Current Research in Biotechnology 11 (2026) 100367

Contents lists available at ScienceDirect

Current Research in Biotechnology

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

Effects of physical properties on the heavy metal adsorption of biochar via a robust approach

Mahran Al-Zyoud [a], Salama A. Mostafa [b], Ibrahim Khersan [c], Gowrishankar J [d], Prabhat Kumar Sahu [e], Siya Singla [f], Sardor Sabirov [g], Islom. Khudayberganov [h], Samim Sherzod [i,*]

[a] Department of Networks and Cybersecurity, Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, Amman, Jordan
[b] Department of Artificial Intelligence, College of Engineering Technology, Alnoor University, Mosul 41012 Nineveh, Iraq
[c] Department of Computers Techniques Engineering, College of Technical Engineering, The Islamic University, Najaf, Iraq
[d] Department of Computer Science Engineering, School of Engineering and Technology, JAIN (Deemed to be University), Bangalore, Karnataka, India
[e] Department of Computer Science and Information Technology, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha 751030, India
[f] Sharda School of Engineering and Sciences, Sharda University, Greater Noida, UP, India
[g] Department of General Professional Sciences, Mamun University, Khiva, Uzbekistan
[h] Department of Chemistry, Urgench State University, Uzbekistan
[i] Faculty of Engineering, Nangarhar University, Nangarhar, Afghanistan

ARTICLE INFO

Keywords:
Biochar
heavy-metal adsorption
machine learning
physicochemical properties
SHAP interpretability

ABSTRACT

Heavy-metal contamination of soils and aqueous environments poses critical ecological and health risks, necessitating efficient sorbents for remediation. This study addresses the problem of unpredictable adsorption behavior of biochar by developing a comprehensive machine-learning approach that relates its physicochemical attributes to metal-uptake efficiency. A robust dataset of 380 experiments encompassing diverse biomass origins and preparation conditions was assembled to quantify this relationship using descriptors including elemental ratios, pH, cation-exchange capacity (CEC), surface area, and structural charge. Eight algorithms (Decision Tree, AdaBoost, Random Forest, K-Nearest Neighbor, Ensemble Learning, Convolutional Neural Network, Support Vector Regression, and Multilayer Perceptron) were evaluated through 5-fold cross-validation and optimized by hyperparameter tuning. Statistical indicators (R^2, MSE, AARE%) and graphical diagnostics confirmed the CNN model as the most reliable predictor ($R^2 = 0.991$, MSE = 0.00148), capturing nonlinear physicochemical patterns with minimal overfitting. SHAP interpretation revealed C_0 and CEC as dominant determinants, while surface pH exerted inverse influence on adsorption. The hierarchical feature effects emphasize charge-controlled and diffusion-dependent mechanisms rather than morphological properties. The approach provides interpretable, transferable insight into how compositional and activation parameters govern heavy-metal retention by biochars under varying conditions. Hence, the developed predictive framework not only advances modeling precision but also supports rational design of tailored biochars for environmental detoxification applications.

1. Introduction

The persistence of heavy-metal pollutants in ecosystems has emerged as one of the most pressing global environmental concerns. Industrial discharge, mining residues, and agricultural runoff continuously elevate concentrations of toxic metals such as lead, cadmium, copper, and chromium in terrestrial and aquatic systems (Zhao et al., 2024; Leng et al., 2025). These metals do not degrade biologically and tend to bioaccumulate through trophic chains, causing severe risks to soil fertility, water quality, and human health. Conventional treatment technologies (chemical precipitation, ion exchange, membrane filtration) are often costly and inefficient for dilute or multi-component metal systems. Consequently, adsorption-based removal using carbonaceous sorbents has gained prominence due to its simplicity, regenerative potential, and suitability for decentralized water purification (Sar et al., 2023; Liu et al., 2017).

Among various low-cost adsorbents, biochar has attracted significant attention for its environmental compatibility and versatile production

* Corresponding author.
E-mail address: samimsherzod@gmail.com (S. Sherzod).

https://doi.org/10.1016/j.crbiot.2026.100367
Received 12 November 2025; Received in revised form 27 December 2025; Accepted 6 January 2026
Available online 7 January 2026

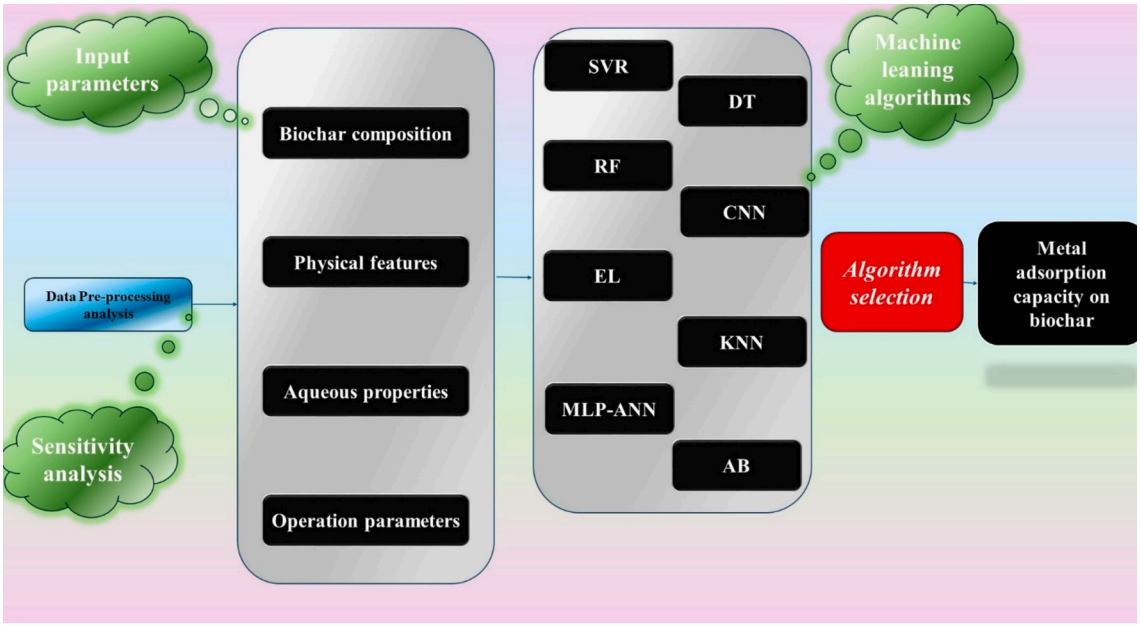

Fig. 1. Overall workflow of the study.

routes. Derived via pyrolysis of diverse biomass residues, biochar presents a porous structure enriched with oxygenated functional groups capable of binding metal ions through ion exchange, complexation, and physical adsorption (Gertsen et al., xxxx; Ahmad and Mirza, 2017). Its fabrication from waste feedstocks aligns with circular-economy strategies and carbon-sequestration goals, enabling conversion of organic waste into remediation materials. Nevertheless, despite the increasing application of biochar in heavy-metal removal, adsorption efficiencies vary drastically depending on feedstock composition, pyrolysis conditions, and surface chemistry. This inconsistency limits predictive control over performance and impedes systematic optimization of biochar design (Chen et al., 2024; Jiang et al., 2025; Hai et al., 2023).

Scientific evidence indicates that physicochemical attributes such as pH, cation-exchange capacity, surface area, and elemental ratios strongly influence metal-ion binding. Still, the extent of their contribution remains ambiguous due to complex coupling among structural, chemical, and environmental factors (Ma et al., 2021; Li et al., 2023; Liu et al., 2022). For instance, changes in surface oxygen content affect the polarity and charge density of adsorption sites, while mineral ash may introduce competing reactions that either promote or hinder uptake. The challenge therefore lies in disentangling these multivariate interactions and quantifying how each parameter governs equilibrium adsorption capacity under different preparation and operating conditions (Qu et al., 2021; Guo et al., 2024; Liang et al., 2023).

Traditional empirical methods analyze adsorption mechanisms using single-variable experiments, yet they fail to capture nonlinear dependencies typical of heterogeneous biochar systems. Linear regression and response-surface models oversimplify multivariate effects and offer limited extrapolative power beyond experimental domain boundaries (Zhao et al., 2024; Qu et al., 2021; Song et al., 2024). Consequently, predictive modeling has shifted toward data-driven techniques that exploit statistical learning to reveal hidden correlations within complex datasets. Machine-learning algorithms in particular provide a powerful platform for mapping nonlinear relationships between physicochemical descriptors and adsorption performance without assuming predefined functional forms. However, the accuracy and interpretability of such models depend critically on dataset quality and transparent evaluation procedures (Hou et al., 2024; Zhang et al., 2023; Dou et al., 2022).

Recent advancements in machine learning have successfully addressed complex environmental engineering challenges, particularly in modeling adsorption systems and production optimization. For instance, Liu et al. (Liu et al., 2025; Liu et al., 2024) demonstrated the efficacy of Gradient Boosting and CatBoost algorithms in predicting dye adsorption on hydrochar and biochar, identifying experimental conditions rather than material properties as the primary drivers of removal efficiency. To address data scarcity in bioenergy and wastewater applications, subsequent studies implemented Generative Adversarial Networks (GANs) to augment limited datasets, significantly improving the prediction of hydrogen yields and methylene blue adsorption capacities (Liu et al., 2025; Liu et al., 2025). Furthermore, Leng et al. (Leng et al., 2024) and Shen et al. (Shen et al., 2024) highlighted the critical role of feature engineering, such as utilizing molar elemental ratios and optimizing biomass mixture recipes, to enhance model interpretability and replace costly characterization metrics like specific surface area in predictive frameworks. Similarly, recent work on ammonia nitrogen removal has confirmed the robustness of tree-based ensembles when coupled with Bayesian optimization (Liu et al., 2025; Liu et al., 2025), establishing a strong foundation for data-driven environmental remediation.

However, a distinct gap remains in applying these advanced methodologies to heavy metal removal, where adsorption mechanisms differ fundamentally from organic contaminants. While dye and nutrient adsorption are often governed by pore-filling, hydrophobic interactions, or molecular size exclusion, heavy metal uptake is predominantly driven by electrostatic interactions, ion exchange, and surface complexation. Existing models often focus on single-contaminant systems or categorical metal types, which limits transferability (Leng et al., 2025; Huang et al., 2023; Huang et al., 2022; Shafizadeh et al., 2023; Wang et al., 2024). This study addresses these limitations by developing a unified, physics-informed framework that models multiple heavy metals simultaneously. Unlike prior studies that rely on generic labels, we employ specific ionic descriptors (electronegativity χ, ionic radius r, and oxidation state Ncharge) to encode the metal type, allowing the model to learn governing physical laws rather than memorizing categories.

Recognizing this scientific need, the present research aims to establish a robust predictive model for heavy-metal adsorption of biochar based on measurable physicochemical parameters. The work is significant because it unifies statistical learning, data curation, and physical interpretability, offering actionable knowledge for designing efficient carbonaceous adsorbents. The workflow (Fig. 1) encompasses dataset

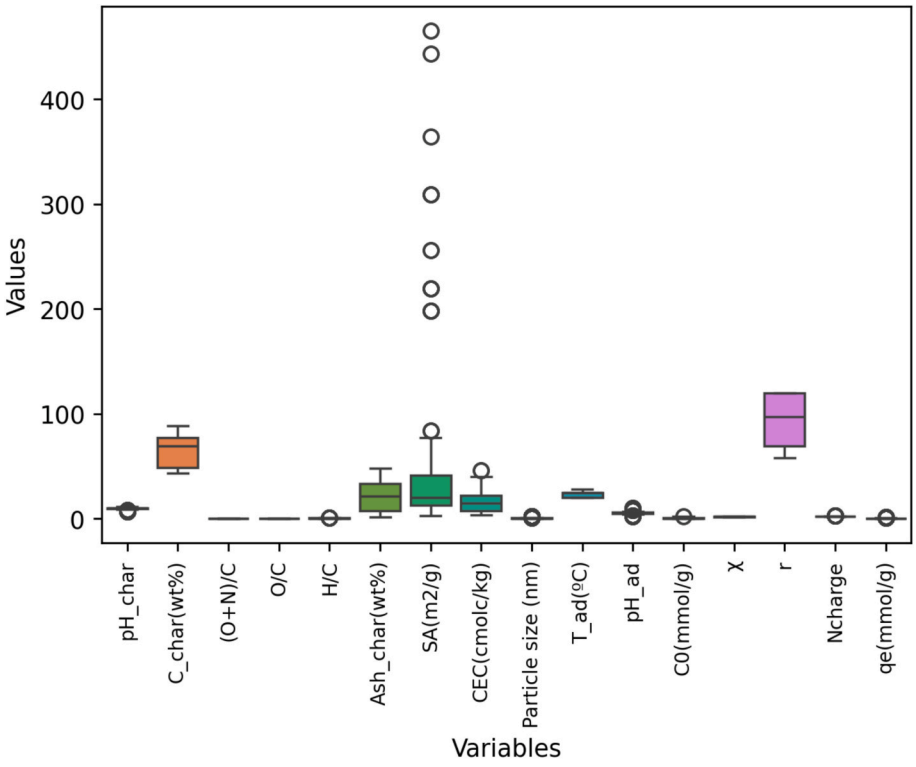

Fig. 2. Distribution and statistical dispersion of input and output variables describing biochar physicochemical properties and adsorption parameters used for model development.

compilation of 380 experimental records, preprocessing through normalization and outlier control, implementation of eight machine-learning algorithms, and evaluation under 5-fold cross-validation. Hyperparameter tuning is conducted for each model to ensure optimized performance, followed by quantitative performance metrics (R^2, MSE, AARE%) and visual diagnostics (parity and residual plots). Finally, SHAP analysis identifies the hierarchy of feature contributions, establishing the scientific hypothesis that adsorption capacity is governed mainly by charge-exchange and concentration-driven mechanisms rather than morphological factors. Collectively, these elements form a coherent framework to test the hypothesis and delineate the physicochemical determinants of metal uptake across biochar types.

2. Methodology and materials

2.1. Machine learning backgrounds

To rigorously model the complex non-linear relationships between biochar physicochemical properties and heavy metal adsorption capacity, eight distinct machine learning algorithms were employed. These models were categorized into three groups based on their learning mechanisms: tree-based ensembles, instance-based/kernel methods, and neural network architectures. Decision Tree (DT), Random Forest (RF), and AdaBoost (AB) were selected to leverage hierarchical decision boundaries and ensemble averaging, which effectively reduce variance and handle high-dimensional feature interactions inherent in biomass variability. Additionally, Ensemble Learning (EL) techniques were applied to integrate diverse weak learners into a robust predictive framework.

Complementing the tree-based approaches, K-Nearest Neighbor (KNN) and Support Vector Regression (SVR) were utilized to capture local data similarities and optimize decision hyperplanes, respectively. KNN relies on the proximity of physicochemical descriptors in the feature space, making it sensitive to local clustering of data points, while SVR employs kernel functions to map input features into higher-dimensional spaces where linear separation becomes feasible. These methods provide a baseline for understanding how geometric relationships in the data influence adsorption predictions."

