scientific reports

Check for updates

OPEN

Heavy metal adsorption efficiency prediction using biochar properties: a comparative analysis for ensemble machine learning models

Zaher Mundher Yaseen✉ & Farah Loui Alhalimi

The contamination of water and soils with heavy metals poses a significant environmental threat, making the development of effective removal strategies a global priority. Hence, the determination of heavy metals can play an essential role in environmental monitoring and assessment. In the current research, ensemble machine learning (ML) models (i.e., Random Forest Regressor (RFR), Adaptive Boosting (Adaboost), Gradient Boosting (GB), HistGradientBoosting, Extreme Gradient Boosting (XGBoost), and Light Gradient-Boosting Machine (LightGBM)) were applied in attempt to predict the adsorption efficiency of several heavy metals (i.e., Pb, Cd, Ni, Cu, and Zn) according to different factors including temperature, pH, and biochar characteristics. Data were collected from open-source literature review including 353 samples. At the first stage, data processing was performed including outliers' removal and scaling for better data modeling applicability; whereas, in the second stage the predictive models were conducted. The results showed that XGBoost model attained the superior accuracy in comparison with other models by achieving the highest determination coefficient ($R^2 = 0.92$). The research was extended to investigate the feature importance analysis which indicated that the initial concentration ratio of metals to biochar and pH were the most influential factors toward the adsorption efficiency followed by Pyrolysis temperature, while other features like physical properties as surface area and pore structure had a minimal effect on efficiency. These findings highlighted the importance of using ensemble ML models in guiding heavy metals removal solutions as it provides an efficient prediction and ease the selection of the environmental application.

Keywords Heavy metals adsorption, Ensemble learning, Environmental applications, Biochar characteristics, Predictive modeling

Abbreviations

Adaboost	Adaptive boosting
RFR	Random Forest Regressor
LightGBM	Light Gradient-Boosting Machine
RF	Random forest
GB	Gradient boosting
DT	Decision trees
ML	Machine learning
XGBoost	Extreme gradient boosting
ANN	Artificial neural network
SVM	Support vector machine
ANFIS	Adaptive neuro-fuzzy inference system
MLR	Multiple linear regression
GBRT	Gradient boosted regression trees
PCA	Analysis
DE	Differential evolution
GP	Genetic programming
ELM	Extreme learning machine

Civil and Environmental Engineering Department, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia. ✉email: z.yaseen@kfupm.edu.sa

PLSR	Partial least squares regression
CEC	Cation exchange capacity
PS	Biochar particle size
C	Total carbon content
(O + N)/C	Molar ratio of oxygen and nitrogen to carbon
O/C	Molar ratio of oxygen to carbon
H/C	Molar ratio of hydrogen to carbon
R^2	Coefficient of determination
MAE	Mean absolute error
MSE	Mean squared error
WI	Willmott's index of agreement
NSE	Nash-sutcliffe efficiency
md	Modified index of agreement
SA	Surface area
PFAS	Per- and polyfluoroalkyl substances

Research background

Metals exist naturally in Earth's crust, but the excessive presence of certain heavy metals due to the industrial processes' discharges can cause severe environmental concerns[1]. These discharges can infiltrate to groundwater and accumulate in aquifers or even carried out into surface waters and soils with runoff causing widespread pollution[2]. Biochar is a sustainable and commonly used method for heavy metals removal in contaminated soils[3]. As a result of its physical and chemical properties such as high cation exchange ability, surface functional groups and pH modifying abilities, biochar becomes effective in reducing heavy metals concentrations in soil[4]. Other factors affect its efficiency such as feedstock material, pyrolysis temperature and interaction with soil properties[5].

Several mechanisms are used depending on specific metal being treated, such as complexation, cation exchange, electrostatic interactions, precipitation and reduction[6]. For example, arsenic adsorbs through complexation and electrostatic interactions, while cadmium and lead rely on cation exchange and precipitation[7]. These mechanisms can be shaped by biochar's characteristics that varies with pyrolysis conditions[8]. Recent advanced modifications such as loading with minerals or nanoparticles and activation with alkali solutions improve its efficiency[9]. Consequently, some challenges present such as the need for cost-effective production and in mixed metal environments competitive sorption mechanisms are required in addition to the feasibility of biochar reuse in field applications[8]. Hence, the feasibility of using computer-aided models can take substantial alternative in biochar modeling.

Literature review

Interactions between biochar characteristics, environmental conditions and heavy metals behavior are very complex and non-linear which presents several challenges for optimizing biochar properties[10–12]. Traditional experimental methods often fail to capture the intricate dynamics of adsorption efficiency as they are resources intensive and limited in scope, same as empirical models like Langmuir and Freundlich rely on linear assumptions that aren't sufficient for these relationships[13,14]. Hence, there is a high motivation and interest to adopt computer based theoretical models for mimicking the internal mechanism of heavy metal adsorption efficiency and the related biochar properties characteristics. Machine learning (ML) models offers an impactful alternative to address these challenges[15,16]. It reduces experimental workload and provide cost-effective approaches[17]. ML models have provided a robust framework to optimize and predict removal efficiency, as they offered insights into parameters effects and enhancing material design[18]. Over the past couple decades, there were several versions of ML models adopted for heavy metal modeling[19–21]. For instant, support vector machine (SVM) model is one of the successful methods adopted on the environmental application due to its high adaptability to non-linear data distribution when using kernels[22]. Specifically, SVM model was proved to be more accurate in estimating Pb adsorption capacity in comparison with other models including neural network and fuzzy logic models[23]. By dealing with multiple variables such as pH, contact time, adsorbent dosage and concentration of heavy metals, model consistency acquired by SVM asserts its suitability in predicting adsorption. In another research, SVM model used successfully to predict Adsorption capacity of Pb (II) using features like contaminant concentration, temperature, pH value, surface area, adsorbent dosage and contact time[24]. Another version of ML models is artificial neural network (ANN), which was predominantly adopted to predict the adsorption process (optimizing/prediction) of biochar heavy metal adsorption[23,25]. Some researchers tested ANN model on experimental parameters like pH, temperature (K), adsorbent dose (g), metal ion concentration (mg/L), and contact time (min) to predict the Adsorption capacity (mg/g) of the nano-silica-coated biochar for Cr (VI) removal[26]. The obtained results were demonstrated an optimistic prediction accuracy for the adopted model.

As per the reported literature review study[27], there were several ML models developed for simulating soil, water bodies and adsorption of heavy metals. Models are including RF, SVM, ANN, Least-Squares Support Vector Machine (LS-SVM), Genetic Algorithm (GA), Hybrid ANN- Particle Swarm Optimization (ANN-PSO), Fuzzy Logic, Decision Trees (DT), Principal Component Analysis (PCA), Differential Evolution (DE), Genetic Programming (GP). Worth to mention, although there were different versions of ML models adopted for the heavy metals' adsorption modeling; however, the concerns of the modeling development are still ongoing with respect to the accuracy, uncertainties, data size, generalization, interpretability and others[28].

To enhance the predictive accuracy of the standalone ML models in heavy metal removal using biochar, researchers have explored hybrid approaches. Researchers have developed 20 hybrid models combining various ML algorithms, including SVM, RF, and Gaussian processes (GP), to predict heavy metal sorption

efficiency in biochar systems[29]. Their findings demonstrated that ensemble models, particularly the SVM-ANN, outperformed standalone models in prediction accuracy. Similarly, a novel hybrid variable cross-layer-based ML model that incorporated domain knowledge and monitoring indicators to enhance interpretability and efficiency in wastewater treatment plant modeling achieved an 8.7% accuracy improvement over conventional ML models, emphasizing the role of hybridization in improving prediction reliability[30,31]. These studies highlight the effectiveness of hybrid ML models in improving model accuracy and optimizing heavy metal removal efficiency using biochar.

Research gap and motivation

Although, the literature review indicated a huge efforts of research establishment on heavy metal adsorption efficiency. However, the research domain of this area is still eager to further investigate the applicability of other versions of ML models. Among several recently explored ML models is ensemble learning algorithms (e.g., RF, GB, XGBoost)[32]. The ensemble models applied in different applications including prediction, optimization and feature selection for environmental engineering[33,34]. (For instant, Random Forest (RF) was applied for feature selection and prediction tasks in different environmental applications[21]. In addition, some researchers modified the RF algorithm with Boruta algorithm, which is a wrapper method around RF[35]. Authors used this algorithm to enhance the feature selection process using the RF model. Another ensemble learning model was used in literature e.g., Gradient Boosting (GB) which was able to predict the removal capacity with 99% accuracy[36]. Hence, to the best knowledge of the current research authors was to explore the applicability of six different ensemble models for heavy metals adsorption efficiency of which are known to be efficient in the analysis and modeling highly stochastic and complex environmental related issues.

Research objectives

This research was adopted to develop newly explored ML models called ensemble models for optimizing biochar properties for heavy metals adsorption and identify key features impacting the process using models such as RF, Adaboost, GB, HistGradientBoosting, XGBoost and LightGBM. The significance of this research is presented by its potential to fill the gap between experiment and practical applications by accelerating the development of biochar as a sustainable solution for heavy metal remediation. The ultimate objective of this study is firstly to develop a predictive ML model for the adsorption efficiency of heavy metals using biochar properties. Secondly, to optimize biochar properties and identify key factors that significantly influence the adsorption efficiency of heavy metals. Finally, to evaluate and compare ensemble ML models performance in handling complex, non-linear relationships in heavy metal adsorption data.

Description of data and processing

Data used in this research were collected and published by several researchers[37–48]. (The dataset collected contains 353 adsorption experiments on heavy metals^{2+} (Cu, Zn^{2+}, Pb^{2+}, Cd^{2+}, Ni^{2+}, As^{2+}), compiled from previous literature sources. The data were extracted directly from tables, graphs, and supplementary materials in the referenced studies using Plot Digitizer 2.6.8 (http://plotdigitizer.sourceforge.net/). The dataset includes 15 influencing factors, categorized into four groups. First, biochar characteristics (pH of biochar in water (pH_{H2O}), Surface area (SA, m^2/g), Cation exchange capacity (CEC, $cmol_{(+)}/kg$), Ash content (%), Biochar particle size (PS, mm), Total carbon content (C, %), Molar ratio of oxygen and nitrogen to carbon [(O + N)/C], Molar ratio of oxygen to carbon (O/C), Molar ratio of hydrogen to carbon (H/C)). Second, adsorption conditions (Solution pH, Adsorption temperature (°C)). Third, initial concentration ratio of heavy metals to biochar (C_0 (mmol/g)). Finally, Heavy metal properties (Charge number (N_{charge})Ion radius (r, nm), Electronegativity (χ)).

