npj | clean water

Article

Published in partnership with King Fahd University of Petroleum & Minerals

https://doi.org/10.1038/s41545-024-00429-z

Machine learning prediction of ammonia nitrogen adsorption on biochar with model evaluation and optimization

Chong Liu[1], Paramasivan Balasubramanian[2] ✉, Jingxian An[1] & Fayong Li[3] ✉

In light of escalating nitrogen pollution in aquatic systems, this study presents a comprehensive machine learning (ML) approach to predict ammonia nitrogen adsorption capacity of biochar and identify optimal conditions. Twelve ML models, including tree-based ensembles, kernel-based methods, and deep learning, were evaluated using Bayesian optimization and cross-validation. Results show tree-based ensemble models excel, with CatBoost performing best (R^2 = 0.9329, RMSE = 0.5378) and demonstrating strong generalization. Using SHAP and Partial Dependence Plots, we found experimental conditions (67.2%) and biochar's chemical properties (18.2%) most influenced adsorption capacity. Moreover, under these experimental conditions (C_0 > 50 mg/L and pH 6–9), a higher adsorption capacity could achieved. A Python-based GUI incorporating CatBoost facilitates practical applications in designing efficient biochar adsorption systems. By merging advanced ML techniques and interpretability tools, this study deepens understanding of biochar's ammonia adsorption and supports sustainable strategies for mitigating nitrogen pollution.

Nitrogen is an essential nutrient for plant growth and development, yet it may be introduced into aquatic environments through agricultural soils, livestock wastewater, urban leachate, and numerous industrial sectors such as fertilizer production, food processing, rubber, textiles, oil refining, and paper manufacturing[1,2]. Ammonia nitrogen (NH_4^+-N) represents one of the primary forms of nitrogen in aquatic systems, constituting the majority of soluble nitrogen in wastewater and serving as a key contributor to eutrophication[3,4]. It has been reported that NH_4^+-N concentrations as low as 3 mg/L can exert toxic effects on certain fish species[5]. Thus, the prompt removal of ammonia nitrogen from water bodies is critical for maintaining a healthy aquatic ecosystem[4]. Presently, the common methods employed for the removal of NH_4^+-N include biological processes, ion exchange, electrochemical treatment, chemical precipitation, breakpoint chlorination, and adsorption[3,5,6]. Among these, adsorption is highly regarded and extensively applied in water treatment technologies due to its simplicity, cost-effectiveness, and the absence of secondary pollution[4,7–10].

Currently, various adsorbents have been extensively studied for the removal of NH_4^+-N, including zeolites, bentonite, polymeric clays, biochar, and activated carbon[5,10–12]. Biochar, in particular, has attracted significant attention due to its large specific surface area, well-developed pore structure, high carbon content, abundant oxygen functional groups, and high cation exchange capacity[13–15]. Additionally, biochar as an adsorbent not only offers favorable cost-effectiveness but can also be applied to soil as a fertilizer after NH_4^+-N adsorption, thereby achieving resource recycling[15,16]. Recent studies have demonstrated that the adsorption of NH_4^+-N by biochar primarily involves physical mechanisms (physical adsorption, electrostatic attraction, and ion exchange) as well as chemical mechanisms (chemical adsorption, surface complexation, and surface precipitation)[5,8,17,18]. Consequently, the adsorption performance of biochar towards NH_4^+-N is significantly influenced by its physical and chemical properties, as well as experimental conditions[19,20]. However, traditional batch adsorption experiments are labor-intensive, time-consuming, and costly, presenting significant challenges to elucidating the complex interrelationships and influencing factors governing adsorption performance[21–23]. Therefore, there is an urgent need to establish comprehensive predictive models to clarify the independent contributions of various factors and enhance understanding of the adsorption capacity of biochar for NH_4^+-N.

Machine Learning (ML) is a technology that learns patterns from data automatically through algorithms and makes predictions or decisions based on these patterns[24]. Recent studies have employed ML to model and predict the removal of NH_4^+-N by adsorbents[25–29]. For instance, Yolcu et al.[27] proposed a hybrid predictive model combining response surface methodology (RSM) with feedforward neural networks and Elman recurrent neural networks to predict the efficiency of zeolite adsorption of NH_4^+-N

[1]Department of Chemical & Materials Engineering, University of Auckland, Auckland, New Zealand. [2]Department of Biotechnology & Medical Engineering, National Institute of Technology Rourkela, Rourkela, India. [3]College of Water Resources and Architectural Engineering, Tarim University, Tarim, China.
✉e-mail: biobala@nitrkl.ac.in; lisen8279@163.com

from landfill leachate, achieving an accuracy of 95%. Additionally, artificial neural networks (ANN) have been used to model NH_4^+-N removal by waste foundry sand, with results indicating high prediction accuracy, as reflected by a correlation coefficient exceeding 0.98[25]. Similarly, Ohale et al.[28] employed ANN and an adaptive neuro-fuzzy inference system (ANFIS) to optimize and predict the adsorption efficiency of NH_4^+-N from abattoir wastewater using iron-functionalized crab shells, with the ANFIS model ($R^2 = 0.9998$) demonstrating superior predictive performance compared to other models. Moreover, a recent study explored the combination of RSM and ANN to optimize NH_4^+-N removal from cattle manure under microwave irradiation, revealing that the RSM-ANN model exhibited higher accuracy in predicting and estimating NH_4^+-N removal efficiency compared to the standalone RSM model[26].

Although ML has been employed to predict NH_4^+-N removal from wastewater using various adsorbents, several challenges remain: (1) studies specifically focusing on the adsorption performance of biochar towards NH_4^+-N using ML are still limited[24]; (2) the heterogeneity of biochar properties and their complex interactions with NH_4^+-N introduce unique modeling challenges that have yet to be fully addressed; (3) the relationships between the physical properties, chemical characteristics, and adsorption conditions of biochar with its equilibrium adsorption capacity need more comprehensive exploration; (4) datasets are often of limited scale, lacking sufficient features and modeling support, generally encompassing only experimental conditions[25–27]. Therefore, this study aims to utilize the physical and chemical properties of biochar, along with experimental conditions, as inputs to estimate the equilibrium adsorption capacity of NH_4^+-N onto biochar. The objectives of this study are: (1) to apply appropriate data processing methods for analyzing data collected from relevant published studies; (2) to compare and select suitable ML models for predicting the adsorption performance of biochar towards NH_4^+-N (The reasons for selecting these models have been listed in Supplementary Note 1); (3) to elucidate the influence of key feature variables on adsorption capacity; and (4) to develop a Python-based application integrating the selected model to evaluate the adsorption performance of biochar.

Methods

Data collection

The data for this study were collected from published literature available in the Web of Science, Google Scholar, and Scopus databases (2014–2024). The search terms used were: "TS: [Biochar AND NH_4^+-N AND (Adsor* OR Remov*)]". Following the preliminary screening, 417 sets of adsorption data were gathered, representing 46 distinct biochar types (Data availability section). It is important to note that, compared to some previous studies, the amount of data collected in this research is sufficient for ML applications[30–33]. The parameters were divided into three categories: chemical properties of the biomass, physical properties of the biochar, and experimental conditions. The chemical properties of the biomass included carbon content (C, wt.%), the molar ratio of oxygen and nitrogen to carbon [$(O+N)/C$], the molar ratio of hydrogen to carbon (H/C), the molar ratio of oxygen to carbon (O/C), and ash content (Ash, %). The physical properties of the biochar comprised specific surface area (SSA, m^2/g), total pore volume (V, cm^3/g), and the pH of the biochar (pH_bio). The experimental conditions were experimental temperature ($Temp$, °C), the pH of the solution (pH), and the ratio of the initial concentration of NH_4^+-N to the dosage of biochar (C_0, mg/g). It is worth noting that all data collected in this study were free of bias. The methodology employed for collecting data can be referenced in a previously published work[19].

Data preprocessing

The proportion of missing values, which has not been reported in previous literature, was found to be 2.88% (pH_bio, BET, V). To address these missing values, the K-Nearest Neighbors (KNN) algorithm was employed for imputation[19,34–36]. By utilizing the most similar neighboring samples to fill in the gaps, the KNN algorithm effectively preserves the intrinsic structure and similarity of the data, avoiding the distortions that may arise from the simple mean or median imputation[37]. Before conducting ML, it is often necessary to enhance the normality of the data and stabilize variance, allowing the model to fit the data better and improve predictive accuracy[32]. In this study, normality was assessed based on the skewness and kurtosis of the data, and the Box-Cox transformation was applied to improve the model's normality[38]. The Box-Cox transformation is a statistical method that converts data to a distribution closer to normality, aiming to reduce skewness and thereby enhance the performance of ML models[38,39]. The formula for the Box-Cox transformation (BCT) is as follows[38]:

$$BCT(x) = \begin{cases} \frac{(x)^{\lambda}-1}{\lambda} & if\ \lambda \neq 0 \\ \log(x) & if\ \lambda = 0 \end{cases} \quad (1)$$

Where x represents the original data, $BCT\ (x)$ denotes the transformed data, and λ is a parameter determined through maximum likelihood estimation.