Finally, deep learning architectures, specifically Multilayer Perceptron (MLP-ANN) and Convolutional Neural Networks (CNN), were implemented to extract high-level abstractions from the dataset. While MLP-ANN offers a universal approximation capability through fully connected layers, the CNN was specifically adapted for this tabular dataset (as detailed in Section 2.6) to detect local correlations between adjacent feature descriptors. Detailed mathematical formulations, architectural diagrams, and theoretical backgrounds for all utilizing algorithms are provided in the Supplementary Information (S1).

2.2. Dataset introduction

A comprehensive dataset comprising 380 data points was compiled from peer-reviewed journal articles (Leng et al., 2025) focused on heavy-metal adsorption using various biochar types. The dataset integrates experimental outcomes from diverse biomass sources, including algae, sewage sludge, manure, agricultural and forestry residues, food waste, and herbal biomass. These records represent a wide range of physicochemical conditions and preparation parameters to ensure the robustness of the adsorption models. The collected data are used to establish the relationship between physicochemical properties of biochar and its adsorption capacity for metal ions. All data were curated systematically, treating each record as a distinct experiment characterized by a set of input descriptors and one adsorption outcome.

The selected input features include pH_char, C_char(wt%), (O+N)/C, O/C, H/C, Ash_char(wt%), SA(m^2/g), CEC(cmolc/kg), Particle size(nm), T_ad(°C), pH_ad, and C_0 (mmol/g). It should be noted that while initial concentrations in literature are often reported in varying units (mg/L, mmol/g or mmol/L), all values in this study were converted to mmol/g (millimoles of metal per gram of solution) to homogenize the dataset and ensure thermodynamic consistency across experiments. Additionally, to model the distinct heavy metal species (Pb(II), Cu(II), Cd(II), Zn

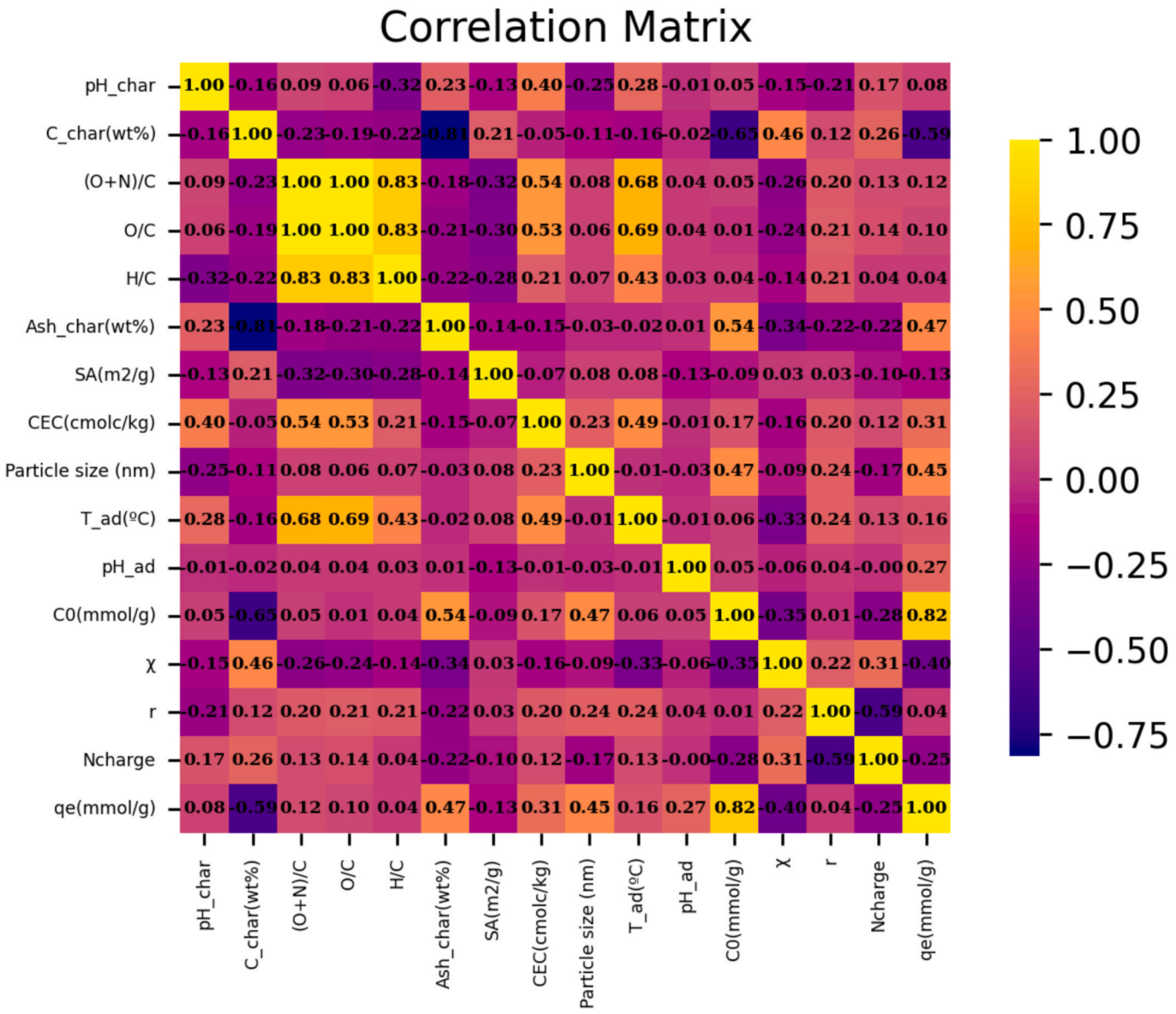

Fig. 3. Heat map of Pearson correlation coefficients illustrating linear relationships between biochar physicochemical parameters and adsorption capacity variables.

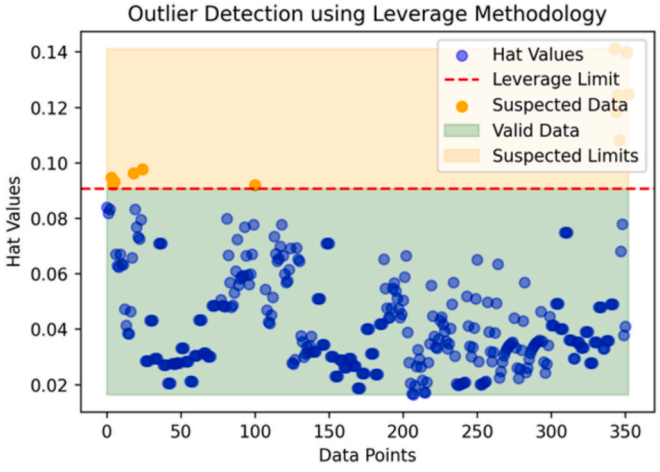

Fig. 4. Scatter plot of Hat values illustrating leverage distribution and detection of influential data points for quality assessment.

(II)) within a single unified framework, the metal type was encoded using intrinsic ionic descriptors: Pauli electronegativity (χ), hydrated ionic radius (r), and oxidation state (N_{charge}). These parameters represent intrinsic composition, surface characteristics, and specific ionic properties, respectively. This physics-informed encoding strategy ensures that the models capture the underlying chemical behavior, such as electrostatic potential and covalent bonding tendencies, rather than treating metals as arbitrary categorical labels. Each descriptor provides specific insight into the physicochemical interaction between the biochar surface and target heavy metal ions. For instance, CEC and surface area reflect the availability of exchangeable sites, whereas particle size and pH influence diffusion and speciation behavior. The single output variable, qe(mmol/g), represents the equilibrium adsorption capacity, serving as an integrative measure of adsorption performance under given experimental conditions.

The data were divided into two sets: 90 % for training and 10 % for validation, allowing unbiased evaluation of machine-learning models. Standard preprocessing steps such as normalization and outlier control were applied to ensure consistent scaling across features. This partitioning aims to provide a statistically reliable foundation for model development and testing, minimizing overfitting and improving

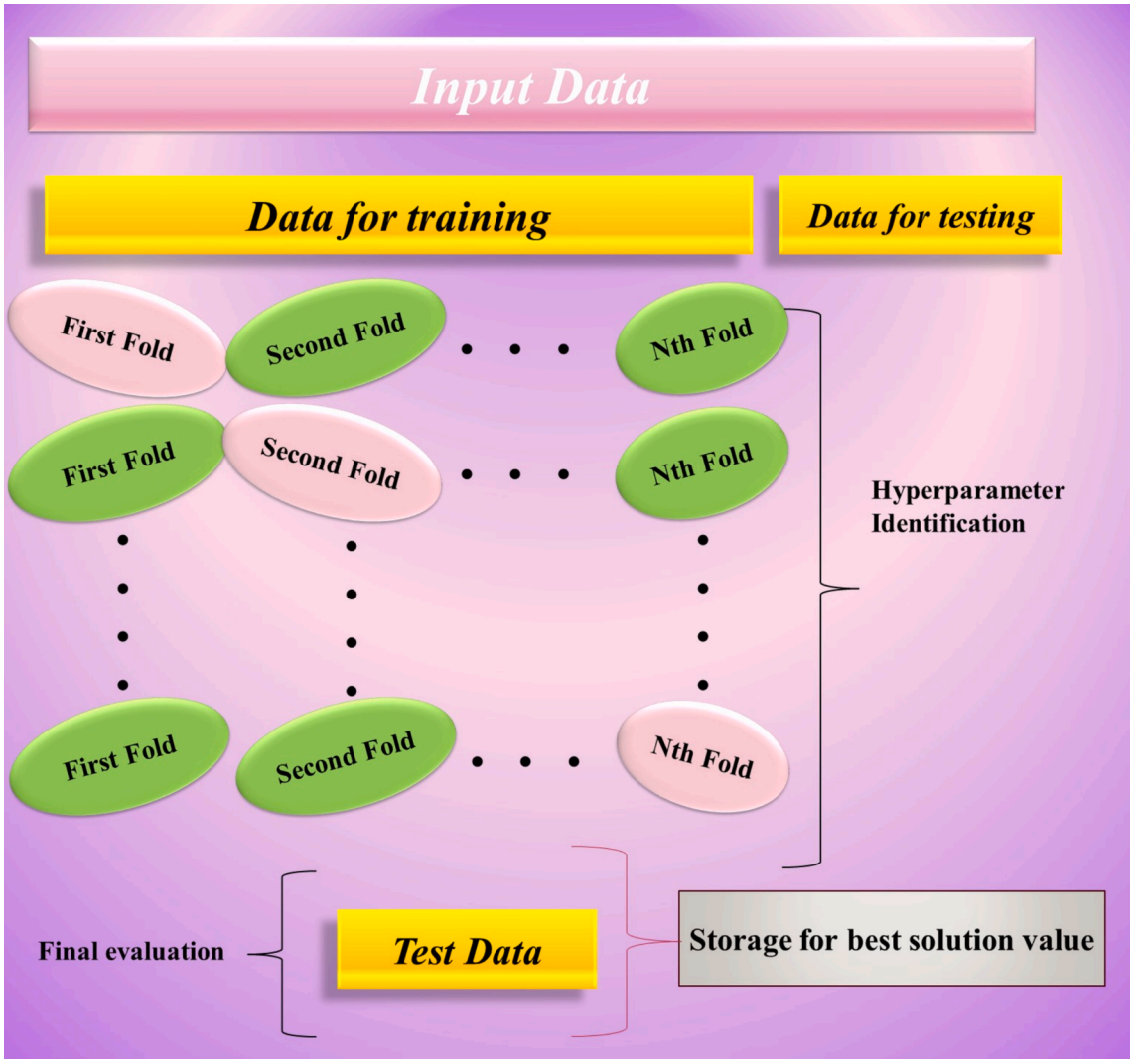

Fig. 5. 5-Fold schematic work flow.

generalization efficiency across different biomass sources and preparation conditions.

Specifically, to address the inherent heterogeneity arising from diverse experimental protocols, a rigorous data harmonization strategy was employed. All concentration data were converted to a unified mass-basis unit (mmol/g) to resolve density-dependent variations found in volumetric reports. Furthermore, to prevent numerical bias where features with larger magnitudes dominate the learning gradient, all input variables were transformed using Min-Max Normalization via the equation:

$$x_{norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

This transformation scales all physicochemical descriptors to the range [0,1], ensuring that the model weights are optimized based on feature correlation rather than raw numerical scale.

Fig. 2 illustrates the distribution, frequency, and range of all variables, presenting a visual summary of heterogeneity within the dataset. The box plot highlights varying scatter intensities among parameters, with some features (such as surface area and ash content) exhibiting wide ranges due to biomass diversity, while others remain relatively stable such as atomic ratios and pH values. The variability observed reflects the complex physicochemical nature of biochar derived from multiple sources and under different activation and pyrolysis conditions, ensuring that the dataset sufficiently captures realistic experimental conditions for predictive modeling of adsorption performance.

2.3. Sensitivity analysis

Fig. 3 presents the Pearson correlation matrix visualized as a heat map, illustrating the degree of linear association between the output variable qe(mmol/g) and each physicochemical descriptor of biochar. The Pearson coefficient ranges from −1 to +1, where positive values indicate direct correlation and negative values indicate inverse correlation. This matrix provides a quantitative view of how individual physical and chemical factors influence the adsorption capacity. Parameters with strong positive coefficients contribute synergistically to enhanced adsorption, while negative correlations signal oppositely directed effects that tend to limit the adsorption efficiency under identical conditions.

