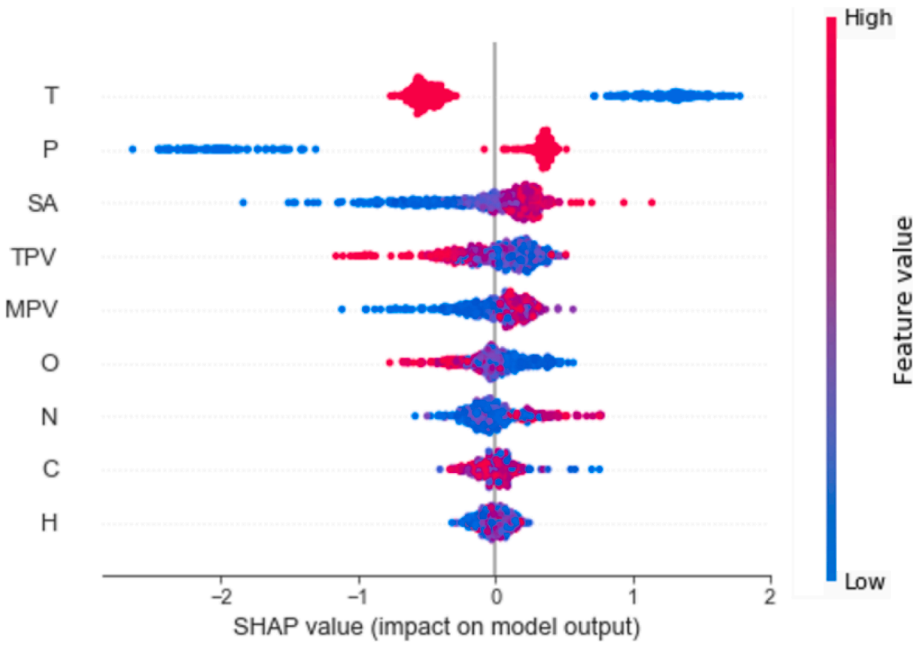
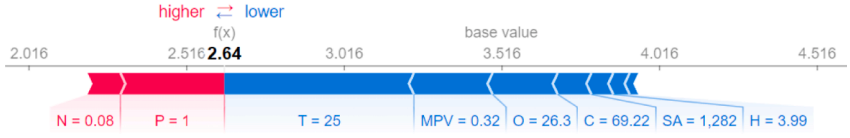
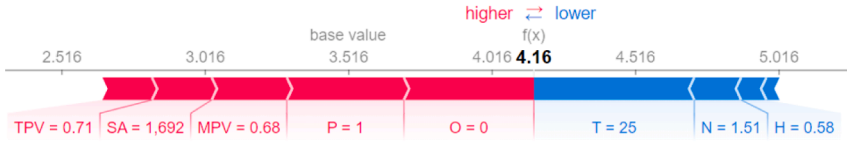


Fig. 6. SHAP global explanation which is based on GBDT model. This explanation represents the feature importance analysis indicating the features that are affecting the CO_2 adsorption uptake onto porous carbons based on the SHAP value. The highest impact on CO_2 uptake came from temperature T (℃) (adsorption condition), and the lowest came from H% which is the property of porous carbon used in this study.



(a) Instance force plot 1: N%, P (bar), T (°C), MPV (cm^3/g), O%, C%, SA (m^2/g), H%.



(b) Instance force plot 2: TPV (cm^3/g), SA (m^2/g), MPV (cm^3/g), P(bar), O%, T (°C), N%, H%.