To mitigate the impact of multicollinearity among features on the ML model, this study employed Pearson correlation coefficients (PCC) to examine the relationships between variables[40]. Moreover, to ensure that features are on comparable scales and to prevent features with larger magnitudes from dominating the training process, Z-score normalization was applied to the data before model training[41]. Compared to other normalization techniques, Z-score normalization effectively addresses disparities in feature magnitudes, reducing the model's sensitivity to these differences and enhancing its robustness and accuracy[42].

Machine learning model construction

This study systematically compared 12 different ML models, encompassing six tree-based ensemble models, three kernel-based models, and three deep-learning models. The specific models included: Random Forest (RF), Gradient Boosting Decision Trees (GBDT), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Categorical Boosting (CatBoost), Extremely Randomized Trees (ET), Support Vector Machines (SVM), Kernel Ridge Regression (KRR), Gaussian Process Regression (GPR), Multilayer Perceptron (MLP), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM). One of the objectives of this study was to evaluate the effectiveness of these models in predicting the adsorption performance of NH_4^+-N on biochar. It should be noted that RF, GBDT, XGBoost, LightGBM, CatBoost, and ET are classified as tree-based ensemble models, while SVM, KRR, and GPR fall under kernel-based ML models. In contrast, MLP, CNN, and LSTM are categorized as deep learning models. The theoretical underpinnings of these algorithms are provided in Supplementary Table 1.

The dataset was split into a training set and a test set in an 8:2 ratio. Bayesian optimization combined with 5-fold cross-validation was employed to determine the optimal hyperparameters for each model[43]. Compared to conventional methods such as grid search and random search, bayesian optimization offers a more intelligent approach to finding the optimal hyperparameters while reducing computational costs[44]. This advantage is particularly notable in high-dimensional or computationally expensive search spaces. The hyperparameters and their respective ranges used in this study are detailed in Supplementary Table 1. It is worth noting that the selection of hyperparameters was based on previous studies, preliminary results, and the specific characteristics of the data in this study. Efforts were made to maintain consistency in the range of selected hyperparameters wherever possible. Additionally, to ensure model stability and avoid the impact of dataset partitioning on the model, a 5-fold cross-validation method was implemented after identifying the optimal hyperparameters, and model performance was thoroughly evaluated across 50 randomly generated training and test sets (resulting in 250 data points). The model performance was assessed using root mean square error (RMSE) and the coefficient of determination (R^2) on both the training and test sets, with the

formulas for RMSE and R^2 provided in Eqs. (2) and (3), respectively[22].

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i - x_i)^2}{n}} \quad (2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - x_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (3)$$

Where y_i and x_i denote the predicted and experimental values, respectively; $\bar{y}_i$ represents the average of the experimental values, and n is the number of samples.

Model interpretation

Feature importance and partial dependence are critical for enhancing the interpretability of ML models, facilitating a deeper understanding of the model's decision-making process and thereby improving its credibility and reliability[45]. In this study, Embedded Feature Importance (EFI) and Shapley Additive Explanations (SHAP) were employed for comprehensive analysis[19,32]. SHAP values quantify the average impact of each feature on the model's predictions and explicitly outline the contribution of individual feature values to specific predictions[46]. It has been reported that SHAP is based on game theory, with its mathematical formulation shown in Eq. (4) [47,48]. This is particularly important for explaining the dynamics of complex tree-based models and identifying key features. Additionally, partial dependence plots (PDPs) were used to illustrate the influence of critical features on adsorption capacity[22,49]. PDPs reveal the relationship between feature values and model predictions by creating copies of the dataset for each value of the selected feature, allowing the model to generate predictions and compute the average predicted values, as shown in Eq. (5)[49,50].

$$\varnothing_i = \sum_{S \subseteq N\{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!}[f(S \cup \{i\}) - f(S)] \quad (4)$$

Where $\varnothing_i$ is the SHAP value for feature i; N represents the set of all features; S is a subset of features not including I; $f(S)$ represents the model output given the subset S.

$$\hat{f}_{PDP}(x_j) = \frac{1}{n}\sum_{i=1}^{n} f(x_j, x_{-j}^{(i)}) \quad (5)$$

Where $\hat{f}_{PDP}(x_j)$ is the partial dependence function for feature x_j; $f(x_j, x_{-j}^{(i)})$ represents the model prediction given feature x_j and the other features x-$_j$; n is the number of data points, and the average is taken over the marginal distribution of the other features x-$_j$.

Results and discussion

Description of datasets

As illustrated in Table 1, the dataset describing NH_4^+-N adsorption by biochar shows substantial variability across several variables, as indicated by their standard deviations and the range between minimum and maximum values. For instance, the *BET* surface area demonstrates considerable variability, ranging from 0.41 to 185 m^2/g, highlighting the heterogeneity of biochar samples, which may be attributed to different feedstocks or processing conditions[51]. The median values for variables such as *C*, *H/C*, *O/C*, *(O + N)/C* and *pH_bio* are close to their mean values, suggesting a symmetric distribution centered around the mean. However, variables like *V* exhibit a noticeable difference between the mean (0.2519) and median (0.0078), indicating a skew in the data.

Skewness is used to assess the symmetry of a data distribution, reflecting both the direction and degree of its asymmetry[52]. Similarly, kurtosis describes the degree of peakedness or the thickness of the tails in the distribution[52]. After applying the Box-Cox transformation (BCT), the skewness for most variables has been significantly reduced, bringing their distributions closer to normality. For example, the skewness of *C* decreased

Table 1 | Distribution characteristics of ammonia nitrogen adsorption by biochar

	Mean	Std	Min	25%	50%	75%	Max	Median	Skew	Skew_BCT	Kurtosis	Kurtosis_BCT
C (wt. %)	64.65	11.88	18.60	59.50	65.23	72.79	80.55	65.23	−2.03	−0.19	5.65	−0.18
H/C	0.0042	0.0018	0.0010	0.0028	0.0042	0.0058	0.0086	0.0042	0.14	−0.11	−0.85	−0.94
O/C	0.34	0.18	0.01	0.21	0.37	0.48	0.85	0.37	0.42	−0.11	0.24	−0.30
(O + N)/C	0.39	0.19	0.06	0.28	0.40	0.50	1.02	0.40	0.92	0.01	2.05	0.37
Ash (%)	14.59	11.96	5.00	6.58	10.86	15.00	71.40	10.86	2.02	0.11	3.80	−0.75
pH_bio	9.69	0.93	7.81	9.17	9.60	10.40	12.50	9.60	0.41	0.00	0.52	−0.01
BET (m^2/g)	25.32	48.36	0.41	0.92	1.66	13.49	185.00	1.66	2.00	0.29	2.71	−1.22
V (cm^3/g)	0.2519	0.5058	0.0014	0.0030	0.0078	0.1105	1.6360	0.0078	2.03	0.23	2.46	−1.34
Temp (°C)	24.48	1.81	15.00	25.00	25.00	25.00	35.00	25.00	−1.30	0.82	9.98	20.86
pH	7.03	1.03	2.00	7.00	7.00	7.00	11.00	7.00	−0.76	0.60	8.89	7.89
C_0 (mg/g)	14.60	25.28	0.00	2.00	6.00	12.50	125.00	6.00	3.04	0.07	9.19	0.61

C carbon content, *H/C* molar ratio of hydrogen to carbon, *O/C* molar ratio of oxygen to carbon, *(O + N)/C* molar ratio of oxygen and nitrogen to carbon, *Ash* Ash content, *SSA* specific surface area, *V* total pore volume, *pH_bio* pH value of biochar, *Temp* experimental temperature, *pH* solution pH, C_0 ratio of the initial concentration of NH_4^+-N to the dosage of biochar.

from −2.03 to −0.9, indicating an improved, more symmetric distribution. Reducing skewness is crucial for subsequent statistical analyses, as it ensures more reliable and unbiased modeling. Kurtosis values provide insight into the shape of the data distribution, especially regarding tail heaviness and outliers[53]. Initially, the kurtosis for *Ash* content is 3.80, suggesting a distribution with heavier tails and potential outliers. Following the *BCT*, the kurtosis reduces to −0.75, indicating a more platykurtic distribution, which is desirable for analysis. Similarly, the *BET* surface area shows a reduction in kurtosis from 2.71 to −1.22, which implies fewer extreme values and a more balanced dataset for analysis. Overall, the application of the BCT has effectively reduced both skewness and kurtosis across the dataset, making the variables more suitable for detailed statistical analysis and predictive modeling.

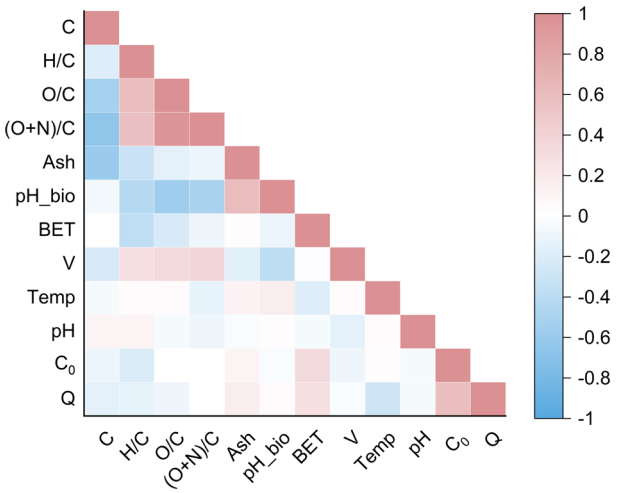

Fig. 1 | Pearson correlation coefficients between input features and the target adsorption capacity (Q).