The dataset includes biochar samples produced from 44 types of lignocellulosic biomass at pyrolysis temperatures ranging from 300 to 700 °C. The diversity in biochar characteristics is due to differences in feedstock type (24 kinds of biomass) and pyrolysis conditions. Missing values were removed rather than imputed to avoid introducing biases into a dataset of only a few hundred data points. A correlation heatmap in Fig. 1 illustrates the interrelationships between biochar characteristics in the dataset. The features contributing to model training were selected using a statistical analysis based on the correlation heatmap in Fig. 1. This was to examine the relationships between different variables, in harmony with the previously published research[49]. The ML models were developed to determine the relative importance of various factors. The results indicated that the most significant factors include the initial concentration ratio of metals to biochar, pH, and Pyrolysis temperature. While surface area played a less dominant role compared to chemical properties. By combining statistical analysis of original data with insights from previous studies, the selection of these factors was validated and their influence on the adsorption process was reinforced. This research aims to expand the original analysis using a comparative model evaluation approach.

Data pre-processing included outlier removal and data normalization using log transformation and for categorical data such as metal type one-hot encoding were used. Boxplots of each variable in the dataset showed the need for the outlier removal process (see Appendix A), which were done using the interquartile range method to minimize the impact of extreme values on model performance. Regarding the normalization process, the normality was represented in the Q–Q plots and high skewness variables were visually detected (see Appendix B). Afterwards, log transformations were applied on the skewed data which improved their normality and resulted in higher accuracy predictions. For categorical variables which have no ordinal relationships the integer encoding can lead to poor performance as the model assumes a natural ordering of the categories[50]. One-hot encoded method is the solution to avoid this misleading ordering by adding a new binary variable for each category.

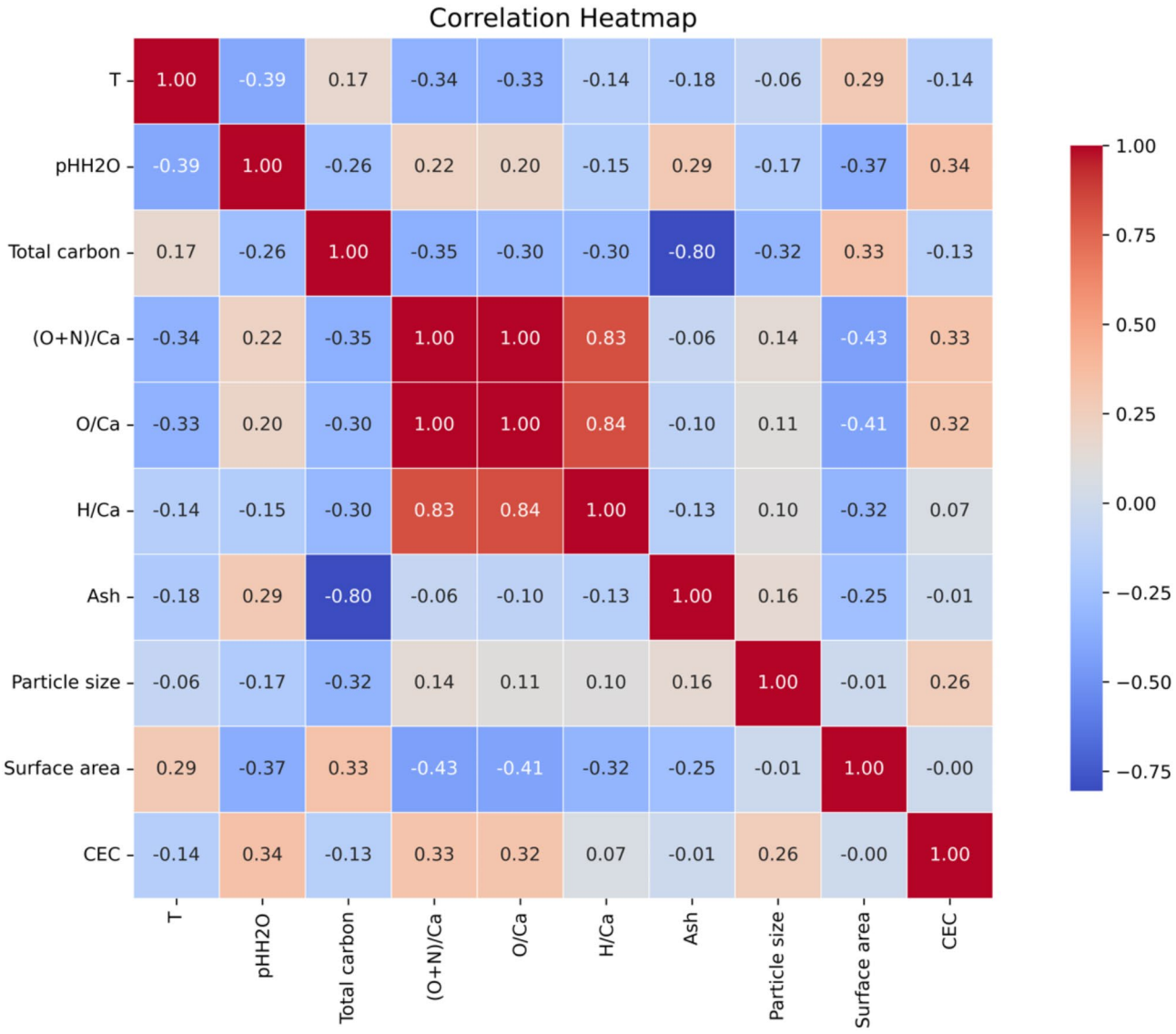

Fig. 1. The correlation heatmap for biochar characteristics and the related parameters.

Applied machine learning

Random forest regressor

Ensemble ML is a powerful approach that constructs a set of classifiers and then classify new data points by taking a (weighted) vote of their predictions[51]. Ensemble learning creates a stronger, more accurate predictor that can handle complex relationships and noisy data[52]. This approach is categorized into two main categories. Bagging and boosting, the main difference between them is that bagging trains multiple models on different subsets of the data, usually created through random sampling with replacement as shown in Fig. 2a. The outputs are combined, typically through averaging (for regression) or voting (for classification)[53]. One of the commonly used models of bagging is RF model. On the other hand, boosting focuses on sequentially improving the model by giving higher weight to the data points that were previously mis predicted[54] (Fig. 2b). It incrementally builds a strong model by combining multiple weak learners. Boosting models include Adaboost, GB, XGBoost, LightGBM.

RF Regression is an ensemble ML technique used for both regression and classification tasks, utilizing multiple decision trees to improve accuracy and control overfitting[55]. This approach is based on the Bootstrap Aggregating (bagging) method, where numerous decision trees are trained on different subsets of the dataset, and their predictions are aggregated to produce the final output, as illustrated in Fig. 2a. When using RF algorithm in regression, the result " y" is the mean of the output of the individual trees used, and can be mathematically presented as:

$$y = \frac{1}{N} \sum {}^{N}_{i=1} T_i\,(x) \tag{1}$$

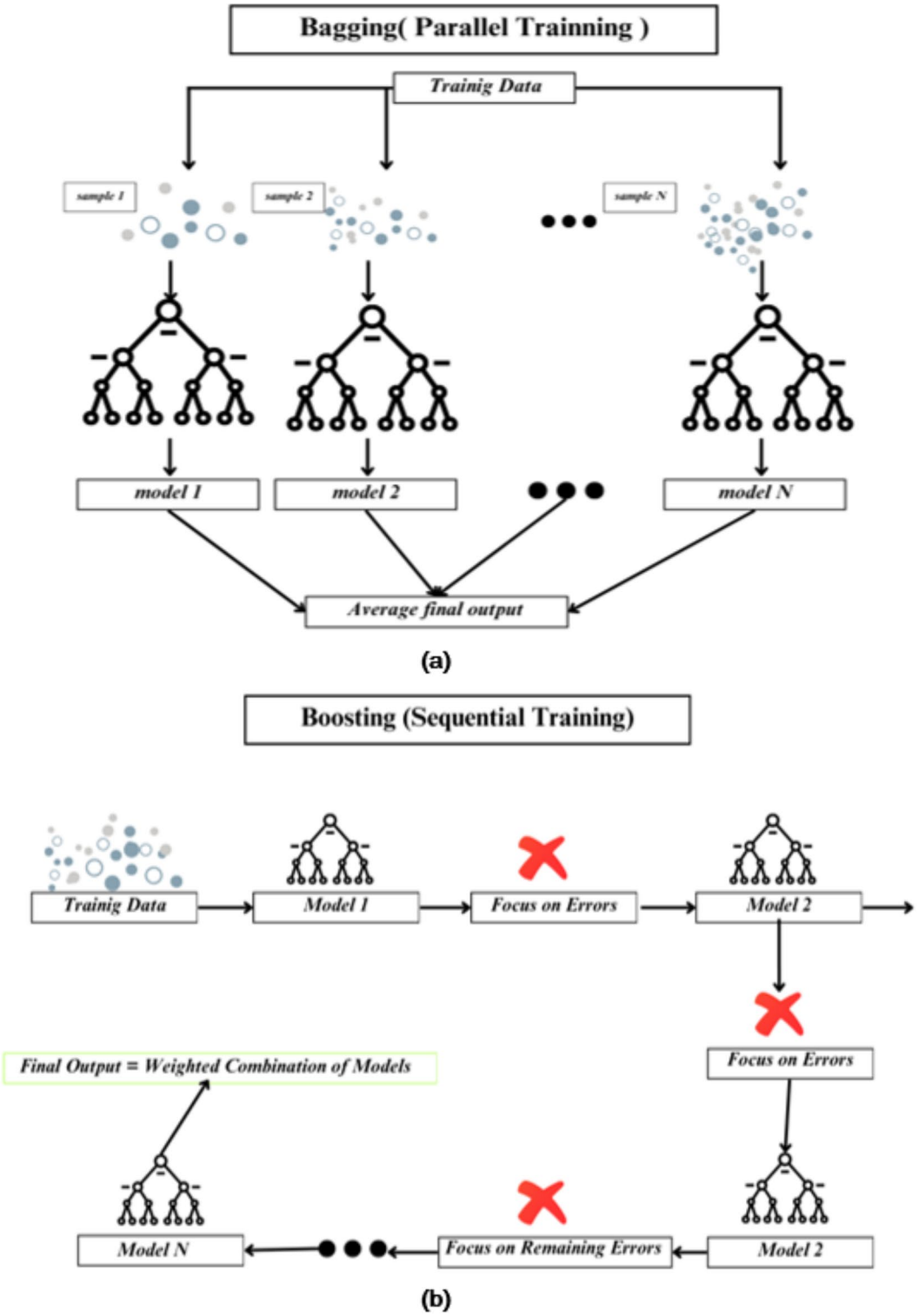

Fig. 2. Ensemble learning framework (**a**) bagging parallel training and (**b**) Boosting sequential training.

