Chemosphere 370 (2025) 144031

Contents lists available at ScienceDirect

Chemosphere

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

Probabilistic prediction of phosphate ion adsorption onto biochar materials using a large dataset and online deployment

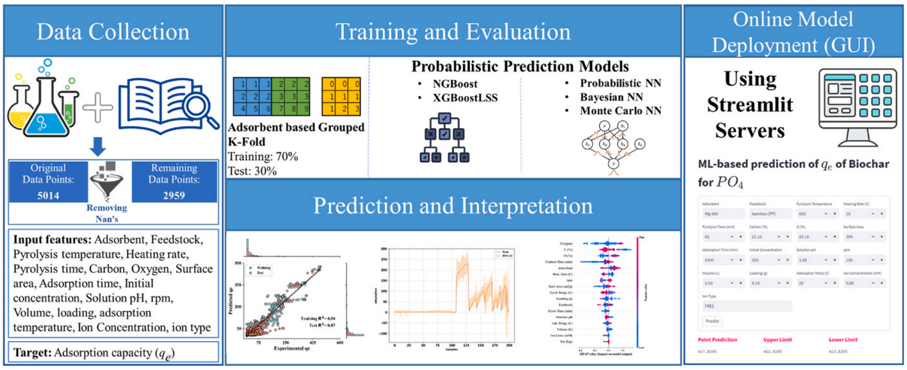
Sara Iftikhar [a,1], Rehan Ishtiaq [b,1], Nallain Zahra [a], Fazila Ruba [a], Sze-Mun Lam [c], Ather Abbas [d,**], Zeeshan Haider Jaffari [e,*]

[a] *Environmental Artificial Intelligence Research Group, Islamabad, Pakistan*
[b] *Department of Environmental Sciences, The University of Lahore, Lahore, 54590, Pakistan*
[c] *Faculty of Engineering and Green Technology, Universiti Tunku Abdul Rahman, Jalan Universiti, Bandar Barat, Perak, Kampar, 31900, Malaysia*
[d] *Physical Science and Engineering Division, 4700, King Abdullah University of Science and Technology, Thuwal, Mecca Province, Saudi Arabia*
[e] *Department of Civil, Environmental and Ocean Engineering, Stevens Institute of Technology, 1 Castle Point Terrace, Hoboken, NJ, 07030, USA*

HIGHLIGHTS

- 5014 data-points with 16 features on PO_4(III) adsorption onto 132 different biochars.
- Five ML models were used to predict the PO_4(III) adsorption capacity onto biochar.
- ML findings suggest that the NGBoostLSS model outperformed other models.
- SHAP analysis proposed that the C_i, C (%) and O (%) were the most controlling inputs.
- A web-based GUI was developed for wider application of developed ML models.

GRAPHICAL ABSTRACT


Data Collection
Original Data Points: 5014
Removing Nan's
Remaining Data Points: 2959
Input features: Adsorbent, Feedstock, Pyrolysis temperature, Heating rate, Pyrolysis time, Carbon, Oxygen, Surface area, Adsorption time, Initial concentration, Solution pH, rpm, Volume, loading, adsorption temperature, Ion Concentration, ion type
Target: Adsorption capacity (q_e)
Training and Evaluation
Adsorbent based Grouped K-Fold
Training: 70%
Test: 30%
Probabilistic Prediction Models
NGBoost
XGBoostLSS
Probabilistic NN
Bayesian NN
Monte Carlo NN
Prediction and Interpretation
Online Model Deployment (GUI)
Using Streamlit Servers
ML-based prediction of q_e of Biochar for PO_4
Point Prediction
Upper Limit
Lower Limit

ARTICLE INFO

Handling editor: Chang Min Park

Keywords:
Biochar
Adsorption
Probabilistic prediction
XGBoostLSS
PO_4(III)