Fig. 7. SHAP force plot on selected instances for local interpretation. Two instances (a) and (b) have been explored to understand the CO_2 adsorption onto porous carbon based on the SHAP value. As a comparison of instance (a) with (b), higher value of SA = 1692 m^2/g is in more favor (red color) of CO_2 adsorption. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and b, respectively) is exactly the same as when the length of all red bars is subtracted from the length of all blue bars. The bolded value is the summation of the SHAP value and base value and represents the prediction obtained while the ML training sequence. Therefore, it elucidates features that have unique contributions and interactions toward CO_2 uptake at any given instance. Fig. 7a represents an instance for the local explanation, where the predicted value or function value is 2.64 and the base value is 3.516. Most of the features (T, MPV, O%, C%, SA%, H%) at their particular values are not in the favour of CO_2 uptake (Fig. 7a). They have a negative influence on the prediction where the largest impact comes from the T = 25℃ then MPV = 0.32 cm^3/g, O% = 26.3, C% = 69.22, SA = 1282 m^2/g, H% = 3.99, respectively, Fig. 7a. For this instance, only two features account in the favour of the CO_2 uptake, P = 1 bar and N% = 0.08. For another instance, Fig. 7b, where, O% = 0, P = 1 bar, MPV = 0.68 cm^3/g, SA = 1692 m^2/g, TPV = 0.71 cm^3/g have the positive impact and T = 25℃, N% = 1.51, H% = 0.58 have the negative impact on the CO_2 uptake. Comparing these two instances, lower SA content (1282 m^2/g) in the porous carbons pushes CO_2 uptake towards lower values while higher SA content (1692 m^2/g) pushes CO_2 uptake base value towards the positive side. Likewise, the interpretation of comparison could be conducted for the remaining features for two instances.

At last, SHAP explanations elucidated the numerous detailed resolutions for each instance and the global interpretation of the model. The inner working of ML models is clearly explained through SHAP explanations without aiding the help of another user to be involved. Therefore, non-technical audiences can also use SHAP in the decision-making section. The local explanations by SHAP acquired the logic behind the ML model prediction in a very significant way. Thus, we can refer to the involvement of ML interpretability methods in the section on decision-making for chemical engineering applications. In the long run, this will keep the faith between various domain experts on ML techniques.

3.6. Study limitation

Besides keeping the fact in mind, there are also some limitations of any study. SHAP explanation is used for this study for which the limitations are explained below:

(i) There are various post-hoc methods (i.e., LIME) available for the model prediction analysis but this study involved SHAP explanation. Therefore, a comparison between different methods of explanation can be introduced for a more lucid explanation. But one should keep in mind that the interpretations obtained from different methods might be different from each other. For example, the feature importance obtained from SHAP can be different from other methods' feature importance.

(ii) This study involved the explanation of the adsorption of CO_2 onto porous materials using ML techniques. And further prediction interpretation is continued through the SHAP explanation. However, this study can also be applied for the explanation of the adsorption of different gases (i.e., CH_4, H_2) onto different adsorbents (i.e., metal–organic frameworks). ML methods accompanied by any post-hoc explanation methods (i.e., SHAP, LIME) will help to predict adsorption efficiency for other adsorbents and adsorbates and can explain the underlying reasoning. Accordingly, the identification of features/parameters for the particular study can be done at any point in the design stage of modelling.

3.7. Relevance of ML in CO_2 adsorption

The application of ML is growing rapidly not only in the field of CO_2 adsorption but also in gas adsorption, along with the discovery of porous materials for carbon capture [66–69] (Table 3). Although ML models are considered black-box in nature, it is still a challenge to find the direct quantitative correlation between the output variable (CO_2 uptake: adsorption characteristic) and the features (structural parameters of porous carbon) [26]. To overcome this issue, two-factor partial dependence plots based on the ML model are implemented in this study to reveal insight into the CO_2 adsorption mechanism. They reveal a potential correlation between CO_2 uptake (adsorption characteristic) in the form of partial dependence, and structural parameters (C%, O%, N%, H%) in the form of three-dimensional graphs (Fig. 5 and section 3.4). For getting the direct correlation between adsorption characteristics and structural parameters, the Pearson correlation matrix has been used (Fig. 2). It provides the quantitative assessment between two features (i.e., CO_2 uptake and surface area) based on their linear relationship.

