Liu *et al. Carbon Research* (2025) 4:46
https://doi.org/10.1007/s44246-025-00213-9

Carbon Research

ORIGINAL ARTICLE

Open Access

# Enhanced machine learning prediction of biochar adsorption for dyes: Parameter optimization and experimental validation

Chong Liu[1,2*], Paramasivan Balasubramanian[3], Xuan Cuong Nguyen[4,5], Jingxian An[2,6], Sai Praneeth[7], Pengyan Zhang[8] and Haiming Huang[1*]

## Abstract

Biochar, as an eco-friendly, carbon-rich,and economical adsorbent, proven effective in removing toxic dyes from aquatic environments. This study evaluated the efficacy of machine learning (ML) models in predicting the adsorption capacity of biochar for dye removal. Nine models, namely CatBoost, XGBoost, Gradient Boosted Decision Trees, Random Forest, Histogram-Based Gradient Boosting, Kernel Extreme Learning Machine, Kriging, Light Gradient Boosting Machine, and AdaBoost, were deployed to ascertain their predictive accuracies. The CatBoost model was highlighted for its exceptional performance, achieving the highest $R^2$ (0.9880) and the lowest RMSE (0.0839). The stability of the model was affirmed through residual analysis and random partitioning dataset. A detailed feature importance analysis revealed that experimental conditions predominantly affect adsorption, accounting for 50.8%, followed by biochar characteristics (34.1%) and dye types (15.1%). The most significant feature impacting dye adsorption was identified as the $C_0$ through SHapley Additive exPlanations. Partial dependence plots were used further to illustrate the influence of features on the predictive model. Additionally, experimental validation of the ML approach yielded $R^2$ of 0.9037, reinforcing the applicability of the model. This study adds to supportive evidence of the use of ML for the prediction of adsorption capacity and encourages the development of user-friendly software, using PySimpleGUI, opening new paths to advanced data-driven methods in environmental engineering.

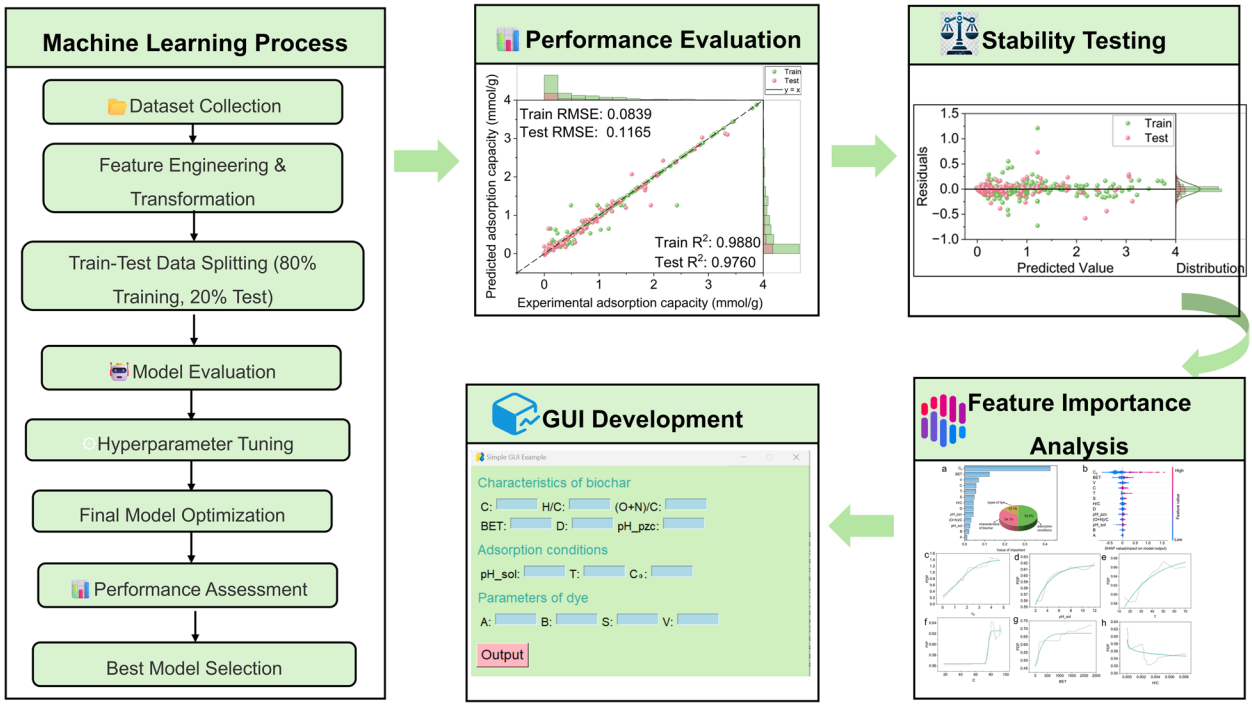
## Highlights

- Nine ML models were tested, with CatBoost standing out for its exceptional performance.
- The experimental verification demonstrated the promising potential of the CatBoost model.
- Advanced data engineering, including ash-free standardization, KNN imputation, and outlier removal, ensured dataset reliability.
- An easy-to-use graphical user interface (PySimpleGUI) was developed for biochar adsorb dyes.

**Keywords** Biochar, Machine learning, Adsorption, Dye, CatBoost, PySimpleGUI

*Correspondence:
Chong Liu
17609858895@163.com
Haiming Huang
huanghaiming52hu@163.com
Full list of author information is available at the end of the article

**Graphical Abstract**




## 1 Introduction

Traditional industries such as paper manufacturing, plastics, textiles, paint, food processing, rubber, cosmetics, and healthcare generate wastewater, the primary source of dye contaminants (Grace Pavithra et al. 2019). According to reports, the commercial dye market sells at least $7 \times 10^5$ tons of dyes annually to various coloring industries (Farhan Hanafi and Sapawe 2020). Due to the complex and stable chemical structures of dyes, conventional wastewater treatment methods often fail to effectively decolorize them (Liu et al. 2024a). In particular, many dyes exhibit toxicity toward microorganisms, which not only disrupts biological treatment processes but also poses environmental risks by inhibiting the metabolic activities of beneficial microbial communities (Liu et al. 2024a). As a result, biological treatments may become less efficient, leading to incomplete dye removal and potential release of toxic byproducts into the environment (Alsukaibi 2022). Wastewater containing dyes discharged into water bodies impedes sunlight penetration, disrupting photosynthesis in aquatic ecosystems (Al-Tohamy et al. 2022). Furthermore, wastewater containing dye poses significant health risks, such as severe skin issues, eye irritation, vomiting, reduced cardiac output, and diminished renal blood flow (Ristea and Zarnescu 2023). Common industrial dye treatment methods include advanced oxidation, ion exchange, membrane processes, adsorption, extraction, and coagulation/flocculation (Shindhal et al. 2021; Liu et al. 2024b, 2025a). Among these, adsorption is considered an economical, non-toxic, highly efficient, and robust technology for removing dyes from wastewater (Liu et al. 2024c). Biochar, as a highly efficient, environmentally friendly, and cost-effective adsorbent, demonstrates substantial potential in dye adsorption and degradation and has been used to eliminate various pollutants (Liu et al. 2024c).

Adsorption of dyes onto biochar could be due to the mass transfer process, which primarily involves chemical reactions, such as ionic and covalent bonds, and physical interactions, including hydrogen bonding and van der Waals forces, between the dye molecules and the biochar (Praveen et al. 2022). Consequently, the adsorption capacity of biochar for dyes is influenced by multiple factors, including the physicochemical properties of biochar (physical structure and chemical composition), adsorption conditions (adsorption time, temperature, relative concentration of adsorbate to adsorbent), and the molecular structure of the dyes (Praveen et al. 2022; Zhang et al. 2023). Previous research employing traditional

adsorption experiments has extensively examined dye adsorption onto biochar, revealing that the adsorption mechanisms include pore diffusion, hydrophobic interactions, hydrogen bonding, cationic and anionic interactions, and partitioning through uncarbonized regions (Praveen et al. 2022; Ouedrhiri et al. 2022; Albanio et al. 2021). However, these studies are generally static and one-dimensional, limiting their ability to identify the relationship between multifaceted factors (such as adsorption conditions and adsorbent properties) and adsorption capacity (Zhang et al. 2023). Moreover, conventional adsorption experiments are often costly and time-consuming, potentially inadequately representing the interactions between parameters that impact adsorption capacity (Yang et al. 2021). Therefore, the development of a universal and stable model that explains the adsorption capacity of biochar for dyes and quantifies the contribution of each factor to adsorption capacity is crucial for synthesizing and screening high-performance biochar materials.

Machine learning (ML), which predicts the state of new data by summarizing the underlying relationships and rules within known datasets, can efficiently tackle complex, non-linear problems (Sarker 2021). For this reason, ML is increasingly employed in environmental studies due to its superior predictive accuracy and efficiency. Applications include predicting ammonia nitrogen fluctuations in wetlands (Nguyen et al. 2024), the adsorption capacity of biochar for tetracycline (Balasubramanian et al. 2024), antibiotic distribution in soil (Wang et al. 2024), and changes in various wastewater treatment plant indicators (Ye et al. 2024). Although some studies have utilized ML to predict dye removal efficiencies using biochar (Bibi et al. 2023; Kaya et al. 2022; Gamboa et al. 2024; Kumari et al. 2024; John et al. 2024), these investigations face several critical challenges: (1) the elemental compositions are not standardized, exhibiting a lack of uniformity between ash and non-ash components; (2) the models employed are often overly simplistic, characterized by insufficient data volumes and an inadequate number of features. Feature collection was limited to either physical or chemical characteristics of biochar rather than both; (3) the focus is restricted to only a select few types of dyes, biochars or models; (4) the primary subjects of study are typically other carbon-based materials, such as activated carbon, rather than biochar; (5) the derived model has not been experimentally validated.

This study integrates biochar characteristics, adsorption conditions, and dye parameters as input features to predict the adsorption capacity of biochar for dyes. During data preprocessing, feature engineering standardized the elemental composition (e.g., converting elemental composition from mass fraction to molar fraction and unifying it to an ash-free basis). Subsequently, nine different ML models were employed to establish predictive models for dye adsorption onto biochar. The optimal model was interpreted using Shapley additive explanations (SHAP) and Partial Dependence Plots (PDP). Finally, experimental validation confirmed the effectiveness of the model, and a graphical user interface program was developed to predict dye adsorption onto biochar.

## 2 Methodologies

### 2.1 Data collection

Data for this study were gathered from the Web of Science™ (WoS) core collection database, Google Scholar and Scopus (from 2013), using the search terms "biochar AND dye AND (adsorption OR sorption)." Following initial scrutiny, the study amassed data encompassing 43 varieties of biochar, 15 categories of dyes, and 685 experimental data sets (Data Availability section). The collected literature is listed in Supplementary Information (SI). All data collected for this study were unbiased. The schematic diagram of this study is shown in Fig. 1.