ABSTRACT

Phosphate (PO_4(III)) contamination in water bodies poses significant environmental challenges, necessitating efficient and accurate methods to predict and optimize its removal. The current study addresses this issue by predicting the adsorption capacity of PO_4(III) ions onto biochar-based materials using five probabilistic machine learning models: eXtreme Gradient Boosting LSS (XGBoostLSS), Natural Gradient Boosting, Bayesian Neural Networks (NN), Probabilistic NN, and Monte-Carlo Dropout NN. Utilizing a dataset of 2952 data points with 16 inputs, XGBoostLSS demonstrated the highest R^2 (0.95) on new adsorbents. SHapely Additive exPlanations analysis showed that adsorption experimental conditions had the most significant impact (43%), followed by synthesis conditions (29%) and adsorbent characteristics (28%). Optimized conditions included an initial PO_4(III) concentration of 125 mg/L, carbon content of 11.5%, oxygen content of 23%, a contact time of 1440 min, a heating rate of 5 °C/min, 200 rpm, and a surface area of 410 m^2/g, using Ra-LDO adsorbent synthesized

[*] Corresponding author.
[**] Corresponding author.
E-mail addresses: ather.abbas@kaust.edu.sa (A. Abbas), jaffarizh@hotmail.com (Z.H. Jaffari).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.chemosphere.2024.144031
Received 7 October 2024; Received in revised form 15 December 2024; Accepted 25 December 2024
Available online 28 December 2024

from rape cabbage feedstock. This study developed and presented a practical online framework for predicting PO_4(III) removal onto biochar using a web-based graphical user interface.

1. Introduction

Human population growth and industrialization have increased the demand for phosphorus (P) for fertilizers and P-based chemicals like surfactants and pesticides (Davison et al., 2023). P is essential for cellular metabolism, energy transport, and gene synthesis in living organisms, yet economically viable P reserves are limited (Grossmann, 2024). Over 90% of P is sourced from non-renewable resources expected to be depleted within 100 years (Zhang et al., 2019). While crucial for plant growth, high P concentrations cause eutrophication, reducing dissolved oxygen and producing toxic compounds harmful to aquatic life (Zhang et al., 2025). To address P shortages and prevent eutrophication, efficient recovery from wastewater is essential (Yang et al., 2022). Struvite precipitation is commonly used for this purpose, though it can cause pipe clogging in treatment plants (Gao et al., 2023). Advanced methods like adsorption show promise but still face challenges in selectivity and separation, spurring the development of more effective materials (Foroutan et al., 2022; Iftikhar et al., 2023; Ishtiaq et al., 2024; Jaffari et al., 2021).

Biochar-based adsorbents, synthesized from pyrolysis of agricultural and food waste, offer superior pore structures, higher thermal stability, and heterogeneous surfaces (Foroutan et al., 2024). Despite these advantages, they are unsuitable for reclaiming phosphate (PO_4(III)) from livestock or municipal wastewater due to negatively charged ions (Zheng et al., 2023). Chemical modifications like post grafting and co-precipitation enhance the physiochemical properties (surface area, pore volume) and elemental composition (H%, C%, O%, N%) of these adsorbents (Qu et al., 2023). Additionally, experimental conditions such as contact time, ad sorbent loading, initial phosphorus concentration, solution volume, and pH significantly influence adsorption performance. Previous studies on PO_4(III) adsorption using biochar-based adsorbents focused on adsorption mechanisms and parameter optimization (Wan et al., 2019). However, the relationship between physiochemical characteristics, experimental conditions, and adsorption capacity (q_e) is complex and non-linear, making experimental optimization costly and time-consuming. A robust method is needed to understand these relationships efficiently.