4. Conclusions

This research expresses an insight into the CO_2 adsorption onto porous carbons using ML approaches, which are based on porous carbon's properties and experimental CO_2 adsorption conditions. GBDT has been found as the best model for the prediction of CO_2 uptake among all models (XGB, LBGM, RF, Catboost, and Adaboost) which is having train and test R^2 values of 0.99 and 0.89, respectively. GBDT model is having the lowest MAPE (12%) among all models used for this study. Two-factor partial dependence plots revealed the effect of the combined feature on CO_2 uptake. These plots state the relationship among the structural parameters (i.e., C%, H%) of porous carbons, adsorption conditions (i.e., T℃), and partial dependence of CO_2 in 3-dimensional graph form (Fig. 5). Moreover, a suitable range of features are explored for getting the highest CO_2 uptake onto porous carbons (i.e., SA range 2000–2300 m^2/g). SHAP, a local model-agnostic method, is used for interpretable ML [26]. According to the SHAP explanation, feature importance analysis is conducted, where the highest impact on CO_2 adsorption onto porous carbons comes from the temperature (T℃) followed by pressure (P bar) which are the experimental conditions of the adsorption process. Another feature of importance comes from the properties of the porous carbons which are used for CO_2 adsorption in this study. In which, the feature impact on CO_2 uptake is given as follows: SA (m^2/g) > TPV (cm^3/g) > MPV (cm^3/g) > O% > N% > C% > H%. SHAP also explained two instances of adsorption for the eloquent understanding of CO_2 adsorption onto porous carbons. This study can potentially help researchers in the selection and rational designing of porous carbons for the adsorption of CO_2 and be insightful to understand the mechanism of similar types of problems.

Table 3
Prediction performance comparison between the previous work and this study. This comparison gives a lucid idea of ML used in the previous literature for CO_2 adsorption.

Purpose	Output variable	Model used	Performance	Reference
Understanding CO_2 adsorption mechanism and prediction	CO_2 uptake	GBDT, XGB, LBGM, RF, Catboost, Adaboost	R^2 = 0.99 MSE = 0.006 MAPE = 12%	This work
CO_2 gas adsorptive control by zeolite using ML models	predicted CO_2 adsorption	artificial intelligence algorithms	R^2 = 0.999 MSE = 0.0012 %AARD = 3.8266	[70]
CO_2 uptake in metal organic frameworks	CO_2 uptake	MLP, MLP-LMA, MLP-BR, ELM, GP	R^2 = 0.999 RMSE = 0.1319 %AARD = 0.2607	[3]
ML for prediction of CO_2 on porous carbon	CO_2 uptake	GBTs	R^2 = 0.84	[10]
ANN for CO_2 prediction in MOF	CO_2 uptake	ANN	R^2 = 0.99	[71]
CO_2 reduction	adsorption energies	ANN, KNN, DNN	–	[72]
ML used for the CO_2 prediction	CO_2 adsorption	RF, SVM, DT	R^2 = 0.91	[73]
CO_2 adsorption on porous carbon at different pressure	CO_2 adsorption	RF	R^2 = 0.95 RMSE = 0.148	[74]
Deep learning models for CO_2 adsorption prediction on MOF	CO_2 adsorption	ANN	R^2 = 0.96	[75]

CRediT authorship contribution statement

Sarvesh Namdeo: Conceptualization, Methodology, Data curation, Writing – original draft. **Vimal Chandra Srivastava:** Supervision, Conceptualization, Writing – review & editing. **Paritosh Mohanty:** Supervision, Data curation, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data is already given in SI.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jcis.2023.05.052.