To forecast the equilibrium adsorption capacity ($Q$, mmol/g) of biochar, seventeen parameters were delineated, informed by a comprehensive review of relevant literature (Yang et al. 2022; Zhu et al. 2021, 2022; Zhao et al. 2022). These parameters are categorized into three groups:

(i) Characteristics of biochar: this category encapsulates the total carbon content ($C$, wt.%), hydrogen to carbon ratio ($H/C$), the sum of oxygen and nitrogen to carbon ratio [$(O+N)/C$], oxygen to hydrogen ratio ($O/H$), ash content ($Ash$, %), specific surface area ($BET$, $m^2/g$), pore diameter ($D$, nm), total pore volume ($PV$, $cm^3/g$), and pH point of zero charge ($pH\_pzc$). It is pertinent to note that, due to discrepancies in measurement methods across different sources, this study adopts the elemental content measurements as opposed to atomic weights (Yang et al. 2022).

(ii) Adsorption conditions: this includes the pH value of the adsorptive solution ($pH\_sol$), the temperature at which adsorption occurs ($T$, °C), and the ratio of the initial concentration of dye to the dosage of biochar ($C_0$, mmol/g).

(iii) Types of dye: it is determined by the Abraham parameters for neutral dye. These parameters encompass hydrogen bond acidity ($A$), hydrogen bond acceptor capability ($B$), polarizability ($S$), excess molar refraction ($E$), and molecular volume ($V$) (Yang et al. 2022; Ruiz et al. 2022).

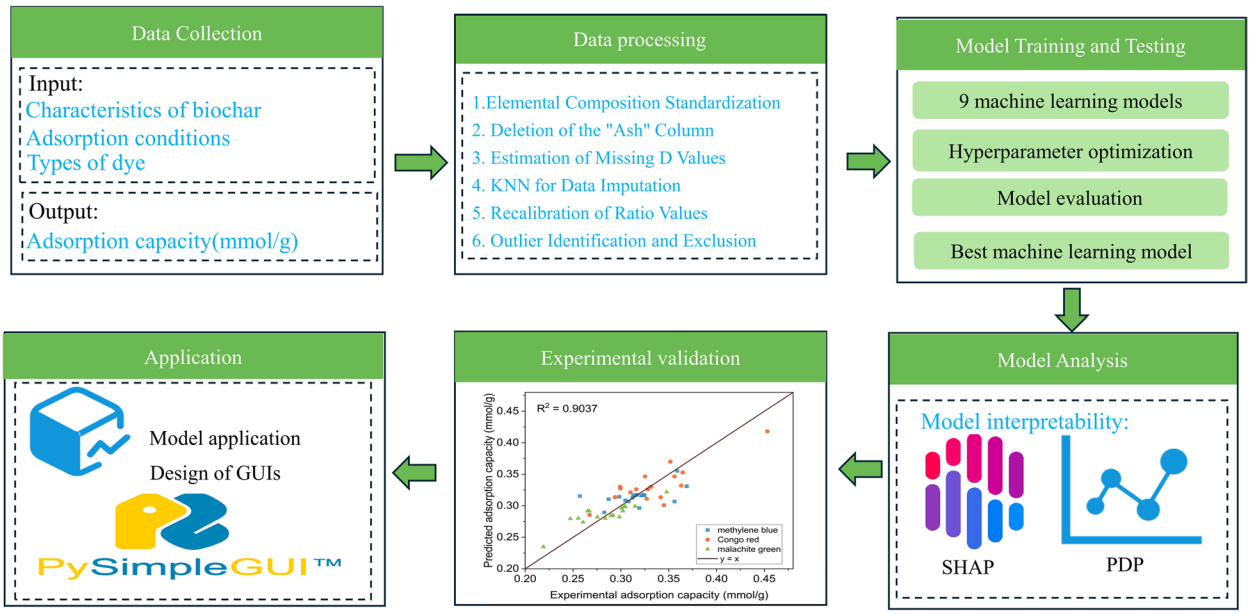



**Fig. 1** Scheme of this study for comprehending dye adsorption capacity of biochar

### 2.2 Data preprocessing and feature engineering

Given the constraints of data capacity and the extensive range of variables under investigation, this study encountered data inadequacies in some features, particularly in the *Ash* and *D* variables, which exhibited missing ratios of 0.731 and 0.266, respectively (Fig. S1). Moreover, inconsistencies in the presentation of certain statistical metrics within the dataset were also detected (some literature element compositions adopt an ash-free basis, while others adopt an ash basis). To effectively address these challenges, the following preprocessing measures were instituted:

(1) Elemental composition standardization: This process entailed standardizing all elemental composition data to an ash-free basis. This step was necessitated by the discovery that different sources reported elemental compositions on either a dry or an ash-free basis (Liu et al. 2024a, 2025b).
(2) Deletion of the "*Ash*" column: Due to the high prevalence of missing values concerning ash content, the "*Ash*" column was removed following the conversion of data to an ash-free basis.
(3) Missing *D* values were estimated utilizing Eq. (1) (Zhu et al. 2022):

$$D = \frac{4 * PV}{BET} \tag{1}$$

where *D* denotes the particle size of the biochar, *PV* signifies the pore volume of the biochar, and *BET* represents the specific surface area.

(4) For repleting a small portion of the remaining missing data (Fig. S1), the *K*-Nearest Neighbours (KNN) algorithm was employed (Yang et al. 2022).
(5) The *H/C*, *(O+N)/C*, and *O/H* values were recalibrated into molar ratios to ensure uniformity.
(6) Outliers were identified by examining the normal distribution of *Q* values. Upon examining the data, it was observed that the majority of Q values were concentrated around ~4 mmol/g. However, 17 rows exhibited values exceeding 4 mmol/g with a considerable range variation, necessitating their removal from the dataset to improve accuracy, thereby refining it to 668 rows.

Additionally, as a preliminary to ML endeavors, the dataset was subjected to Pearson Correlation Coefficient (PCC) analysis. This was undertaken to discern highly correlated features, whereupon one was retained to circumvent multicollinearity. The PCC is calculated as Eq. (2):

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

in which, $\rho_{xy}$ symbolizes the PCC between two variables, and $\overline{x}$ and $\overline{y}$ represent the mean values of $x$ and value $y$, respectively.

Furthermore, to attenuate significant discrepancies amongst the input variables, a $Z$-score normalization method (mean = 0, standard deviation = 1) will be enacted ahead of deploying ML techniques.

### 2.3 Machine learning models

In order to compare the predictive accuracy of some common tree-based models and kernel-based models, nine ML models were implemented to ascertain the optimal adsorption capacity prediction model according to several previous literature (Wang et al. 2024; Haider Jaffari et al. 2023; Shen et al. 2024; Zhao et al. 2021). These ML models encompass CatBoost (CB), eXtreme Gradient Boosting (XGB), Gradient Boosted Decision Trees (GBDT), Random Forest (RF), Histogram-Based Gradient Boosting (HGB), Kernel Extreme Learning Machine (KELM), Kriging, Light Gradient Boosting Machine (LGBM), and AdaBoost (AB). The operational principles of these nine ML models are delineated in Table S1(SI). It is noteworthy that KELM and Kriging rely on kernel methodologies. Besides, the Kriging operates by interpolating or predicting based on spatial correlation, whereas the remaining seven methods employ tree-based ML techniques (Janga et al. 2023).

During the ML modelling process, the dataset underwent random partitioning into two distinct subsets: a training set comprising 80% of the data and a test set comprising the remaining 20%. The selection and range of hyperparameters in this study were guided by several previous studies, with appropriate expansions made to the scope (Shen et al. 2024; Zhao et al. 2021; Zhu et al. 2019a; Zhou et al. 2024). The chosen hyperparameters and their respective ranges for all models are catalogued in Table S1 (SI). Critical parameters of each model underwent fine-tuning through Bayesian optimization techniques. It is noteworthy that Bayesian optimization demonstrates superior efficiency in discovering optimal hyperparameter combinations within a comparatively smaller number of iterations when contrasted with traditional grid search cross-validation methodologies (Xia et al. 2017). This optimization approach leverages information more effectively, rendering the search process adaptive and intelligent (Xia et al. 2017; Yang et al. 2024). To ascertain the robustness of the models, a five-fold cross-validation technique was implemented, and the models were subjected to evaluation across 1000 random train-test partitions. Moreover, this investigation employed Root Mean Square Error (RMSE) and the coefficient of determination ($R^2$) as evaluative metrics to gauge model performance across both training and test sets.

### 2.4 Model interpretation

The interpretation of ML models plays a pivotal role in unraveling the opaque nature of algorithms, thereby shedding light on the intricate workings of their internal mechanisms (Ali et al. 2023). In this investigation, the SHAP method is utilized to gauge the significance of each input feature in shaping the ultimate outcome (Ekanayake et al. 2022). Anchored in cooperative game theory, SHAP quantifies marginal contributions and weights, endeavoring to achieve equity between contribution and acquisition (Kim et al. 2023). Additionally, this study employs PDP to visually elucidate how these features impact the target variable, delineating relationships as linear, monotonic, or complex in nature (Zhang et al. 2023).

### 2.5 ML experimental validation

The culminating model was employed to predict the adsorptive capabilities of biochar available in our laboratory for various dyes, iterating through diverse adsorptive conditions. It is noteworthy that due to the scarcity of data, this experiment was merely a preliminary validation of the model's accuracy, with changes made only in the experimental conditions (Shen et al. 2024; Liu et al. 2024d; Guo et al. 2024). Cotton straw (Xinjiang, China) biochar was carbonized at 600 °C for a duration of three hours, subsequently cooled overnight, and designated as CSB. Then, 0.2 g of the biochar is added to 25 mL of three distinct dyes—Methylene Blue, Congo Red, and Malachite Green—with dye concentrations set between 10 and 100 mg/L. Initial pH levels were meticulously adjusted to 2, 4, 6, 8, 10, and 12. The mixtures underwent oscillation in the dark for 24 h at varying temperatures of 25, 35, and 45 °C, generating a total of 51 experimental cases for validation (Data Availability section). Moreover, elemental analysis of the biochar was conducted using an element analyzer (Elementar Vario, Elementar, Germany), while the structural composition of dry biochar was examined under nitrogen via Brunauer–Emmett–Teller (QuantaChrome, NOVA-200E, USA). The *pH_pzc* was measured using the pH drift method (Xu et al. 2021). The equilibrium concentration of dyes in aqueous solutions is measured using spectrophotometry (Reis et al. 2023). The variation in dye concentration in solutions pre- and post-adsorption was utilized to calculate the adsorptive capacity, which was then juxtaposed with the predicted $Q$ from the optimum ML model.