Machine learning (ML) as a data-driven technique can model highly complex and non-linear relationship between input and output efficiently (Iftikhar et al., 2023; Jaffari et al., 2023; Kim et al., 2024). Due to this, several studies have adopted this route to simulate the adsorption phenomenon. Zhu et al. (2019) modeled the adsorption of six heavy metals using a random forest technique. Da et al. (2022) evaluated the performance of linear regression, random forest, and support vector machine for the prediction of the q_e of uranium on biochar in radioactive wastewater. In a previous study, tree-based ML models were applied for the prediction and optimization of the q_e of various emerging contaminants on biochar (Jaffari et al., 2023). Recently, advanced ML methods such as artificial neural networks (ANNs) and neuro-fuzzy inference system have also been adopted for simulation of the environmental purification (Iftikhar et al., 2023; Nasab et al., 2023). Turan et al. (2011) used ANNs for prediction of adsorption efficiency for the removal of Cu (II) from industrial leachate by pumice. Pauletto et al. (2020) employed ANNs for forecasting multicomponent adsorption of nimesulide and paracetamol. Despite the abundant use of ML modeling for adsorption process, no study has employed ML for simulating adsorption of PO_4(III) from wastewater. Moreover, all the previous studies for adsorption prediction have applied ML techniques for point prediction. This method involves predicting specific outcome values without quantifying the associated uncertainties. Consequently, these ML models fail to capture and represent the inherent uncertainty in their predictions, which is crucial for comprehensive risk assessment and decision-making (Charnock et al., 2022). A probabilistic prediction model on the other hand, can accommodate inherent variability and uncertainty of real-world systems, which provides a more realistic representation of forecasts in different scenarios (Dürr et al., 2020). This allows a more accurate assessment of risks and opportunities for the decision-makers.

To the best of our knowledge, this is the first report on the probabilistic ML modeling to predict PO_4(III) ion q_e on biochar surfaces. Two tree-based models, Natural Gradient Boosting (NGBoost) and eXtreme Gradient Boosting for location scale and shape (XGBoostLSS), and three DL models, probabilistic neural networks (PNN), Bayesian neural networks (BNN), and Monte-Carlo (MC) dropout, were developed using 2959 data points from 132 unique adsorbents. The dataset was split 70:30 by adsorbent types, with 92 adsorbents (2109 samples) for training and 40 adsorbents (823 samples) for testing. Results from the test data showed effective prediction performance for new adsorbents. Additionally, the SHapely Additive exPlanations (SHAP) method optimizes various physiochemical and experimental conditions for best adsorption capacity. The models are accessible through a user-friendly interface that provides point estimates and prediction bounds, informing users about prediction uncertainty. Practical implications and limitations are also detailed, aiding environmental engineers in wastewater treatment. The overall framework of this study is shown in Fig. 1.

2. Materials and methods

2.1. Data collection

In the data collection process, a comprehensive literature search on the adsorption of PO_4(III) onto biochar surfaces from 2000 to 2024 was conducted using the Web of Science and Scopus. This involved keyword searches like ("biochar" AND "adsorption" AND "PO_4(III)"). The initial search yielded over 227 works, which were meticulously screened, resulting in the shortlisting of 71 most relevant articles for data

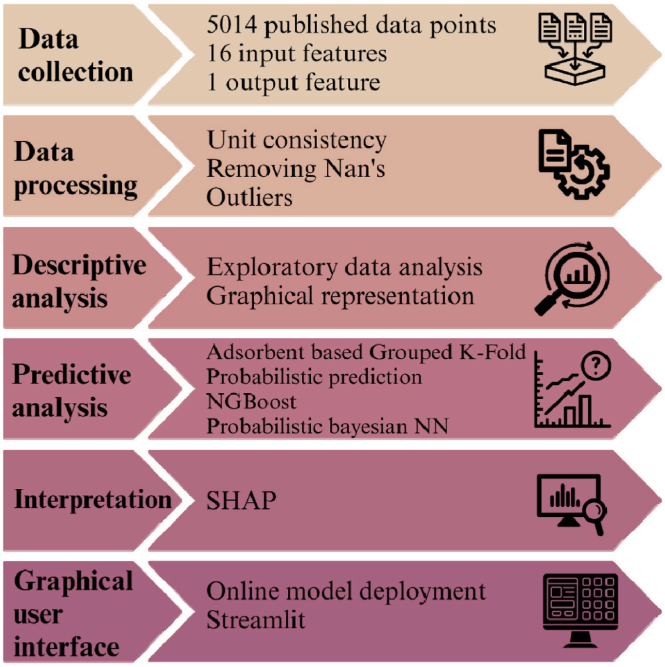


Fig. 1. Schematic illustration of the entire study.

collection. Throughout this process, certain assumptions and strategies were employed to ensure the relevance and reliability of the data gathered for the study.