The correlation between qe and specific features such as C_char(wt %), Ash_char(wt%), SA(m^2/g), and C_0(mmol/g) appears notably positive, suggesting that higher carbon content, surface area, and initial concentration facilitate greater binding affinity toward heavy metals. From a physicochemical perspective, elevated carbon fraction and surface heterogeneity improve surface functionality and pore accessibility, enabling improved diffusion and ion exchange processes. Similarly, a larger surface area creates more active sites, and higher initial concentration increases adsorption driving force, both contributing to

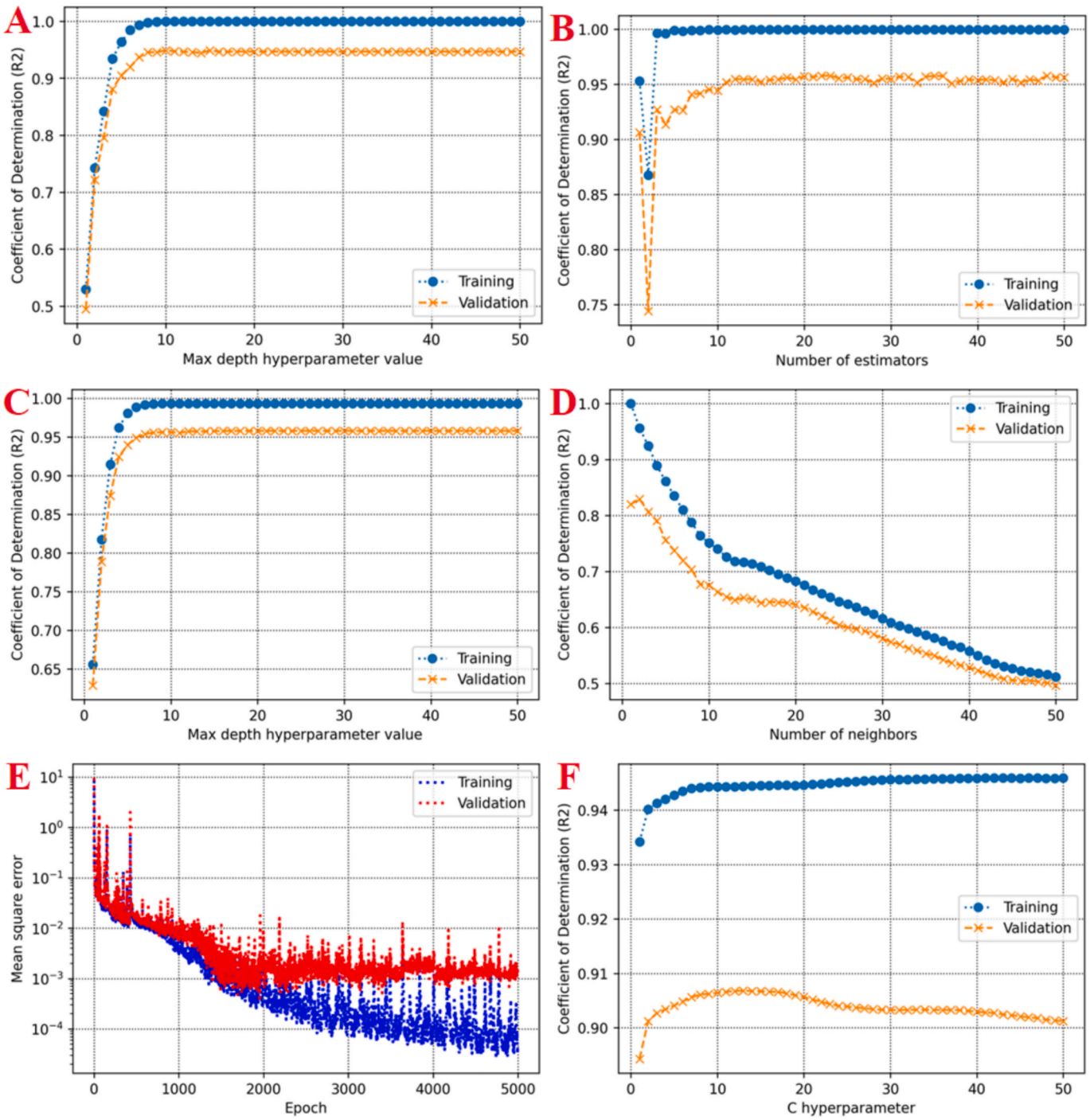

Fig. 6. Hyperparameter optimization profiles for the applied machine learning algorithms: (A) Decision Tree max depth, (B) AdaBoost estimators, (C) Random Forest max depth, (D) KNN neighbors, (E) CNN training epochs (MSE), (F) SVR C-parameter, and (G) MLP-ANN training iterations (MSE). Blue dotted lines represent training performance; orange/red dashed lines represent validation performance.

enhanced qe values through mass-transfer and equilibrium mechanisms.

Conversely, mild negative correlations are observed between qe and attributes such as r, Ncharge, and pH_char, implying that specific electrostatic or structural constraints can suppress metal uptake. Theoretically, higher surface negative charge or excessive mineralization may decrease available active sites or block fine pores, reducing accessible sorption domains. Moreover, the reduction trend linked with surface charge density indicates possible competition between metal cations and protonated species, diminishing adsorption potential under certain conditions. Thus, the correlation matrix not only distinguishes favorable and unfavorable physicochemical contributors but also underscores that adsorption efficiency reflects the delicate balance among surface composition, charge, and structural traits governing the ion–biochar interaction.

2.4. Outlier detection

Assessing data integrity is crucial for building reliable prediction models, particularly in complex systems such as heavy-metal adsorption by biochar. In this study, outlier detection was carried out using leverage

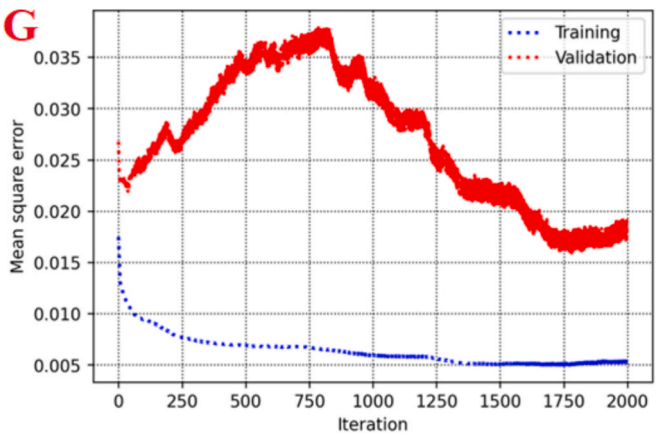

Fig. 6. (*continued*).

Table 1
Statistical performance metrics (R^2, MSE, and AARE%) for all machine-learning models during training and testing phases used in adsorption capacity prediction.

Model	R2			MSE			AARE%		
	Training	Test	Total	Training	Test	Total	Training	Test	Total
Decision Tree	1.000	0.646	0.950	0.00004	0.05736	0.00589	3.9	104.2	14.2
AdaBoost	1.000	0.965	0.995	0.00002	0.00560	0.00059	23.8	72.9	28.8
Random Forest	0.996	0.933	0.987	0.00042	0.01092	0.00149	8.3	61.3	13.7
KNN	0.958	0.959	0.958	0.00483	0.00667	0.00502	16.7	24.9	17.6
Ensemble Learning	0.996	0.929	0.987	0.00040	0.01148	0.00153	14.4	66.9	19.7
CNN	1.000	0.991	0.998	0.00004	0.00148	0.00018	25.8	60.4	29.3
SVR	0.943	0.921	0.940	0.00643	0.01284	0.00708	1579.2	2336.8	1656.5
MLP-ANN	0.954	0.892	0.945	0.00527	0.01743	0.00651	422.3	1008.3	482.0

statistics derived from Hat values, which quantify the influence of each observation on the model's overall fit. The Hat matrix originates from linear regression theory and is expressed as (Hemmati-Sarapardeh et al., 2013; Hemmati-Sarapardeh et al., 2018; Bemani et al., 2023; Bemani et al., 2023)

$$H = X(X^T X)^{-1} X^T, \qquad h_i = H_{ii} \tag{10}$$

where X is the feature matrix. Each diagonal element of H, called a Hat value, measures the leverage that the corresponding data point exerts on model estimation. By monitoring these leverage values, influential or aberrant examples that can distort learning outcomes are identified and controlled before model training.

Fig. 4 presents a visualization of Hat values for all data points. A leverage threshold of 0.092 was defined based on the empirical formula

$$h_{limit} = \frac{3(p+1)}{n} \tag{11}$$

where k is the number of predictors and n is the total data size. Points exceeding this limit are potential outliers with disproportionately high influence on regression coefficients or decision boundaries in machine-learning models. In this dataset, only approximately 2 percent of the samples surpass that cutoff, implying that the dataset is largely homogeneous with minimal leverage distortion. The small fraction of high-leverage observations reflects specific biochars produced under extreme pyrolysis conditions or possessing unusual surface features but does not compromise the reliability of the full dataset for predictive analysis.

High-quality data, characterized by consistent measurement scales and low influence outliers, ensures stability and generalizability of machine-learning predictions. The limited presence of high-Hat instances confirms the absence of severe data bias and supports robust outcomes across algorithms such as RF, SVR, and MLP-ANN. This preprocessing step mitigates numerical instability, prevents model overfitting to extreme values, and maintains proper representation of the natural variability inherent in biochar physicochemical characteristics. Hence, the quality control through Hat-value examination underpins the statistical soundness and reproducibility of the predictive framework.

2.5. Model construction and evaluation

Robust evaluation of model performance is essential to ensure predictive reliability and generalization beyond the training data. In this study, the K-fold cross-validation approach was employed with K = 5, providing a systematic methodology to assess consistency and minimize overfitting. This technique divides the dataset randomly into five equal subsets (folds). During each iteration, four folds are used for training and one fold for validation, rotating until each subset has served once as a validation set. This process yields five independent performance estimates that are averaged to produce a statistically representative measure of model accuracy, error, and stability. The method is preferred for datasets of moderate size, such as the present 380-point set, because it balances computational efficiency with reliable error estimation.

The theoretical basis of K-fold cross-validation arises from variance reduction through repeated re-sampling. By using mutually exclusive folds, the model encounters diverse training and testing configurations, revealing its capability to adapt to unseen data while maintaining structural integrity. The predicted output for each fold y_i^{pred} is compared to the observed value y_i^{obs} using predefined metrics such as R^2, RMSE, MAE, and MAPE, which quantify precision, deviation, and consistency. These metrics are then averaged across folds according to

$$M_{avg} = \frac{1}{K}\sum_{k=1}^{K} M_k \tag{12}$$

where M_k represents the performance measure for the k^{th} fold. This aggregation reflects overall model robustness and helps identify

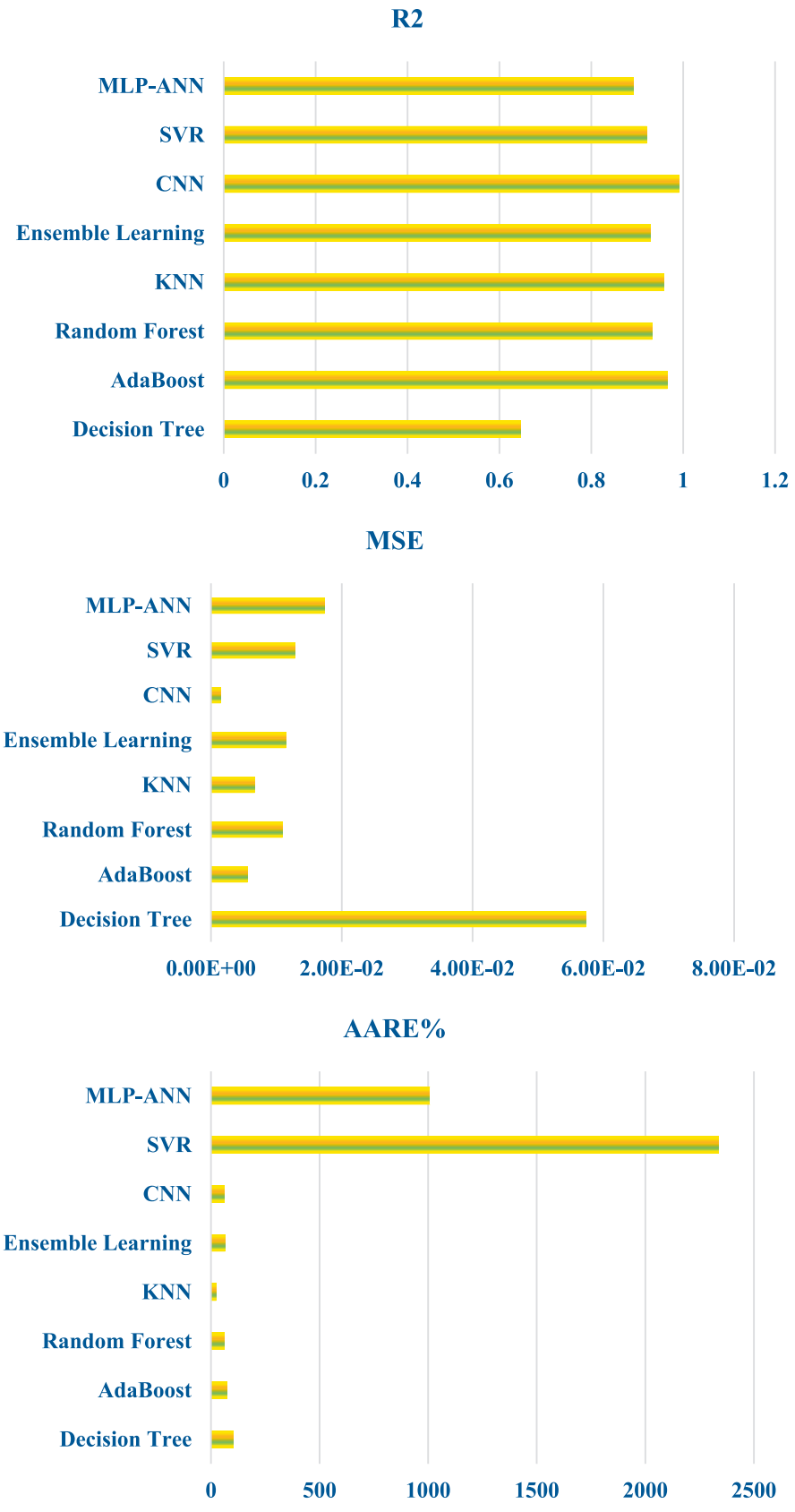

Fig. 7. Schematic representation of the test-phase performance comparison among machine-learning models based on statistical evaluation metrics.