## 3 Results and discussion

### 3.1 Dataset description

In this study, fifteen distinct types of dyes were examined, comprising cationic dyes (malachite green, crystal violet, methylene blue, rhodamine B), anionic dyes (acid red 18, acid orange 7, acid blue 9, congo red), non-ionic dyes (reactive yellow, reactive orange 16, Remazol brilliant blue R), and others (food red 17). As illustrated in Table 1, the mean *pH_pzc* of the biochar is 6.8, with a standard deviation of 3.01, spanning from 2.30 to 12.31, indicative of notable diversity in surface charge neutrality across various materials. Furthermore, the average *C* stands at 77.83%, albeit with a relatively wide variability (standard deviation of 14.40%), with the minimum and maximum values recorded at 17.69% and 94.04%, respectively, underscoring the heterogeneous nature of carbonaceous components among the materials. The *H/C* ratio denotes a modest hydrogen content on average, coupled with a high skewness of 2.80, suggestive of the presence of outliers and uneven data distribution. The examination of *O/C* and *(O+N)/C* ratios further accentuates this heterogeneity, with skewness values of 4.64 and 4.94, respectively, implying the existence of samples with notably elevated oxygen and nitrogen constituents. Analyses of *BET* and *PV* reveal a diverse array of physical adsorption properties among biochars, with an average *BET* of 621.16 ±751.09 $m^2$/g and an average *PV* of 0.38 ±0.39 $cm^3$/g. Nonetheless, the maximum values extend to 2301.61 $m^2$/g for *BET* and 1.21 $cm^3$/g for *PV*, signifying the presence of materials with remarkably high porosity and surface activity. The *D* value stands at 4.94 nm, characterized by a broad distribution indicated by a standard deviation of 6.31 nm and a skewness of 3.26. The average value for the target variable is measured at 0.62 mmol/g.

The PCC among various input and output features can be seen in Fig. 2. Negative correlations are denoted by green cells, while positive correlations are represented by pink cells. The number of asterisks indicates the level of statistical significance, with a greater number of asterisks signifying higher significance. Several conclusions emerge from the analysis: (i) the adsorption capacity (*Q*) exhibits relatively strong correlations with biochar components (*BET* and *PV*) and the initial concentration ratio ($C_0$), with respective PCC of 0.46, 0.48, and 0.57. This implies a notable influence of *BET*, *PV*, and $C_0$ on the adsorption process. (ii) Most correlation coefficients between input and output features are below 0.6, suggesting that each feature can independently contribute to the performance of ML models. (iii) Consistent with prior research, the biochar composition variables *O/C* and *(O+N)/C* display a robust correlation. To mitigate collinearity, it is recommended to exclude one of these variables. Given that the correlation between adsorption capacity (*Q*) and *O/C* is weaker than that with *(O+N)/C*, the *O/C* variable is selected for removal. (iv) The correlation between *BET* and *PV* is notably high at 0.96. Similarly, as the correlation between *Q* and *PV* is weaker compared to that with *BET*, the *PV* variable is chosen for exclusion. (v) Among types of dye, the correlation coefficients between *E* and *S*, and between *E* and *V*, exceed 0.95, indicating collinearity. Therefore, variable *E* is removed to mitigate this issue. Following these adjustments, the dataset has been structured into 668 rows and 13 columns.

### 3.2 Evaluation of machine learning model performance

This study has developed and rigorously assessed the efficacy of nine distinct ML models, specifically CB, XGB, GBDT, RF, HGB, KELM, Kriging, LGBM, and AB. Employing the entire dataset comprising 668 data points, all models underwent training facilitated by Bayesian optimization. The optimal hyperparameters for each model are detailed in Table S1 (SI). Figure 3 illustrates the actual adsorption capacity of biochar for dyes alongside the predicted values computed by various ML models. Within Fig. 3, the green and pink markers denote the

**Table 1** The statistical information of biochar characteristics and predicted values for machine learning algorithms

| Parameter | Mean | Std | Min | 25% | 50% | 75% | Max | Skew |
|---|---|---|---|---|---|---|---|---|
| *pH_pzc* | 6.80 | 3.01 | 2.30 | 4.30 | 5.06 | 9.59 | 12.31 | 0.57 |
| *C* | 77.83 | 14.40 | 17.69 | 73.88 | 80.25 | 86.50 | 94.04 | − 1.99 |
| *H/C* | $7.62^{-4}$ | $1.64^{-3}$ | 0.00 | $1.00^{-5}$ | $6.90^{-5}$ | $2.40^{-4}$ | $8.33^{-3}$ | 2.80 |
| *O/C* | 0.40 | 0.48 | 0.08 | 0.20 | 0.30 | 0.39 | 3.20 | 4.64 |
| *(O+N)/C* | 0.43 | 0.47 | 0.13 | 0.23 | 0.35 | 0.45 | 3.26 | 4.94 |
| *BET* | 621.16 | 751.09 | 0.01 | 72.70 | 136.20 | 1156.25 | 2301.61 | 1.07 |
| *PV* | 0.38 | 0.39 | $4.35^{-3}$ | 0.07 | 0.25 | 0.67 | 1.21 | 0.81 |
| *D* | 4.94 | 6.31 | $3.46^{-3}$ | 2.09 | 2.53 | 4.08 | 36.00 | 3.26 |
| *Q* | 0.62 | 0.76 | 0.00 | 0.06 | 0.27 | 0.92 | 3.88 | 1.74 |

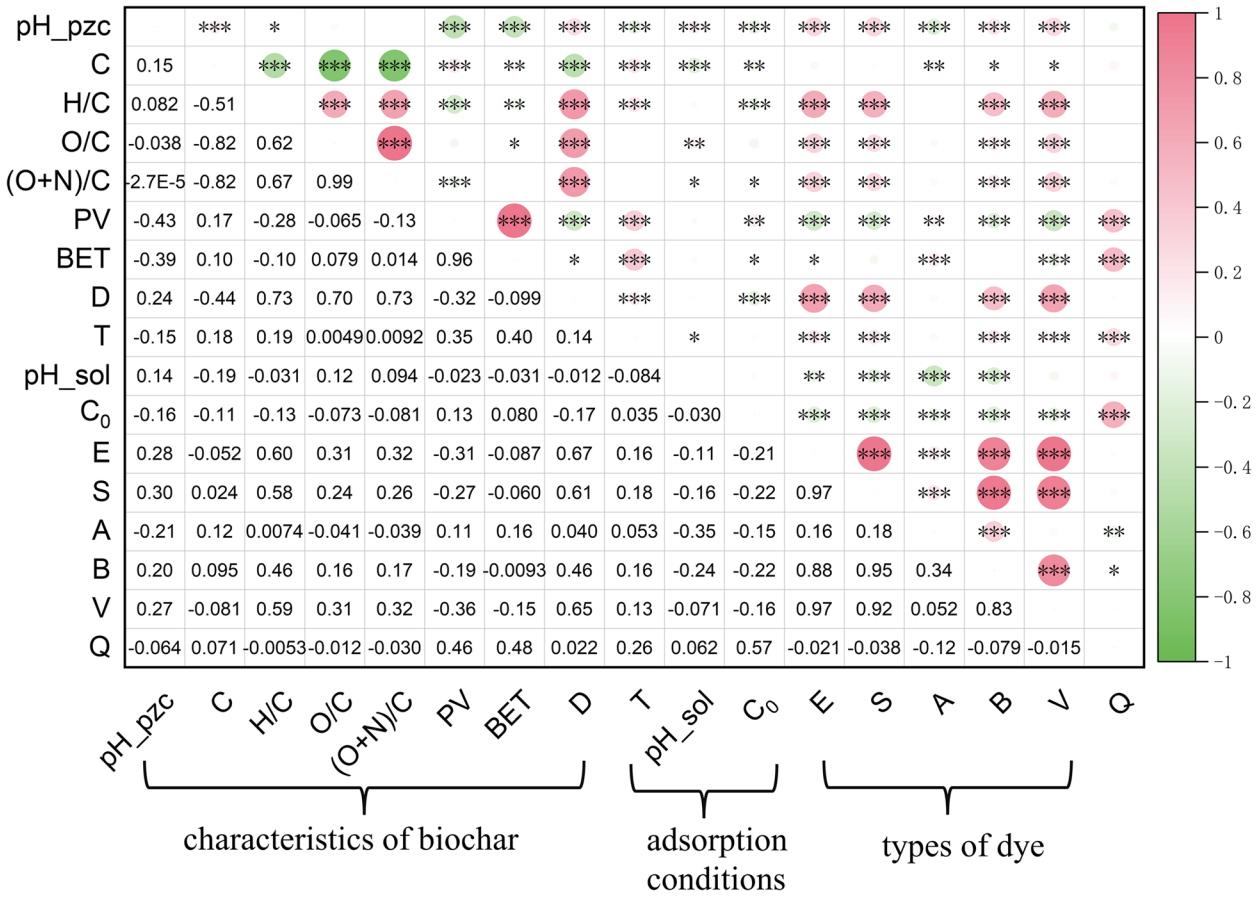



**Fig. 2** Pearson correlation coefficient analysis of input and output features in dye adsorption by biochar study

training and testing samples, respectively. A dashed line representing $y=x$ signifies parity between experimental and predicted values, with proximity to this line indicating enhanced predictive accuracy. Histograms adjacent to the $X$ and $Y$ axes depict the distribution of model data allocated for training (80%) and testing (20%) datasets.