During the screening process, all data were initially accepted, without any preconceived judgment or bias regarding data validity. For information not explicitly presented in tables or text, the web-based WebPlotDigitizer software (https://plotdigitizer.com/app) was utilized to extract the necessary data from figures. The primary criteria for assessing adsorption performance encompassed removal efficiency (%), q_e, and catalyst stability. Among these, q_e emerged as the most frequently reported parameter in the literature, indicating its prominence in evaluating the effectiveness of the adsorption processes studied.

Thus, data on q_e was collected and selected as the target output. The data on descriptors were broadly classified into three categories: (1) synthesis conditions, (2) adsorbent characteristics, and (3) adsorption experimental conditions. The data collection on the synthesis conditions included adsorbents, feedstock, pyrolysis temperature, and heating rate. Carbon (C), oxygen (O), and surface area of biochar materials estimated using Brunauer-Emmett-Teller method (S_{BET}) were used as adsorbent characteristics. All these three features can highly influence the adsorbents' surface properties, and thus, were selected in the dataset. Finally, the experimental conditions included solution pH, initial concentration, solution volume, adsorption temperature, and contact time. A total of 5014 data points were meticulously collected from the 71 papers, and careful screening was conducted to eliminate duplicate or multiple entries. The details of input features are presented in Table S1.

2.2. Data processing

The importance of data processing, particularly in the context of applying ML to adsorption, is crucial. Experimental data are often reported with differences in the adsorbent characterization techniques and adsorption experimental conditions, leading to the incomplete dataset. This variability is notably pronounced in wastewater treatment due to a lack of standardized data reporting guidelines and formats. This study encountered similar challenges during the data collection process. Out of 5014 collected datapoints, 46, 146, 380, 380, 915, 72, 65, 10, and 41 data points for pyrolysis temperature, pyrolysis time, S_{BET}, C, O, rpm, solution pH, contact time, and ion concentration, respectively, were not reported, nor was there any means to calculate them. The exact number of missing values for each feature and the distribution of missing values in complete dataset is illustrated in Fig. S1 and Fig. S2. Consequently, these data points were deemed redundant and were removed from the dataset. The data processing steps resulted in an effective dataset of 2959 data points (Fig. 2), hereafter referred to as the dataset. To put into context, approximately 40.98% of the raw data had to be discarded due to missing values, while reporting experimental data.

Three categorical features—"Adsorbents," "Feedstock," and "Ion type"—were converted to numerical values using label encoding, which assigns a unique number to each category (Rodriguez et al., 2018). To ensure the model's robustness across different adsorbents, the dataset was divided into training (92 adsorbents, 2109 data-points) and test sets (40 adsorbents, 823 data-points). This split helps evaluate the model's performance on unseen adsorbents.

2.3. Machine learning models and hyperparameter optimization

Herein, we selected five state-of-the-art probabilistic machine learning models—XGBoostLSS, NGBoost, BNN, PNN, and MC Dropout Neural Networks (MC Dropout NN)—to address the complex dynamics of PO_4(III) adsorption onto biochar surfaces. These models were specifically chosen for their robust capabilities in modeling complex, non-linear relationships and quantifying the inherent uncertainties typical of environmental processes. XGBoostLSS and NGBoost are particularly valued for their ability to model entire distributions of target variables, providing not just predictions but also insights into the likely range and variability of those predictions. Meanwhile, BNN, PNN, and MC Dropout NN extend these capabilities by directly modeling the uncertainty in the predictions, which is crucial for applications where decision-making depends on the reliability and confidence of the predicted outcomes.

Hyperparameter tuning was performed to optimize the performance of each model using a 5-fold cross-validation approach on the training set. The training data was systematically divided based on adsorbent groups to ensure a robust evaluation across unseen adsorbent types. This cross-validation process involved splitting the training dataset into five distinct subsets, training the model on four subsets, and using the fifth as the validation set, with each subset taking turns as the validation set