where $T_i(x)$ are the results from each tree. N is the number of samples. Averaging multiple predictions also minimizes the variance that improves the performance of the model because it does not rely heavily on a single decision tree. To generalize the model, features are selected randomly. The tool used for this algorithm is Python's ML libraries, that have built-in functions for the model to create it, for data processing and graphical representation of the results. Libraries such as Pandas and NumPy were adopted for data processing, Matplotlib and Seabo were used for results plots[56]. However, it is worth highlighting that the limitation of the RF regression

is the model needs some parameter tuning like number of trees and depth of each tree to avoid overfitting so that the model remains efficient[57].

Adaboost

Adaptive Boosting is also an ensemble technique, its algorithm combines weak learners, each one of it is fitted on different weighted versions of the data and creates a strong predictive model. To be represented in a mathematical way, the weight of training instances must be based on the error of the previous model prediction[58]. It can be written as:

$$w_i \leftarrow w_i \cdot \exp\left(\alpha_t \cdot L\left(y_0, y_{\mathrm{p}}\right)\right) \tag{2}$$

where w_i is the weight of the i-th instance, α_t is the model weight, y_0 is the actual value, y_{p} is the predicted value, L is the loss function used to measure the error in prediction.

Adaboost works best for the regression problem of modeling as opposed to the classification problems of computation, as it can reduce the mistakes gradually and come to appropriate findings regarding different patterns[59]. As shown in Fig. 2b each weak learner contributes on the model performance based on its accuracy, as a result more accurate weights have more influence on the predictors. This process improves the performance of the algorithm where there is underlaying complex relationships between the data which are difficult to capture by weak learners[58].

Gradient boosting

Gradient boosting is a versatile and robust ML algorithm used primarily in regression and classification tasks[60]. Its core principle is to combine multiple weak predictive models, typically decision trees, into a strong predictive model through a process called boosting[61]. This sequential construction of models aims to improve the mistakes of preceding models, which improves the prediction accuracy[62]. In GB for regression, the algorithm begins the process with training an initial model with a prediction base model such as mean of the target values and then adds subsequent models that are preferentially trained on the residuals. A new model in the sequence is fitted on the residuals of the last prediction and this makes the technique important in minimizing the prediction errors achieved[62]. One of the important characteristics of GB is the utilization of the loss function, which measures the distance between y_i (the actual value) and y' (the predicted value). Subsequently, for regression tasks, a popular metric is the mean square error (MSE). This loss function is then fed to the algorithm to determine gradients—partial derivatives that define both the direction and rate of steepest incline or decline in error. These gradients dictate on how new models are integrated. Mathematically, the GB model updates its predictions through the formula:

$$y_{\mathrm{p}} = f_0(x) + \eta \cdot \sum_{m=1}^{M} h_m(x) \tag{3}$$

where y_{p} is the predicted value, $f_0(x)$ is the initial model's prediction, η is the learning rate, a factor that scales the contribution of each tree, M is the number of boosting stages, $h_m(x)$ represents the contribution of the m-th tree. As per Eq. (3), it could be seen that each tree adds an adjustment $\eta \cdot h_m(x)$ derived from the learning rate and the gradient of the loss function, making the model better step by step.

Another critical parameter influencing this method is "learning rate (η)" which defines how each tree attempts to learn from the mistakes made by the previous trees. A small learning rate means that more trees are needed for modeling all the interactions but generally results in stronger model. In fact, hyperparameters can fine-tune GB depending on different indicator settings such as the number of trees and depths and learning rate, so that it has a vast application in different datasets and regression problems[62].

HistGradientBoosting

The HistGradientBoosting Regressor from Scikit-Learn is a highly efficient tool tailored for large datasets, leveraging histogram-based techniques to expedite GB processes[61]. Introduced initially as an experimental feature in Scikit-Learn v0.21.0, this estimator has since evolved into a stable and robust option for regression tasks, especially when faced with data scales larger than 10,000 samples[63]. Histogram-based GB simplifies and speeds up the traditional GB method by discretizing continuous features into bins, which dramatically reduces the computational complexity. Specifically, the model is structured based of the following mathematical concept:

$$y_i' = \sum_{k=1}^{K} f_k(x_i) \tag{4}$$

where K is the number of boosting stages and f_k are the individual regression trees, to predict the target variable y. Besides, this approach able to handle large datasets and includes native support for missing values. The model chooses the direction (left or right child in the tree) where the samples with missing values in the tree to provide better performance during training. This model algorithm is inspired by another model called LightGBM (were also used in this research) which is a GB framework that is known for its ability to function with large datasets[64].

The HistGradientBoosting regressor uses many loss functions that include squared error, absolute error and others that cover various regression types. Moreover, it includes specific strategies such as early stopping, which prevents overfitting. The use of the learning rate, tree depth and other parameters amongst others is another advantage of this model. These features make the HistGradientBoosting Regressor an ideal tool and powerful

for solving other forms of regression problems especially in complex domains as compared to normal prediction problems[65].

XGBoost

XGBoost: Extreme gradient boosting, it is a ML algorithm developed by researchers at the university of Washington[66]. The unique thing about this model is its generalization and ability to outperform other models with large datasets and that it could be applied for various tasks as regression and classification and still perform with the same strength[67].This model enhances the traditional boosting methods in the following enhanced features. First, it features a powerful tool for managing missing values[66]. And even greater advantage arises from the fact of its ability to work with several tasks at the same time, which significantly saves time for training models: especially in cases when working with large data sets. The basic working approach includes the consecutive inclusion of predictors into the ensemble, each one fixing up the preceding one as illustrated in Fig. 2b, using the formula:

$$y_p^{(t)} = y_0^{(t-1)} + \eta \cdot f_t(x_i) \tag{5}$$

where $y_p^{(t)}$ is the prediction at iteration t, f_t is the decision tree added at iteration t, and η is the learning rate, enhancing the model's ability to generalize. XGBoost performance depends on a wide range of hyperparameter to be optimized such as n_estimators, max_depth, learning_rate and subsample, which gives the user the freedom to control the results. The accuracy of the model based on the nature of the data analyzed because of that it is widely used in many areas from recommendation systems to finical modeling and environmental research. In addition, XGBoost has features such as cross validation, regularization to avoid overfitting as well as containing a tree pruning mechanism that makes XGBoost model simple and effective. Another method implemented for the treatment of weighted datasets involves the quantile sketch algorithm. These features make XGBoost an essential tool for data scientists who want to maximize the performance of algorithms in predictive models[68]. As a result, XGBoost is the tool of choice for its solidity, and flexibility to solutions to many problems in the field of ML.

LightGBM

Light gradient boosting Machine is an ensemble learning farmwork, made to be a faster version of GB. This tool was developed by Microsoft[69]. Their main interest when developing it was to reduce training time and memory usage using Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), that helped the algorithm to handle large datasets easily. The easy-to-understand aspect of LightGBM is that it constructs trees based on leaves (leaf-wise) as opposed to the traditional boosting tree construction based on levels (level-wise)[70]. This makes it possible to better minimize the loss at every split of the tree, as represented by the formula:

$$\text{Loss} = \sum_{i=1}^{n} l(y_0, y_p) + \sum \Omega(f_i) \tag{6}$$

where l is the loss function, y_0 are the true labels, y_p are the predicted labels, and Ω represents the regularization term applied to the function f. This makes it faster and more accurate, especially when managing data with high dimensionality.

LightGBM similar to XGBoost model works on both types of regression and classification systems. In addition, LightGBM users aren't required to perform preprocessing on the data such as one-hot encoding because this model can handle both continuous and categorical features. Another similarity with XGBoost is the freedom of customization of a variety of hyperparameters (like n_estimators, num_leaves, learning_rate, max_depth, subsample, colsample_bytree and min_child_samples)[70]. Therefore, due to the ability of using practical and efficient data structures and algorithms as well as its adjustability and applicability in different kinds of data and ML tasks LightGBM is a strong tool in the field of advanced analytics, which provides a decent perspective to substitute more conventional methods in a broad range of real-life applications starting from finances through Environmental development to other industries[71].

Modeling development

Hyperparameter tuning

For each model used in this study, the hyperparameter tuning was performed to ensure optimal performance using RandomizedSearchCV technique with three-fold cross validation to help find the best combination of hyperparameter for each individual model. The key hyperparameters tuned are listed in Table 1. Selection of hyperparameter ranges happened through a combination of default library settings in Scikit-learn, LightGBM, and XGBoost and practical experimental testing. Development of the parameter ranges occurred through experimental procedures that used grid search/random search to test various value combinations for maximizing model validation outcomes.

Train-test split

For models' development, 80% of the data were allocated for training the model to ensure that there are sufficient examples for it to learn robust patterns, and using the remaining 20% for testing was sufficient to estimate model performance. The data vision was in harmony to several research adopted over literature. In addition to randomization for each time the model will run, using random_state = 42 which ensured that the model was reliable under various conditions and reproducible.