References

[1] Global Monitoring Laboratory, Earth System Research Laboratories, Trends Atmos. Carbon Dioxide, (2022) (https://gml.noaa.gov/ccgg/trends/global.html).
[2] F. Chen, J. Wang, L. Guo, X. Huang, Z. Zhang, Q. Yang, Y. Yang, Q. Ren, Z. Bao, Carbon dioxide capture in gallate-based metal-organic frameworks, Sep. Purif. Technol. 292 (2022), 121031.
[3] M.N. Amar, H. Ouaer, M.A. Ghriga, Robust smart schemes for modeling carbon dioxide uptake in metal– organic frameworks, Fuel 311 (2022), 122545.
[4] R.G. Watts, Global Warming and the Future of the Earth, Synth. Lect. Energy. Environ. Tech., sci., soc. 1 (1) (2007) 1–114.
[5] M. Chaudhary, R. Muhammad, C.N. Ramachandran, P. Mohanty, Nitrogen amelioration-driven carbon dioxide capture by nanoporous polytriazine, Langmuir 35 (14) (2019) 4893–4901.
[6] A. Dai, Drought under global warming: a review, Wires. Clim. Change. 2 (1) (2011) 45–65.
[7] K.J. Walsh, J.L. McBride, P.J. Klotzbach, S. Balachandran, S.J. Camargo, G. Holland, T.R. Knutson, J.P. Kossin, T.C. Lee, A. Sobel, M. Sugi, Tropical cyclones and climate change, Wires. Clim. Change. 7 (1) (2016) 65–89.
[8] M.A. Bender, T.R. Knutson, R.E. Tuleya, J.J. Sirutis, G.A. Vecchi, S.T. Garner, I. M. Held, Modeled impact of anthropogenic warming on the frequency of intense Atlantic hurricanes, Sci. 327 (5964) (2010) 454–458.
[9] J.P. Ciferno, T.E. Fout, A.P. Jones, J.T. Murphy, Capturing carbon from existing coal-fired power plants, Chem. Eng. Prog. 105 (4) (2009) 33.
[10] X. Yuan, M. Suvarna, S. Low, P.D. Dissanayake, K.B. Lee, J. Li, X. Wang, Y.S. Ok, Applied machine learning for prediction of CO_2 adsorption on biomass waste-derived porous carbons, Env. Sci. Technol. 55 (17) (2021) 11925–11936.
[11] X. Yuan, J.G. Lee, H. Yun, S. Deng, Y.J. Kim, J.E. Lee, S.K. Kwak, K.B. Lee, Solving two environmental issues simultaneously: Waste polyethylene terephthalate plastic bottle-derived microporous carbons for capturing CO_2, Chem. Eng. J. 397 (2020), 125350.
[12] Z. Deng, Y.i. Liu, M. Wan, S. Ge, Z. Zhao, J. Chen, S. Chen, S. Deng, J. Wang, Breaking trade-off effect of Xe/Kr separation on microporous and heteroatoms-rich carbon adsorbents, Sep. Purif. Technol. 308 (2023) 122942.
[13] X. Ma, W. Xu, R. Su, L. Shao, Z. Zeng, L. Li, H. Wang, Insights into CO_2 capture in porous carbons from machine learning, experiments and molecular simulation, Sep. Purif. Technol. 306 (2023), 122521.
[14] J. Wang, Z. Guo, S. Deng, R. Zhao, L. Chen, J. Xue, A rapid multi-objective optimization of pressure and temperature swing adsorption for CO_2 capture based on simplified equilibrium model, Sep. Purif. Technol. 279 (2021), 119663.
[15] C. Trinh, D. Meimaroglou, S. Hoppe, Machine Learning in Chemical Product Engineering: The State of the Art and a Guide for Newcomers, Process. 9 (8) (2021) 1456.
[16] F. Cichos, K. Gustavsson, B. Mehlig, G. Volpe, Machine learning for active matter, Nat. Mach. Intell. 2 (2) (2020) 94–103.
[17] M.R. Dobbelaere, P.P. Plehiers, R. Van de Vijver, C.V. Stevens, K.M. Van Geem, Machine learning in chemical engineering: strengths, weaknesses, opportunities, and threats, Eng. 7 (9) (2021) 1201–1211.
[18] Z. Hajjar, S. Tayyebi, M.H.E. Ahmadi, Application of AI in chemical engineering, Artif. Intell. Emerg. Trends Appl. (2018) 399–415.
[19] V.M. Nagulapati, H.M.R.U. Rehman, J. Haider, M.A. Qyyum, G.S. Choi, H. Lim, Hybrid machine learning-based model for solubilities prediction of various gases in deep eutectic solvent for rigorous process design of hydrogen purification, Sep. Purif. Technol. 298 (2022), 121651.