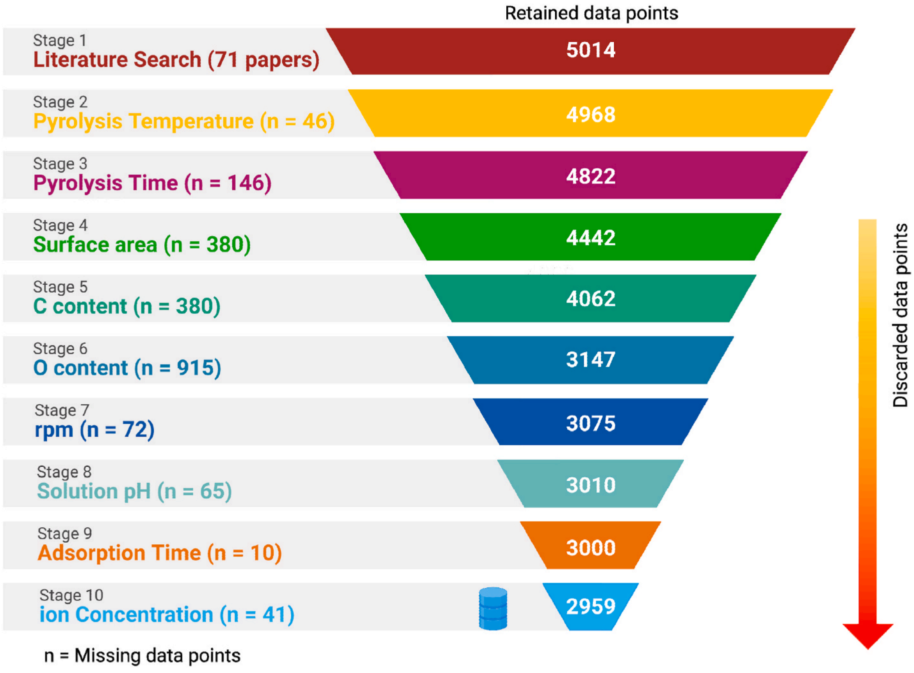


Fig. 2. Schematic presentation of data collection and processing pipeline devised in this study.

(Fig. S3). This approach not only fine-tunes the models for optimal performance but also reduces overfitting, thereby improving the models' generalizability to new, real-world data (Krstajic et al., 2014).

2.4. Neural network-based probabilistic models

NNs consist of interconnected layers that process inputs through non-linear transformations to predict outputs (Iftikhar et al., 2023). Each NN architecture in this study incorporates input, hidden, and output layers. The input layers consisted of 16 nodes corresponding to 16 features while the output layer was used to predict phosphate ion adsorption. The following equation can be used to represent the relationship between inputs (x), hidden layers (W and b) and outputs (y) in a NN:

$$\widetilde{y}(x) = \phi(\ldots\phi(xW_1 + b_1)W_2 + b_2)W_L + b_L \tag{1}$$

Building an NN requires definition of several hyperparameters such as number of hidden layers, type of activation (ϕ) and the size of W.

2.4.1. Probabilistic neural networks

Standard NN predictions lack uncertainty quantification which is crucial for adsorption processes due to inherent data variability (Hertel et al., 2023). This study employed PNNs which replace the output layer with a distribution-based layer, specifically a Gaussian distribution, to model the uncertainty (aleatoric) directly from the data (Guth et al., 2024). This method involves outputting the mean (μ) and standard deviation (σ) of the predicted adsorption capacity. Training of PNNs utilizes the negative log-likelihood (Eq. (2)) to align model predictions with observed distributions, enhancing reliability under varied operational conditions.

$$N_{LL} = -\sum_{i=1}^{n} \log\left(\mathrm{P}(y_i)\right) \tag{2}$$

2.4.2. Bayesian neural networks

The parameters (weights and biases) of a regular neural network are deterministic in nature and therefore consist of scaler values. Hence, the prediction of a regular NN is deterministic because its prediction is derived from these deterministic parameters. The training process of these deterministic NNs consists of learning the values of these parameters. A Bayesian Neural Network (BNN) is a kind of NN in which each parameter is a probability distribution rather than having a fixed value (Charnock et al., 2022). For gaussian distribution, therefore, each parameter in BNN is represented by mean and standard deviation of probability distribution. In this way, the BNNs estimate the uncertainty inherent in model parameters which is also called epistemic uncertainty (Hertel et al., 2023). The training process of a BNN consists of learning the distribution parameters for each weight and bias (Charnock et al., 2022). Therefore, the total number of trainable parameters in a BNN exceeds far more than those of a regular NN. This, however, makes BNN computationally intractable.