It is evident from Fig. 3 that all models exhibit no signs of overfitting when comparing the value of $R^2$ between testing and training. It can be clearly seen that those models based on the tree—CB (0.9880), AB (0.9816), GBDT (0.9869), HGB (0.9828), LGBM (0.9833), RF (0.9832), XGB (0.9820)—demonstrated superior performance in the training dataset compared to the other two kernel-based models (KELM (0.7110) and Kriging (0.9246)), with similar outcomes observed within the testing dataset. Potential reasons why tree-based models (selected) may outperform kernel-based (selected) models include: (1) tree-based models address non-linear relationships directly by partitioning the data space, thereby typically enhancing their efficacy in managing complex and highly non-linear data relationships compared to models based on linear kernels. (2) Tree models inherently capture interactions among features. At each node, the model splits the data based on a feature, allowing tree structures to consider complex interrelations among features during the decision-making process. (3) Tree models, particularly ensemble trees, often provide robustness and excellent generalization capabilities by constructing multiple trees and aggregating their predictions. This approach enhances performance on unseen data by reducing the variance of the model. Besides, the CB model can reach the lowest RMSE (Train RMSE = 0.0839; Test RMSE = 0.1165) among the tree-based models. This result is similar to that in a previous study (Haider Jaffari et al. 2023). This can be explained by the fact that CB can automatically process variables without the need for extensive data preprocessing or conversion, which is particularly effective when dealing with datasets that contain a large number of features (Zhang and Jánošík 2024; Fujimoto et al. 2022). Additionally, CB's ordered boosting technique trains each tree using different data subsamples rather than the entire dataset (Zhang and Jánošík 2024; Fujimoto et al. 2022; Hancock and

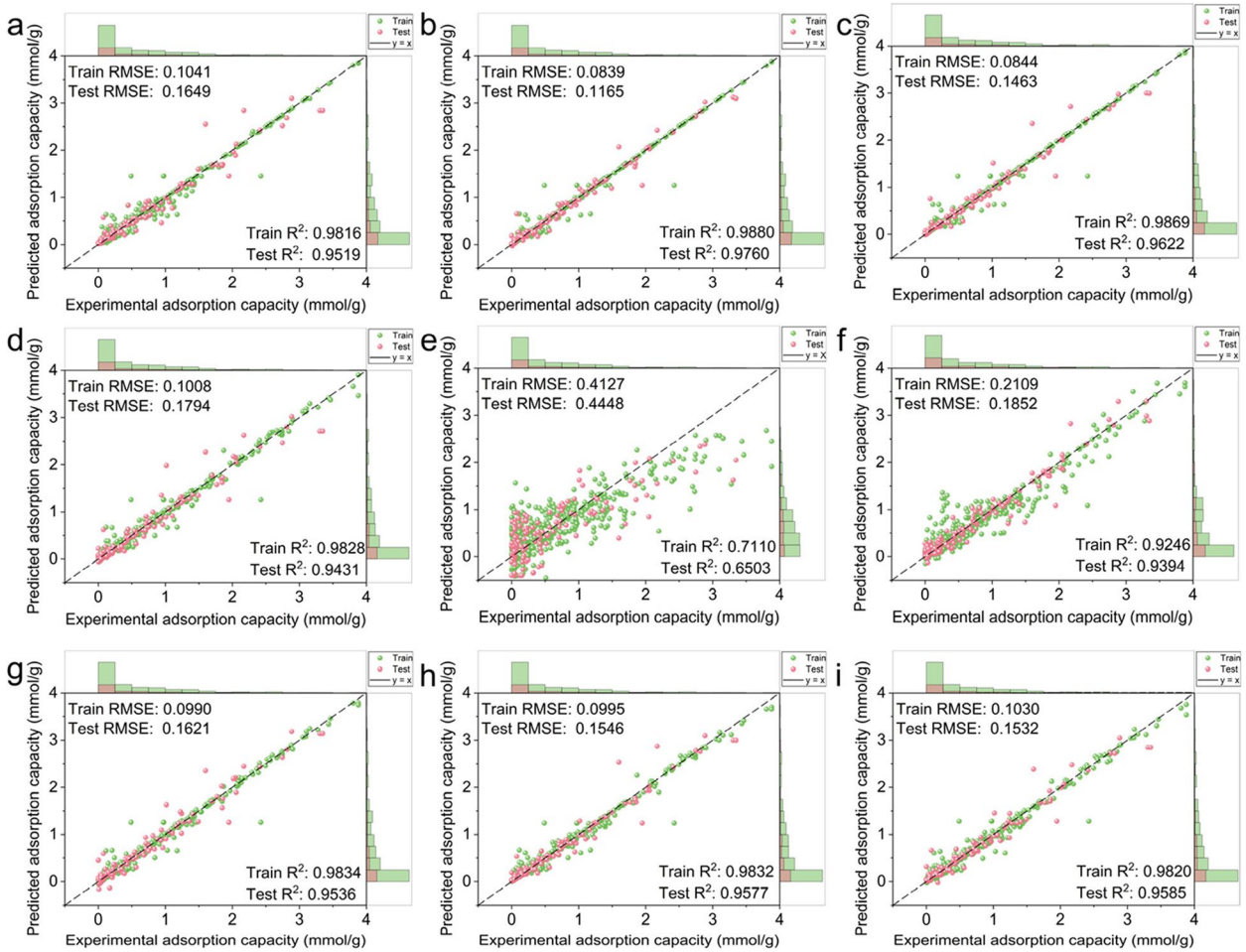



**Fig. 3** Comparative analysis of actual vs. predicted dye adsorption capacities by various ML models: **a** AdaBoost, **b** CatBoost, **c** Gradient Boosted Decision Trees, **d** Histogram-Based Gradient Boosting, **e** Kernel Extreme Learning Machine, **f** Kriging, **g** Light Gradient Boosting Machine, **h** Random Forest, **i** eXtreme Gradient Boosting

Khoshgoftaar 2020; Karbasi et al. 2022). This method effectively reduces target leakage and label noise, thereby enhancing the accuracy of the model.

To verify the accuracy of the model, this study employed residual analysis. Figure 4 illustrates the residual plots and their corresponding normal distributions for the test datasets using the nine selected models. Defined as the difference between observed and predicted values, residuals centered at residual = 0 and conforming to a normal distribution signify a model's reliability. Consistent with the earlier findings, the majority of residuals and their histograms for tree-based models—CB, AB, GBDT, HGB, LGBM, RF, XGB—are clustered near the zero line, while KELM and Kriging (kernel-based models) display more dispersed residuals. Additionally, as predictive values increase, the residuals tend to concentrate around the zero line, underscoring the exemplary performance of the developed models. Notably, the CB model stands out for its tightly concentrated residual distribution, primarily clustering around residual = 0.

To verify the stability of the model, each model underwent 1000 iterations of random validation and fivefold cross-validation to evaluate their stability, resulting in 5000 data results per model. The outcomes of this analysis have been systematically documented on GitHub (Data Availability section). Figure 5 provide a comprehensive comparison of the accuracy and stability of the aforementioned models using test set $R^2$, RMSE and MSE. It is evident that the average $R^2$ values on the test dataset for CB, XGB, GBDT, RF, HGB, KELM, Kriging, LGBM, and AB models are 0.9568, 0.9479, 0.9519, 0.9470, 0.9426, 0.5661, 0.8905, 0.9505, and 0.9398, respectively (Fig. 5b), which has the similar result of Fig. 5a. These results further corroborate the superior accuracy encapsulated in the CB model. Similarly, in terms of the RMSE and MSE metric, the CB model exhibits markedly low RMSE and

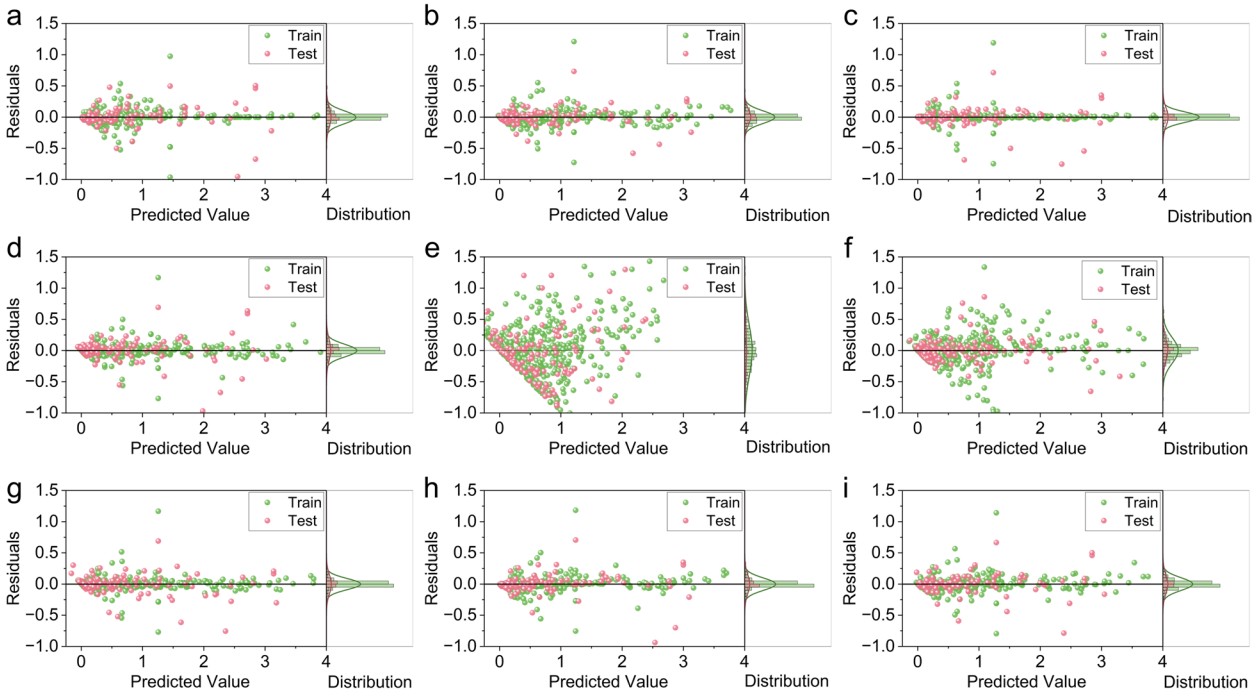



**Fig. 4** Residual plots of various ML models: **a** AdaBoost, **b** CatBoost, **c** Gradient Boosted Decision Trees, **d** Histogram-Based Gradient Boosting, **e** Kernel Extreme Learning Machine, **f** Kriging, **g** Light Gradient Boosting Machine, **h** Random Forest, **i** eXtreme Gradient Boosting

MSE, indicating minimal discrepancy between the model's prediction and empirical values. Consequently, the CB model was chosen for further analysis. An interesting observation is the comparative stability of the HGB model compared to other models, as indicated by the $R^2$ and RMSE scores. The enhanced stability of the HGB model could be attributed to its method of distributing the values of continuous features across a finite number of bins (Costa et al. 2019). This histogram-based approach significantly reduces computational complexity and enhances the model's stability when managing large-scale datasets.

Table 2 presents a comprehensive comparison of $R^2$ values derived from this study with existing literature utilizing ML to predict the adsorption efficiency of biochar on dyes.

It is evident that most studies have utilized only one model and one type of dye, whereas this study employed multiple models and dyes. Furthermore, although a recent study utilized more data than this one (Iftikhar et al. 2023), it may have the following issues: (i) the majority of the materials are activated carbon (90%), not biochar; (ii) only the physical properties of the materials are considered in the input features, neglecting their chemical characteristics; (iii) despite the large volume of data, it is relatively concentrated, which to some extent compromises the reliability of the model; (iv) the derived algorithm has not been experimentally validated; (v) artificial neural networks (ANN) are more complex for users compared to integrated algorithms. Besides, our study stands out by conducting a broader comparison of ML models than many previous investigations. Additionally, the inclusion of readily accessible ML codes and raw data underscores the study's uniqueness, fostering a constructive discourse on reproducibility and transparency in scientific inquiry.