Pearson correlation coefficient analysis

To mitigate the impact of collinearity among the various features on ML outcomes, this study conducted a *PCC* analysis to eliminate highly collinear variables[54]. Figure 1 and Supplementary Fig. 1 presents the *PCC* analysis, delineating the critical interactions between several input features and the target value of equilibrium adsorption capacity (Q). Notably, some features demonstrate significant correlations with Q. For example, the initial concentration (C_0) exhibits a strong positive correlation of 0.58 with Q, indicating that higher initial concentrations can enhance adsorption capacity. This effect can be attributed to the more significant concentration gradient, which promotes a higher driving force for mass transfer, thereby increasing the rate and extent of adsorption[55]. Additionally, the molar ratio of hydrogen to carbon (H/C) shows a weak negative correlation with Q (−0.13), which implies that a higher H/C ratio may slightly reduce the adsorption capacity. A more detailed discussion will be provided in Section (Partial dependence plots (PDP) analysis).

Based on previous research, acorrelation coefficient threshold of $|r| > 0.7$ between variables is considered an appropriate indicator of when collinearity begins to distort model estimation and subsequent predictions[52] significantly. Notably, the O/C and $(O+N)/C$ ratios exhibit a strong correlation (0.94), possibly because the nitrogen content in the collected biochar data is relatively low, resulting in a minimal impact on the ratio. This strong correlation suggests that these features could lead to collinearity issues. With O/C having a more significant influence on Q compared to $(O+N)/C$, $(O+N)/C$ was removed prior to ML to mitigate the adverse effects of collinearity.

Performance evaluation of machine learning models

Figure 2 presents the scatter plots and data distributions for the experimental and predicted adsorption capacities across 12 different ML models. Parameter optimization was executed using Bayesian optimization, and the optimal parameters are listed in Supplementary Table 1. The RMSE and R^2 values for training and test sets for each model are shown. From Fig. 2, it is evident that all models exhibit relatively consistent performance between the training and test sets, with no significant signs of overfitting. This indicates that the models possess strong generalization capabilities for

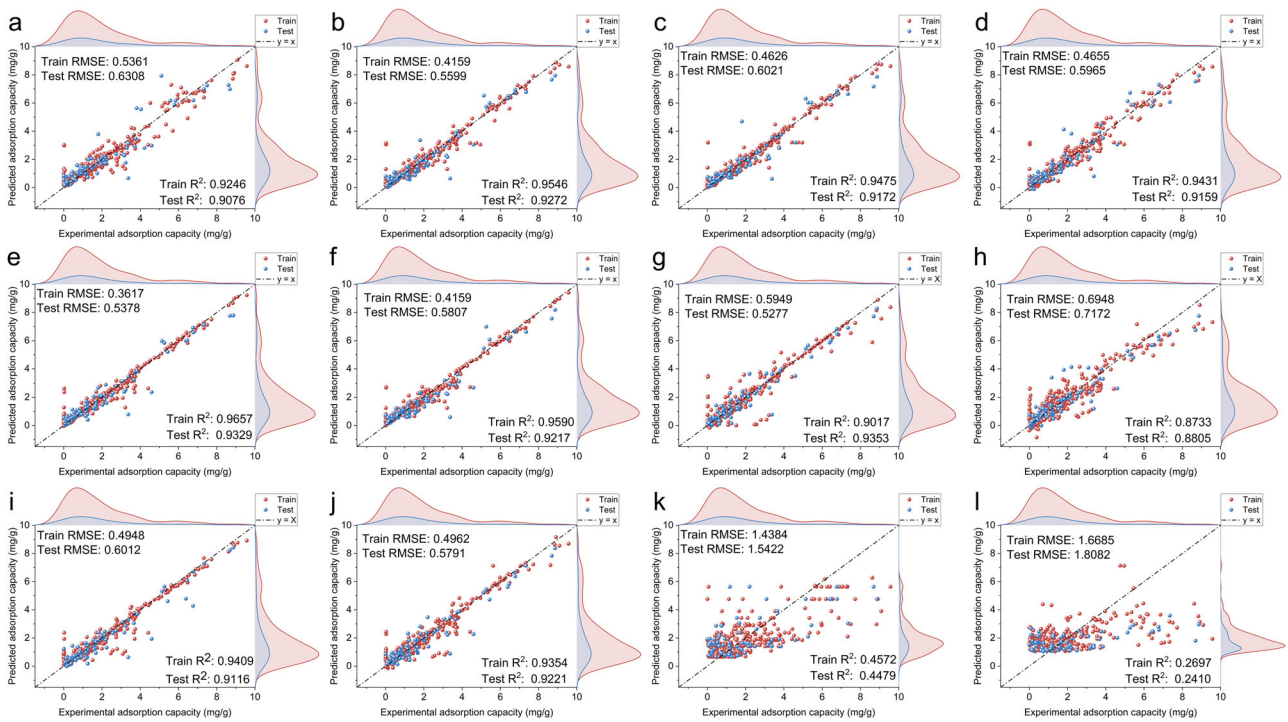

Fig. 2 | Scatter plot and data distribution of experimental and predicted values. a RF, **b** GBDT, **c** XGBoost, **d** LightGBM, **e** CatBoost, **f** ET, **g** SVM, **h** KRR, **i** GPR, **j** MLP, **k** CNN, **l** LSTM.

predicting adsorption performance. Among these models, CatBoost (Fig. 2e) stands out with exceptional performance, achieving an RMSE of 0.3617 on the training and 0.5378 on the test set. Additionally, its R^2 value for the test set reaches 0.9329, underscoring its superiority in predictive accuracy and model robustness. In contrast, other models perform slightly less well. For instance, LightGBM (Fig. 2d) and GBDT (Fig. 2b) have test set RMSE values of 0.5965 and 0.5599, respectively, which, although commendable, fall short of CatBoost in terms of precision.

The ensemble learning models, as a whole, demonstrate robust predictive performance, particularly CatBoost, LightGBM, and GBDT, which show close RMSE and R^2 values between the training and test sets, highlighting their consistency and generalization ability. Notably, CatBoost not only yields the lowest RMSE on the test set but also achieves a high R^2 value, indicating its remarkable capability in handling the complexity of the data. This superior performance can likely be attributed to CatBoost's unique mechanism for handling missing values and categorical features, allowing it to more effectively extract features in high-dimensional, complex datasets[56]. Furthermore, CatBoost employs a symmetric tree structure and a gradient-based balancing mechanism during training, which helps mitigate overfitting and enhances the model's robustness and generalization[57].

In contrast, deep learning models (such as CNN and LSTM) perform poorly on this task. For example, LSTM (Fig. 2l) records an RMSE of 1.8082 and an R^2 value of only 0.2410 on the test set, indicating insufficient capability in capturing the data patterns. This may be because deep learning models typically excel with larger datasets, where their nonlinear mapping abilities are more pronounced, but underperform with smaller datasets and limited feature tasks[58]. Similarly, CNN (Fig. 2k) follows this trend, with a test set RMSE of 1.5422 and a relatively low R^2 value, further reinforcing the limited applicability of deep learning models for this type of task. Traditional models like KRR (Fig. 2h) and SVM (Fig. 2g) also exhibit commendable stability, but their overall performance remains somewhat inferior to that of ensemble learning models, particularly in terms of the test set RMSE and R^2 values. This discrepancy may be because single kernel functions often struggle to capture the complex nonlinear relationships in more intricate datasets[59]. By contrast, ensemble learning models, especially tree-based methods, can automatically adapt to the feature distribution and nonlinear relationships within the data without the need for manually specified feature transformations or kernel functions[60].

To evaluate the stability of the ML models, this study conducted stability tests across various models. The resulting data have been uploaded to GitHub for reference (Data availability section). Figure 3 illustrates the stability assessment outcomes for the 12 ML models, showcasing CatBoost's superior performance across several metrics. Throughout both the training and test phases, the CatBoost model consistently exhibited low RMSE (Fig. 3a, b) and low MSE (Fig. 3c, d), as well as the highest R^2 values (Fig. 3e, f), demonstrating its exceptional fitting capability and predictive accuracy. Among all models, CatBoost displayed consistently low errors and high precision on the test set, outperforming most other models such as RF, GBDT, XGBoost, and LightGBM. In contrast, deep learning models like LSTM and CNN exhibited greater instability, particularly regarding RMSE and MSE on the test set (Fig. 3b, d), failing to maintain their strong performance from the training set. Notably, LSTM's performance on the test set was significantly inferior, potentially due to its complex reliance on input data features, which, in the context of this study's dataset, could not fully exploit its advantages given the smaller sample size and limited feature set. CatBoost's outstanding performance can be attributed to several built-in regularization mechanisms, such as L2 regularization, and its stochastic treatment of training data, both of which effectively prevent overfitting[61]. Moreover, CatBoost benefits from well-optimized default parameter settings, which yield stable results across a wide range of tasks. This feature further contributes to the model's reliability and stability in terms of performance.