Model	Hyperparameter	Range tested	Best value chosen
RF	n_estimators	[100, 200, 300]	300
	max_depth	[10, 20, None]	20
	min_samples_split	[2, 5, 10]	2
	min_samples_leaf	[1, 2, 4]	1
LightGBM	n_estimators	[300, 500, 700]	700
	num_leaves	[31, 70, 100]	70
	learning_rate	[0.05, 0.1]	0.05
	max_depth	[10, 20, 30]	20
	subsample	[0.8, 1.0]	0.8
	colsample_bytree	[0.8, 1.0]	0.8
	min_child_samples	[5, 10, 20]	10
	lambda_l1	[0.0, 1.0]	1
	lambda_l2	[0.0, 1.0]	1
HistGradientBoosting	learning_rate	[0.01, 0.05, 0.1]	0.05
	max_iter	[100, 200, 300]	200
	max_depth	[None, 10, 20]	10
	min_samples_leaf	[10, 20, 30]	20
	max_leaf_nodes	[31, 50, 70]	50
	l2_regularization	[0.0, 1.0, 10.0]	1
AdaBoost	n_estimators	[500, 1000, 1500]	500
	learning_rate	[0.01, 0.05, 0.1]	0.1
	estimator__max_depth	[3, 5, 7]	5
	loss	['linear', 'square', 'exponential']	'linear'
XGBoost	n_estimators	[100, 150]	150
	max_depth	[3, 5, 7]	7
	learning_rate	[0.1, 0.05]	0.1
	subsample	[0.8, 1.0]	0.8

Table 1. Hyperparameter tuning values.

Performance metrics

There are various well recognized performance metrics that can be used for modeling evaluation[28], such as determination coefficient (R^2), mean square error (MSE), mean absolute error (MAE), Willmott's Index of Agreement (WI), Nash-Sutcliffe Efficiency (NSE), Modified index of agreement (md). All metrics were reported to provide insights into the model's accuracy. The mathematical formulations for the calculated PMs over the study are:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_0 - y_p)^2}{\sum_{i=1}^{N} (y_o - y_p)^2} \tag{7}$$

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_p - y_o)^2 \tag{8}$$

$$MAE = \frac{\sum_{i=1}^{N} |y_p - y_o|}{N} \tag{9}$$

$$WI = 1 - \frac{\sum |y_p - y_o|}{\sum (|y_p - y_o^*| + |y_o - y_o'|)} \tag{10}$$

$$NSE = 1 - \frac{\sum_{i=1}^{N} (y_p - y_o)^2}{\sum_{i=1}^{N} (y_p - y_p')^2} \tag{11}$$

$$md = 1 - \frac{\sum_{i=1}^{N} (y_o - y_p)^j}{\sum_{i=1}^{N} (|y_p - y_o'| + |y_o - y_o'|)^j} \tag{12}$$

where y_o is the observed values, y_p is the Predicted values, y_o' is the mean of observed values, y_p' is the mean of predicted values, j is the exponent (typically 1 or 2), N is the total number of data points.

Results and discussion

Statistical and graphical models' assessment

The current sub-section presents the modelling results evaluation of various ensemble ML algorithms adopted for predicting the adsorption efficiency of heavy metals onto biochar. The selected models included simply structured models like RFs that used ensemble learning to combine the results of multiple decision trees and Adaboost which used a straightforward boosting approach by combining weak learners. In addition, complex models that involved iterative boosting, where trees are added sequentially to achieve a higher accuracy by minimizing prediction errors from previous iterations, such as Gradient Boosting, XGBoost, LightGBM, and HistGradientBoosting. Each one of these algorithms tends to show high R^2 that fluctuated between 0.86 and 0.93 as the data contained strong straightforward relationships between features and absorption efficiency. It was demonstrated in the feature importance analysis, which illustrated that the initial concentration ratio of metals to biochar, pH, and Pyrolysis temperature consistently ranked as the most influential features on efficiency. These findings align with the previously published research[49], the authors showed the demonstration of chemical properties such as cation exchange capacity and functional group availability in addition to the temperature over physical attributes such as porosity and surface area in the process.

XGBoost model emerged as the most accurate model, achieving the lowest MSE = 0.0057, MAE = 0.0428, and the highest $R^2 = 0.921$, demonstrating its superior predictive performance. This is followed by GB model, which achieved $R^2 = 0.9120$, a slightly higher MSE = 0.0063, and a comparable MAE = 0.0426. The results indicated the capacity of the model to effectively capture the internal relationship between the inputs and output of the dataset. On the other hand, RF regressor model performed closely behind, with an $R^2 = 0.9097$, MSE = 0.0065, and an MAE = 0.0428, showing robust performance similar to GB model. LightGBM achieved moderate performance, with an $R^2 = 0.8836$, MSE = 0.0089, and MAE = 0.0562, indicating a slight increase in prediction error compared to the top-performing models. HistGradientBoosting model followed with $R^2 = 0.8755$, MSE = 0.0095, and MAE = 0.0541, catch the eye with a comparable but slightly lower performance relative to LightGBM. The lowest-performing model was AdaBoost, with $R^2 = 0.8694$, MSE = 0.0161, and MAE = 0.0646. AdaBoost model revealed a limitation to capture the relationships in the data as effectively as the other models. However, it is worth noting that AdaBoost still achieved a reasonable WI = 0.9671, which indicates its predictions align moderately well with the observed values, albeit with greater deviation. Using the performance metrics (i.e., WI and NSE), XGBoost also led with values of 0.9757 and 0.9209, respectively, further reinforcing its superior agreement with observed values. RF and GB models similarly showed high WI values of 0.9733 and 0.9732, alongside with strong NSE values of 0.9097 and 0.9129, reflecting their strong predictive consistency. On the other hand, HistGradientBoosting and AdaBoost recorded the lowest NSE values at 0.8755 and 0.8694, respectively. Indicating their comparatively reduced performance. Finally, the Modified index of agreement (md) values for all models remained low, with XGBoost (md = 0.0144) and RF (md = 0.0105), further emphasizing their accuracy in minimizing the average prediction error. The performance metrics of all remaining models used were listed in Table 2. XGBoost model was consistently outperforming other models when dealing with environmental applications for instance this algorithm was robust predictor for PFAS: per- and polyfluoroalkyl substances removal efficiency[72]. Which was due to its capacity of capturing complex interactions between features. In addition, its advanced regularization mechanism prevents overfitting in predictive modeling tasks the involves complicated interactions between features.

To visually present the performance the predicted vs. actual, scatter plots were generated in Fig. 3. The XGBoost model showed a tight clustering of points along the ideal prediction line reflecting minimal deviation between actual and predicted values, which indicates efficient generalization across various test samples and reliability of this model. Similarly, GB and RF models showed high performance. In contrast, models like Adaboost showed a higher deviation from the ideal line. Regression formulas for each model were also presented in Fig. 3 to quantify the linear relationships between predicted and actual values of adsorption efficiency.

The residual error plots were presented in Fig. 4, in order to provide further justification of each model's performance. For instance, XGBoost model were distributed symmetrically around zero, which suggests effectiveness of the model and its ability to capture relationships between the features without underfitting or overfitting. These graphs align with the statistical results (R^2, MSE, MAE, WI, NSE, md) (see Table 2), and reinforce the reliability and robustness of the models' claims in this research.

The violin diagrams in Fig. 5 illustrate the distribution of predicted and actual values for each ensemble ML model, providing insights into their consistency and performance[73]. XGBoost distribution in the diagram is narrow and centralized, which indicates high consistency of its predictions and a close alignment with the actual values. This also demonstrates the generalization ability of this model and aligns with the numerical matrices like R2 which had a value of 0.921and the low MSE and MAE. In contrast, AdaBoost shows a wider

Model	R^2 (test)	MSE	MAE	WI	NSE	md
RF regressor	0.9097	0.0065	0.0428	0.9733	0.9097	0.0105
LightGBM	0.8836	0.0089	0.0562	0.9660	0.8836	0.0094
HistGradientBoosting	0.8755	0.0095	0.0541	0.9646	0.8755	0.0174
AdaBoost	0.8694	0.0161	0.0646	0.9671	0.8694	0.0164
GB	0.9120	0.0063	0.0426	0.9732	0.9129	0.0129
Xgboost	0.9209	0.0056	0.0428	0.9757	0.9209	0.0144

Table 2. The performance metrics for all the developed ensemble ML models.