### 3.3 Impact of features on dye adsorption

The adsorption of dyes onto biochar surfaces is undoubtedly influenced by various factors whose impacts are distinct from one another. Figure 6a, b employs bar charts and the SHAP technique to study the significance of different input features on the adsorptive capacity for dyes. These inputs are divided into three groups to explore their individual impacts: experimental conditions, characteristics of biochar, and types of dye. The pie chart in Fig. 6a reveals that experimental conditions have the greatest influence on the predictive model for adsorption capacity, accounting for 50.8% of the impact, followed by biochar properties (35.1%), and dye parameters (15.1%). Within experimental conditions, $C_0$ has the greatest influence, reflecting the adsorbate-to-adsorbent ratio in the reaction. The significant impact of the ratio of initial dye concentration to biochar dosage (*C0*) on the

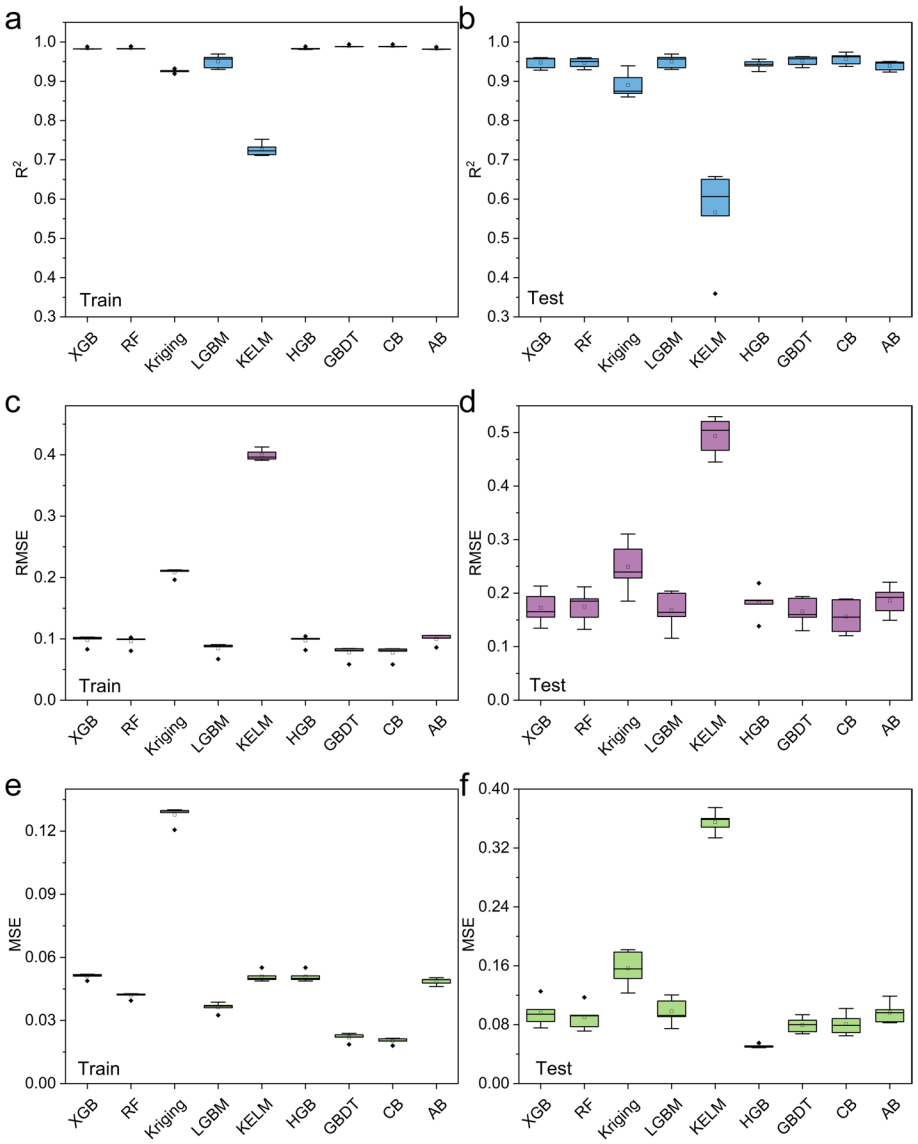



**Fig. 5** Comparison of stability across different models

adsorption reaction is mainly because it directly affects the number of available adsorption sites on the biochar (Ganguly et al. 2020). When the dye concentration is higher relative to the amount of biochar, more dye molecules compete for the limited adsorption sites (Ganguly et al. 2020; Zhang et al. 2020). This may lead to these sites becoming saturated more quickly, which could reduce the efficiency of dye removal per unit of biochar. Conversely, if the dosage of biochar is higher relative to the dye concentration, there are more available adsorption sites, which can enhance the adsorption efficiency and increase the overall removal of dye from the solution. The balance between dye concentration and biochar dosage is crucial for optimizing the adsorption process.

Following this, temperature ($T$) can affect the adsorptive capacity of dyes on biochar, thereby influencing the position and strength of the adsorption equilibrium. Generally, an increase in $T$ tends to accelerate the adsorption process (Toor and Jin 2012). This is due to the increased kinetic energy of the molecules, which allows them to diffuse more rapidly to the surface of the adsorbent, thereby speeding up the adsorption process. The significance of *pH_sol* in experimental conditions is not high. A possible reason is that the effect of pH is overshadowed by other factors, such as temperature and reaction time. The structural features of biochar also have a notable influence on dye adsorption, whereas, for instance, the Brunauer–Emmett–Teller (*BET*) surface

**Table 2** A comparison of machine learning model predictions of dyes adsorption by biochar in various studies

| Number of dyes | ML algorithms | Best model | Number of data (collected) | Number of features (collected) | $R^2$ | ML codes availability | Raw data availability | Ref |
|---|---|---|---|---|---|---|---|---|
| 1(Congo red) | SVR, RF, XGB,PSO | PSO-XGB | 67 | 4 | 0.998 | No | No | (Bibi et al. 2023) |
| 1(Congo red) | ANN | ANN | - | 5 | 0.999 | No | No | (Kaya et al. 2022) |
| 1(Congo Red) | HGB | HGB | 105 | 5 | 0.9583 | No | No | (Gamboa et al. 2024) |
| 1(crystal violet) | ANN, KNN, ANFIS | ANN | 30 | 4 | 0.968 | No | No | (Kumari et al. 2024) |
| 1(Congo Red) | ANN | ANN | 220 | 3 | 0.999 | No | No | (Karaman et al. 2022) |
| 1(Methyl Orange) | ANN | ANN | - | - | 0.975 | No | No | (Adsorption of Methyl Orange on Bentonite: Design, Modeling, and Analysis of Experiments, (n.d.). 2024) |
| 1(acid black 172) | ANN | ANN | - | 5 | 0.996 | No | No | (Yang et al. 2014) |
| 16 | ANN | ANN | 1160 | 12 | 0.98 | Yes | Yes | (Iftikhar et al. 2023) |
| **15** | **CB, XGB, GBDT, RF, HGB, KELM, Kriging, LGBM, AB** | **CB** | **685** | **17** | **0.988** | **Yes** | **Yes** | **This study** |

*Abbreviation: SVR* Support Vector Regression, *PSO* Particle Swarm Optimization, *KNN* k-Nearest Neighbors, *ANN* Artificial Neural Network, *CB* CatBoost, *XGB* eXtreme Gradient Boosting, *GBDT* Gradient Boosted Decision Trees, *RF* Random Forest, *HGB* Histogram-Based Gradient Boosting, *KELM* Kernel Extreme Learning Machine, *ANFIS* Adaptive Neuro-Fuzzy Inference System, *LGBM* Light Gradient Boosting Machine, and *AB* AdaBoost

area of biochar provides more active sites for dye adsorption. More detailed explanations will be given in the PDP analysis.

To further validate the significance of each factor, this study conducted a SHAP analysis. It can be seen from Fig. 6b that the SHAP importance of all input characteristics, with color coding indicating the value of each input trait, where red denotes high and blue denotes lower values. This means that the red on the left side of the central line indicates a negative correlation with the predicted adsorption capacity, and the red on the right indicates a positive correlation. Here, $C_0$ is clearly the most influential feature, followed by the surface area of the biochar, as expected. Simultaneously, $C_0$, *BET*, and *C* correlate positively with adsorption capacity. However, parameters in the dye, such as *V* and *A*, are negatively correlated with adsorption capacity. According to the SHAP analysis, *pH_sol*, *B*, and *A* rank as the least essential features. This feature importance analysis can also provide clues for researching the dye adsorption mechanism of biochar and offer suggestions for future synthesis and screening improvements for biochar.

PDP can analyze the influence of each characteristic on the predicted values (Zhu et al. 2019b). Selected here are the three most significant features of experimental conditions and biochar properties. The gray lines in Fig. 6c-h are the actual PDP lines, and the green lines are the fitted curves. For dye adsorption on biochar, results suggest that $C_0$ is the most significant factor in the change in adsorption capacity (Fig. 6c). Below 4 mmol/g, the partial dependence increases with $C_0$. It can be explained by the fact that an increased concentration gradient of dye between the adsorbate and adsorbent benefits dye adsorption on biochar because raising the initial dye concentration provides a significant driving force to overcome all mass transfer resistances between the liquid and solid phases (Praveen et al. 2022; Dwivedi and Dey 2023). Moreover, previous research indicates that increasing the initial dye concentration can increase the number of collisions between dye anions and adsorbents (Praveen et al. 2022; Zhu et al. 2021). It is clear from Fig. 6c that when $C_0$ reaches 4-5 mmol/g, the adsorption capacity tends toward equilibrium. A possible explanation for this might be biochar has a limited number of adsorption sites for dyes, and as the dye concentration reaches a certain level, the amount of biochar's adsorption sites saturates; hence, adsorption efficiency does not increase with higher dye concentrations (Praveen et al. 2022; Zhu et al. 2021; Dwivedi and Dey 2023).