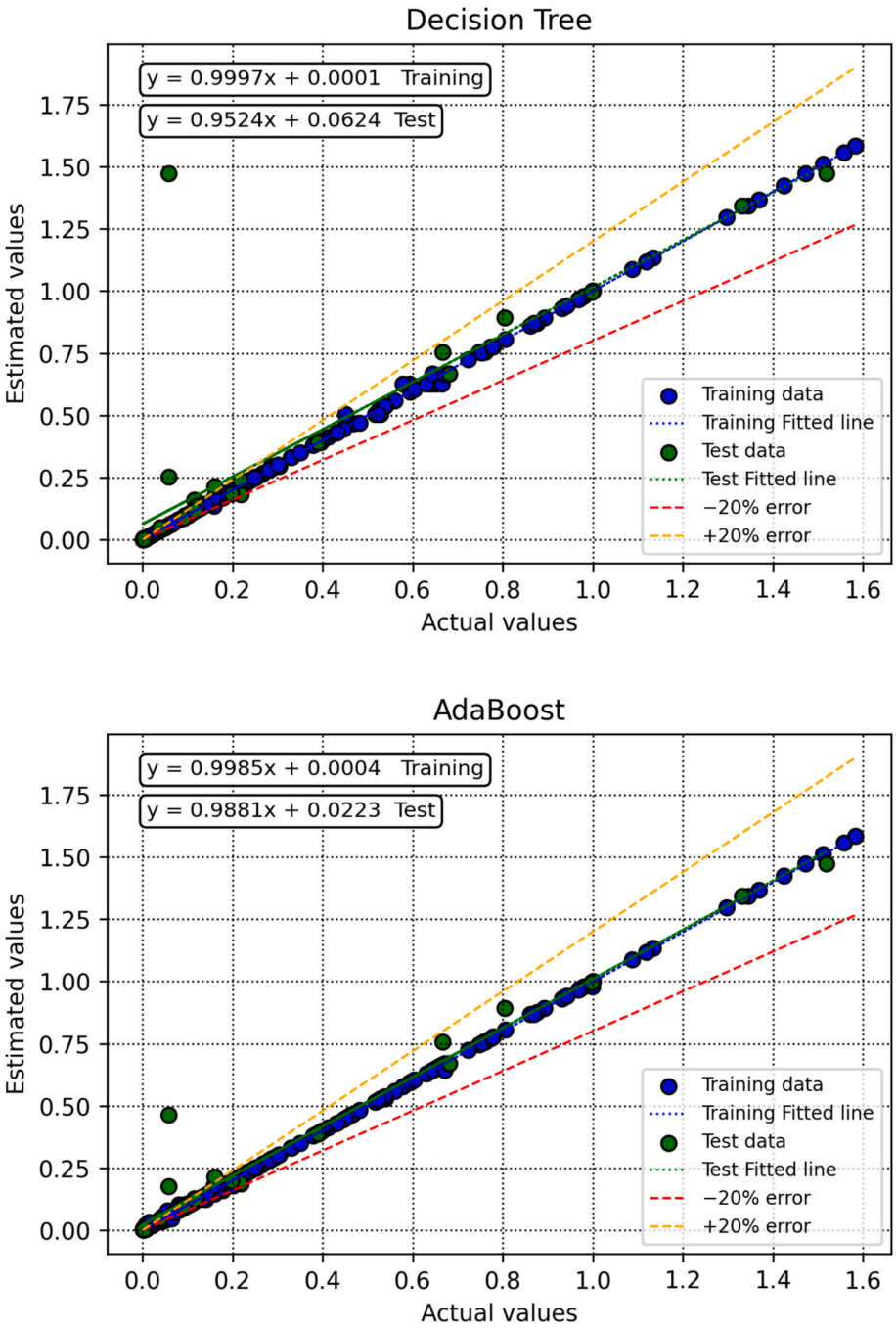

Fig. 8. Cross-plots of predicted versus experimental adsorption capacities for all machine-learning models during training and test phases.

algorithms that maintain low variance across folds.

Implementing the 5-fold validation ensures that all samples contribute both to training and evaluation, reducing bias linked to particular subsets. The balanced partitioning provides a reliable estimate of generalization capacity under varying data distributions. This approach forms the foundation of the comparative assessment among algorithms such as RF, SVR, AdaBoost, and MLP-ANN, confirming their predictive validity in modeling the heavy-metal adsorption capacity of biochar. Through repeated partitioning and error averaging, the K-fold method verifies that the developed models learn intrinsic physicochemical relationships rather than memorizing specific experimental configurations. (Fig. 5).

Accurate evaluation of model outputs is indispensable in validating prediction capabilities and determining the quality of regression modeling for heavy-metal adsorption estimation. Three statistical metrics were employed for this purpose: R^2, MSE, and AARE%. These indicators collectively measure how effectively the predicted adsorption capacities (qe_{pred}) match the experimental values (qe_{exp}). R^2 quantifies overall goodness of fit, MSE represents the average magnitude of prediction error, and AARE% evaluates relative deviation in percentage form. Together, they provide a balanced evaluation across precision, accuracy, and consistency criteria for both training and validation datasets.

The coefficient of determination (R^2) measures the proportion of variance in experimental data that is explained by the model. An R^2 value close to 1 indicates strong alignment between predictions and

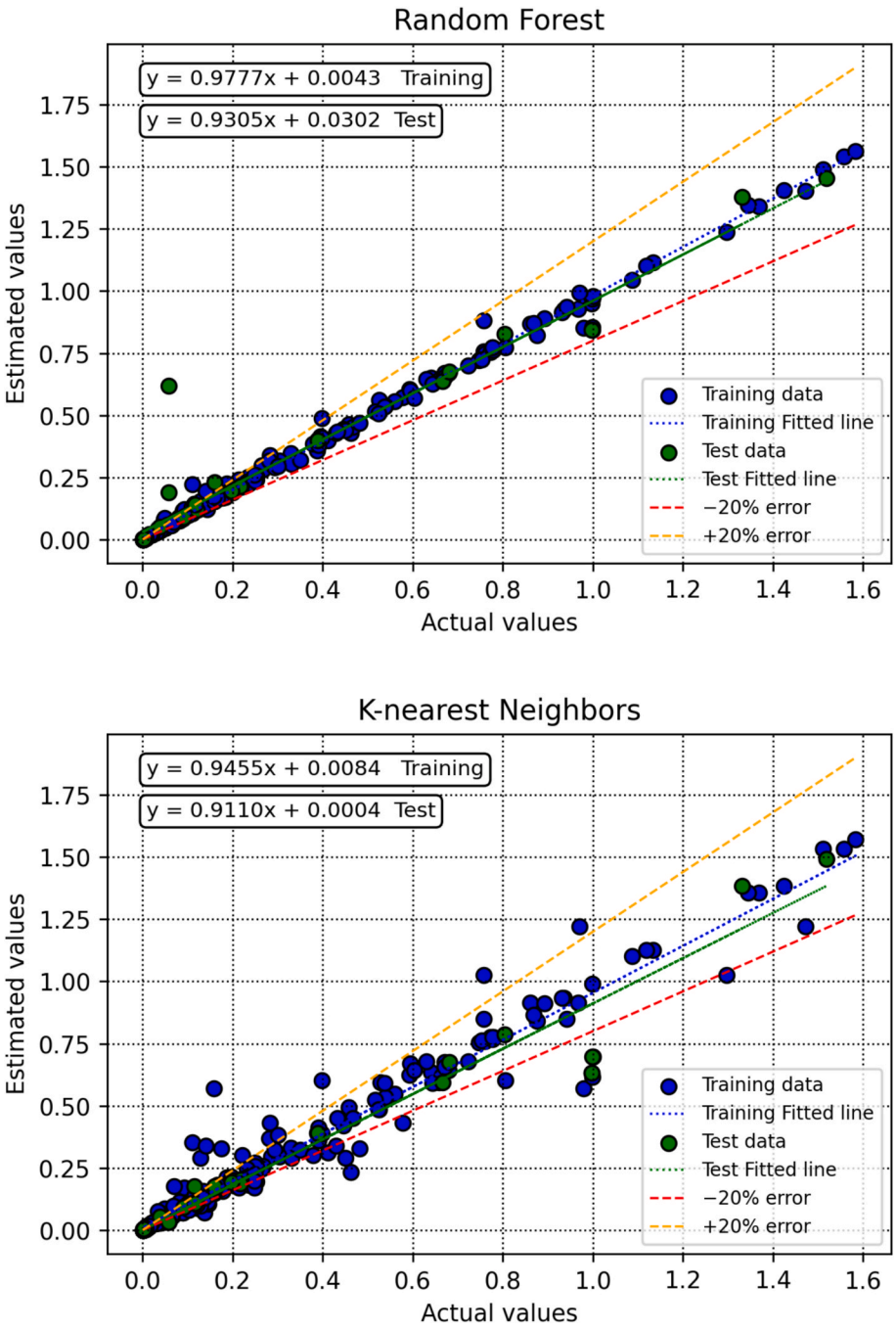

Fig. 8. (*continued*).

observations, while lower values signify inadequate representation of underlying relationships. The theoretical formula is expressed as (Daryasafar et al., 2018):

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

where n is the total number of data points, $qe_{exp,i}$ and $qe_{pred,i}$ are experimental and predicted adsorption capacities, and qe_{exp} denotes their mean. R^2 reflects how well the variation in outputs is captured by the model and is widely used as a reliability indicator for regression performance.

The Mean Squared Error (MSE) quantifies the absolute magnitude of prediction deviation without regard to direction. It is highly sensitive to large residuals and provides a measure of average model accuracy by penalizing greater discrepancies more heavily. The formulation is given by (Meybodi et al., 2015):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(qe_{\exp,i} - qe_{pred,i})^2 \tag{14}$$

Smaller MSE values reflect greater proximity between predicted and observed qe values, indicating improved numerical precision of the model. MSE is particularly valuable for comparing algorithms with different complexity levels, as it emphasizes stability and numerical accuracy across varying data distributions.

The Average Absolute Relative Error (AARE%) represents the mean percentage deviation between predicted and observed adsorption outcomes, providing normalized interpretability across different scales. This metric presents a straightforward yet sensitive indicator of relative

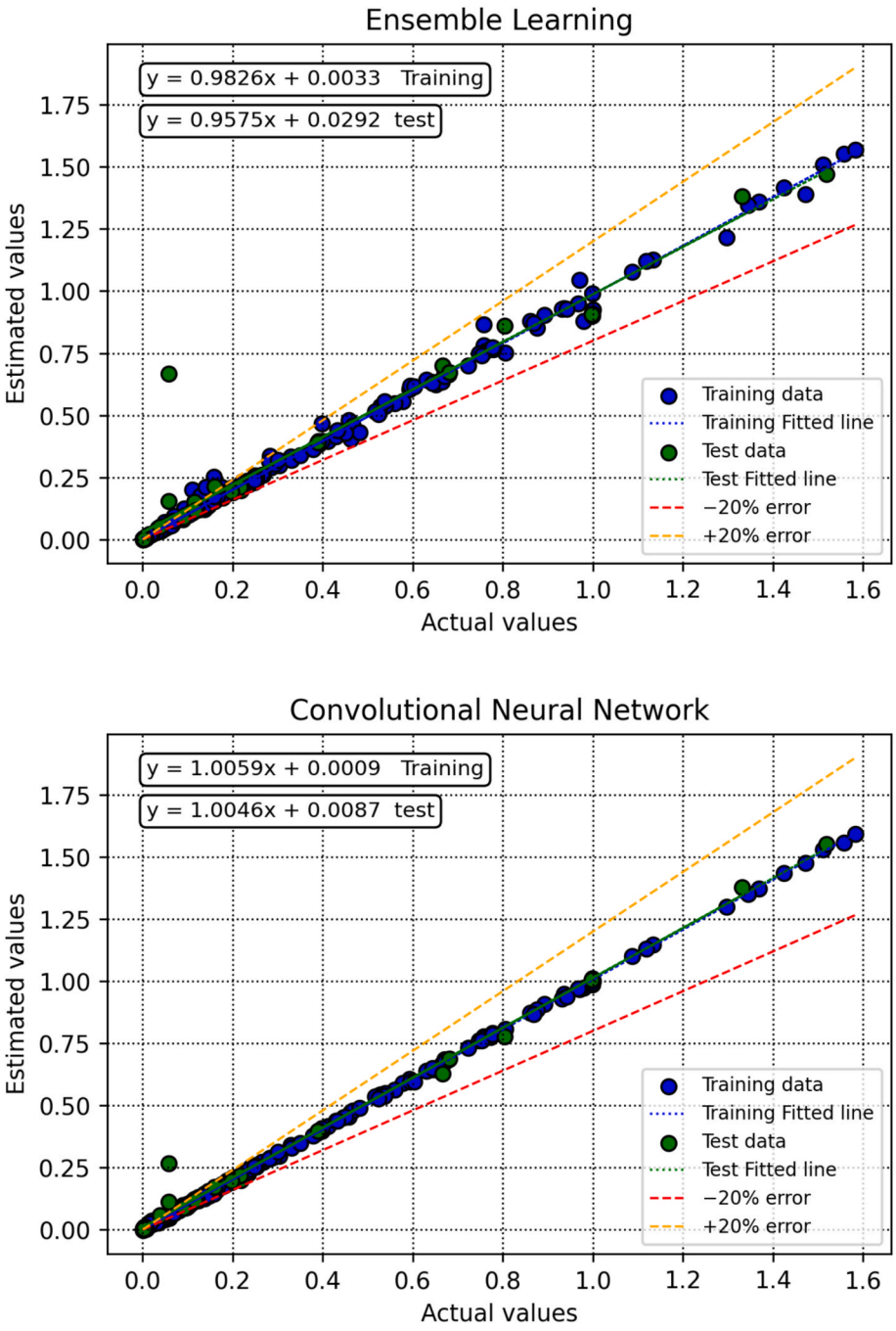

Fig. 8. (*continued*).

predictive error that captures local discrepancies more clearly than absolute measures. It is expressed as (Abbasi et al., 2023):

$$AARE\% = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{qe_{\exp,i} - qe_{pred,i}}{qe_{\exp,i}}\right| \quad (15)$$

Lower AARE% values demonstrate improved agreement between simulation and experimental adsorption performance, highlighting the confidence level of each algorithm's predictive capability. The combined use of R2, MSE, and AARE% delivers a comprehensive quantitative evaluation of regression quality, balancing goodness-of-fit with real-world consistency and numerical robustness for assessing the prediction of heavy-metal adsorption efficiency by biochar.