[20] X.C. Nguyen, Q.V. Ly, T.T.H. Nguyen, H.T.T. Ngo, Y. Hu, Z. Zhang, Potential application of machine learning for exploring adsorption mechanisms of pharmaceuticals onto biochars, Chemosphere 287 (2022), 132203.
[21] L. Li, S. Rong, R. Wang, S. Yu, Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: A review, Chem. Eng. J. 405 (2021), 126673.
[22] M. Feng, M. Cheng, X. Ji, L. Zhou, Y. Dang, K. Bi, Z. Dai, Y. Dai, Finding the optimal CO_2 adsorption material: Prediction of multi-properties of metal-organic frameworks (MOFs) based on DeepFM, Sep. Purif. Technol. 302 (2022), 122111.
[23] L. He, L. Bai, D.D. Dionysiou, Z. Wei, R. Spinney, C. Chu, Z. Lin, R. Xiao, Applications of computational chemistry, artificial intelligence, and machine learning in aquatic chemistry research, Chem. Eng. J. 426 (2021), 131810.
[24] M. Meng, Z. Qiu, R. Zhong, Z. Liu, Y. Liu, P. Chen, Adsorption characteristics of supercritical CO_2/CH_4 on different types of coal and a machine learning approach, Chem. Eng. J. 368 (2019) 847–864.
[25] X. Zhu, Z. Wan, D.C. Tsang, M. He, D. Hou, Z. Su, J. Shang, Machine learning for the selection of carbon-based materials for tetracycline and sulfamethoxazole adsorption, Chem. Eng. J. 406 (2021), 126782.
[26] C. Molnar, Interpretable machine learning, A guide for making black box models explainable, (2019) (https://christophm.github.io/interpretable-ml-book/).
[27] N.M. Shahani, X. Zheng, X. Guo, X. Wei, Machine learning-based intelligent prediction of elastic modulus of rocks at thar coalfield, Sustainability 14 (6) (2022) 3689.
[28] S. Czarnecki, M. Hadzima-Nyarko, A. Chajec, Ł. Sadowski, Design of a machine learning model for the precise manufacturing of green cementitious composites modified with waste granite powder, Sci. Rep. 12 (1) (2022) 13242.
[29] Scikit-learn, Machine Learning in Python, (https://scikit-learn.org/stable/).
[30] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[31] X. Zhu, X. Wang, Y.S. Ok, The application of machine learning methods for prediction of metal sorption onto biochars, J. Hazard. Mater. 378 (2019), 120727.
[32] N. Dehghanian, M. Ghaedi, A. Ansari, A. Ghaedi, A. Vafaei, M. Asif, S. Agarwal, I. Tyagi, V.K. Gupta, A random forest approach for predicting the removal of Congo red from aqueous solutions by adsorption onto tin sulfide nanoparticles loaded on activated carbon, Desalin. Water Treat. 57 (20) (2016) 9272–9285.
[33] K. Were, D.T. Bui, Ø.B. Dick, B.R. Singh, A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape, Ecol. Indic. 52 (2015) 394–403.
[34] J. Zhou, E. Li, H. Wei, C. Li, Q. Qiao, D.J. Armaghani, Random forests and cubist algorithms for predicting shear strengths of rockfill materials, Appl. Sci. 9 (8) (2019) 1621.
[35] R. Sun, G. Wang, W. Zhang, L.T. Hsu, W.Y. Ochieng, A gradient boosting decision tree-based GPS signal reception classification algorithm, Appl. Soft Comput. 86 (2020), 105942.
[36] H. Rao, X. Shi, A.K. Rodrigue, J. Feng, Y. Xia, M. Elhoseny, X. Yuan, L. Gu, Feature selection based on artificial bee colony and gradient boosting decision tree, Appl. Soft Comput. 74 (2019) 634–642.
[37] H. Wang, Y. Meng, P. Yin, J. Hua, A Model-Driven Method for Quality Reviews Detection: An Ensemble Model of Feature Selection, WHICEB (2016, May).
[38] X. Yuan, M. Abouelenien, A multi-class boosting method for learning from imbalanced data, International Journal of Granular Computing, Rough Sets Intell. Syst. 4 (1) (2015) 13–29.
[39] J. Fan, X. Wang, L. Wu, H. Zhou, F. Zhang, X. Yu, X. Lu, Y. Xiang, Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China, Energ. Convers Manage. 164 (2018) 102–111.