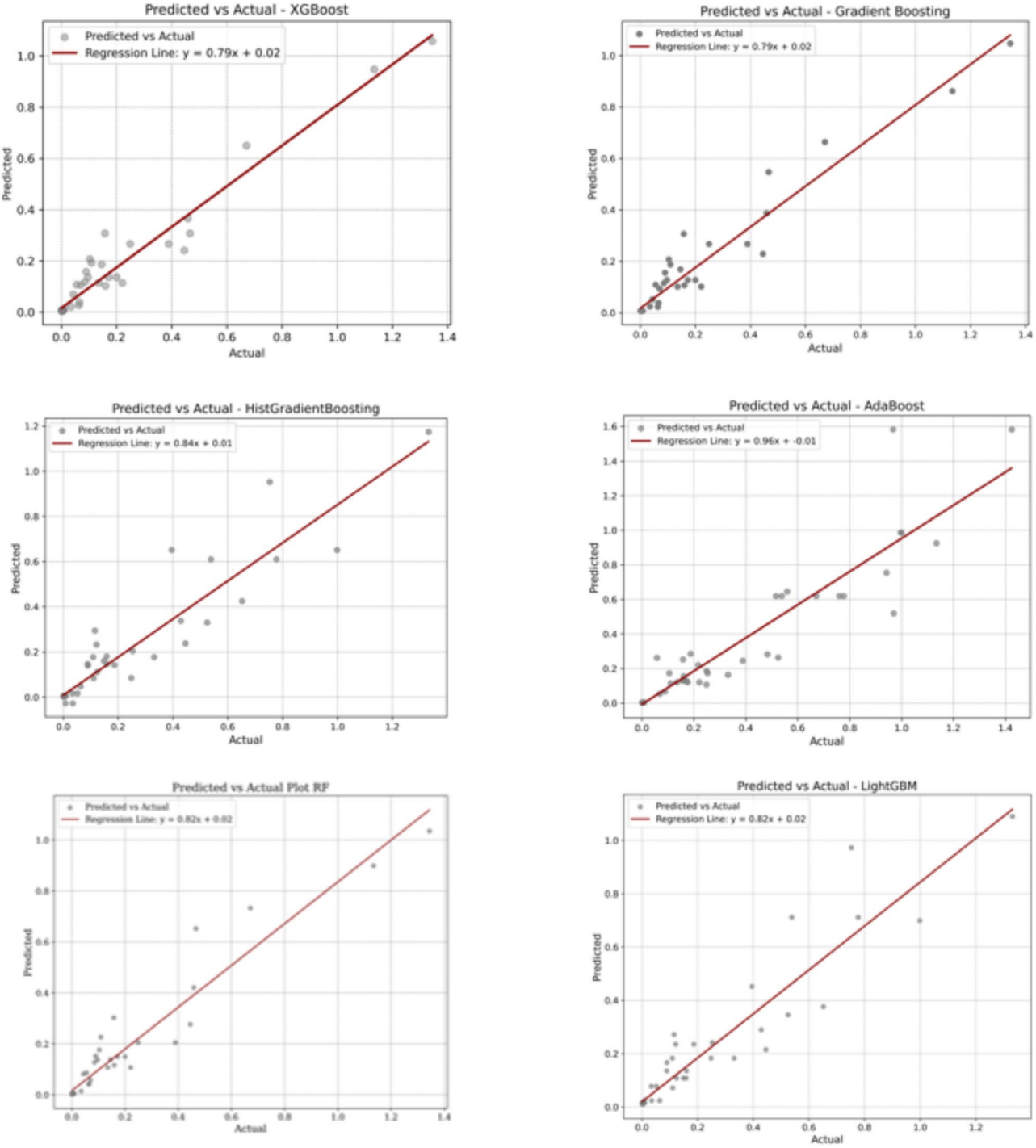

Fig. 3. The predicted vs. actual for the adsorption efficiency using each developed ensemble ML model.

distribution in addition to a noticeable spread and variability in the predicted values reflecting inconsistency in its predictions. A broader shape in the violin diagram illustrates struggling in capturing relationships between features and the target[74]. Which also aligns with the R^2 (0.8694) obtained by the model. Intermediate behavior was shown by GB and RF with distributions that are slightly wider than XGBoost, as they both achieved high R^2 of 0.9120 and 0.9097, respectively. LightGBM and HistGradientBoosting showed a similar result with a wider distributions and moderate alignment between actual and predicted values.

To provide a comprehensive visualization of the models' outcomes, Taylor Diagram was plotted in Fig. 6. This diagram provides a valuable tool for assessing model performance because it presents multiple statistical indicators (i.e., standard deviation, root mean square error, and correlation coefficient) in a single plot[75]. The diagram demonstrated that XGBoost model showed the highest correlation coefficient of 0.9745 and standard

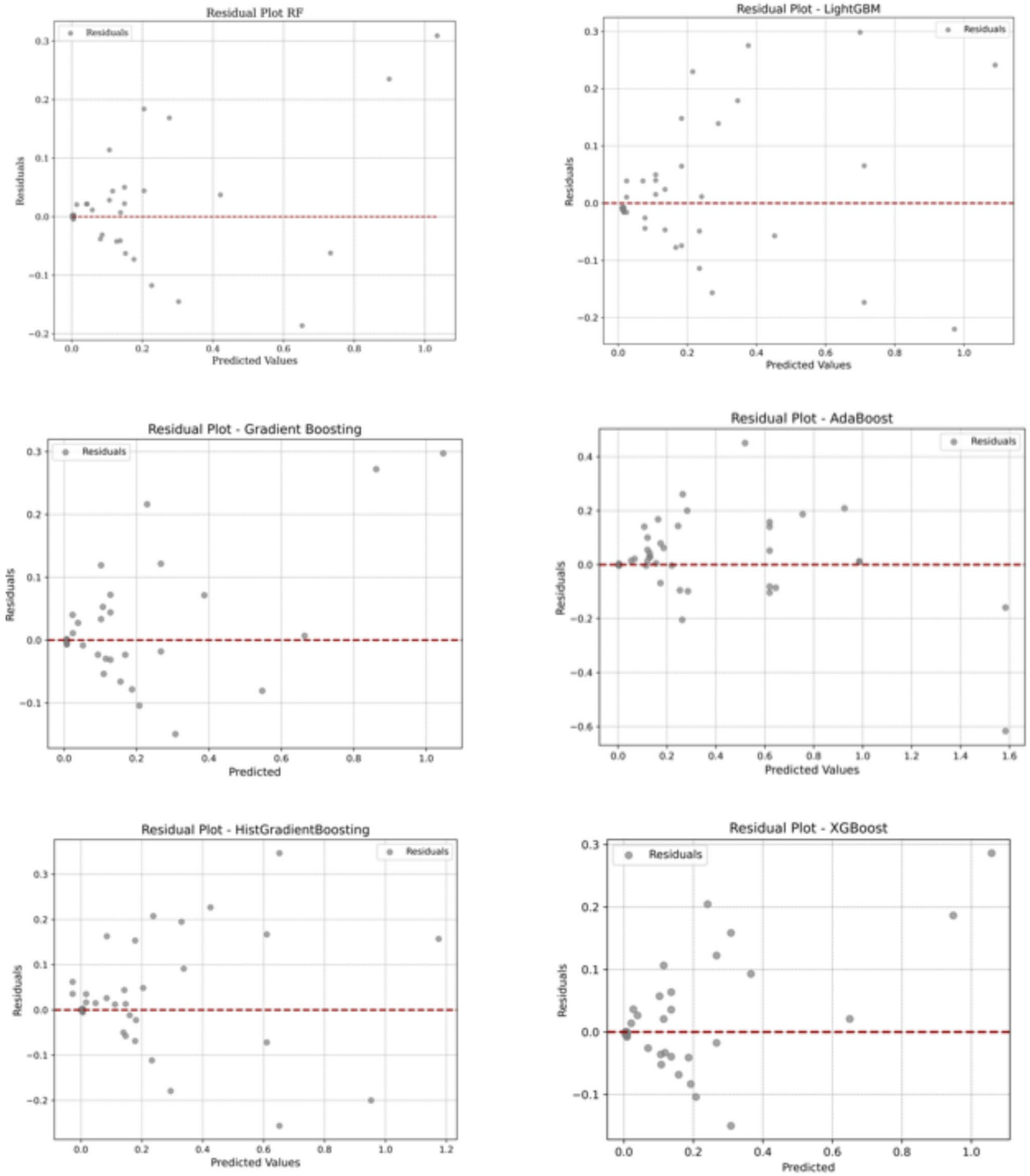

Fig. 4. Residual error for each developed ML model.

deviation of 0.2185 which aligns with the actual adsorption efficiency closely standard deviation of 0.2684. In addition, XGBoost model has centered RMSE of 0.0741 which supports the enhanced efficiency of model prediction. Nevertheless, AdaBoost and HistGradientBoosting models have significant deviations from the actual absolute error values as demonstrated by their large standard deviations and correlation risks. This verifies their low forecasting efficiency and low ability to generalize. These results align with the higher tendency of their error measures noted before, including MSE and less accuracy found in their lower R^2.

The RF and GB models satisfactorily preserve the correlation/standard deviation balance of the dataset and satisfactorily achieve results which are slightly lower than XGBoost's. It can be observed that the two models have

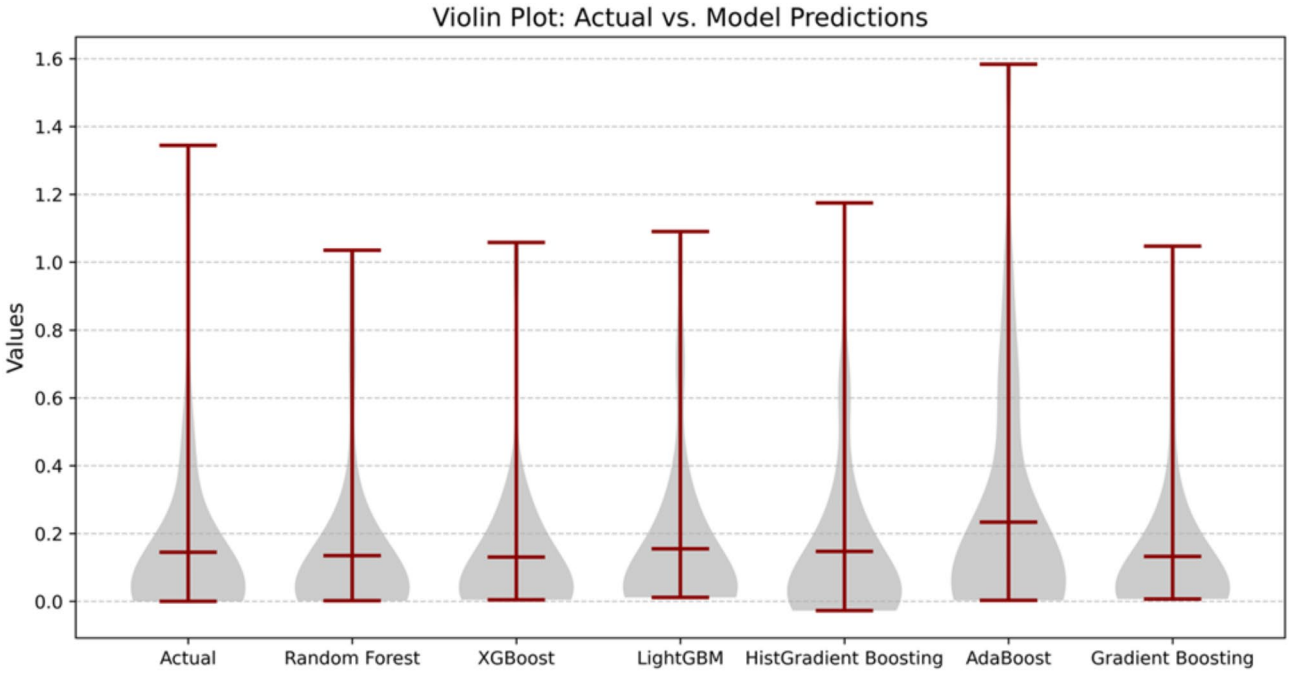

Fig. 5. Violin diagrams performance of each developed ML model.

high correlation coefficients, and the standard deviations inferred by the models are close to average absolute error values. Similarly, LightGBM showed high correlation coefficient, but its standard deviation deviates from actual absolute error values slightly, pointing moderately good generalization capacity but slightly lower accuracy in comparison with the most accurate models. In general, the form of Taylor Diagram captures more positively in performance for XGBoost as it most closely mimics the correlation, standard deviation, and error minimization model. The other models show greater variances and lesser generalizability in terms of predictive accuracy by having lower absolute error values closer to the reference absolute error values.