3. Results and discussions

3.1. Hyperparameter optimization and structural sensitivity analysis

The optimization of model hyperparameters is critical for balancing the bias-variance trade-off and preventing overfitting. Fig. 6 (A–C) illustrates the tuning process for tree-based and ensemble algorithms. For DT and RF, increasing the maximum tree depth initially yields a sharp improvement in R^2 for the validation set, reflecting the capture of nonlinear adsorption patterns. However, performance plateaus beyond a depth of 10, while the training R^2 approaches 1.0, indicating that further complexity only increases computational cost and overfitting risk without aiding generalization. Similarly, the AdaBoost model (Fig. 6-B) exhibits a saturation point around 36 estimators, confirming that a finite

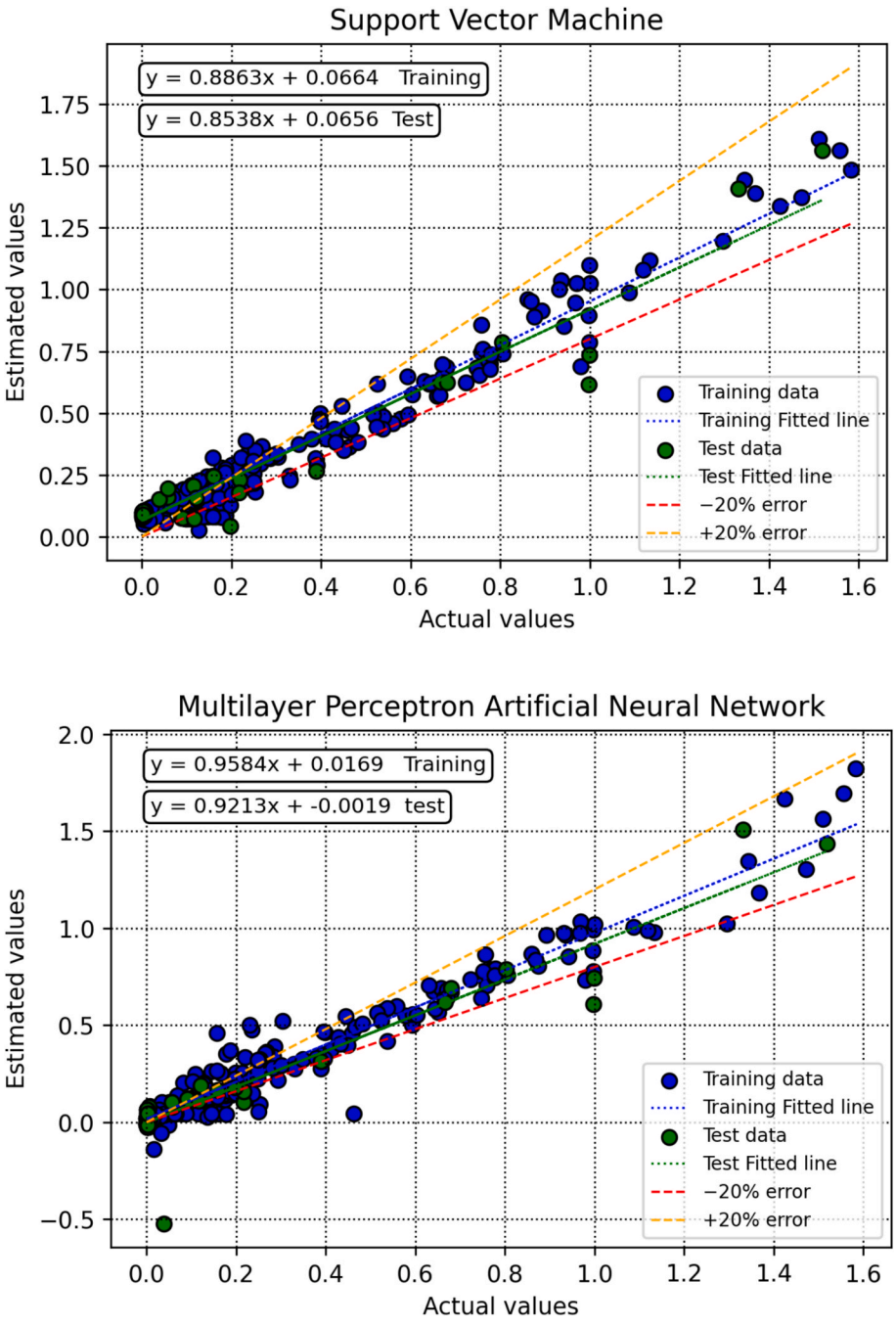

Fig. 8. (*continued*).

ensemble of weak learners is sufficient to stabilize the boosting error gradient."

In the case of instance-based and kernel methods, the hyperparameter influence reflects the underlying geometric structure of the data (Fig. 6 D, F). KNN model shows a rapid decline in accuracy as the number of neighbors (k) increases beyond 2. This high sensitivity to local proximity suggests that heavy metal adsorption is governed by specific, localized physicochemical matches rather than broad regional averages in the feature space. Conversely, SVR model (Fig. 6 F) demonstrates an optimal regularization parameter (C) around 14. Lower C values result in underfitting (excessive margin stiffness), while values exceeding this threshold introduce noise sensitivity, slightly degrading the validation R^2.

Finally, the iterative learning dynamics of neural network architectures are depicted in Fig. 6 (E, G) using MSE metrics. CNN (Fig. 6 E) displays a classic learning curve where training error continuously diminishes, but validation error stabilizes and exhibits stochastic noise around 1800 epochs, marking the optimal stopping point to prevent memorization. Interestingly, the MLP-ANN (Fig. 6 G) reveals a complex validation trajectory that minimizes around 1750 iterations after overcoming initial fluctuations. This divergence between the monotonic decrease in training error (blue lines) and the non-monotonic behavior of validation error (red lines) underscores the necessity of Early Stopping techniques to secure the most robust weights for predicting adsorption capacity."

3.2. *Comparative evaluation of predictive performance and generalization*

The comparative evaluation presented in Table 1 demonstrates notable variations in predictive efficiency among all tested algorithms.

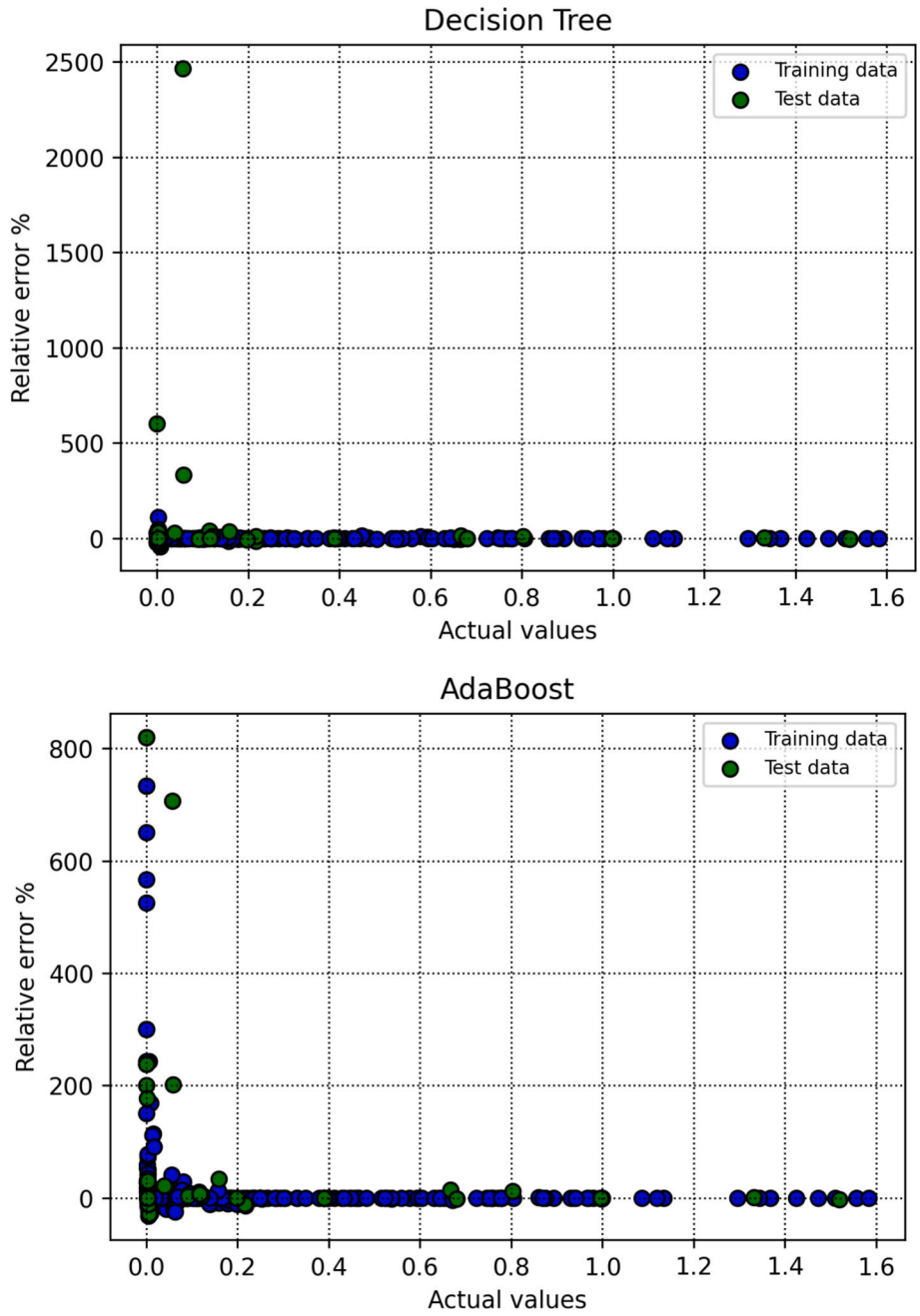

Fig. 9. Relative error (%) distribution of predicted adsorption capacities across algorithms for training and test datasets.

The deep learning CNN model achieved the highest test $R^2 = 0.991$ and maintained the lowest test MSE = 0.00148, indicating strong generalization and minimal overfitting. Its training $R^2 = 1.000$ closely matches the test value, confirming numerical stability and effective convergence. Similarly, AB also performed exceptionally well, with test $R^2 = 0.965$ and very low MSE = 0.00560, revealing robust adaptation to complex nonlinearities and relatively balanced learning behavior. In contrast, traditional models such as DT exhibited substantial overfitting, evidenced by an extreme training $R^2 = 1.000$ but notably reduced test $R^2 = 0.646$, accompanied by a sharp rise in test AARE% = 104.2. This gap demonstrates excessive reliance on training relationships and loss of predictive generality. Addressing the potential concern of overfitting given the high training accuracy ($R^2 \approx 1.0$), the robustness of the CNN model is evidenced by the minimal performance gap between the training and testing phases ($R^2_{test} = 0.991$). A true case of overfitting would exhibit a sharp decline in testing accuracy. This stability was achieved through Early Stopping (halting optimization at ~1800 epochs when validation error minimized) and cross-validation, confirming that the model has successfully captured the deterministic physicochemical relationships governing adsorption rather than memorizing statistical noise.

Among the ensemble algorithms, RF and combined EL approaches maintained balanced behavior between training and testing datasets. The RF model achieved test $R^2 = 0.933$ with moderate MSE = 0.01092, accompanied by reasonable AARE% = 61.3, suggesting stable generalization across varied input physicochemical attributes. The EL yielded comparable results ($R^2 = 0.929$, MSE = 0.01148), confirming the benefits of aggregation in minimizing prediction variance. Although KNN showed consistent training and test $R^2 \approx 0.958$, indicating no major overfitting, its test AARE% = 24.9 implies weaker accuracy under

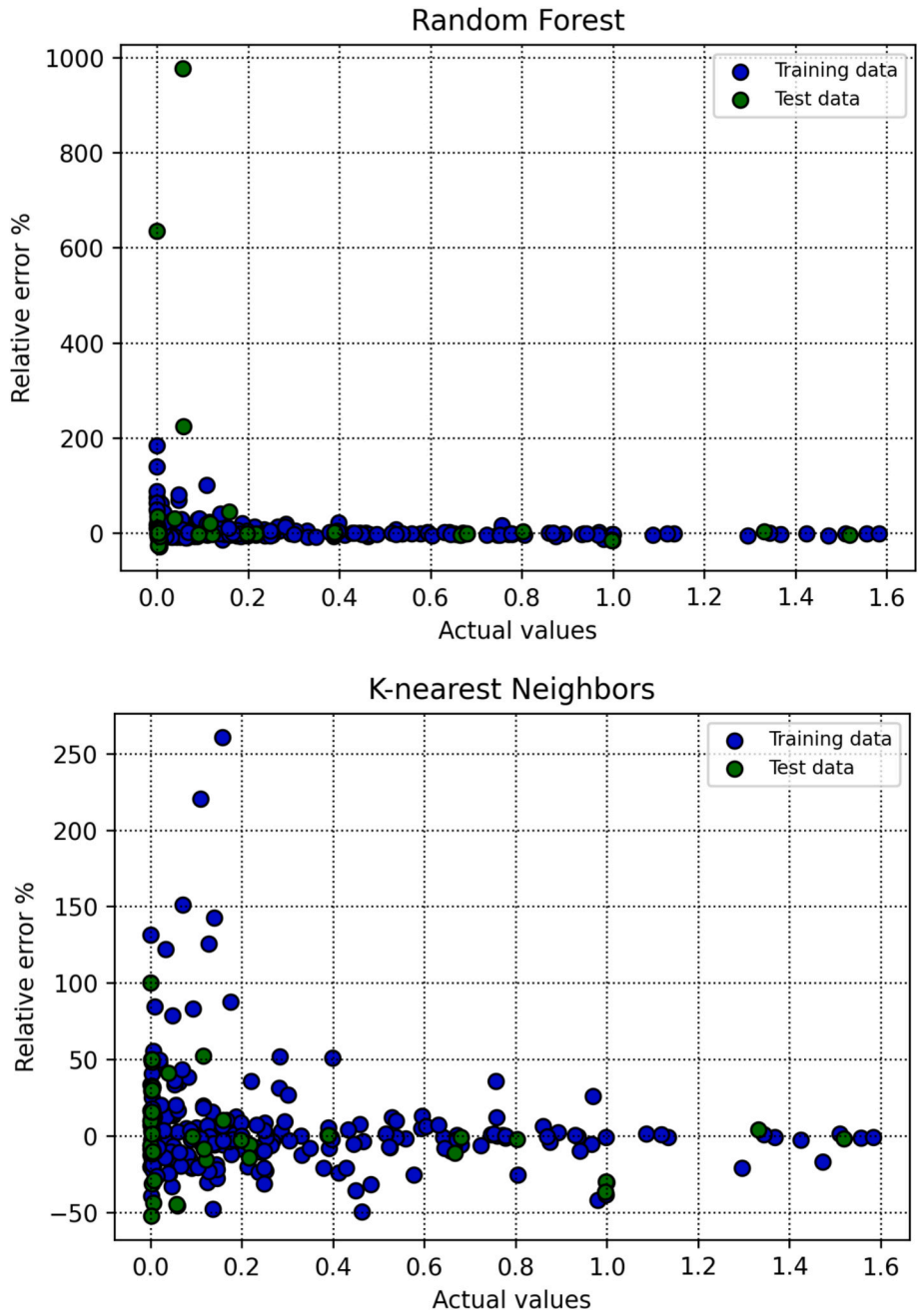

Fig. 9. (*continued*).

high-complexity adsorption conditions. Both RF and EL thus exhibit reliable predictive robustness but fall slightly behind AB and CNN in capturing nonlinear interactions with the same precision level.

Models based on kernel or fully connected architectures displayed contrasting tendencies. The SVR model, despite acceptable test $R^2 = 0.921$, recorded exceptionally high AARE% > 2000, which signals poor scale normalization and instability in predicted adsorption magnitudes. Likewise, the MLP-ANN exhibited diminishing test performance ($R^2 = 0.892$) and elevated AARE% > 1000, reflecting overfitting beyond 1750 iterations and insufficient regularization. Comparatively, CNN stands out as the best-performing model, combining maximal accuracy and minimal overfitting, followed closely by AB and RF. Conversely, DT, SVR, and MLP-ANN are categorized as weaker predictors with varying degrees of bias and generalization loss, emphasizing the importance of precise hyperparameter calibration and architecture selection for reliable adsorption modeling.

Fig. 7 provides a schematic visualization of the test-phase performance comparison among all models listed in Table 1, highlighting their relative predictive accuracy and error magnitudes across the evaluated metrics (R^2, MSE, and AARE%). This figure effectively summarizes the overall statistical behavior and stability patterns of the algorithms in a unified graphical framework.