[40] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, 22nd ACM SIGKDD Int. Conf. on Knowledge Discov. Data Min. (2016, August) 785-794.
[41] Y.C. Chang, K.H. Chang, G.J. Wu, Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions, Appl. Soft Comput. 73 (2018) 914–920.
[42] S. Wang, P. Dong, Y. Tian, A novel method of statistical line loss estimation for distribution feeders based on feeder cluster and modified XGBoost, Energies 10 (12) (2017) 2067.
[43] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, Adv. Neural Inf. Process. Syst. 30 (2017).
[44] J. Fan, X. Ma, L. Wu, F. Zhang, X. Yu, W. Zeng, Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data, Agr. Water Manage. 225 (2019), 105758.
[45] I.U. Ekanayake, D.P.P. Meddage, U. Rathnayake, A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP), Case Stud. Constr. Mater. 16 (2022) 01059.
[46] D. Chakraborty, H. Elhegazy, H. Elzarka, L. Gutierrez, A novel construction cost prediction model using hybrid natural and light gradient boosting, Adv. Eng. Inform. 46 (2020), 101201.
[47] R.K. Vinayak, R. Gilad-Bachrach, Dart: Dropouts meet multiple additive regression trees, In Artif. Intell. Stat. (2015, February) 489-497.
[48] X. Sun, M. Liu, Z. Sima, A novel cryptocurrency price trend forecasting model based on LightGBM, Financ. Res. Lett. 32 (2020), 101084.
[49] L. Prokhorenkova, G. Gusev, A. Vorobev, A.V. Dorogush, A. Gulin, CatBoost: unbiased boosting with categorical features, Adv. Neural Inf. Process. Syst. 31 (2018).
[50] C. Lu, S. Zhang, D. Xue, F. Xiao, C. Liu, Improved estimation of coalbed methane content using the revised estimate of depth and CatBoost algorithm: A case study from southern Sichuan Basin, China, Comput. Geosci. 158 (2022), 104973.
[51] J. Wang, R. Ding, F. Cao, J. Li, H. Dong, T. Shi, L. Xing, J. Liu, Comparison of state-of-the-art machine learning algorithms and data-driven optimization methods for mitigating nitrogen crossover in PEM fuel cells, Chem. Eng. J. 442 (2022) 136064.
[52] J. Lu, H. Hu, Y. Bai, Generalized radial basis function neural network based on an improved dynamic particle swarm optimization and AdaBoost algorithm, Neurocomputing 152 (2015) 305–315.
[53] T. Peng, J. Zhou, C. Zhang, Y. Zheng, Multi-step ahead wind speed forecasting using a hybrid model based on two-stage decomposition technique and AdaBoost-extreme learning machine, Energ. Convers, Manage. 153 (2017) 589–602.
[54] H. Zhao, H. Yu, D. Li, T. Mao, H. Zhu, Vehicle accident risk prediction based on AdaBoost-so in vanets, IEEE Access 7 (2019) 14549–14557.
[55] L. Wang, Y. Guo, M. Fan, X. Li, Wind speed prediction using measurements from neighboring locations and combining the extreme learning machine and the AdaBoost algorithm, Energ. Rep. 8 (2022) 1508–1518.
[56] WebPlotDigitizer 4.6, (https://apps.automeris.io/wpd/).
[57] S.W. Choi, J. Tang, V.G. Pol, K.B. Lee, Pollen-derived porous carbon by KOH activation: Effect of physicochemical structure on CO_2 adsorption, J. CO_2 Util. 29 (2019) 146–155.
[58] E.A. Hirst, A. Taylor, R. Mokaya, A simple flash carbonization route for conversion of biomass to porous carbons with high CO_2 storage capacity, J. Mater. Chem. A 6 (26) (2018) 12393–12403.
[59] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).
[60] E. Elgeldawi, A. Sayed, A.R. Galal, A.M. Zaki, Hyperparameter Tuning for Machine Learning Algorithms Used for Arabic Sentiment Analysis, Inform. 8 (2021, December) 4–79.
[61] M. Shahhosseini, G. Hu, H. Pham, Optimizing ensemble weights and hyperparameters of machine learning models for regression problems, Mach. Learn. Appl. 7 (2022) 100251.