Residuals are defined as the differences between actual and predicted values, and analyzing their distribution allows for an assessment of the model's fit and predictive accuracy[19]. The residual analysis in Fig. 4 demonstrates that most models, such as RF, CatBoost, and GBDT, exhibit relatively symmetric residual distributions, indicating minimal bias and good predictive performance. The residuals for these models are centered around zero, with their kernel density curves approximating a normal distribution. Among them, the CatBoost model's residuals are tightly clustered around zero, with the normal distribution of residuals demonstrating a sharp peak, indicating minimal bias and consistent performance. It can also be noted that a small number of data points exhibit larger errors at low concentrations, which may be due to limitations in model sensitivity or data variability at these lower levels. This suggests that CatBoost maintains a high degree of fit accuracy on both the training and test sets, thereby effectively mitigating the risk of overfitting. Moreover, the concentrated and normal distribution of the residuals further corroborates the reliability and robustness of the CatBoost model. However, certain models, particularly CNN and LSTM, show noticeable skewness in their residuals, with deviations from symmetry and bias evident in the test sets. This suggests systematic over- or under-predictions, highlighting the need for further optimization of these models or preprocessing adjustments to improve their accuracy and generalization. Therefore, based on the before-mentioned analysis and the findings illustrated in Fig. 4, CatBoost has been chosen as the suitable ML model for subsequent analyses.

Model interpretation

This study used embedded feature importance (EFI) and Shapley additive explanations (SHAP) to analyze the factors affecting NH_4^+-N adsorption in biochar. Figure 5a shows the EFI analysis, which quantifies the relative importance of different features in the ML model. The initial concentration of NH_4^+-N to the dosage of biochar (C_0) is the most influential factor, followed by experimental temperature (*Temp*), carbon content of biochar (*C*), and total pore volume (*V*). These results indicate that experimental conditions, particularly C_0 and *Temp*, play a significant role in determining adsorption capacity. Regarding the adsorption process, the dominant role of C_0 is consistent with previous studies[22,62]. The likely reason is that the relative quantities of the adsorbate (NH_4^+-N) and the adsorbent (biochar) play a critical role in the mass transfer process. Additionally, *Temp* significantly impacts the adsorption process in aqueous phases by affecting adsorption equilibrium, diffusion rates, solubility, adsorbent surface characteristics, and the adsorption enthalpy[63]. The pie chart inset in Fig. 5a demonstrates that experimental conditions account for the most significant contribution to NH_4^+-N adsorption (67.2%), followed by the chemical properties of biochar (18.2%) and its physical properties (14.7%). This might be because experimental conditions directly affect the main forces and interactions involved in adsorption, such as electrostatic forces, hydrogen bonding, and Van der Waals forces.

The SHAP analysis in Fig. 5b provides further insights into how each feature influences the model output. SHAP values indicate the magnitude and direction of each feature's impact on the prediction. The larger the SHAP value, the greater the positive impact of the feature on the prediction result; conversely, the smaller the SHAP value, the greater the negative impact. For example, C_0 and *Temp* have a strong positive influence on adsorption, as higher SHAP values are associated with higher feature values, suggesting enhanced adsorption capacity. The variability in SHAP values for features such as *Ash* content indicates nonlinear interactions, where adsorption behavior does not consistently increase or decrease with feature values. A more detailed discussion of the impact of each feature on NH_4^+-N adsorption will be provided in the next section.

Partial dependence plots (PDP) analysis

Partial dependence plots (PDP) provide insights into the relationship between specific features and the prediction outcomes of ML models[64]. The gray lines in the plots represent actual model predictions, while the blue line indicates the fitted values. Figure 6a shows that the adsorption capacity of biochar for NH_4^+-N increases with the initial concentration (C_0) up to ~50 mg/g, driven by an enhanced concentration gradient and greater interaction between NH_4^+-N molecules and the biochar surface. Beyond

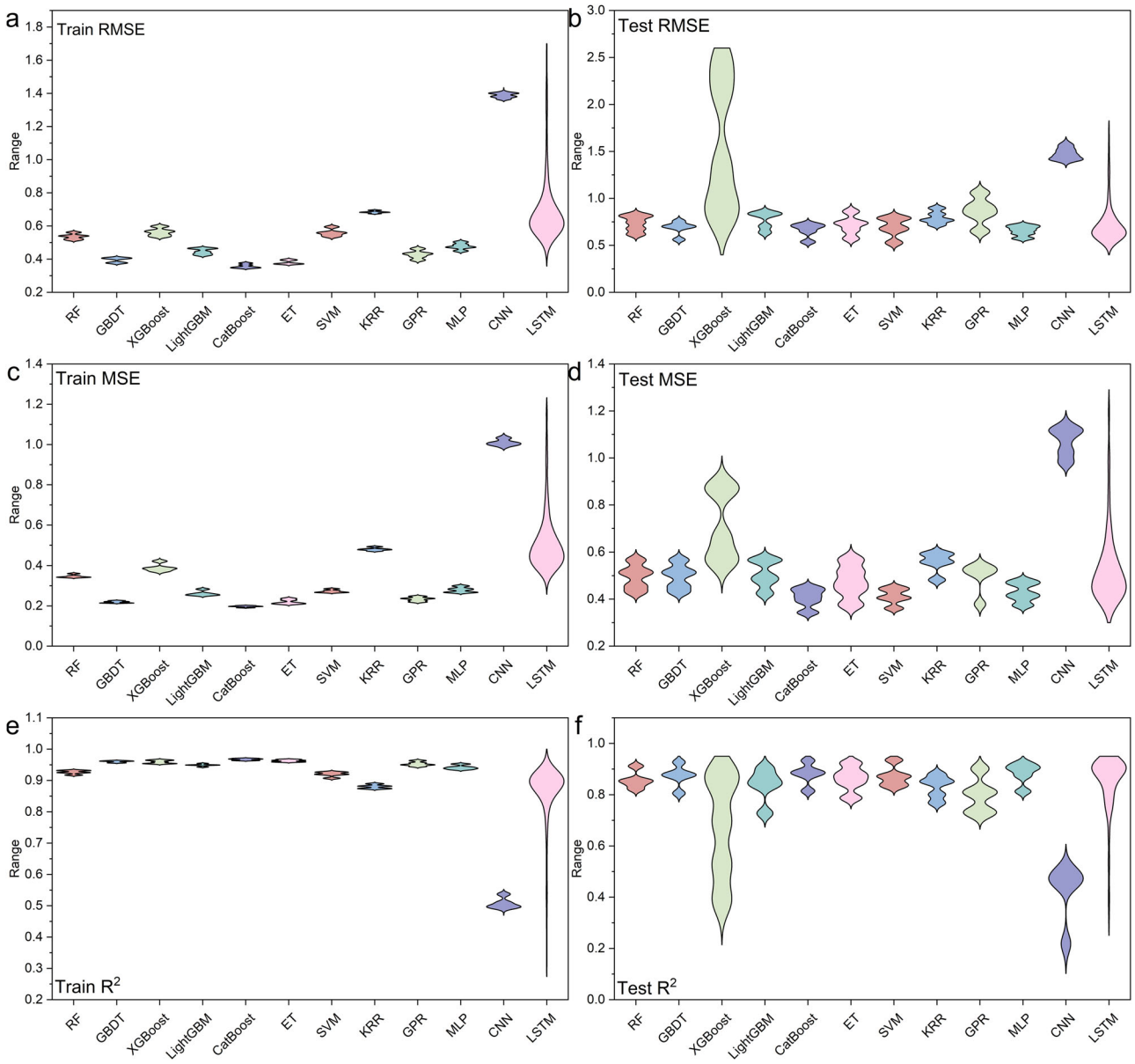

Fig. 3 | The evaluation results of the stability of different ML models. a Train RMSE, **b** Test RMSE, **c** Train MSE, **d** Test MSE, **e** Train R^2, **f** Test R^2.

50 mg/g, the adsorption stabilizes, likely due to saturation of adsorption sites and potential blockage of biochar pores by excess NH_4^+-N[65,66]. As shown in Fig. 6b, increasing the temperature from 25 °C to 35 °C enhances adsorption capacity, attributed to higher kinetic energy of NH_4^+-N molecules and faster diffusion to biochar surfaces. However, experimental data indicate that higher temperatures and pH values (>8.6) lead to NH_4^+-N volatilization into gaseous ammonia, limiting adsorption[67]. Similarly, Fig. 6c reveals an optimal pH range of 6 ~ 9 for adsorption, influenced by competition between H^+ and NH_4^+ for adsorption sites at low pH and $NH_3{\cdot}H_2O$ formation at high pH. When the pH > 9, NH_4^+-N undergoes conversion into its inorganic form ($NH_3{\cdot}H_2O$), as indicated by reaction (6), which is unfavorable for adsorption[68].

$$NH_3 + H_2O \leftrightarrow NH3{\cdot}H_2O \leftrightarrow NH_4^+ + OH^- \quad (6)$$