Fig. 8 illustrates the parity plots of actual versus predicted adsorption capacities, reflecting the calibration and generalization efficiency of each algorithm. The blue and green markers represent training and test phases, respectively, revealing how closely predictions follow the ideal 1:1 line. The CNN model displays the tightest clustering of both training and test points along this line, denoting excellent uniformity and minimal bias between phases. Similarly, AB and RF exhibit nearly aligned patterns, confirming their ability to capture nonlinear relationships within the physicochemical descriptors while maintaining robust generalization. In contrast, DT shows a visible divergence between blue

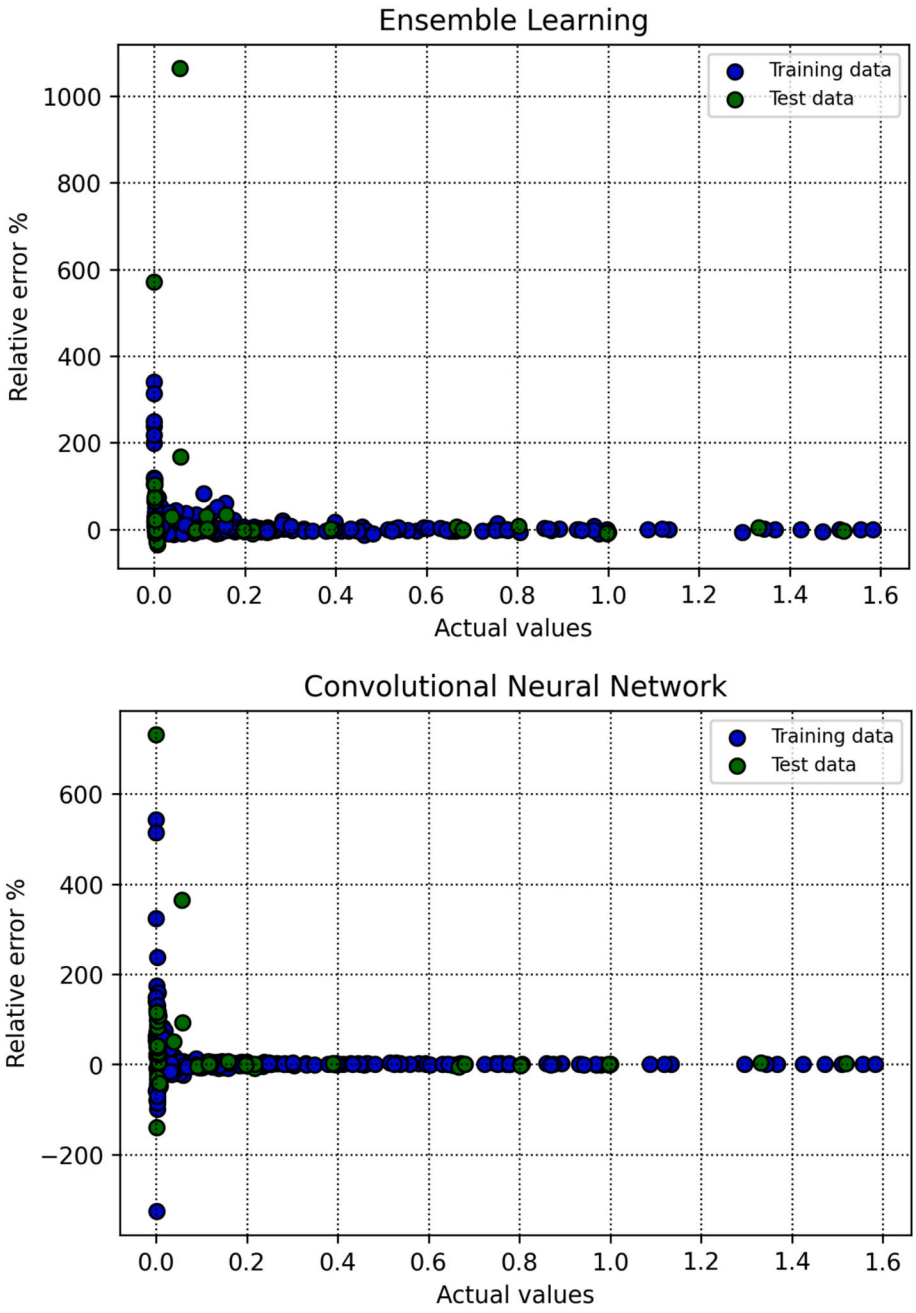

Fig. 9. (*continued*).

and green distributions, marking the typical presence of overfitting where the training fit remains perfect but test predictions scatter widely.

Further inspection of Fig. 8 reveals that models such as KNN and EL sustain balanced distributions across both data splits, demonstrating moderate but consistent performance, with points concentrated near the parity line though slightly more dispersed than the top-performing models. Conversely, SVR and MLP-ANN display significant scatter, especially in the test data, suggesting weak adaptability and potential scale mismatch. The wide dispersion of predicted values from the 1:1 trend signifies under- and over-prediction behaviors typical of poorly optimized architectures. Collectively, the comparative observation emphasizes that CNN remains the most stable and reliable model, followed by AB and RF, while SVR, MLP-ANN, and DT show limitations in maintaining consistency between training and validation outcomes.

Fig. 9 presents the distribution of relative error percentages for the adsorption output predicted by different algorithms, illustrating the dispersion and symmetry of residual behavior between training and test phases. The blue points (training) and green points (testing) indicate model consistency and stability under independent data evaluation. The CNN model shows the narrowest error distribution for both phases, confirming exceptional reliability and minimal deviation from experimental adsorption values. Similarly, AB and RF exhibit dense clustering of relative errors around the zero line, evidencing accurate fitting and low random fluctuation. In contrast, the DT and EL models reveal substantial separation between blue and green points, where training errors remain near zero but test errors extend widely, an explicit indicator of overfitting and reduced generalization control under unseen conditions.

An overview of the figure indicates that algorithms with balanced architecture and ensemble averaging, such as RF and AB, maintain confined error bands, consistent with the desired residual pattern for robust regression. Conversely, KNN shows moderate error variability yet maintains a stable form compared with deep and kernel-based models.

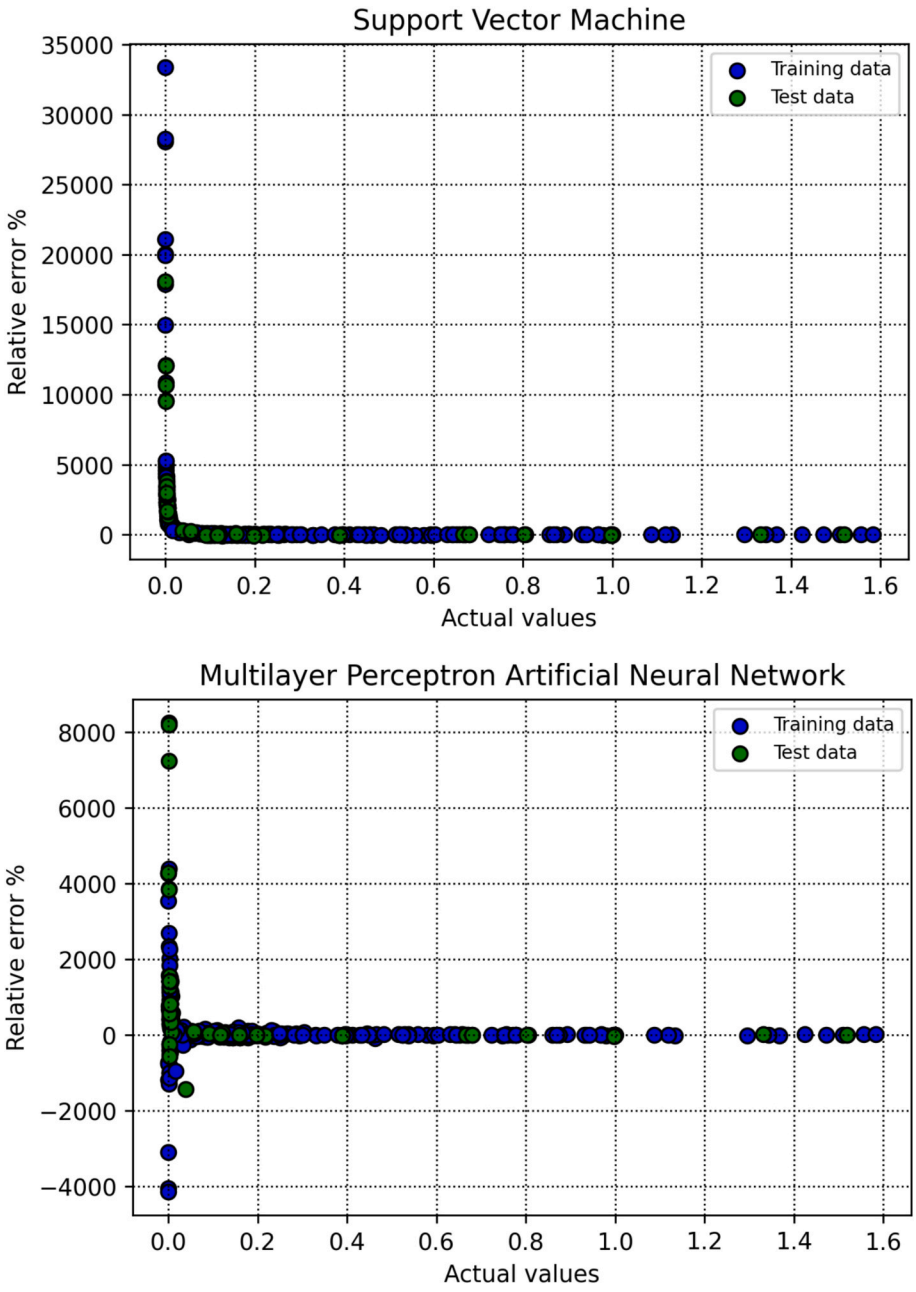

Fig. 9. (*continued*).

The SVR and MLP-ANN models present extensive error scattering and asymmetry, suggesting sensitivity to data scaling and network instability under validation. These fluctuations imply that their predictive structures fail to internalize the dominant adsorption relationships effectively. Overall, Fig. 9 underscores that the lowest and most symmetric error patterns belong to CNN, followed by AB and RF, while DT, SVR, and MLP-ANN exhibit non-uniform error expansions typical of overfitted or under-regularized models.

The integrated assessment of predictive performance combining quantitative metrics and graphical diagnostics confirms distinct contrasts among the examined algorithms. The CNN model consistently demonstrates superior accuracy, stability, and generalization, its test phase yields the highest statistical fit with negligible discrepancy between actual and predicted adsorption capacities, as evidenced by dense clustering along the 1:1 line and minimal relative error dispersion around zero. AB and RF follow closely, showing similarly narrow error bands and balanced parity plots indicative of robust ensemble learning and reduced variance propagation. In contrast, DT exhibits pronounced overfitting, where training precision appears perfect but test results deviate widely, accompanied by asymmetric residual distributions. SVR and MLP-ANN record the highest error magnitudes and weakest correlation with experimental data, denoting poor convergence and sensitivity to input scaling, while EL and KNN maintain intermediate performance across all tests. Collectively, the combined evidence from quantitative indices and visual diagnostics unambiguously identifies CNN, AB, and RF as the most reliable approaches for modeling heavy-metal adsorption on biochar.

3.3. *Mechanistic interpretation of adsorption drivers via SHAP analysis*

To ensure that the physical interpretation aligns with the highest predictive precision achieved in this study, the SHAP analysis was explicitly conducted on the optimized CNN model. By applying the SHAP method to the deep learning architecture, it was aimed to

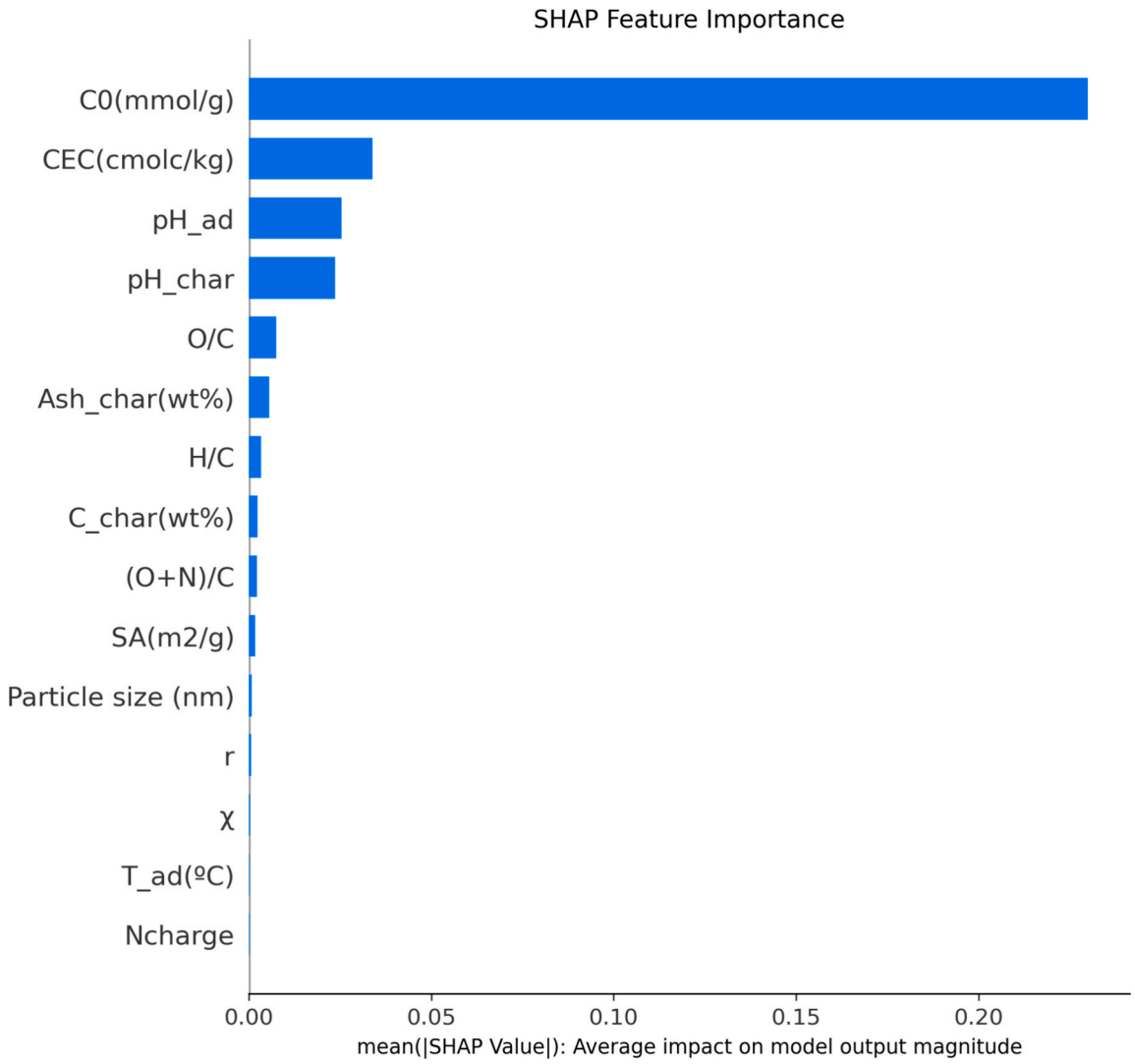

Fig. 10. Global SHAP feature-importance ranking of biochar physicochemical and adsorption variables used for predictive model interpretation.

deconvolute the black-box decision-making process of the study's best performer. Fig. 10 presents the feature-ranking outcome based on SHAP impact values, providing an interpretable quantification of each physicochemical variable's contribution to adsorption capacity prediction. The observed sequence reflects the dominance of adsorption-related parameters (C_0 (mmol/g) and CEC(cmolc/kg)) as the most influential descriptors. Their strong SHAP importances signify that both the initial concentration of metal ions and the cation-exchange capacity are primary drivers for defining equilibrium uptake, linking directly to adsorption thermodynamics and charge-neutralization behavior on the biochar surface. The middle-rank factors such as pH_ad, pH_char, O/C, and Ash_char(wt%) further highlight the effect of chemical surface activation and ash-bound mineral interactions. Variables toward the lower ranks (Particle size(nm), r, χ, T_ad(°C), and Ncharge) display minimal influence, suggesting that under studied conditions, morphological and numerical charge parameters were secondary to interfacial electrochemical mechanisms.