[62] R. Maleki, S.M. Shams, Y.M. Chellehbari, S. Rezvantalab, A.M. Jahromi, M. Asadnia, R. Abbassi, T. Aminabhavi, A. Razmjou, Materials discovery of ion-selective membranes using artificial intelligence, Commun. Chem. 5 (1) (2022) 1–13.
[63] M.J. Ahmed, Adsorption of quinolone, tetracycline, and penicillin antibiotics from aqueous solution using activated carbons, Environ. Toxicol. Phar. 50 (2017) 1–10.
[64] J. Li, B. Michalkiewicz, J. Min, C. Ma, X. Chen, J. Gong, E. Mijowska, T. Tang, Selective preparation of biomass-derived porous carbon with controllable pore sizes toward highly efficient CO_2 capture, Chem. Eng. J. 360 (2019) 250–259.
[65] M.J. Kim, S.W. Choi, H. Kim, S. Mun, K.B. Lee, Simple synthesis of spent coffee ground-based microporous carbons using K_2CO_3 as an activation agent and their application to CO_2 capture, Chem. Eng. J. 397 (2020), 125404.
[66] C. Zhang, Y. Xie, C. Xie, H. Dong, L. Zhang, J. Lin, Accelerated discovery of porous materials for carbon capture by machine learning: A review, MRS Bulletin 47 (4) (2022) 432–439.
[67] H. Mashhadimoslem, A. Ghaemi, Machine learning analysis and prediction of N_2, N_2O, and O_2 adsorption on activated carbon and carbon molecular sieve, Environ. Sci. Pollut. Res. 30 (2) (2023) 4166–4186.
[68] J. Abdi, F. Hadavimoghaddam, M. Hadipoor, A. Hemmati-Sarapardeh, Modeling of CO_2 adsorption capacity by porous metal organic frameworks using advanced decision tree-based models, Sci. Rep. 11 (1) (2021) 24468.
[69] Z. Zhang, J.A. Schott, M. Liu, H. Chen, X. Lu, B.G. Sumpter, J. Fu, S. Dai, Prediction of carbon dioxide adsorption via deep learning, Angewandte Chemie 131 (1) (2019) 265–269.
[70] M. Raji, A. Dashti, M.S. Alivand, M. Asghari, Novel prosperous computational estimations for greenhouse gas adsorptive control by zeolites using machine learning methods, J. Environ. Manage. 307 (2022), 114478.
[71] F. Yulia, I. Chairina, A. Zulys, Multi-objective genetic algorithm optimization with an artificial neural network for CO_2/CH_4 adsorption prediction in metal–organic framework, Therm. Sci. and Eng. Prog. 25 (2021), 100967.
[72] N. Zhang, B. Yang, K. Liu, H. Li, G. Chen, X. Qiu, W. Li, J. Hu, J. Fu, Y. Jiang, M. Liu, J. Ye, Machine Learning in screening high performance electrocatalysts for CO_2 reduction, Small Methods 5 (11) (2021) 2100987.
[73] G.S. Fanourgakis, K. Gkagkas, E. Tylianakis, G. Froudakis, A generic machine learning algorithm for the prediction of gas adsorption in nanoporous materials, J. Phys. Chem. C 124 (13) (2020) 7117–7126.
[74] X. Zhu, D.C. Tsang, L. Wang, Z. Su, D. Hou, L. Li, J. Shang, Machine learning exploration of the critical factors for CO_2 adsorption capacity on porous carbon materials at different pressures, J. Clean. Prod. 273 (2020), 122915.
[75] J. Burner, L. Schwiedrzik, M. Krykunov, J. Luo, P.G. Boyd, T.K. Woo, High-performing deep learning regression models for predicting low-pressure CO_2 adsorption properties of metal–organic frameworks, J. Phys. Chem. C 124 (51) (2020) 27996–28005.