Carbon content (*C*) significantly impacts biochar's adsorption performance, as depicted in Fig. 6d. When C < 0.60, the adsorption capacity increases due to the larger specific surface area. However, at C > 0.75, the decline in surface functional groups reduces adsorption efficiency[69]. The surface of biochar contains various oxygen-containing functional groups, such as carbonyl, carboxyl, hydroxyl, and phenolic hydroxyl groups, which play a critical role in the adsorption of NH_4^+-N[70]. Figure 6e highlights the critical role of oxygen-containing functional groups in NH_4^+-N adsorption, facilitated by strong electrostatic interactions with positively charged NH_4^+-N molecules[70]. Additionally, studies have suggested that NH_4^+ can act as a Brønsted or Lewis acid, and the adsorption of NH_4^+-N on biochar surfaces generally involves reactions with these oxygen functional groups, resulting in the formation of amines or amides[10,70]. The H/C molar ratio represents the aromaticity characteristics of biochar, which may directly influence its adsorption performance[71]. The H/C molar ratio (Fig. 6f) also affects adsorption, with higher ratios indicating more hydrophilic groups that enhance adsorption efficiency through hydrogen bonding and electrostatic interactions[72]. Ash content shows a nonlinear effect on NH_4^+-N adsorption (Fig. 6g), as alkaline minerals in ash influence biochar surface alkalinity. However, excessive ash can block micropores, reducing effective adsorption sites[73]. Finally, the porous structure (Fig. 6h) and large surface area (Fig. 6i) of biochar facilitate mass transfer and adsorption, although physical

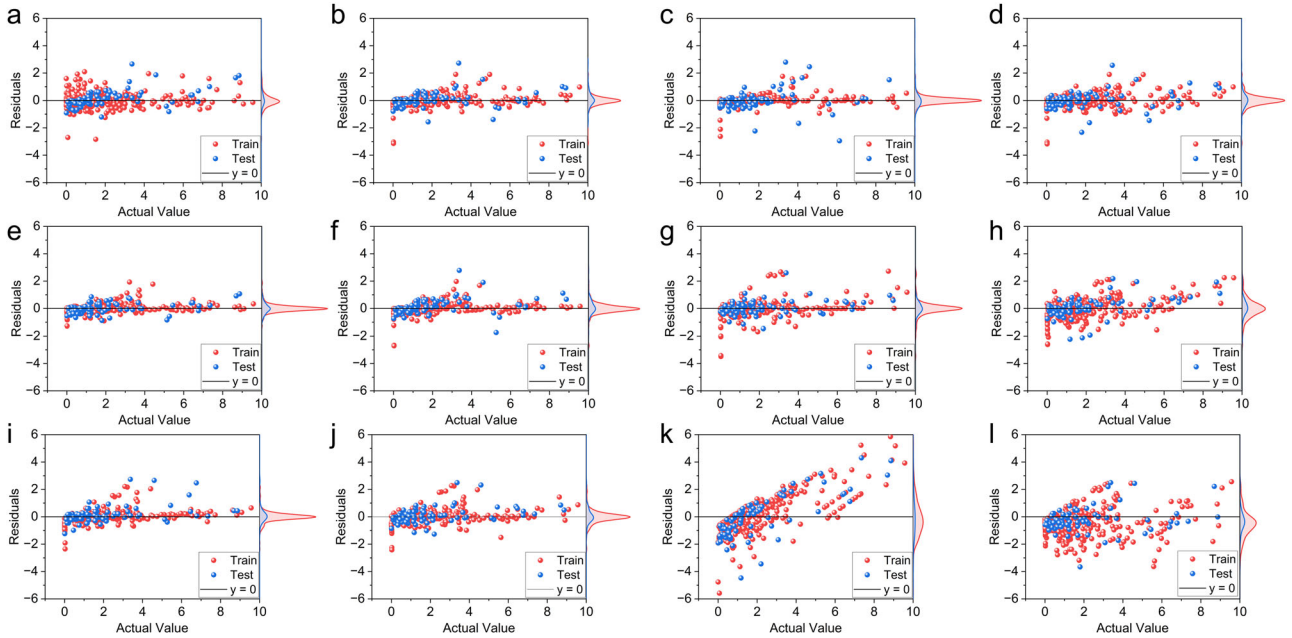

Fig. 4 | Residual analysis and normal distribution of ML model. **a** RF, **b** GBDT, **c** XGBoost, **d** LightGBM, **e** CatBoost, **f** ET, **g** SVM, **h** KRR, **i** GPR, **j** MLP, **k** CNN, **l** LSTM.

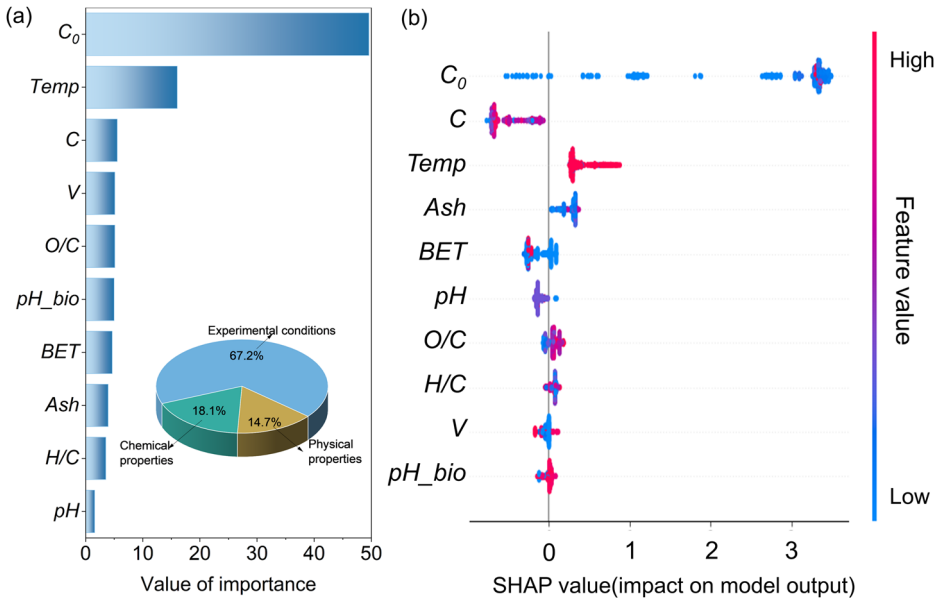

Fig. 5 | EFI analysis and SHAP value visualization. **a** EFI analysis (**b**) SHAP value visualization.

adsorption is not the predominant mechanism, given the complexity of electrostatic and ion-exchange interactions[10,73,74].

GUI application development

To facilitate the advancement of the CatBoost model, this study has developed a graphical user interface (GUI) in Python, integrating the optimized CatBoost model as illustrated in Fig. 7. This GUI allows for direct prediction of the NH_4^+-N adsorption capacity of biochar by entering values for ten essential features, thus aiding in optimizing experimental design and supporting practical implementation. The interface includes input fields for parameters related to the chemical properties of biochar, physical characteristics of biochar, and experimental conditions. All relevant source code is available in the Data Availability section on GitHub. Users can open the file titled "GUI.ipynb" provided in this study, input the necessary parameters, and click the "Predict" button to obtain the predicted adsorption capacity displayed at the bottom of the screen.

Challenges and future directions

Table 2 compares various studies using ML models to predict the adsorption performance of adsorbents for NH_4^+-N. This study stands out due to its significantly larger dataset of 417 data points and 11 features, compared to earlier works that used smaller datasets (14–60 data points) and fewer features (3–7). The larger dataset and wider range of features provide a more robust and generalizable model, leading to a deeper understanding of the complex relationships involved. This study employed a variety of ML models, including RF, GBDT, XGBoost, LightGBM, CatBoost, ET, SVM, KRR, GPR, and MLP, with CatBoost emerging as the best-performing model. The R^2 value of 0.9657 achieved by CatBoost is comparable to or better than those of previous studies (R^2 values ranging from 0.9629 to 0.9998), indicating high predictive capability. Another advantage of this study is the availability of code and data, which enhances transparency and reproducibility. Unlike many previous studies, this study provides both, ensuring that the findings can be independently verified and built upon by future researchers.