The pH of the solution is also a crucial factor in affecting the adsorption of dyes on biochar because it directly affects the surface chemistry of biochar and the accessibility of binding sites for dye molecules (Dwivedi and Dey 2023; Faheem et al. 2019). As shown in Fig. 6d, with the increase in *pH_sol*, the PDP values first increase rapidly and then level off. It seems possible that these results occured because the optimal *pH_sol* for dye removal is typically neutral or slightly alkaline, which concurs with the findings of this study (Praveen et al. 2022; Dwivedi and Dey 2023; Ambaye et al. 2021). It has been reported that temperature considerably impacts the adsorption equilibrium rate (Aksu 2001). As depicted in Fig. 6(e),

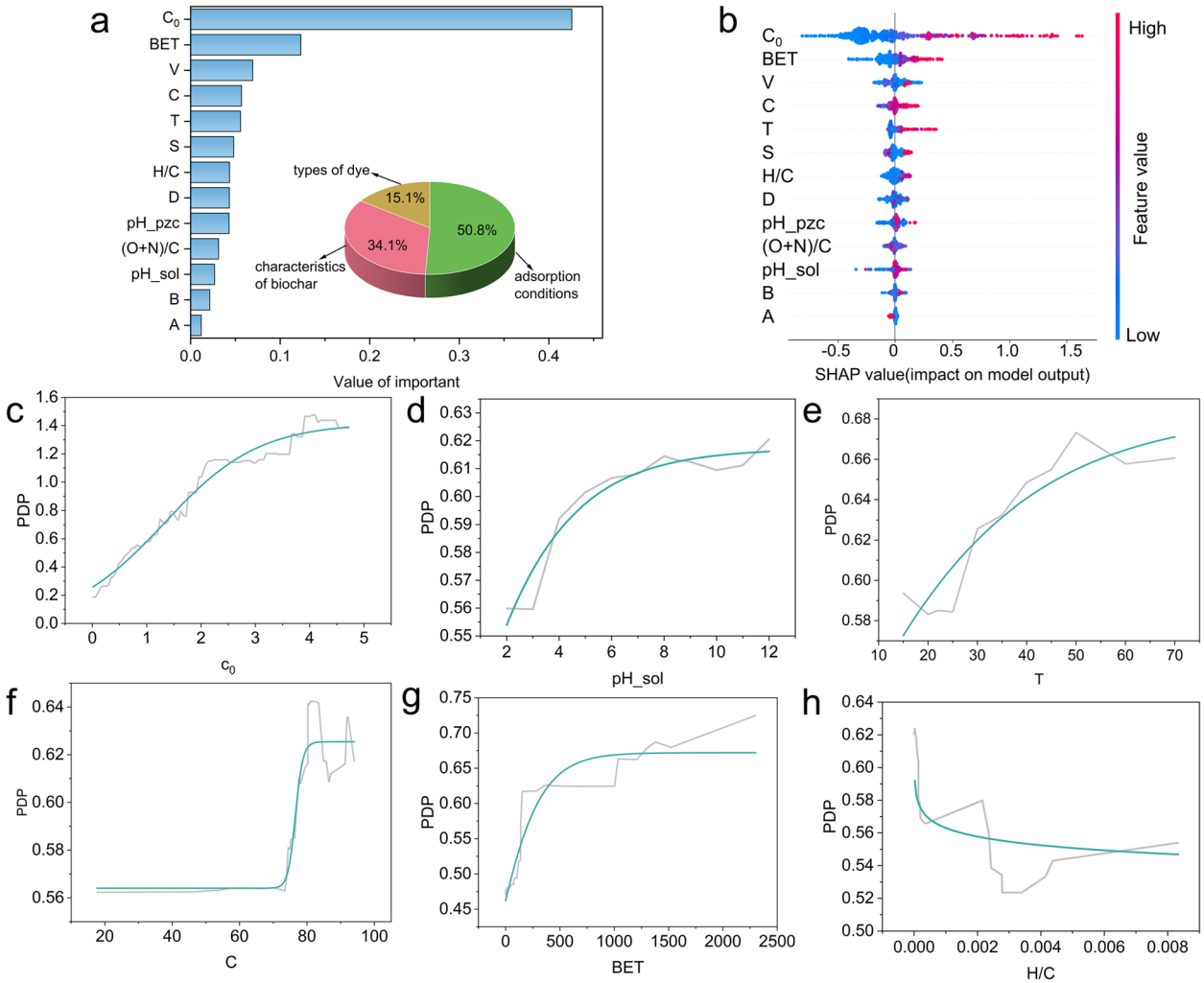



**Fig. 6** Analysis of the SHapley Additive exPlanations value **a-b** and partial dependency plot **c-h**

with *T* rising from 10 °C to 70 °C, PDP values also ascend markedly. This may be due to the fact that raising the temperature accelerates the adsorption process (Toor and Jin 2012). This acceleration occurs because the molecules gain more kinetic energy, enabling them to diffuse more quickly to the adsorbent's surface, thus hastening the adsorption.

The *C* of biochar is another substantial factor affecting its dye adsorption capacity. Figure 6f shows that biochar with a carbon content greater than 80% exhibits improved dye adsorption. It seems possible that these results are due to biochar with higher carbon content typically having a porous structure and larger surface area, offering more sites for dye adsorption (Zhu et al. 2021; Ambaye et al. 2021). Another possible explanation is that higher carbon content usually translates to hydrophobic properties, which could enhance the adsorption capacity for hydrophobic dyes (Praveen et al. 2022; Zhu et al. 2021; Dwivedi and Dey 2023). Hydrophobic biochar easily combines with hydrophobic groups within dyes, thus elevating adsorption efficiency.

It can be seen from Fig. 6g that the impact of biochar on dye adsorption continues to increase up to a *BET* of approximately 750 $m^2/g$; however, beyond this point, a trend toward equilibrium is observed. There are several reasons that may explain this phenomenon, including limited micropore accessibility, sluggish mass transfer, and reduced surface functional groups, all affecting the overall adsorptive performance of biochar for dyes (Zhang et al. 2023). The influence of chemical composition is relatively minor, but the *H/C* notably impacts dye adsorption, correlating with a well-developed carbon structure. Figure 6h illustrates that as the H/C ratio increases, the dependency on adsorption capacity drops sharply, yet it is important to note that when the H/C ratio exceeds 0.006 (molar ratio), the dependency almost

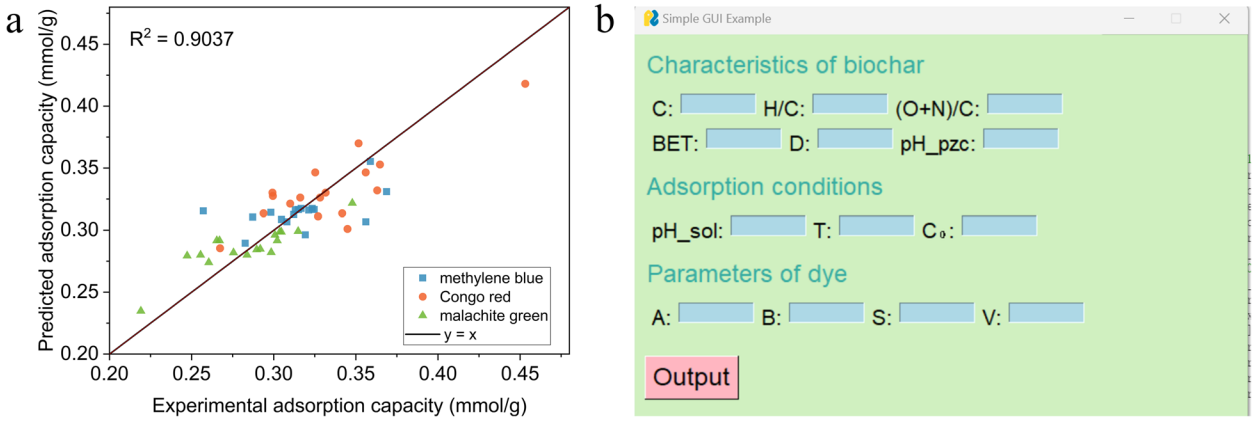



**Fig. 7** Verification performance for CB model **a** and the screenshot of program graphical user interface (GUI) **b**

remains constant. This can be due to an increase in the *H/C* being associated with a decrease in the aromatization level of biochar, making the structure more unstable (Oginni and Singh 2020).

### 3.4 Experimental validation

Experimental validation is depicted in Fig. 7a, where the data aligns with the predictions of the CB model, yielding a validation $R^2$ of 0.9037. This indicates the model's feasibility within a certain range. Overall, the type of dye used has a minimal impact on adsorption capacity, although the deviation for congo red is slightly greater compared to the other two dyes (methylene blue and malachite green). Possible reasons for discrepancies between experimental data and predictions may include (1) the training constraints of the model, which limit predicted values to $Q < 4$ mmol/g, yet some experimental values exceeded this range; (2) the continuous nature of input features, coupled with data collection that only encompassed 43 types of biochar. Thus, incorporating more cases could enhance the accuracy of the model. Additionally, more related research will be done through further study.

The CB model developed in this study is poised to facilitate future research on the biochar adsorption of dyes. To this end, a visual interface based on the CB model and PySimpleGUI (https://www.pysimplegui.com/) was designed, as shown in Fig. 7b. It features 13 input fields and one output, offering a comprehensive and flexible framework for future studies.

### 3.5 Conclusion

This study employed nine ML models—CB, XGB, GBDT, RF, HGB, KELM, Kriging, LGBM, and AB—to predict the adsorption capacity of biochar for dyes. The CB model demonstrated superior performance with the highest $R^2$ value of 0.9880 and the lowest RMSE of 0.0839, significantly outperforming the other models. Residual plots indicate that most errors are located near the zero line, suggesting robust stability of the model. An analysis of feature importance revealed that experimental conditions (50.8%) exert a greater influence than the characteristics of the biochar (34.1%) or the type of dye (15.1%). Furthermore, SHAP results identified the $C_0$ as having the most significant impact on dye adsorption. PDP was utilized to illustrate the effects of six selected features on the model. Lastly, the feasibility of ML was corroborated through experimental validation with an $R^2 = 0.9037$, and a predictive program for the CB model can be developed using PySimpleGUI.

Despite these promising findings, several critical challenges must be addressed before ML-based approaches can be seamlessly integrated into large-scale or industrial contexts. Real-world wastewater typically contains a diverse array of co-occurring constituents and exhibits dynamic water chemistry, introducing complexities not fully replicated in controlled laboratory settings. Moreover, expanding and diversifying training datasets with additional real-world scenarios is paramount to bolster both the generalizability and robustness of these models. Consequently, future efforts should concentrate on validating model performance under variable field conditions, elucidating the long-term influences of key operational parameters (e.g., pH, ionic strength, and competing ions), and exploring the feasibility of scaling up biochar-based remediation systems. By surmounting these challenges, ML-guided strategies can progress from proof-of-concept tools in the laboratory to indispensable assets in sustainable wastewater treatment and broader environmental remediation endeavors.