This figure thus indicates that the adsorption capacity is controlled predominantly by physicochemical parameters governing ion exchange, surface charge density, and solution chemistry, rather than particle morphology or thermal preparation metrics. The measurable decline of SHAP impact along the ordered series implies a hierarchical relationship where surface functional composition outweighs bulk structural features. The ranking pattern also implies considerable coupling among pH conditions and elemental ratios (O/C, H/C, (O+N)/C), supporting their combined modulation of surface polarity and binding affinity. Overall, Fig. 10 provides a direct, data-driven prioritization of descriptors responsible for adsorption performance, allowing physical insight consistent with well-established mechanisms of ion exchange and Lewis acid–base surface complexation.

Fig. 11 elucidates the local SHAP contribution behaviors through a color-gradient scatter representation, distinguishing between high (red) and low (blue) input values on the positive or negative SHAP scale. For C_0(mmol/g) and CEC(cmolc/kg), red points appear predominantly on the positive side, meaning higher values of these attributes strengthen predicted adsorption capacity, a direct reflection of increased driving force and enhanced ion-exchange potential. Conversely, for pH_char, the concentration of red points on the negative side indicates that higher char pH reduces adsorption propensity, consistent with surface deprotonation and decreased cation affinity in alkaline conditions.

Across other descriptors, such as O/C, Ash_char(wt%), and C_char(wt%), both red and blue values scatter on both sides of the SHAP axis, showing non-linear or conditional effects shaped by interaction terms and competing surface processes. This mixed pattern demonstrates that the influence of elemental and compositional ratios varies depending on co-existent parameters such as solution pH and initial concentration. The balanced distribution for r, χ, T_ad(°C), and Ncharge further supports their lower predictive leverage, where high or low numerical

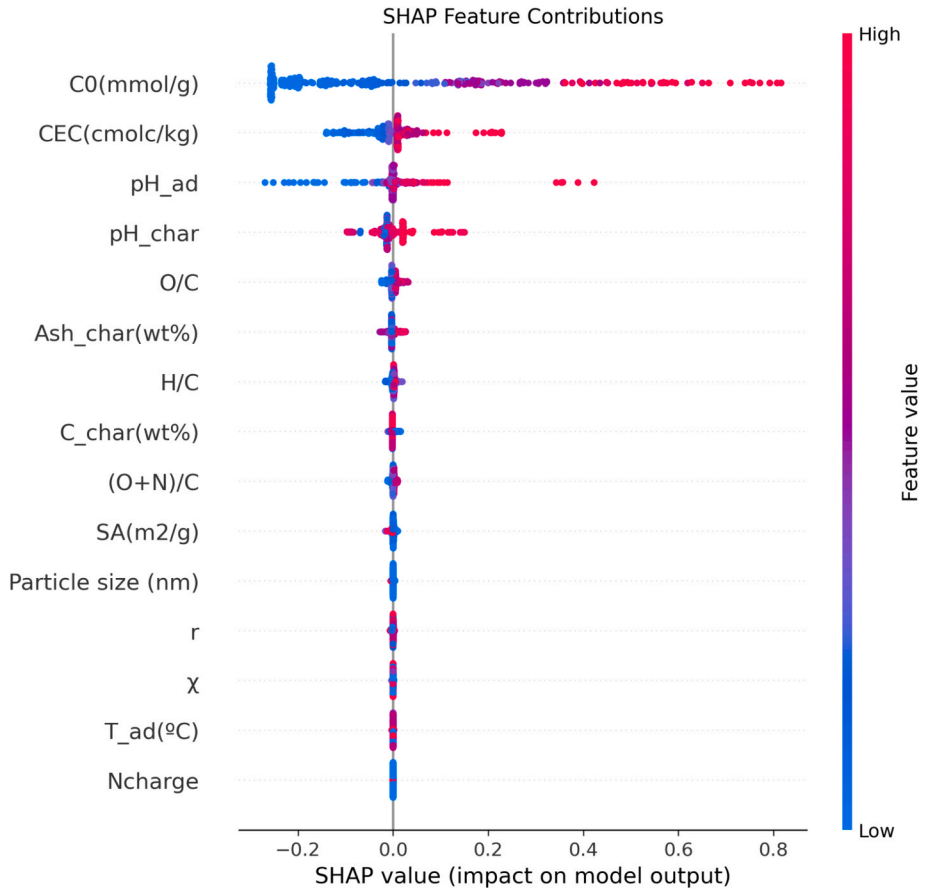

Fig. 11. SHAP value distribution showing individual feature contributions across input ranges for model interpretability analysis.

magnitudes yield comparable SHAP contributions, implying limited direct control over the adsorption outcome.

From a scientific and theoretical perspective, these SHAP-derived trends reflect physicochemical principles governing metal–biochar interactions. High C_0 values increase the mass-transfer gradient between fluid and solid phase, enhancing sorption through greater contact probability, while elevated CEC amplifies electrostatic binding sites available for ion replacement. The inverse impact of pH_char demonstrates surface charge reversal; alkaline chars exhibit more negatively charged sites, which impede cationic metal attachment, aligning with established surface complexation theory.

To further deconstruct the non-linear interplay between surface chemistry and adsorption drivers, Fig. 12 presents SHAP dependence plots for four critical surface descriptors: pH_char, O/C, (O+N)/C, and C_char. Unlike global importance rankings, these plots reveal how the impact of a specific feature changes as its value varies, while simultaneously mapping the interaction with initial concentration (C_0) via the color gradient. As observed in Fig. 12B and Fig. 12C, there is a distinct positive trend for the atomic ratios of O/C and (O+N)/C. As these ratios increase, indicating a richer density of oxygenated and nitrogenous functional groups (carboxyl, hydroxyl, and amine groups), the SHAP value rises significantly. Theoretically, this validates the dominance of surface complexation mechanisms; the model has learned that polar biochars act as superior Lewis bases, providing abundant active sites for heavy metal coordination, whereas low-ratio, highly aromatic chars lack the necessary chelating handles.

In contrast, the structural indicators exhibit an inverse relationship with adsorption efficacy. Fig. 12D shows that as Carbon content (C_char) increases, the SHAP value trends negatively. From a pyrolysis perspective, extremely high carbon content typically signals high-temperature carbonization, which leads to the volatization of functional groups and the formation of turbostratic graphitic structures. While this increases surface area, it reduces the ion-exchange potential required for heavy metals, confirming that chemical functionality is more critical than mere physical structure for this specific application. Similarly, Fig. 12A demonstrates that highly alkaline chars (pH > 10) generally yield negative SHAP contributions. This aligns with surface chemistry principles where extreme alkalinity may alter metal speciation or cause electrostatic repulsion depending on the specific metal ion charge, effectively hindering the adsorption process.

Crucially, the vertical dispersion in these plots, colored by C_0 (initial concentration), uncovers the thermodynamic governing force of the system. For almost every physicochemical feature, the red points (high C_0) cluster at the top of the vertical spread, while blue points (low C_0) sit at the bottom. This pattern, particularly visible in the O/C plot, implies a strong interaction effect: even a biochar with mediocre functional properties (moderate O/C) can achieve a high positive contribution to the prediction if the concentration gradient (C_0) is sufficiently steep to overcome surface resistance. This indicates that the CNN model has successfully encoded Fickian diffusion principles, recognizing that a strong mass-transfer driving force can partially compensate for a lack of active surface sites. This ability to disentangle the coupled effects of material properties (biochar chemistry) and environmental conditions (concentration) highlights the superior mechanistic interpretability of the proposed deep learning.

3.4. Comparisons to literature, implications and limitations

To validate the reliability of the proposed CNN framework, it is essential to compare its performance against existing generalized modeling approaches. For instance, Hafsa et al. (Hafsa et al., xxxx) achieved robust adsorption predictions ($R^2 > 0.96$) using Random Forest

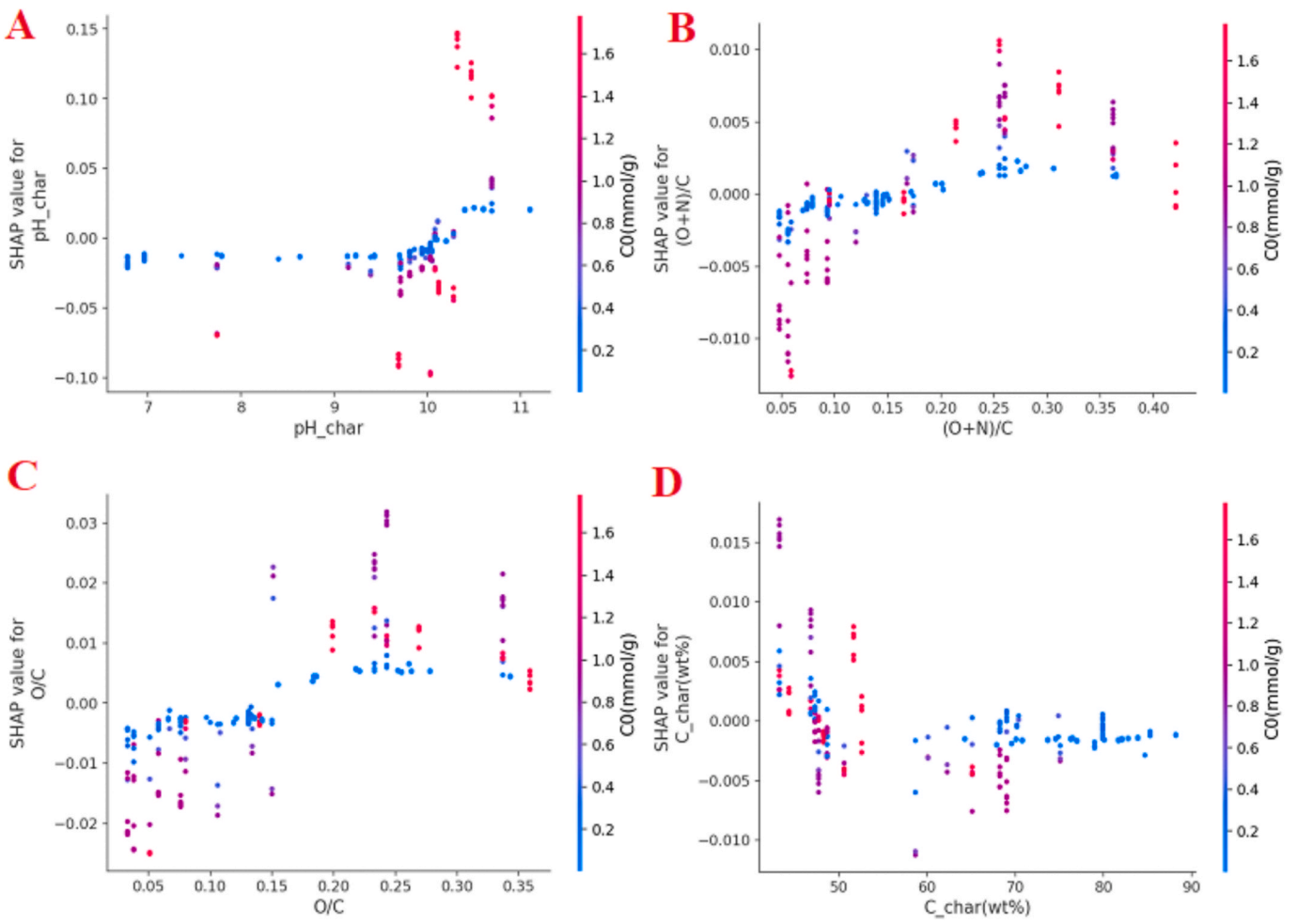

Fig. 12. SHAP dependence plots illustrating the non-linear marginal effect of (A) char pH, (B) (O+N)/C ratio, (C) O/C ratio, and (D) Carbon content on adsorption prediction. The vertical axis represents the SHAP value (impact on model output), while the color scale represents the interaction with initial concentration (C_0, mmol/g), revealing how thermodynamic driving forces modulate the influence of surface properties.

and BART algorithms for multiple heavy metal-adsorbent pairs, while Yang et al. (Jang et al., 2005) successfully mapped global soil adsorption capacities using Gradient Boosting. Our results align with these findings regarding the efficacy of ensemble methods but demonstrate that the 1D-CNN architecture yields superior stability ($R^2 = 0.991$) by effectively capturing the non-linear coupling between ionic descriptors and surface properties. Unlike traditional models that often treat different metals as categorical labels, our physics-informed encoding (utilizing electronegativity and ionic radius) allows for a more fundamental generalization, mirroring the universal kinetic models proposed by Shi et al. (Shi et al., 2013) for soil organic matter.

Furthermore, the mechanistic insights derived from our SHAP analysis are quantitatively consistent with traditional adsorption theories reported in the literature. The dominant influence of initial concentration (C_0) and CEC observed in our feature ranking mirrors the findings of Jang et al. (Jang et al., 2005), who reported that Langmuir constants for Cu, Pb, and Zn on mulch are directly governed by the driving force of concentration gradients (reaching saturation capacities ~0.3 mmol/g). Similarly, Li et al. (Li et al., 2025; Li et al., 2007) demonstrated that phosphorus recovery by FeMgCu-LDH is controlled by interlayer ion exchange and surface complexation, a mechanism quantitatively reflected in our model by the high predictive weight of charge-related features. This confirms that the black-box neural network has successfully learned the physical laws of mass transfer and electrostatic attraction, rather than merely memorizing statistical patterns.