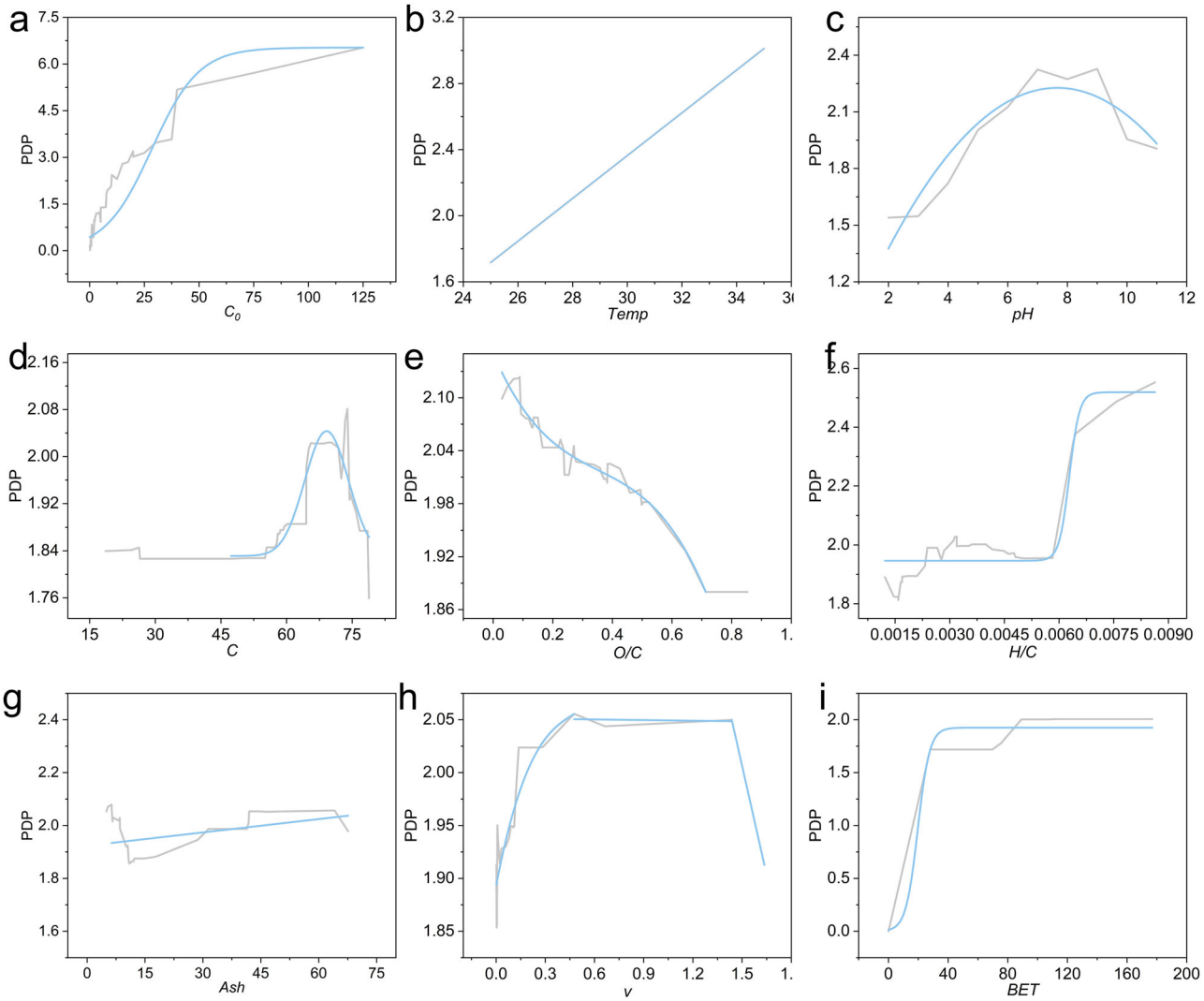

Fig. 6 | Partial dependence plots of the selected features on ammonia nitrogen adsorption predictions by biochar. The gray line in the figure shows the actual values predicted by the ML model, while the blue line indicates the fitted values. **a** C_0, **b** *Temp*, **c** *pH*, **d** *C*, **e** *O/C*, **f** *H/C*, **g** *Ash*, **h** *V*, **i** *BET*.

This study presents an innovative approach for predicting biochar adsorb NH_4^+-N through ML, which facilitates efficient adsorbent selection and reduces the need for labour-intensive experimental trials. This methodology can also serve as a reference for environmental studies applying ML to optimize adsorbents, potentially enhancing water treatment processes across different sectors. Additionally, another major advantage of this study is the comprehensive comparison of various ML models, offering insights into the best-performing models for predicting adsorption efficiency and thereby improving the accuracy and reliability of treatment outcomes. Moreover, the integration of the physical and chemical properties of biochar with experimental conditions provides a holistic perspective on adsorption behavior.

However, the study does have some limitations. The availability and quality of data used for model training present challenges; although 417 adsorption datasets were used, there is still limited coverage of certain conditions and features, which may restrict the model's applicability in diverse environmental scenarios. Besides, in real wastewater, multiple contaminants are often present, and the interactions among these contaminants, as well as their effects on the adsorption of NH_4^+-N by biochar, have not been adequately discussed. Moreover, the deep learning models demonstrated weaker performance, suggesting the need for further improvements to enhance their adaptability to smaller datasets.

In this research, twelve ML models were systematically evaluated for predicting the equilibrium adsorption capacity of NH_4^+-N onto biochar. Among these models, tree-based ensemble methods consistently outperformed kernel-based and deep learning models, with the CatBoost algorithm achieving the highest accuracy ($R^2 = 0.9329$, RMSE = 0.5378). The CatBoost model's robust performance is attributed to its ability to handle complex datasets, mitigate overfitting, and process missing or categorical data effectively. Stability and residual analyses further validated CatBoost's reliability, making it the optimal model for this study.

The feature importance analysis revealed that experimental conditions contributed most significantly to NH_4^+-N adsorption (67.2%), followed by the chemical properties (18.2%) and physical characteristics (14.7%) of biochar. PDPs indicated that experimental conditions such as $C_0 > 50$ mg/g and a pH range of 6–9 are more favorable for adsorption, providing deeper insights into the nonlinear effects of key variables like temperature, carbon content, and pH on adsorption capacity. To facilitate practical applications, a Python-based GUI integrating the CatBoost model was developed. This tool allows researchers to efficiently predict NH_4^+-N adsorption by inputting experimental parameters, thereby optimizing experimental design and enhancing decision-making in environmental engineering.

This study demonstrates the potential of ML in advancing the understanding and optimization of adsorption processes. The proposed framework can be extended to other adsorption systems or environmental scenarios, promoting more efficient and sustainable water treatment solutions. Future work should focus on incorporating real wastewater scenarios

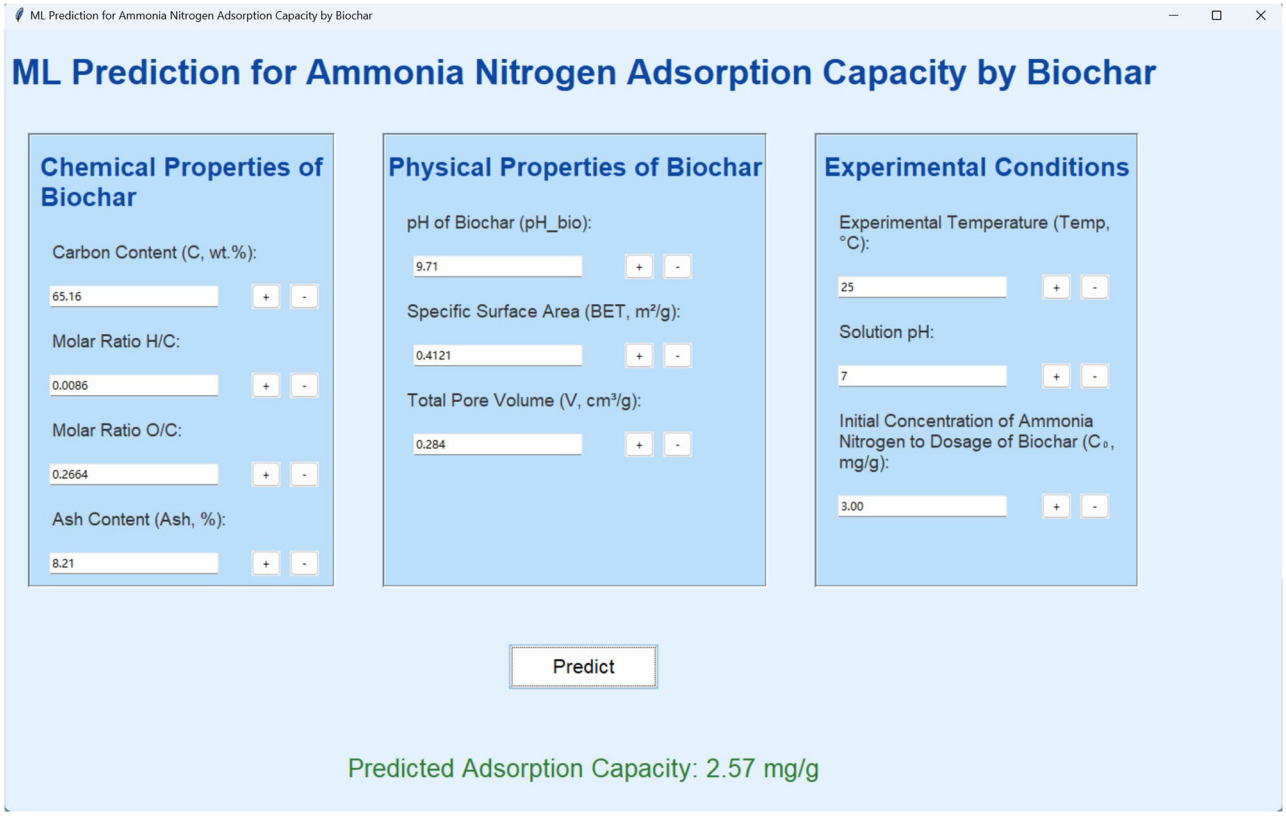

Fig. 7 | The graphical user interface for biochar adsorption capacity based on the CatBoost model.