**Abbreviations**

| | |
|---|---|
| ML | Machine Learning |
| CB | CatBoost |
| XGB | EXtreme Gradient Boosting |
| GBDT | Gradient Boosted Decision Trees |
| RF | Random Forest |
| HGB | Histogram-Based Gradient Boosting |
| KELM | Kernel Extreme Learning Machine |
| LGBM | Light Gradient Boosting Machine |
| AB | AdaBoost |
| SVR | Support Vector Regression |
| PSO | Particle Swarm Optimization |
| KNN | K-Nearest Neighbors |
| ANN | Artificial Neural Network |
| SHAP | SHapley Additive exPlanations |
| PDP | Partial Dependence Plots |
| PCC | Pearson Correlation Coefficient |
| Q | Equilibrium Adsorption Capacity |
| BET | Brunauer-Emmett-Teller (surface area) |
| PV | Total Pore Volume |
| pH_pzc | PH point of zero charge |
| T | Temperature |
| $C_0$ | Initial concentration of dye to dosage of biochar |
| WoS | Web of Science |
| SI | Supplementary Information |
| E | Excess Molar Refraction |
| S | Polarizability |
| A | Hydrogen Bond Acidity |
| B | Hydrogen Bond Acceptor Capability |
| V | Molecular Volume |
| R2 | Coefficient of Determination |
| RMSE | Root Mean Square Error |
| MSE | Mean Square Error |
| H/C | Hydrogen to Carbon Ratio |
| (O + N)/C | Oxygen and Nitrogen to Carbon Ratio |
| O/H | Oxygen to Hydrogen Ratio |
| D | Pore Diameter |
| H | Hydrogen |
| N | Nitrogen |

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s44246-025-00213-9.

Additional file 1

**Acknowledgements**
The authors thank their institutions for providing the necessary research facilities for this work.

**Authors' contributions**
Chong Liu designed the research framework, conducted data preprocessing, and built the machine learning models. Paramasivan Balasubramanian supervised the overall project and provided critical guidance in model optimization. Xuan Cuong Nguyen collected and curated the dataset, ensuring data quality and relevance to the study. Jingxian An contributed to the experimental design and carried out laboratory experiments to generate training data. Sai Praneeth assisted in coding and computational analysis, supporting model validation and statistical evaluation. Pengyan Zhang provided manuscript writing and editing, improving the clarity and coherence of the presentation. Haiming Huang performed the final review, offering insightful revisions to enhance the quality of the manuscript.

**Funding**
This work was financially supported by The Guangdong Basic and Applied Basic Research Foundation (Grant No. 2021B1515020099) and National Natural Science Foundation of China (Grant No. 42277233; 42477228).

**Data availability**
The Python codes used to build the models, raw data used in this study, model stable test results, and PySimpleGUI program code are available at the following GitHub link (https://github.com/17609858895/ML-predict-biochar-adsorb-dye).

## Declarations

**Competing interests**
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.

**Author details**
[1]School of Environment and Civil Engineering, Dongguan University of Technology, Dongguan 523808, China. [2]Department of Chemical & Materials Engineering, University of Auckland, Auckland 1010, New Zealand. [3]Department of Biotechnology & Medical Engineering, National Institute of Technology Rourkela, Rourkela 769008, India. [4]Center for Advanced Chemistry, Institute of Research and Development, Duy Tan University, Da Nang 550000, Viet Nam. [5]Faculty of Environmental and Chemical Engineering, Duy Tan University, Da Nang 550000, Viet Nam. [6]Department of Food Science, Purdue University, West Lafayette, IN 47907, USA. [7]Department of Civil and Environmental Engineering, Wayne State University, 5050 Anthony Wayne Drive, Detroit, MI 48202, USA. [8]College of Water Resources and Architectural Engineering, Tarim University, Xinjiang 843300, China.

Received: 3 November 2024 Revised: 30 March 2025 Accepted: 12 April 2025
Published online: 03 June 2025

**References**

Adsorption of Methyl Orange on Bentonite: Design, Modeling, and Analysis of Experiments, (n.d.). https://www.ijcce.ac.ir/article_704557.html (accessed July 7, 2024).
Aksu Z, Equilibrium and kinetic modelling of cadmium(II) biosorption by *C. vulgaris* in a batch system: effect of temperature, Separation and Purification Technology 21 (2001) 285–294. https://doi.org/10.1016/S1383-5866(00)00212-4.
Albanio II, Muraro PCL, da Silva WL (2021) Rhodamine B Dye Adsorption onto Biochar from Olive Biomass Waste. Water Air Soil Pollut 232:214. https://doi.org/10.1007/s11270-021-05110-6
Ali S, Akhlaq F, Imran AS, Kastrati Z, Daudpota SM, Moosa M (2023) The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review. Comput Biol Med 166:107555. https://doi.org/10.1016/j.compbiomed.2023.107555
Alsukaibi AKD (2022) Various Approaches for the Detoxification of Toxic Dyes in Wastewater. Processes 10:1968. https://doi.org/10.3390/pr10101968
Al-Tohamy R, Ali SS, Li F, Okasha KM, Mahmoud YA-G, Elsamahy T, Jiao H, Fu Y, Sun J (2022) A critical review on the treatment of dye-containing wastewater: Ecotoxicological and health concerns of textile dyes and possible remediation approaches for environmental safety. Ecotoxicol Environ Saf 231:113160. https://doi.org/10.1016/j.ecoenv.2021.113160
Ambaye TG, Vaccari M, van Hullebusch ED, Amrane A, Rtimi S (2021) Mechanisms and adsorption capacities of biochar for the removal of organic and inorganic pollutants from industrial wastewater. Int J Environ Sci Technol 18:3273–3294. https://doi.org/10.1007/s13762-020-03060-w
Balasubramanian P, Prabhakar MR, Liu C, Zhang P, Li F (2024) Predictive capability of rough set machine learning in tetracycline adsorption using biochar. Carbon Res 3:48. https://doi.org/10.1007/s44246-024-00129-w
Bibi A, Khan H, Hussain S, Arshad M, Wahab F, Usama M, Khan K, Akbal F (2023) Sustainable wastewater purification with crab shell-derived biochar: Advanced machine learning modeling & experimental analysis. Biores Technol 390:129900. https://doi.org/10.1016/j.biortech.2023.129900
Costa MA, Wullt B, Norrlöf M, Gunnarsson S (2019) Failure detection in robotic arms using statistical modeling, machine learning and hybrid gradient