This study fundamentally differs from single-contaminant models by offering an integrated framework that serves as a universal predictor. As reviewed by Zhao et al. (Zhao et al., 2025) and Akpomie et al. (Akpomie et al., 2022), most current ML applications in adsorption are constrained by specific adsorbate-adsorbent pairs or limited datasets. In contrast, our approach unifies diverse experimental conditions into a single predictive engine capable of handling multiple metals simultaneously. By validating these machine learning outcomes against the rigorous thermodynamic parameters (ΔG, ΔH) discussed by Qi et al. (Qi et al., 2019) and Weerasooriya et al. (Weerasooriya et al., 2003), we establish that data-driven models can achieve the same level of mechanistic transparency as classical isotherm modeling while offering significantly broader applicability for industrial design.

The outcomes of this research carry salient industrial implications for sustainable waste conversion and environmental remediation strategies. Accurate prediction of adsorption performance enables rational selection of feedstocks and pyrolysis parameters to produce biochar with tailored charge density and surface area, minimizing experimental trial-and-error. Such predictive design supports scalable manufacturing in wastewater treatment, soil-stabilization, and resource-recovery operations while reducing chemical reagent usage. Interpretable machine-learning models further provide process engineers with quantifiable descriptors, pH, CEC, and initial concentration, to guide reactor optimization and enhance sorbent regeneration efficiency.

Nevertheless, several research limitations remain. Despite extensive data aggregation, the dataset still represents laboratory-scale conditions and does not fully capture dynamic multi-metal systems typical of real

industrial effluents. Thermodynamic variations and kinetic adsorption parameters were not directly modeled. Future studies should integrate transient adsorption data and multicomponent interactions within hybrid neural architectures. Expanding databases with temperature, ionic-strength, and competing-ion descriptors will strengthen universality. Incorporating life-cycle and techno-economic analyses may also transform this predictive framework into a practical design tool for green-manufacturing sectors and circular-economy applications.

6. Conclusion

This study developed a statistically rigorous and physically interpretable machine-learning framework for predicting heavy-metal adsorption capacity of biochar using its measurable physicochemical properties. From a dataset of 380 records encompassing diverse biomass sources, eight algorithms were tested following 5-fold cross-validation and extensive hyperparameter optimization. Among them, the CNN model delivered superior accuracy (R^2 = 0.991, MSE = 0.00148), demonstrating exceptional consistency between training and testing phases. Comprehensive comparison affirmed that deep and ensemble methods (CNN, AdaBoost, and Random Forest) achieve robust generalization, whereas traditional decision-tree and kernel models exhibit overfitting or scale instability. SHAP interpretability clarified underlying causality: initial metal concentration (C_0) and cation-exchange capacity (CEC) dominate adsorption outcomes, while high char pH negatively influences uptake, confirming charge-dependent ion-exchange mechanisms over morphological control. The workflow integrating quality-controlled data, multi-algorithm validation, and post-hoc explainability provides reproducible insight into how elemental composition and surface chemistry interact under real conditions. Though constrained by single-metal laboratory datasets, the framework establishes a transferable foundation for multi-component and dynamic adsorption modeling. Future expansions incorporating kinetic descriptors and environmental factors could yield comprehensive predictive systems for industrial wastewater-treatment design and optimization. Overall, this research delineates a robust pathway connecting data-centric learning with physicochemical mechanisms, advancing both theoretical understanding and applied engineering of biochar-based sorbents.

Funding

None.

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.crbiot.2026.100367.

Data availability

Data is available on request.

References

Hafsa, N., Rushd, S., Al-Yaari, M., Rahman, M. A Generalized Method for Modeling the Adsorption of Heavy Metals with Machine Learning Algorithms, Water, 12(12) pp. 3490 doi: 10.3390/w12123490.

Gertsen, M., Perelomov, L., Kharkova, A., Burachevskaya, M., Hemalatha, S., Atroshchenko, Y., Removal of Lead Cations by Novel Organoclays Derived from Bentonite and Amphoteric and Nonionic Surfactants, Toxics, 12(10) doi: 10.3390/toxics12100713.

Abbasi, P., Aghdam, S.-K.-Y., Madani, M., 2023. Modeling subcritical multi-phase flow through surface chokes with new production parameters. Flow Meas. Instrum. 89, 102293.

Ahmad, R., Mirza, A., 2017. Heavy metal remediation by Dextrin-oxalic acid/Cetyltrimethyl ammonium bromide (CTAB) – Montmorillonite (MMT) nanocomposite. Groundw. Sustain. Dev. 4, 57–65. https://doi.org/10.1016/j.gsd.2017.01.001.

Akpomie, K.G., et al., 2022. Adsorption mechanism and modeling of radionuclides and heavy metals onto ZnO nanoparticles: a review. Appl. Water Sci. 13 (1), 20. https://doi.org/10.1007/s13201-022-01827-9.

Bemani, A., Madani, M., Kazemi, A., 2023. Machine learning-based estimation of nano-lubricants viscosity in different operating conditions. Fuel 352, 129102.

Bemani, A., Kazemi, A., Ahmadi, M., 2023. An insight into the microorganism growth prediction by means of machine learning approaches. J. Petrol. Sci. Eng. 220, 111162.

Chen, L., Hu, J., Wang, H., He, Y., Deng, Q., Wu, F., 2024. Predicting Cd(II) adsorption capacity of biochar materials using typical machine learning models for effective remediation of aquatic environments. Sci. Total Environ. 944, 173955. https://doi.org/10.1016/j.scitotenv.2024.173955.

Daryasafar, A., Daryasafar, N., Madani, M., Kalantari Meybodi, M., Joukar, M., 2018. Connectionist approaches for solubility prediction of n-alkanes in supercritical carbon dioxide. Neural Comput. & Applic. 29 (1), 295–305.

Dou, S., Ke, X.-X., Shao, Z.-D., Zhong, L.-B., Zhao, Q.-B., Zheng, Y.-M., 2022. Fish scale-based biochar with defined pore size and ultrahigh specific surface area for highly efficient adsorption of ciprofloxacin. Chemosphere 287, 131962. https://doi.org/10.1016/j.chemosphere.2021.131962.

Guo, H., Liu, Y., Lv, Y., Liu, Y., Lin, Y., Liu, M., 2024. Nitrogen doped sinocalamus oldhami lignin-based activated biochar with high specific surface area: Preparation and its adsorption for malachite green contaminant. Process Saf. Environ. Prot. 183, 992–1001. https://doi.org/10.1016/j.psep.2023.12.034.

Hai, A., et al., 2023. Machine learning models for the prediction of total yield and specific surface area of biochar derived from agricultural biomass by pyrolysis. Environ. Technol. Innovation 30, 103071. https://doi.org/10.1016/j.eti.2023.103071.

Hemmati-Sarapardeh, A., et al., 2013. Asphaltene precipitation due to natural depletion of reservoir: Determination using a SARA fraction based intelligent model. Fluid Phase Equilib. 354, 177–184.

Hemmati-Sarapardeh, A., Varamesh, A., Husein, M.M., Karan, K., 2018. On the evaluation of the viscosity of nanofluid systems: Modeling and data assessment. Renew. Sustain. Energy Rev. 81, 313–329.

Hou, D., et al., 2024. Degradation of trichloroethylene by biochar supported nano zero-valent iron (BC-nZVI): the role of specific surface area and electrochemical properties. Sci. Total Environ. 908, 168341. https://doi.org/10.1016/j.scitotenv.2023.168341.

Huang, M., et al., 2022. Fast prediction of methane adsorption in shale nanopores using kinetic theory and machine learning algorithm. Chem. Eng. J. 446, 137221. https://doi.org/10.1016/j.cej.2022.137221.

Huang, M., Yu, H., Xu, H., Zhang, H., Hong, X., Wu, H., 2023. Fast and accurate calculation on CO2/CH4 competitive adsorption in shale nanopores: from molecular kinetic theory to machine learning model. Chem. Eng. J. 474, 145562. https://doi.org/10.1016/j.cej.2023.145562.

Jang, A., Seo, Y., Bishop, P.L., 2005. The removal of heavy metals in urban runoff by sorption on mulch. Environ. Pollut. 133 (1), 117–127.

Jiang, Z., et al., 2025. Machine learning prediction of biochar-specific surface area based on plant characterization information. Renew. Energy 243, 122633. https://doi.org/10.1016/j.renene.2025.122633.

Leng, L., et al., 2024. Machine-learning-aided hydrochar production through hydrothermal carbonization of biomass by engineering operating parameters and/or biomass mixture recipes. Energy 288, 129854.

Leng, L., et al., 2025. Engineering biochar from biomass pyrolysis for effective adsorption of heavy metal: an innovative machine learning approach. Sep. Purif. Technol. 361, 131592. https://doi.org/10.1016/j.seppur.2025.131592.

Li, H., et al., 2023. Machine learning assisted predicting and engineering specific surface area and total pore volume of biochar. Bioresour. Technol. 369, 128417. https://doi.org/10.1016/j.biortech.2022.128417.

Li, X., Song, H., Cao, W., Li, M., 2025. Regenerable FeMgCu-LDHs with a memory effect for Sustainable Phosphorus Recovery and Sludge Dewatering. Arab. J. Sci. Eng. 1–16.

Li, Y., Yang, F.H., Yang, R.T., 2007. Kinetics and Mechanistic Model for Hydrogen Spillover on Bridged Metal−Organic Frameworks. J. Phys. Chem. C 111 (8), 3405–3411. https://doi.org/10.1021/jp065367q.

Liang, Y., et al., 2023. Preparation and hydrogen storage performance of poplar sawdust biochar with high specific surface area. Ind. Crop. Prod. 200, 116788. https://doi.org/10.1016/j.indcrop.2023.116788.

Liu, Z., et al., 2017. Recent advances in the environmental applications of biosurfactant saponins: a review. J. Environ. Chem. Eng. 5 (6), 6030–6038. https://doi.org/10.1016/j.jece.2017.11.021.

Liu, Z., et al., 2022. Nanoporous biochar with high specific surface area based on rice straw digestion residue for efficient adsorption of mercury ion from water. Bioresour. Technol. 359, 127471. https://doi.org/10.1016/j.biortech.2022.127471.

Liu, C., et al., 2025. Enhanced machine learning prediction of biochar adsorption for dyes: Parameter optimization and experimental validation. Carbon Research 4 (1), 46.

Liu, C., Balasubramanian, P., Li, F., Huang, H., 2024. Machine learning prediction of dye adsorption by hydrochar: Parameter optimization and experimental validation. J. Hazard. Mater. 480, 135853.
Liu, C., Li, F., Zhang, P., Balasubramanian, P., 2025. Augmented machine learning with limited data for hydrogen yield prediction in wastewater dark fermentation. npj Clean Water 8 (1), 101.
Liu, C., An, J., Nguyen, X.C., Balasubramanian, P., 2025. Machine learning prediction of hydrochar adsorption capacity for methylene blue with limited data: inspired by generative adversarial network-based augmentation. Energy & Environmental Sustainability 1 (4), 100043.
Liu, C., Balasubramanian, P., An, J., Li, F., 2025. Machine learning prediction of ammonia nitrogen adsorption on biochar with model evaluation and optimization. npj Clean Water 8 (1), 13.
Ma, D., et al., 2021. Zero-valent iron and biochar composite with high specific surface area via K2FeO4 fabrication enhances sulfadiazine removal by persulfate activation. Chem. Eng. J. 408, 127992. https://doi.org/10.1016/j.cej.2020.127992.
Meybodi, M.K., Naseri, S., Shokrollahi, A., Daryasafar, A., 2015. Prediction of viscosity of water-based Al2O3, TiO2, SiO2, and CuO nanofluids using a reliable approach. Chemom. Intel. Lab. Syst. 149, 60–69.
Qi, J., Li, Y., Majeed, H., Goff, H.D., Rahman, M.R.T., Zhong, F., 2019. Adsorption mechanism modeling using lead (Pb) sorption data on modified rice bran-insoluble fiber as universal approach to assess other metals toxicity. Int. J. Food Prop. 22 (1), 1397–1410. https://doi.org/10.1080/10942912.2019.1650764.
Qu, J., et al., 2021. Magnetic porous biochar with high specific surface area derived from microwave-assisted hydrothermal and pyrolysis treatments of water hyacinth for Cr (VI) and tetracycline adsorption from water. Bioresour. Technol. 340, 125692. https://doi.org/10.1016/j.biortech.2021.125692.
Qu, J., et al., 2021. KOH-activated porous biochar with high specific surface area for adsorptive removal of chromium (VI) and naphthalene from water: Affecting factors, mechanisms and reusability exploration. J. Hazard. Mater. 401, 123292. https://doi.org/10.1016/j.jhazmat.2020.123292.
Sar, P., Kundu, S., Ghosh, A., Saha, B., Natural surfactant mediated bioremediation approaches for contaminated soil, *RSC Advances,* 10.1039/D3RA05062A. 13(44) (2023) pp. 30586-30605, doi: 10.1039/D3RA05062A.
Shafizadeh, A., et al., 2023. Machine learning-based characterization of hydrochar from biomass: Implications for sustainable energy and material production. Fuel 347, 128467. https://doi.org/10.1016/j.fuel.2023.128467.
Shen, T., et al., 2024. Feature engineering for improved machine-learning-aided studying heavy metal adsorption on biochar. J. Hazard. Mater. 466, 133442.
Shi, Z., Di Toro, D.M., Allen, H.E., Sparks, D.L., 2013. A general model for kinetics of heavy metal adsorption and desorption on soils. Environ. Sci. Technol. 47 (8), 3761–3767. https://doi.org/10.1021/es304524p.
Song, Z., et al., 2024. Machine learning assisted prediction of specific surface area and nitrogen content of biochar based on biomass type and pyrolysis conditions. J. Anal. Appl. Pyrol. 183, 106823. https://doi.org/10.1016/j.jaap.2024.106823.
Wang, J., Chen, X., Liu, L., Du, T., Webley, P.A., Li, G.K., 2024. Vacuum pressure swing adsorption intensification by machine learning: hydrogen production from coke oven gas. Int. J. Hydrogen Energy 69, 837–854.
Weerasooriya, R., Tobschall, H.J., Wijesekara, H., Arachchige, E., Pathirathne, K.A.S., 2003. On the mechanistic modeling of As (III) adsorption on gibbsite. Chemosphere 51 (9), 1001–1013.
Zhang, Z., Huang, G., Zhang, P., Shen, J., Wang, S., Li, Y., 2023. Development of iron-based biochar for enhancing nitrate adsorption: Effects of specific surface area, electrostatic force, and functional groups. Sci. Total Environ. 856, 159037. https://doi.org/10.1016/j.scitotenv.2022.159037.
Zhao, S., Guo, J., Tang, Y., Zhou, Y., 2025. Applications of machine learning in heavy metal adsorption modeling: a review. Sep. Purif. Technol., 134168
Zhao, F., Tang, L., Song, W., Jiang, H., Liu, Y., Chen, H., 2024. Predicting and refining acid modifications of biochar based on machine learning and bibliometric analysis: specific surface area, average pore size, and total pore volume. Sci. Total Environ. 948, 174584. https://doi.org/10.1016/j.scitotenv.2024.174584.