Table 2 | Comparison of studies on using ML to predict the adsorption performance of adsorbents for ammonia nitrogen

Adsorbent	ML models	Best model	Number of data	Number of Features	Best R^2	Data availability	Code availability	Ref.
Waste foundry Sand	ANN	ANN	14	5	0.9800	No	No	68
Zeolite	FF-NN, ER-NN, MLP	-(hybrid models)	60	4	0.9773	No	No	27
Crab shell	ANFIS, ANN,RSM	ANFIS	32	5	0.9998	No	No	28
Bentonite	GA-BPNN, CCD-RSM	GA-BPNN	30	3	0.9629	Yes	No	75
Zeolite	ANN, MLP	MLP	60	7	-	No	No	76
Saline soil	RSM, ANN	ANN	29	5	0.9991	Yes	No	77
Activated Carbon	BP-ANN,GA,RSM	BP-ANN	29	4	0.9959	Yes	No	78
Biochar	RF, GBDT, XGBoost, LightGBM, CatBoost, ET, SVM, KRR, GPR, MLP	CatBoost	417	11	0.9657	Yes	Yes	This study

ANN artificial neural networks, *FF-NN* feedforward neural network, *ER-NN* Elman recurrent neural network, *MLP* multilayer perceptron, *ANFIS* adaptive neuro-fuzzy inference system, *RSM* response surface methodology, *GA* genetic algorithm, *BPNN* backpropagation neural network, *CCD-RSM* central composite design - response surface methodology, *BP-ANN* backpropagation artificial neural network, *RF* random forest, *GBDT* gradient boosting decision tree, *XGBoost* eXtreme gradient boosting, *LightGBM* light gradient boosting machine, *CatBoost* categorical boosting, *ET* extra trees, *SVM* support vector machine, *KRR* Kernel ridge regression, *GPR* Gaussian process regression.

and exploring the influence of multiple contaminants to enhance the model's applicability.

Data availability

The Python codes used to build the models, raw data used in this study, model stable test results, and GUI application code are available at the following GitHub link (https://github.com/17609858895/Ammonia-nitrogen).

Received: 16 November 2024; Accepted: 17 December 2024;
Published online: 22 February 2025

References

1. Su, Y. Revisiting carbon, nitrogen, and phosphorus metabolisms in microalgae for wastewater treatment. *Sci. Total Environ.* **762**, 144590 (2021).
2. Qin, L. et al. Application of encapsulated algae into MBR for high-ammonia nitrogen wastewater treatment and biofouling control. *Water Res.* **187**, 116430 (2020).
3. Ren, Z. et al. Study on adsorption of ammonia nitrogen by iron-loaded activated carbon from low temperature wastewater. *Chemosphere* **262**, 127895 (2021).

4. Shao, Q. et al. Phosphorus and nitrogen recovery from wastewater by ceramsite: adsorption mechanism, plant cultivation and sustainability analysis. *Sci. Total Environ.* **805**, 150288 (2022).
5. Han, B., Butterly, C., Zhang, W., He, J. & Chen, D. Adsorbent materials for ammonium and ammonia removal: a review. *J. Clean. Prod.* **283**, 124611 (2021).
6. Isik, Z., Saleh, M. & Dizge, N. Adsorption studies of ammonia and phosphate ions onto calcium alginate beads. *Surf. Interfaces* **26**, 101330 (2021).
7. Rashid, R., Shafiq, I., Akhter, P., Iqbal, M. J. & Hussain, M. A state-of-the-art review on wastewater treatment techniques: the effectiveness of adsorption method. *Environ. Sci. Pollut. Res.* **28**, 9050–9066 (2021).
8. Lv, Y. et al. Preparation of ceramsite adsorbent and its influencing factors on ammonia nitrogen adsorption: a review. *Mater. Today Commun.* **41**, 110522 (2024).
9. Liu, C., Crini, G., Wilson, L. D., Balasubramanian, P. & Li, F. Removal of contaminants present in water and wastewater by cyclodextrin-based adsorbents: a bibliometric review from 1993 to 2022. *Environ. Pollut.* **348**, 123815 (2024).
10. Zhang, M. et al. Evaluating biochar and its modifications for the removal of ammonium, nitrate, and phosphate in water. *Water Res.* **186**, 116303 (2020).
11. Hoang-Minh, T. et al. Removal of ammonium from water by a KOH-treated bentonite biochar composite. *Colloid Polym. Sci.* https://doi.org/10.1007/s00396-024-05335-x (2024).
12. Huang, J. et al. Removing ammonium from water and wastewater using cost-effective adsorbents: a review. *J. Environ. Sci.* **63**, 174–197 (2018).
13. Liu, C. et al. Critical review of biochar for the removal of emerging inorganic pollutants from wastewater. *Chin. Chem. Lett.* **36**, 109960 (2024).
14. Balasubramanian, P., Prabhakar, M. R., Liu, C., Zhang, P. & Li, F. Predictive capability of rough set machine learning in tetracycline adsorption using biochar. *Carbon Res.* **3**, 48 (2024).
15. Xue, S. et al. Food waste based biochars for ammonia nitrogen removal from aqueous solutions. *Bioresour. Technol.* **292**, 121927 (2019).
16. Marcińczyk, M., Ok, Y. S. & Oleszczuk, P. From waste to fertilizer: nutrient recovery from wastewater by pristine and engineered biochars. *Chemosphere* **306**, 135310 (2022).
17. Xiang, S. et al. New progress of ammonia recovery during ammonia nitrogen removal from various wastewaters. *World J. Microbiol. Biotechnol.* **36**, 144 (2020).
18. Feng, L. et al. Removal of ammonia nitrogen from aqueous media with low-cost adsorbents: a review. *Water. Air. Soil Pollut.* **234**, 280 (2023).
19. Liu, C., Balasubramanian, P., Li, F. & Huang, H. Machine learning prediction of dye adsorption by hydrochar: parameter optimization and experimental validation. *J. Hazard. Mater.* **480**, 135853 (2024).
20. Zhang, P. et al. Unveiling the drives behind tetracycline adsorption capacity with biochar through machine learning. *Sci. Rep.* **13**, 11512 (2023).
21. Balasubramanian, P. et al. Predictive capability of phosphate recovery from wastewater using a rough set machine learning model. *ACS EST Eng.* https://doi.org/10.1021/acsestengg.4c00255 (2024).
22. Zhu, X. et al. Machine learning for the selection of carbon-based materials for tetracycline and sulfamethoxazole adsorption. *Chem. Eng. J.* **406**, 126782 (2021).
23. Zhu, X. et al. 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 (2022).
24. Sarker, I. H. Machine learning: algorithms, real-world applications and research directions. *SN Comput. Sci.* **2**, 160 (2021).
25. Faisal, A. A. H. & Naji, L. A. Simulation of ammonia nitrogen removal from simulated wastewater by sorption onto waste foundry sand using artificial neural network. *Assoc. Arab Univ. J. Eng. Sci.* **26**, 28–34 (2019).
26. Reza, A., Chen, L. & Kruger, K. Microwave irradiated ammonia nitrogen removal from anaerobically digested liquid dairy manure: a response surface methodology and artificial neural network-based optimization and modeling. *J. Environ. Chem. Eng.* **10**, 108279 (2022).
27. Cagcag Yolcu, O., Aydın Temel, F. & Kuleyin, A. New hybrid predictive modeling principles for ammonium adsorption: the combination of response surface methodology with feed-forward and Elman-recurrent neural networks. *J. Clean. Prod.* **311**, 127688 (2021).
28. Ohale, P. E. et al. A comparative optimization and modeling of ammonia–nitrogen adsorption from abattoir wastewater using a novel iron-functionalized crab shell. *Appl. Water Sci.* **12**, 193 (2022).
29. Mashhadimoslem, H. et al. Development of predictive models for activated carbon synthesis from different biomass for CO_2 adsorption using artificial neural networks. *Ind. Eng. Chem. Res.* https://doi.org/10.1021/acs.iecr.1c02754 (2021).
30. Zhu, X., Wang, X. & Ok, Y. S. The application of machine learning methods for prediction of metal sorption onto biochars. *J. Hazard. Mater.* **378**, 120727 (2019).
31. Zhao, Y., Li, Y., Fan, D., Song, J. & Yang, F. Application of kernel extreme learning machine and Kriging model in prediction of heavy metals removal by biochar. *Bioresour. Technol.* **329**, 124876 (2021).
32. Fu, W. et al. Machine learning-driven prediction of phosphorus removal performance of metal-modified biochar and optimization of preparation processes considering water quality management objectives. *Bioresour. Technol.* **403**, 130861 (2024).
33. Zaffar, A., Prabhakar, M. R., Liu, C., Sivaraman, J. & Balasubramanian, P. Machine learning assisted prediction and process validation of electrochemically induced phosphorus recovery from wastewater. *J. Environ. Chem. Eng.* **12**, 114271 (2024).
34. Yang, X. et al. Machine learning-assisted evaluation of potential biochars for pharmaceutical removal from water. *Environ. Res.* **214**, 113953 (2022).
35. Kong, W., Hui, H. W. H., Peng, H. & Goh, W. W. B. Dealing with missing values in proteomics data. *Proteomics* **22**, 2200092 (2022).
36. Mahanty, B., Gharami, M. & Haldar, D. Machine learning modelling for predicting the efficacy of ionic liquid-aided biomass pretreatment. *BioEnergy Res.* **17**, 1569–1583 (2024).
37. Josse, J., Chen, J. M., Prost, N., Varoquaux, G. & Scornet, E. On the consistency of supervised learning with missing values. *Stat. Pap.* https://doi.org/10.1007/s00362-024-01550-4 (2024).
38. Yu, H., Sang, P. & Huan, T. Adaptive Box–Cox transformation: a highly flexible feature-specific data transformation to improve metabolomic data normality for better statistical analysis. *Anal. Chem.* **94**, 8267–8276 (2022).
39. Vélez, J. I., Correa, J. C. & Marmolejo-Ramos, F. A new approach to the Box–Cox transformation. *Front. Appl. Math. Stat.* 1 https://doi.org/10.3389/fams.2015.00012 (2015).
40. Marzi, C. et al. Collinearity and dimensionality reduction in radiomics: effect of preprocessing parameters in hypertrophic cardiomyopathy magnetic resonance T1 and T2 mapping. *Bioengineering* **10**, 80 (2023).
41. Singh, D. & Singh, B. Investigating the impact of data normalization on classification performance. *Appl. Soft Comput.* **97**, 105524 (2020).
42. Kim, Y.-S. et al. Investigating the Impact of Data Normalization Methods on Predicting Electricity Consumption in a Building Using different Artificial Neural Network Models. *Sustain. Cities Soc.* **118**, 105570 (2024).
43. Wu, J. et al. Hyperparameter optimization for machine learning models based on Bayesian optimization. *J. Electron. Sci. Technol.* **17**, 26–40 (2019).
44. Victoria, A. H. & Maragatham, G. Automatic tuning of hyperparameters using Bayesian optimization. *Evol. Syst.* **12**, 217–223 (2021).