boosting. Measurement 146:425–436. https://doi.org/10.1016/j.measurement.2019.06.039
dos Reis G.S., Bergna D, Grimm A, Lima E.C., Hu T, Naushad Mu, Lassi U. Preparation of highly porous nitrogen-doped biochar derived from birch tree wastes with superior dye removal performance, Colloids and Surfaces A: Physicochemical and Engineering Aspects 669 (2023) 131493. https://doi.org/10.1016/j.colsurfa.2023.131493.
Dwivedi S, Dey S (2023) Review on biochar as an adsorbent material for removal of dyes from waterbodies. Int J Environ Sci Technol 20:9335–9350. https://doi.org/10.1007/s13762-022-04364-9
Ekanayake IU, Meddage DPP, Rathnayake U (2022) A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP). Case Studies in Construction Materials 16:e01059. https://doi.org/10.1016/j.cscm.2022.e01059
Faheem J, Du J, Bao MA, Hassan S, Irshad MA (2019) Talib. Multi-Functional Biochar Novel Surface Chemistry for Efficient Capture of Anionic Congo Red Dye: Behavior and Mechanism, Arab J Sci Eng 44:10127–10139. https://doi.org/10.1007/s13369-019-04194-x
Farhan Hanafi M, Sapawe N. A review on the water problem associate with organic pollutants derived from phenol, methyl orange, and remazol brilliant blue dyes, Materials Today: Proceedings 31 (2020) A141–A150. https://doi.org/10.1016/j.matpr.2021.01.258.
Fujimoto S, Mizuno T, Ishikawa A (2022) Interpolation of non-random missing values in financial statements' big data using CatBoost. J Comput Soc Sc 5:1281–1301. https://doi.org/10.1007/s42001-022-00165-9
Gamboa DMP, Abatal M, Lima E, Franseschi FA, Ucán CA, Tariq R, Elías MAR, Vargas J (2024) Sorption Behavior of Azo Dye Congo Red onto Activated Biochar from Haematoxylum campechianum Waste: Gradient Boosting Machine Learning-Assisted Bayesian Optimization for Improved Adsorption Process. Int J Mol Sci 25:4771. https://doi.org/10.3390/ijms25094771
Ganguly P, Sarkhel R, Das P (2020) Synthesis of pyrolyzed biochar and its application for dye removal: Batch, kinetic and isotherm with linear and non-linear mathematical analysis. Surfaces and Interfaces 20:100616. https://doi.org/10.1016/j.surfin.2020.100616
Grace Pavithra K, Senthil Kumar P, Jaikumar V, Sundar Rajan P (2019) Removal of colorants from wastewater: A review on sources and treatment strategies. J Ind Eng Chem 75:1–19. https://doi.org/10.1016/j.jiec.2019.02.011
Guo L, Xu X, Niu C, Wang Q, Park J, Zhou L, Lei H, Wang X, Yuan X (2024) Machine learning-based prediction and experimental validation of heavy metal adsorption capacity of bentonite. Sci Total Environ 926:171986. https://doi.org/10.1016/j.scitotenv.2024.171986
Haider Jaffari Z, Jeong H, Shin J, Kwak J, Son C, Lee Y.-G., Kim S, Chon K, Hwa Cho K. Machine-learning-based prediction and optimization of emerging contaminants' adsorption capacity on biochar materials, Chemical Engineering Journal 466 (2023) 143073. https://doi.org/10.1016/j.cej.2023.143073.
Hancock J, Khoshgoftaar T.M. Performance of CatBoost and XGBoost in Medicare Fraud Detection, in: 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), 2020: pp. 572–579. https://doi.org/10.1109/ICMLA51294.2020.00095.
Iftikhar S, Zahra N, Rubab F, Sumra RA, Khan MB, Abbas A, Jaffari ZH (2023) Artificial neural networks for insights into adsorption capacity of industrial dyes using carbon-based materials. Sep Purif Technol 326:124891. https://doi.org/10.1016/j.seppur.2023.124891
Janga JK, Reddy KR, Raviteja KVNS (2023) Integrating artificial intelligence, machine learning, and deep learning approaches into remediation of contaminated sites: A review. Chemosphere 345:140476. https://doi.org/10.1016/j.chemosphere.2023.140476
John J, Gandhimathi R, Sillanpää M, Chellam PV (2024) Fe3O4-functionalised biochar for persulphate systems towards the removal of Remazol Brilliant Orange 3R: machine learning–based approach and toxicity analysis. Biomass Conv Bioref 14:10319–10334. https://doi.org/10.1007/s13399-022-03056-1
Karaman C, Karaman O, Show P-L, Karimi-Maleh H, Zare N (2022) Congo red dye removal from aqueous environment by cationic surfactant modified-biomass derived carbon: Equilibrium, kinetic, and thermodynamic modeling, and forecasting via artificial neural network approach. Chemosphere 290:133346. https://doi.org/10.1016/j.chemosphere.2021.133346
Karbasi M, Jamei M, Ali M, Abdulla S, Chu X, Yaseen ZM (2022) Developing a novel hybrid Auto Encoder Decoder Bidirectional Gated Recurrent Unit model enhanced with empirical wavelet transform and Boruta-Catboost to forecast significant wave height. J Clean Prod 379:134820. https://doi.org/10.1016/j.jclepro.2022.134820
Kaya N, Yıldız Uzun Z, Altuncan C, Uzun H. Adsorption of Congo red from aqueous solution onto KOH-activated biochar produced via pyrolysis of pine cone and modeling of the process using artificial neural network, Biomass Conv. Bioref. 12 (2022) 5293–5315. https://doi.org/10.1007/s13399-021-01856-5.
Kim JY, Shin UH, Kim K (2023) Predicting biomass composition and operating conditions in fluidized bed biomass gasifiers: An automated machine learning approach combined with cooperative game theory. Energy 280:128138. https://doi.org/10.1016/j.energy.2023.128138
Kumari S, Chowdhry J, Choudhury A, Agarwal S, Narad P, Garg MC (2024) Machine learning approaches for the treatment of textile wastewater using sugarcane bagasse (Saccharum officinarum) biochar. Environ Sci Pollut Res. https://doi.org/10.1007/s11356-024-31826-z
Liu C, Balasubramanian P, Li F, Huang H (2024a) Machine learning prediction of dye adsorption by hydrochar: Parameter optimization and experimental validation. J Hazard Mater 480:135853. https://doi.org/10.1016/j.jhazmat.2024.135853
Liu C, Crini G, Wilson LD, Balasubramanian P, Li F (2024b) Removal of contaminants present in water and wastewater by cyclodextrin-based adsorbents: A bibliometric review from 1993 to 2022. Environ Pollut 348:123815. https://doi.org/10.1016/j.envpol.2024.123815
Liu C, Bolan N, Rajapaksha A.U, Wang H, Balasubramanian P, Zhang P, Nguyen X.C., Li F. Critical review of biochar for the removal of emerging inorganic pollutants from wastewater, Chinese Chemical Letters (2024) 109960. https://doi.org/10.1016/j.cclet.2024.109960.
Liu Q, Zhang G, Yu J, Kong G, Cao T, Ji G, Zhang X, Han L (2024d) Machine learning-aided hydrothermal carbonization of biomass for coal-like hydrochar production: Parameters optimization and experimental verification. Biores Technol 393:130073. https://doi.org/10.1016/j.biortech.2023.130073
Liu C, Crini G, Lichtfouse E, Wilson LD, Picos-Corrales LA, Balasubramanian P, Li F (2025a) Chitosan-based materials for emerging contaminants removal: Bibliometric analysis, research progress, and directions. Journal of Water Process Engineering 71:107327. https://doi.org/10.1016/j.jwpe.2025.107327
Liu C, Balasubramanian P, An J, Li F (2025b) Machine learning prediction of ammonia nitrogen adsorption on biochar with model evaluation and optimization. Npj Clean Water 8:1–12. https://doi.org/10.1038/s41545-024-00429-z
Nguyen XC, Nguyen TP, Lam VS, Le P-C, Vo TDH, Hoang T-HT, Chung WJ, Chang SW, Nguyen DD (2024) Estimating ammonium changes in pilot and full-scale constructed wetlands using kinetic model, linear regression, and machine learning. Sci Total Environ 907:168142. https://doi.org/10.1016/j.scitotenv.2023.168142
Oginni O, Singh K (2020) Influence of high carbonization temperatures on microstructural and physicochemical characteristics of herbaceous biomass derived biochars. J Environ Chem Eng 8:104169. https://doi.org/10.1016/j.jece.2020.104169
Ouedrhiri A, Ait Himi M, Youbi B, Lghazi Y, Bahar J, El Haimer C, Aynaou A, Bimaghra I. Biochar material derived from natural waste with superior dye adsorption performance, Materials Today: Proceedings 66 (2022) 259–267. https://doi.org/10.1016/j.matpr.2022.04.928.
Praveen S, Jegan J, Bhagavathi Pushpa T, Gokulan R, Bulgariu L. Biochar for removal of dyes in contaminated water: an overview, Biochar 4 (2022) 10. https://doi.org/10.1007/s42773-022-00131-8.
Ristea M-E, Zarnescu O (2023) Indigo Carmine: Between Necessity and Concern. Journal of Xenobiotics 13:509–528. https://doi.org/10.3390/jox13030033
Ruiz R, Zamora WJ, Ràfols C, Bosch E (2022) Molecular characteristics of several drugs evaluated from solvent/water partition measurements: Solvation parameters and intramolecular hydrogen bond indicator. Eur J Pharm Sci 168:106066. https://doi.org/10.1016/j.ejps.2021.106066
Sarker IH (2021) Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT SCI 2:160. https://doi.org/10.1007/s42979-021-00592-x
Shen T, Peng H, Yuan X, Liang Y, Liu S, Wu Z, Leng L, Qin P (2024) Feature engineering for improved machine-learning-aided studying heavy metal adsorption on biochar. J Hazard Mater 466:133442. https://doi.org/10.1016/j.jhazmat.2024.133442

Shindhal T, Rakholiya P, Varjani S, Pandey A, Ngo HH, Guo W, Ng HY, Taherzadeh MJ (2021) A critical review on advances in the practices and perspectives for the treatment of dye industry wastewater. Bioengineered 12:70–87. https://doi.org/10.1080/21655979.2020.1863034
Toor M, Jin B (2012) Adsorption characteristics, isotherm, kinetics, and diffusion of modified natural bentonite for removing diazo dye. Chem Eng J 187:79–88. https://doi.org/10.1016/j.cej.2012.01.089
Wang J, Huang R, Liang Y, Long X, Wu S, Han Z, Liu H, Huangfu X (2024) Prediction of antibiotic sorption in soil with machine learning and analysis of global antibiotic resistance risk. J Hazard Mater 466:133563. https://doi.org/10.1016/j.jhazmat.2024.133563
Xia Y, Liu C, Li Y, Liu N (2017) A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Syst Appl 78:225–241. https://doi.org/10.1016/j.eswa.2017.02.017
Xu Y, Bai T, Li Q, Yang H, Yan Y, Sarkar B, Lam SS, Bolan N (2021) Influence of pyrolysis temperature on the characteristics and lead(II) adsorption capacity of phosphorus-engineered poplar sawdust biochar. J Anal Appl Pyrol 154:105010. https://doi.org/10.1016/j.jaap.2020.105010
Yang Y, Lin X, Wei B, Zhao Y, Wang J (2014) Evaluation of adsorption potential of bamboo biochar for metal-complex dye: equilibrium, kinetics and artificial neural network modeling. Int J Environ Sci Technol 11:1093–1100. https://doi.org/10.1007/s13762-013-0306-0
Yang H, Huang K, Zhang K, Weng Q, Zhang H, Wang F (2021) Predicting Heavy Metal Adsorption on Soil with Machine Learning and Mapping Global Distribution of Soil Adsorption Capacities. Environ Sci Technol 55:14316–14328. https://doi.org/10.1021/acs.est.1c02479
Yang X, Nguyen X.C, Tran Q.B, Huyen Nguyen T.T., Ge S, Nguyen D.D, Nguyen V.-T, Le P.C, Rene E.R., Singh P, Raizada P, Ahamad T, Alshehri S.M, Xia C, Kim S.Y., Le Q.V. Machine learning-assisted evaluation of potential biochars for pharmaceutical removal from water, Environmental Research 214 (2022) 113953. https://doi.org/10.1016/j.envres.2022.113953.
Yang K, Liu L, Wen Y (2024) The impact of Bayesian optimization on feature selection. Sci Rep 14:3948. https://doi.org/10.1038/s41598-024-54515-w
Ye G, Wan J, Deng Z, Wang Y, Chen J, Zhu B, Ji S (2024) Prediction of effluent total nitrogen and energy consumption in wastewater treatment plants: Bayesian optimization machine learning methods. Biores Technol 395:130361. https://doi.org/10.1016/j.biortech.2024.130361
Zhang L, Jánošík D (2024) Enhanced short-term load forecasting with hybrid machine learning models: CatBoost and XGBoost approaches. Expert Syst Appl 241:122686. https://doi.org/10.1016/j.eswa.2023.122686
Zhang B, Wu Y, Cha L (2020) Removal of methyl orange dye using activated biochar derived from pomelo peel wastes: performance, isotherm, and kinetic studies. J Dispersion Sci Technol 41:125–136. https://doi.org/10.1080/01932691.2018.1561298
Zhang P, Liu C, Lao D, Nguyen XC, Paramasivan B, Qian X, Inyinbor AA, Hu X, You Y, Li F (2023) Unveiling the drives behind tetracycline adsorption capacity with biochar through machine learning. Sci Rep 13:11512. https://doi.org/10.1038/s41598-023-38579-8
Zhao Y, Li Y, Fan D, Song J, Yang F (2021) Application of kernel extreme learning machine and Kriging model in prediction of heavy metals removal by biochar. Biores Technol 329:124876. https://doi.org/10.1016/j.biortech.2021.124876
Zhao Y, Fan D, Li Y, Yang F (2022) Application of machine learning in predicting the adsorption capacity of organic compounds onto biochar and resin. Environ Res 208:112694. https://doi.org/10.1016/j.envres.2022.112694
Zhou B, Li H, Wang Z, Huang H, Wang Y, Yang R, Huo R, Xu X, Zhou T, Dong X (2024) Prediction of phosphate adsorption amount, capacity and kinetics via machine learning: A generally physical-based process and proposed strategy of using descriptive text messages to enrich datasets. Chem Eng J 479:147503. https://doi.org/10.1016/j.cej.2023.147503
Zhu X, Wang X, Ok YS (2019a) The application of machine learning methods for prediction of metal sorption onto biochars. J Hazard Mater 378:120727. https://doi.org/10.1016/j.jhazmat.2019.06.004
Zhu X, Li Y, Wang X (2019b) Machine learning prediction of biochar yield and carbon contents in biochar based on biomass characteristics and pyrolysis conditions. Biores Technol 288:121527. https://doi.org/10.1016/j.biortech.2019.121527
Zhu X, Wan Z, Tsang DCW, He M, Hou D, Su Z, Shang J (2021) Machine learning for the selection of carbon-based materials for tetracycline and sulfamethoxazole adsorption. Chem Eng J 406:126782. https://doi.org/10.1016/j.cej.2020.126782
Zhu X, He M, Sun Y, Xu Z, Wan Z, Hou D, Alessi DS, Tsang DCW (2022) Insights into the adsorption of pharmaceuticals and personal care products (PPCPs) on biochar and activated carbon with the aid of machine learning. J Hazard Mater 423:127060. https://doi.org/10.1016/j.jhazmat.2021.127060

## Publisher's Note