Models interpretation and discussion

The main motive of the current investigation was to develop a data-intelligence models based on ensemble ML models to predict the efficiency of biochar properties for heavy metals adsorption. From the engineering point of view, the essential features that configure the prediction matrix were the initial concentration ratio, pH, and Pyrolysis temperature. The initial concentration ratio determines the driving force for mass transfer, influencing the adsorption capacity. pH affects the surface charge of biochar and the speciation of heavy metals, altering adsorption interactions. Pyrolysis temperature impacts the biochar's surface area, porosity, and functional groups, which in turn influence the adsorption efficiency. Based on the designed ML models for the selected size of the dataset experiment (353 experiment observations), the models were performed an acceptable prediction accuracy based on the reported statistical and graphical presentation which give the credit to the applied ML models to mimic the physical and chemical relationship between predictors and predictand based on those limited dataset.

The dataset analysis revealed XGBoost produced the high R^2 score of 0.921 alongside a minimum MSE of 0.0057 which proved its ability to identify complex nonlinear data interactions. XGBoost applies GB algorithm with regularization functions to keep high accuracy rates alongside decreased overfitting risks. The tree-based method of XGBoost continues to refine prediction errors from previous models as it enhances performance.

It is very important to validate the obtained results with the previously published literature of the related works. By comparing the outcomes of the present study with the established work adopted in the same manners on the development of ML models for the prediction of adsorption efficiency. The current study outperformed previous studies by demonstrating better predictive accuracy in its results. Some scholars examined dye adsorption onto agricultural waste-activated carbon using RF, GB and DT[76]. Research findings showed that RF delivered the best result with R^2 of 0.90. In adsorption-related datasets ensemble learning models specifically XGBoost demonstrate better prediction quality than other models. Similarly, another study applied RF and CatBoost to predict the efficiency of biochar in pesticide remediation[77]. The RF model achieved an R^2 of 0.796, significantly lower than the present study's results, highlighting the robustness of XGBoost in capturing complex adsorption behaviors. This comparison further demonstrates that XGBoost is a highly effective model for adsorption prediction, outperforming RF in multiple studies.

It can be seen that XGBoost model was at the top of the prediction accuracy ranking, while RF giving a slightly lower accuracy level. These models were able to capture complex, non-linear relationships in the data effectively. The application of ensemble ML models in this study offers several practical benefits for chemical and environmental engineering. Firstly, by identifying key features influences of the process like the initial

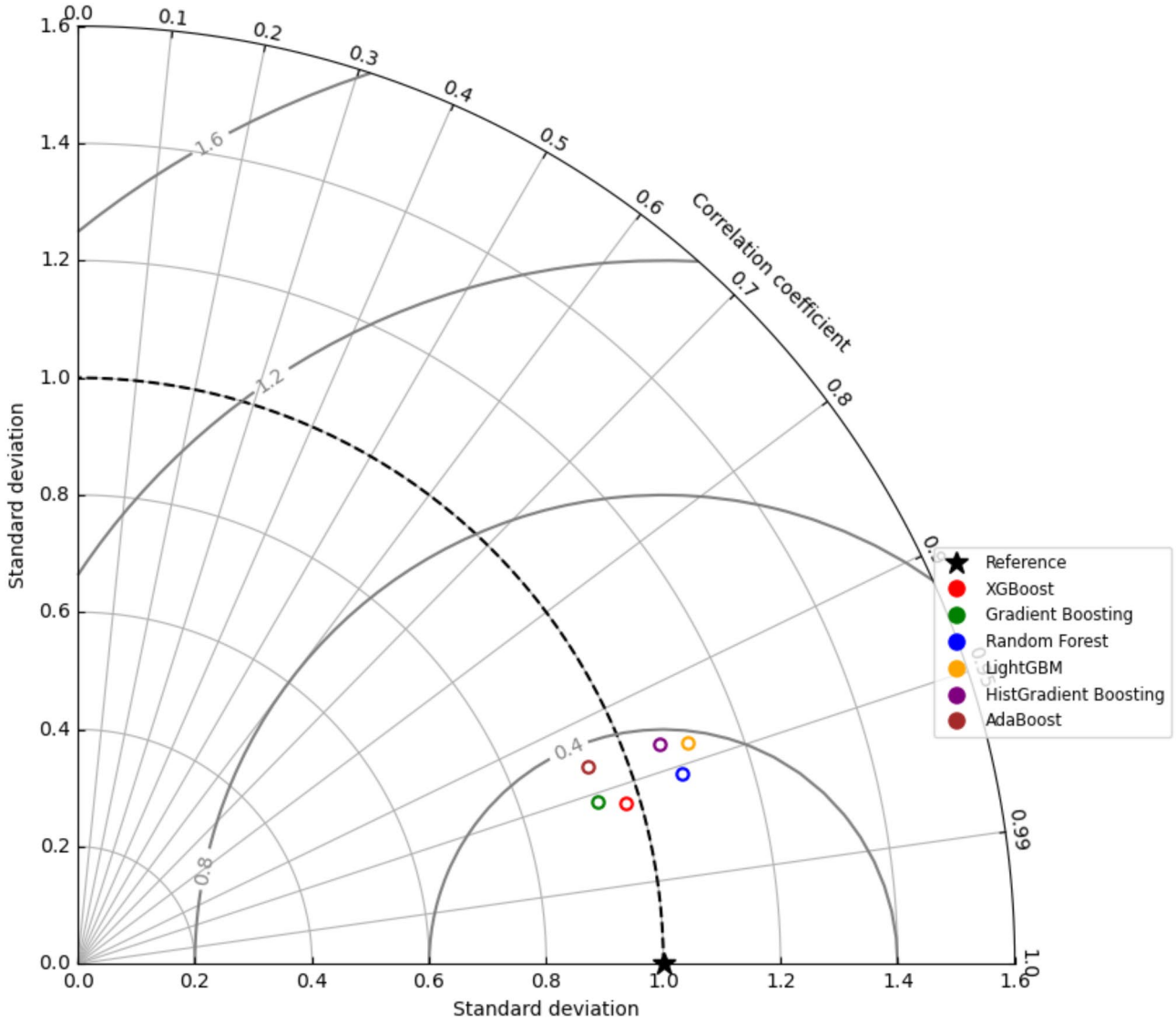

Fig. 6. Taylor diagram presentation for all the developed ensemble ML model.

concentration and pH which enhance the adsorption efficiency. Secondly, by accelerating the optimization process and reducing the reliance on time consuming and costly experiments. Finally, these models enable accurate predictions under various environmental conditions supporting pollution control and remediation efforts.

In conclusion, although the developed ensemble ML models gave superior prediction accuracy. However, the common fact that those ML models are based on black-box. Thus, adopting more advance models as a secondary stage procedure to evaluate the interpretability and explainability using Shapley Additive Explainable (SHAP) model can give a substantial credit on the better understanding for the physical and chemical interactions between the input features and the aimed target efficiency of biochar properties for heavy metals adsorption[78]. By analyzing SHAP values, the study can provide deeper insights into how each input variable (e.g., initial concentration ratio, pH, pyrolysis temperature) influences the adsorption efficiency, thereby enhancing model interpretability and bridging the gap between ML predictions and physical-chemical understanding.

Conclusion

In this research the efficiency of ensemble ML models was investigated in predicting heavy metal adsorption. The modeling results indicated the ability of developed models to capture the complex relationships between chemical and physical features of the biochar and with adsorption efficiency. The adopted models accelerated the optimization of the adsorption process and gave prediction results that guide biochar selection and design. Key observations revealed that chemical properties such as initial concentration, cation exchange capacity and pH were the most influential factors, this shows the importance of optimizing chemical processes in biochar design for adsorption. The results of this work were promising; however, some limitations were concluded and to be further studied. To enhance performance and generalize of the applied algorithms, the dataset can be

expanded and should include more types of biochar, environmental conditions and contamination scenarios. As a result, future research should also test ML models under various and complex real-world circumstances to ensure their robustness and applicability. Further, future studies recommend attempting a hybrid ML models version that combines more than one algorithm to enhance accuracy of predictions. In conclusion, this study offered a preliminary computer-aided model for the further study of the enhancing biochar use for heavy metal adsorption with data analysis.

Data availability

Data can be shared upon request from the corresponding author.