45. Naser, M. Z. An engineer's guide to eXplainable artificial intelligence and interpretable machine learning: navigating causality, forced goodness, and the false perception of inference. *Autom. Constr.* **129**, 103821 (2021).
46. Sahlaoui, H., Alaoui, E. A. A., Nayyar, A., Agoujil, S. & Jaber, M. M. Predicting and interpreting student performance using ensemble models and shapley additive explanations. *IEEE Access* **9**, 152688–152703 (2021).
47. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions, in: Proc. 31st Int. Conf. Neural Inf. Process. Syst., Curran Associates Inc., Red Hook, NY, USA, 2017: pp. 4768–4777.
48. Baptista, M. L., Goebel, K. & Henriques, E. M. P. Relation between prognostics predictor evaluation metrics and local interpretability SHAP values. *Artif. Intell.* **306**, 103667 (2022).
49. Ullah, H. et al. Machine learning approach to predict adsorption capacity of Fe-modified biochar for selenium. *Carbon Res.* **2**, 29 (2023).
50. Goldstein, A., Kapelner, A., Bleich, J. & Pitkin, E. Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. *J. Comput. Graph. Stat.* **24**, 44–65 (2015).
51. Qian, K. et al. Effects of biomass feedstocks and gasification conditions on the physiochemical properties of char. *Energies* **6**, 3972–3986 (2013).
52. Critchley, F. & Jones, M. C. Asymmetry and gradient asymmetry functions: density-based skewness and kurtosis. *Scand. J. Stat.* **35**, 415–437 (2008).
53. Boylan, G. L. & Cho, B. R. The normal probability plot as a tool for understanding data: a shape analysis from the perspective of skewness, kurtosis, and variability. *Qual. Reliab. Eng. Int.* **28**, 249–264 (2012).
54. Solanki, A., Gupta, V. & Joshi, M. Application of machine learning algorithms in landslide susceptibility mapping, Kali Valley, Kumaun Himalaya, India. *Geocarto Int.* **37**, 16846–16871 (2022).
55. Mary Ealias, A. & Saravanakumar, M. P. A critical review on ultrasonic-assisted dye adsorption: Mass transfer, half-life and half-capacity concentration approach with future industrial perspectives. *Crit. Rev. Environ. Sci. Technol.* **49**, 1959–2015 (2019).
56. Hussain, S. et al. A novel feature engineered-CatBoost-based supervised machine learning framework for electricity theft detection. *Energy Rep.* **7**, 4425–4436 (2021).
57. Banik, R. & Biswas, A. Improving solar PV prediction performance with RF-CatBoost ensemble: a robust and complementary approach, renew. *Energy Focus* **46**, 207–221 (2023).
58. Malhotra, G., Dujmović, M. & Bowers, J. S. Feature blindness: a challenge for understanding and modelling visual object recognition. *PLOS Comput. Biol.* **18**, e1009572 (2022).
59. Zhou, W., Yan, Z. & Zhang, L. A comparative study of 11 non-linear regression models highlighting autoencoder, DBN, and SVR, enhanced by SHAP importance analysis in soybean branching prediction. *Sci. Rep.* **14**, 5905 (2024).
60. Papadopoulos, S., Azar, E., Woon, W.-L. & Kontokosta, C. E. Evaluation of tree-based ensemble learning algorithms for building energy performance estimation. *J. Build. Perform. Simul.* **11**, 322–332 (2018).
61. Zhang, L. & Jánošík, D. Enhanced short-term load forecasting with hybrid machine learning models: CatBoost and XGBoost approaches. *Expert. Syst. Appl.* **241**, 122686 (2024).
62. Nguyen, X. C. et al. Potential application of machine learning for exploring adsorption mechanisms of pharmaceuticals onto biochars. *Chemosphere* **287**, 132203 (2022).
63. ten Hulscher, Th. E. M. & Cornelissen, G. Effect of temperature on sorption equilibrium and sorption kinetics of organic micropollutants —a review. *Chemosphere* **32**, 609–626 (1996).
64. Kashem, A. et al. Hybrid data-driven approaches to predicting the compressive strength of ultra-high-performance concrete using SHAP and PDP analyses. *Case Stud. Constr. Mater.* **20**, e02991 (2024).
65. Wang, M. et al. Enhanced nitrogen and phosphorus adsorption performance and stabilization by novel panda manure biochar modified by CMC stabilized nZVZ composite in aqueous solution: mechanisms and application potential. *J. Clean. Prod.* **291**, 125221 (2021).
66. Hassaan, M. A. et al. Isotherm and kinetic investigations of sawdust-based biochar modified by ammonia to remove methylene blue from water. *Sci. Rep.* **13**, 12724 (2023).
67. Qu, J. et al. 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 (2021).
68. Li, L., Lollar, B. S., Li, H., Wortmann, U. G. & Lacrampe-Couloume, G. Ammonium stability and nitrogen isotope fractionations for NH4+–NH3(aq)–NH3(gas) systems at 20–70 °C and pH of 2–13: applications to habitability and nitrogen cycling in low-temperature hydrothermal systems. *Geochim. Cosmochim. Acta* **84**, 280–296 (2012).
69. Zhang, X. et al. Effect of pyrolysis temperature on composition, carbon fraction and abiotic stability of straw biochars: correlation and quantitative analysis. *Carbon Res.* **1**, 17 (2022).
70. Yin, Q., Zhang, B., Wang, R. & Zhao, Z. Biochar as an adsorbent for inorganic nitrogen and phosphorus removal from water: a review. *Environ. Sci. Pollut. Res.* **24**, 26297–26309 (2017).
71. Singh, S., Khan, N. A., Shehata, N., Singh, J. & Ramamurthy, P. C. Insight into biochar as sustainable biomass: production methods, characteristics, and environmental remediation. *J. Clean. Prod.* **475**, 143645 (2024).
72. Kumar, A. et al. Biochar modification methods for augmenting sorption of contaminants. *Curr. Pollut. Rep.* **8**, 519–555 (2022).
73. Vijayaraghavan, K. The importance of mineral ingredients in biochar production, properties and applications. *Crit. Rev. Environ. Sci. Technol.* **51**, 113–139 (2021).
74. Dai, Y., Wang, W., Lu, L., Yan, L. & Yu, D. Utilization of biochar for the removal of nitrogen and phosphorus. *J. Clean. Prod.* **257**, 120573 (2020).
75. Lu, M. et al. Optimization of adsorption performance of cerium-loaded intercalated bentonite by CCD-RSM and GA-BPNN and its application in simultaneous removal of phosphorus and ammonia nitrogen. *Chemosphere* **336**, 139241 (2023).
76. Aydın Temel, F., Cağcağ Yolcu, Ö. & Kuleyin, A. A multilayer perceptron-based prediction of ammonium adsorption on zeolite from landfill leachate: Batch and column studies. *J. Hazard. Mater.* **410**, 124670 (2021).
77. Wu, J., Bian, J. & Sun, X. Comparative assessment on ammonia nitrogen adsorption onto a saline soil–groundwater environment: distribution, multi-factor interaction, and optimization using response surface methodology and artificial neural network. *Environ. Geochem. Health* **45**, 3743–3758 (2023).
78. Yu, A. et al. Modeling and optimizing of NH4+ removal from stormwater by coal-based granular activated carbon using RSM and ANN coupled with GA. *Water* **13**, 608 (2021).

Acknowledgements

This research was supported by the President's foundation of Tarim University (TDZKCX202404), the National Natural Science Foundation of China (42275014), and the Bingtuan Science and Technology Program (2021DB019; 2022CB001-01). Besides, the authors acknowledge the assistance of Instrumental Analysis Center of Tarim University.

Author contributions

C.L. contributed to methodology, investigation, formal analysis, conceptualization, data curation, visualization, and writing—original draft preparation. P.B. was responsible for methodology, supervision, and writing—review and editing. J.A. contributed to methodology, supervision, and writing—review and editing. F.L. oversaw conceptualization, funding acquisition, project administration, supervision, and writing—review and editing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41545-024-00429-z.

Correspondence and requests for materials should be addressed to Paramasivan Balasubramanian or Fayong Li.