Received: 4 January 2025; Accepted: 27 March 2025
Published online: 18 April 2025

References

1. Timothy, N. & Williams, E. T. Environmental pollution by heavy metal: An overview. *Int. J. Environ. Chem.* **3**(2), 72–82. https://doi.org/10.11648/J.IJEC.20190302.14 (2019).
2. Mishra, S. et al. Heavy metal contamination: an alarming threat to environment and human health. *Environ. Biotechnol. Sustain. Future*. 103–125. https://doi.org/10.1007/978-981-10-7284-0_5 (2019).
3. Peng, X., Deng, Y., Peng, Y. & Yue, K. Effects of biochar addition on toxic element concentrations in plants: A meta-analysis. *Sci. Total Environ.* **616**, 970–977. https://doi.org/10.1016/j.scitotenv.2017.10.222 (2018).
4. Liu, C., Lin, J., Chen, H., Wang, W. & Yang, Y. Comparative study of biochar modified with different functional groups for efficient removal of Pb(II) and Ni(II). *Int. J. Environ. Res. Public. Health*. **19** (18), 11163. https://doi.org/10.3390/IJERPH191811163/S1 (2022).
5. Wang, X., Chi, Q., Liu, X. & Wang, Y. Influence of pyrolysis temperature on characteristics and environmental risk of heavy metals in pyrolyzed biochar made from hydrothermally treated sewage sludge. *Chemosphere*. **216**, 698–706. https://doi.org/10.1016/J.CHEMOSPHERE.2018.10.189 (2019).
6. Liu, C. & Zhang, H. X. Modified-biochar adsorbents (MBAs) for heavy-metal ions adsorption: A critical review. *J. Environ. Chem. Eng.* **10**(2), 107393. https://doi.org/10.1016/J.JECE.2022.107393 (2022).
7. Sulaymon, A. H., Mohammed, A. A. & Al-Musawi, T. J. Competitive biosorption of lead, cadmium, copper, and arsenic ions using algae. *Environ. Sci. Pollut. Res.* **20**(5), 3011–3023. https://doi.org/10.1007/S11356-012-1208-2/TABLES/4 (2013).
8. Yuan, C. et al. A meta-analysis of heavy metal bioavailability response to biochar aging: importance of soil and biochar properties. *Sci. Total Environ.* **756**, 144058. https://doi.org/10.1016/J.SCITOTENV.2020.144058 (2021).
9. Li, H. et al. Mechanisms of metal sorption by Biochars: Biochar characteristics and modifications. *Chemosphere*. **178**, 466–478. https://doi.org/10.1016/J.CHEMOSPHERE.2017.03.072 (2017).
10. Moreno-Pérez, J. et al. Artificial neural network-based surrogate modeling of multi-component dynamic adsorption of heavy metals with a biochar. *J. Environ. Chem. Eng.* **6**(4), 5389–5400. https://doi.org/10.1016/J.JECE.2018.08.038 (2018).
11. Mansoor, S. et al. Biochar as a tool for effective management of drought and heavy metal toxicity. *Chemosphere*. **271** https://doi.org/10.1016/J.CHEMOSPHERE.2020.129458 (2021).
12. Shakoor, M. B. et al. A review of biochar-based sorbents for separation of heavy metals from water. *Int. J. Phytorem.* **22** (2), 111–126. https://doi.org/10.1080/15226514.2019.1647405 (2020).
13. Hu, Q., Xiao, Z., Xiong, X., Zhou, G. & Guan, X. Predicting heavy metals' adsorption edges and adsorption isotherms on MnO2 with the parameters determined from Langmuir kinetics. *J. Environ. Sci.* **27**, 207–216. https://doi.org/10.1016/J.JES.2014.05.036 (2015).
14. Özer, A., Özer, D. & Ekiz, H. I. Application of Freundlich and Langmuir models to multistage purification process to remove heavy metal ions by using schizomeris leibleinii. *Process Biochem.* **34** (9), 919–927. https://doi.org/10.1016/S0032-9592(99)00011-4 (1999).
15. Duan, Q., Yan, P., Feng, Y., Wan, Q. & Zhu, X. Machine learning assisted adsorption performance evaluation of Biochar on heavy metal. *Front. Environ. Sci. Eng.* **18** (5), 1–14. https://doi.org/10.1007/S11783-024-1815-4/METRICS (2024).
16. Chen, M. W., Chang, M. S., Mao, Y., Hu, S. & Kung, C. C. Machine learning in the evaluation and prediction models of Biochar application: A review. *Sci. Prog.* **106** (1). https://doi.org/10.1177/00368504221148842/ASSET/IMAGES/LARGE/10.1177_00368504221148842-FIG4.JPEG (2023).
17. Haider, J. et al. Machine-learning-based prediction and optimization of emerging contaminants' adsorption capacity on Biochar materials. *Chem. Eng. J.* **466** https://doi.org/10.1016/J.CEJ.2023.143073 (2023).
18. Ukoba, K., Jen, T. C., Ukoba, K. & Jen, T. C. Biochar and application of machine learning: A review. *Biochar Prod. Technol. Prop. Appl.*. https://doi.org/10.5772/INTECHOPEN.108024 (2022).
19. Ucun Ozel, H. et al. Application of artificial neural networks to predict the heavy metal contamination in the Bartin River. *Environ. Sci. Pollut. Res. Int.* **27**(34), 42495–42512. https://doi.org/10.1007/S11356-020-10156-W (2020).
20. Yuan, X. et al. Machine learning for heavy metal removal from water: Recent advances and challenges. *ACS ES T Water*. **4**(3), 820–836. https://doi.org/10.1021/ACSESTWATER.3C00215/ASSET (2024).
21. Palansooriya, K. N. et al. Prediction of soil heavy metal immobilization by Biochar using machine learning. *Environ. Sci. Technol.* **56** (7), 4187–4198. https://doi.org/10.1021/ACS.EST.1C08302 (2022).
22. Aryafar, A., Gholami, R., Rooki, R. & Doulati Ardejani, F. Heavy metal pollution assessment using support vector machine in the Shur river, Sarcheshmeh copper mine, Iran. *Environ. Earth Sci.* **67** (4), 1191–1199. https://doi.org/10.1007/S12665-012-1565-7/METRICS (2012).
23. Talebkeikhah, F. et al. Investigation of effective processes parameters on lead (II) adsorption from wastewater by Biochar in mild air oxidation pyrolysis process. *Int. J. Environ. Anal. Chem.* **102** (16), 3975–3995. https://doi.org/10.1080/03067319.2020.1777291 (2022).
24. Li, M. et al. EDTA functionalized magnetic Biochar for Pb(II) removal: adsorption performance, mechanism and SVM model prediction. *Sep. Purif. Technol.* **227**, 115696. https://doi.org/10.1016/J.SEPPUR.2019.115696 (2019).
25. Wong, Y. J., Arumugasamy, S. K., Chung, C. H., Selvarajoo, A. & Sethu, V. Comparative study of artificial neural network (ANN), adaptive neuro-fuzzy inference system (ANFIS) and multiple linear regression (MLR) for modeling of Cu (II) adsorption from aqueous solution using biochar derived from rambutan (Nephelium lappaceum) peel. *Environ. Monit. Assess.* **192**(7), 1–20. https://doi.org/10.1007/S10661-020-08268-4/TABLES/7 (2020).
26. Chakraborty, V. & Das, P. Synthesis of nano-silica-coated biochar from thermal conversion of sawdust and its application for Cr removal: kinetic modelling using linear and nonlinear method and modelling using artificial neural network analysis. *Biomass Convers. Biorefin.* **13**(2), 821–831. https://doi.org/10.1007/S13399-020-01024-1 (2023).
27. Bhagat, S. K., Tung, T. M. & Yaseen, Z. M. Development of artificial intelligence for modeling wastewater heavy metal removal: state of the art, application assessment and possible future research. *J. Clean. Prod.* **250**, 119473. https://doi.org/10.1016/J.JCLEPRO.2019.119473 (2020).

28. Yaseen, Z. M. An insight into machine learning models era in simulating soil, water bodies and adsorption heavy metals: review, challenges and solutions. *Chemosphere*. **277**, 130126. https://doi.org/10.1016/J.CHEMOSPHERE.2021.130126 (2021).
29. Ke, B. et al. Predicting the sorption efficiency of heavy metal based on the Biochar characteristics, metal sources, and environmental conditions using various novel hybrid machine learning models. *Chemosphere*. **276**, 130204. https://doi.org/10.1016/J.CHEMOSPHERE.2021.130204 (2021).
30. Li, Y. et al. A novel hybrid variable cross layer-based machine learning model improves the accuracy and interpretation of energy intensity prediction of wastewater treatment plant. *J. Environ. Manag.* **371**, 123209. https://doi.org/10.1016/J.JENVMAN.2024.123209 (2024).
31. Li, J. et al. Biochar design for antibiotics adsorption via a hybrid machine-learning-based optimization framework. *Sep. Purif. Technol.* **348**, 127666. https://doi.org/10.1016/J.SEPPUR.2024.127666 (2024).
32. Wanyonyi, F. S., Fidelis, T. T., Mutua, G. K., Orata, F. & Pembere, A. M. S. Role of pore chemistry and topology in the heavy metal sorption by zeolites: From molecular simulation to machine learning. *Comput. Mater. Sci.* **195**. https://doi.org/10.1016/J.COMMATSCI.2021.110519 (2021).
33. Sun, Y. et al. The application of machine learning methods for prediction of metal immobilization remediation by Biochar amendment in soil. *Sci. Total Environ.* **829**, 154668. https://doi.org/10.1016/J.SCITOTENV.2022.154668 (2022).
34. Gou, J. et al. Optimizing Biochar yield and composition prediction with ensemble machine learning models for sustainable production. *Ain Shams Eng. J.* **16** (1), 103209. https://doi.org/10.1016/J.ASEJ.2024.103209 (2025).
35. Shaheen, A., Iqbal, J. & Algorithm, B. Spatial distribution and mobility assessment of carcinogenic heavy metals in soil profiles using geostatistics and random forest. *Sustain.* **10**(3), 799. https://doi.org/10.3390/SU10030799 (2018).
36. Hanandeh, A. E., Mahdi, Z. & Imtiaz, M. S. Modelling of the adsorption of Pb, Cu and Ni ions from single and multi-component aqueous solutions by date seed derived Biochar: comparison of six machine learning approaches. *Environ. Res.* **192**, 110338. https://doi.org/10.1016/J.ENVRES.2020.110338 (2021).
37. Cui, X., Hao, H., Zhang, C., He, Z. & Yang, X. Capacity and mechanisms of ammonium and cadmium sorption on different wetland-plant derived biochars. *Sci. Total Environ.* **539**, 566–575. https://doi.org/10.1016/J.SCITOTENV.2015.09.022 (2016).
38. Ding, Z. et al. Sorption of lead and methylene blue onto hickory biochars from different pyrolysis temperatures: importance of physicochemical properties. *J. Ind. Eng. Chem.* **37**, 261–267. https://doi.org/10.1016/J.JIEC.2016.03.035 (2016).
39. Jiang, S. et al. Copper and zinc adsorption by softwood and hardwood biochars under elevated sulphate-induced salinity and acidic pH conditions. *Chemosphere*. **142**, 64–71. https://doi.org/10.1016/J.CHEMOSPHERE.2015.06.079 (2016).
40. Shen, Z., Zhang, Y., Jin, F., McMillan, O. & Al-Tabbaa, A. Qualitative and quantitative characterisation of adsorption mechanisms of lead on four biochars. *Sci. Total Environ.* **609**, 1401–1410. https://doi.org/10.1016/J.SCITOTENV.2017.08.008 (2017).
41. Shen, Z., Zhang, Y., McMillan, O., Jin, F. & Al-Tabbaa, A. Characteristics and mechanisms of nickel adsorption on biochars produced from wheat straw pellets and rice husk. *Environ. Sci. Pollut Res. Int.* **24** (14), 12809–12819. https://doi.org/10.1007/S11356-017-8847-2 (2017).
42. Sun, Y. et al. Effects of feedstock type, production method, and pyrolysis temperature on Biochar and hydrochar properties. *Chem. Eng. J.* **240**, 574–578. https://doi.org/10.1016/J.CEJ.2013.10.081 (2014).
43. Trakal, L., Bingöl, D., Pohořelý, M., Hruška, M. & Komárek, M. Geochemical and spectroscopic investigations of Cd and Pb sorption mechanisms on contrasting biochars: Engineering implications. *Bioresour. Technol.* **171**, 442–451. https://doi.org/10.1016/J.BIORTECH.2014.08.108 (2014).
44. Zama, E. F., Zhu, Y. G., Reid, B. J. & Sun, G. X. The role of biochar properties in influencing the sorption and desorption of Pb(II), Cd(II) and As(III) in aqueous solution. *J. Clean. Prod.* **148**, 127–136. https://doi.org/10.1016/J.JCLEPRO.2017.01.125 (2017).
45. Gao, L. Y. et al. Relative distribution of Cd2 + adsorption mechanisms on biochars derived from rice straw and sewage sludge. *Bioresour. Technol.* **272**, 114–122. https://doi.org/10.1016/J.BIORTECH.2018.09.138 (2019).
46. Shen, Z., Jin, F., Wang, F., McMillan, O. & Al-Tabbaa, A. Sorption of lead by Salisbury Biochar produced from British broadleaf hardwood. *Bioresour. Technol.* **193**, 553–556. https://doi.org/10.1016/J.BIORTECH.2015.06.111 (2015).
47. Li, Y. et al. Qualitative and quantitative correlation of physicochemical characteristics and lead sorption behaviors of crop residue-derived chars. *Bioresour. Technol* **270**, 545–553. https://doi.org/10.1016/J.BIORTECH.2018.09.078 (2018).
48. Cui, X. et al. Potential mechanisms of cadmium removal from aqueous solution by *Canna indica* derived Biochar. *Sci. Total Environ.* **562**, 517–525. https://doi.org/10.1016/J.SCITOTENV.2016.03.248 (2016).
49. 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. https://doi.org/10.1016/J.JHAZMAT.2019.06.004 (2019).
50. Jason, B. Why One-Hot Encode Data in Machine Learning?—MachineLearningMastery.com. https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/ (accessed 09 Nov 2024).
51. Yaseen, Z. M. & Alawi, O. A. Artificial intelligence models development for profitability factor prediction in concentrated solar power with dual backup systems. *Sci. Rep.*. **15**(1), 1–27. https://doi.org/10.1038/s41598-025-87584-6 (2025).
52. Sagi, O. & Rokach, L. Ensemble learning: A survey. *Wiley Interdiscip Rev. Data Min. Knowl. Discov.* **8** (4). https://doi.org/10.1002/WIDM.1249 (2018).
53. Galar, M., Fernández, A., Barrenechea, E., Bustince, H. & Herrera, F. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. https://doi.org/10.1109/TSMCC.2011.2161285 (2011).
54. Ferreira, A. J. & Figueiredo, M. A. T. Boosting algorithms: A review of methods, theory, and applications. *Ensemble Mach. Learn.* 35–85. https://doi.org/10.1007/978-1-4419-9326-7_2 (2012).
55. Wang, H. et al. Prediction models of soil heavy metal(loid)s concentration for agricultural land in Dongli: A comparison of regression and random forest. *Ecol. Indic.* **119**, 106801. https://doi.org/10.1016/J.ECOLIND.2020.106801 (2020).
56. Breiman, L. Random forests. *Mach. Learn.* **45**(1), 5–32. https://doi.org/10.1023/A:1010933404324 (2001).
57. Tan, K., Ma, W., Wu, F. & Du, Q. Random forest-based estimation of heavy metal concentration in agricultural soils with hyperspectral sensor data. *Environ. Monit. Assess.* **191**, 1–14. https://doi.org/10.1007/S10661-019-7510-4/FIGURES/6 (2019).
58. Schapire, R. E. Explaining adaboost, empirical inference: festschrift in honor of Vladimir N. Vapnik. 37–52. https://doi.org/10.1007/978-3-642-41136-6_5/TABLES/2 (2013).
59. Lin, N. et al. Estimating the heavy metal contents in farmland soil from hyperspectral images based on stacked adaboost ensemble learning. *Ecol. Indic.* **143**, 109330. https://doi.org/10.1016/J.ECOLIND.2022.109330 (2022).
60. Bentéjac, C., Csörgő, A. & Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. *Artif. Intell. Rev.* **54**(3), 1937–1967. https://doi.org/10.1007/S10462-020-09896-5/TABLES/12 (2021).
61. Tao, H., Awadh, S. M., Salih, S. Q., Shafik, S. S. & Yaseen, Z. M. Integration of extreme gradient boosting feature selection approach with machine learning models: application of weather relative humidity prediction. *Neural Comput. Appl.* **34** (1), 515–533. https://doi.org/10.1007/S00521-021-06362-3/METRICS (2022).
62. Natekin, A. & Knoll, A. Gradient boosting machines, a tutorial. *Front. Neurorobot.* **7**, 63623. https://doi.org/10.3389/FNBOT.2013.00021/BIBTEX (2013).
63. scikit-learn developers HistGradientBoostingClassifier—scikit-learn 1.5.2 documentation. https://scikit-learn.org/1.5/modules/generated/sklearn.ensemble.HistGradientBoostingClassifier.html (accessed 08 Dec 2024).
64. Padhy, N., Dharmireddi, S., Padhy, D. K., Saikrishna, R. & Raju, K. S. Stock market prediction performance analysis by using machine learning regressor techniques. *Commun. Comput. Inform. Sci.* **1892 CCIS**, 39–50. https://doi.org/10.1007/978-3-031-56998-2_4 (2024).

65. Ahmed, Y. et al. Optimizing photocatalytic dye degradation: A machine learning and metaheuristic approach for predicting methylene blue in contaminated water. *Results Eng.* 103538. https://doi.org/10.1016/J.RINENG.2024.103538 (2024).
66. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17, 785–794. https://doi.org/10.1145/2939672.2939785/SUPPL_FILE/KDD2016_CHEN_BOOSTING_SYSTEM_01-ACM.MP4 (2016).
67. Tao, H. et al. An intelligent evolutionary extreme gradient boosting algorithm development for modeling scour depths under submerged weir. *Inf. Sci. (N. Y.)* **570**, 172–184. https://doi.org/10.1016/J.INS.2021.04.063 (2021).
68. Ramraj, S., Nishant, U., Sunil, R. & Shatadeep, B. Experimenting XGBoost algorithm for prediction and classifi cation of different datasets. *Int. J. Control Theory Appl.* **9**, 650–662 (2016).
69. Ke, G. et al. LightGBM: A highly efficient gradient boosting decision tree. https://github.com/Microsoft/LightGBM (accessed 19 Feb 2025).
70. Microsoft Welcome to LightGBM's documentation!—LightGBM 4.5.0 documentation. https://lightgbm.readthedocs.io/en/stable/ (accessed 08 Dec 2024).
71. Fan, J. et al. Light gradient boosting machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. *Agric. Water Manag.* **225**, 105758. https://doi.org/10.1016/J.AGWAT.2019.105758 (2019).
72. Karbassiyazdi, E. et al. XGBoost model as an efficient machine learning approach for PFAS removal: effects of material characteristics and operation conditions. *Environ. Res.* **215**, 114286. https://doi.org/10.1016/J.ENVRES.2022.114286 (2022).
73. Yu, L. et al. Simulation monitoring of tetracyclines in wastewater based on fluorescence image processing and machine learning classifier. *Sens. Actuators B Chem.* **385**, 133678. https://doi.org/10.1016/J.SNB.2023.133678 (2023).
74. Huang, G. et al. A hybrid data-driven framework for diagnosing contributing factors for soil heavy metal contaminations using machine learning and Spatial clustering analysis. *J. Hazard. Mater.* **437**, 129324. https://doi.org/10.1016/J.JHAZMAT.2022.129324 (2022).
75. Taşan, M., Demir, Y., Taşan, S. & Öztürk, E. Comparative analysis of different machine learning algorithms for predicting trace metal concentrations in soils under intensive paddy cultivation. *Comput. Electron. Agric.* **219**, 108772. https://doi.org/10.1016/J.COMPAG.2024.108772 (2024).
76. Moosavi, S. et al. A study on machine learning methods' application for dye adsorption prediction onto agricultural waste activated carbon. *Nanomaterials* **11** (10), 2734. https://doi.org/10.3390/NANO11102734/S1 (2021).
77. Nighojkar, A. et al. Using machine learning to predict the efficiency of biochar in pesticide remediation. *NPJ Sustain. Agric.* **1**(1), 1–7. https://doi.org/10.1038/s44264-023-00001-1 (2023).
78. Ibrahim, B., Ewusi, A. & Ahenkorah, I. Assessing the suitability of boosting machine-learning algorithms for classifying arsenic-contaminated waters: A novel model-explainable approach using shapley additive explanations. *Water*. **14**(21), 3509. https://doi.org/10.3390/W14213509 (2022).

Acknowledgements

The authors would like to thank the reviewers and editors for their comprehensive and constructive comments for improving the manuscript. In addition, Zaher Mundher Yaseen would like to thank the Civil and Environmental Engineering Department, King Fahd University of Petroleum & Minerals, Saudi Arabia.

Author contributions

Z.M.Y.: Conceptualization, project leader, supervision, writing and review, investigation. F.L.A.: Conceptualization, writing and review, software, modeling, analysis, investigation.

Declarations

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/s41598-025-96271-5.

Correspondence and requests for materials should be addressed to Z.M.Y.