2.4.3. Monte Carlo dropout neural network

This technique involves switching off random subset of nodes in NN during training and evaluation (Scala et al., 2023). This leads to training ensemble of thin versions of NNs which share weights (Huang et al., 2023). Subsequently these NNs are evaluated multiple times for the same input thereby resulting in an ensemble of prediction results. These ensemble of prediction results can then be used to estimate uncertainty in the predictions (Gal and Ghahramani, 2016). The mean and standard deviation of these ensemble prediction results were then used for comparison with true efficiency data (q_e) and confidence interval estimation respectively.

2.5. Decision tree based probabilistic models

This section discusses two probabilistic ML models based on decision trees. Unlike previously discussed models, where the parameters of distribution are learned by neural networks, in these models, an ensemble of decision trees is used to represent distribution parameters.

2.5.1. Natural gradient boosting decision tree

NGBoost is a decision tree-based algorithm that forecasts the target variable's probability distribution to estimate predictive uncertainty (Duan et al., 2020). NGBoost learns the conditional probability distribution of y (qe) given x (input features), as $P_\theta(y\,|x)$, where P_θ stands for distribution parameters. The θ consists of mean (μ) and standard deviation (σ) of the target distribution (Gneiting and Raftery, 2007). The development of NGBoost requires definition of hyperparameters such as the depth of decision trees, number of decision trees or learning rate. The NGBoost model was trained through maximum likelihood which measures the degree of conformity between the estimated probability distribution of qe and the observed qe. It can be mathematically defined by the following equation

$$\mathscr{L}\,(\theta, y) = -\log P_\theta(y) \tag{3}$$

2.5.2. XGBoost for location scale and shape

XGBoostLSS employs state of the art gradient boosting algorithm (XGBoost) to model parameters of any probability distribution which can be anyone of the location, scale and shape, hence the name XGBoostLSS (März, 2022). In doing so, it combines the accuracy and speed of XGBoost with the estimation and prediction of distribution parameters. XGBoostLSS also allows modeling of target variables by a mixture of distributions (Ziel, 2022). This is based on the idea that the target variable is the result of complex multiple underlying processes. This is especially true for the case of q_e, which is a complex process and depends on a wide range of experimental and physical parameters which act independently as well as in conjunction with each other (Ziel, 2022). It has been shown that the XGBoostLSS is capable of modeling data generating process and provide predictions with high accuracy in many real world examples (Kirichenko and Lavrynenko, 2023).

2.6. Performance evaluation

The predictive performance of the ML models for point estimation was evaluated using root mean squared error (RMSE) and coefficient of determination (R^2). The RMSE considers magnitude of error between experimentally calculated values and ML predicted values. Therefore, higher RMSE signifies higher error between true and predicted values and vice versa. The R^2 on the other hand signifies goodness of fit. Its values vary between 0 and 1 with 0 being poor and 1 being the best fit. The probabilistic prediction performance of the ML models was evaluated using negative log-likelihood. This performance metric quantifies how well the predicted distribution fits the true data (Iftikhar et al., 2023; Jaffari et al., 2023). The equations to calculate RMSE and R^2 are given below.

$$RMSE = \sqrt{\frac{\left[\sum\limits_{i=1}^{n}(o_i - p_i)^2\right]}{n}}, \tag{4}$$

$$R^2 = 1 - \frac{\sum (p_i - o_i)^2\,(p_i - \overline{p})}{\sum (o_i - \overline{o})^2 \sum (p_i - \overline{p})^2}, \tag{5}$$

where, o_i and p_i are the calculated and predicted adsorption capacities, respectively. n is the number of samples and $\overline{p}$ is the average of all predicted adsorption capacities.

3. Results and discussion

3.1. Descriptive statistics

A thorough review of 16 input features provides an overview of biochar synthesis, characteristics, and environmental factors influencing PO_4(III) removal from aqueous solutions (Fig. 3). Key synthesis parameters include pyrolysis temperature (550 °C–800 °C, mean around 750 °C), heating rate (mostly below 10 °C), and heating time (30–120 min). Variations in these parameters across 44 feedstocks resulted in significant alterations in 132 synthesized biochars. The S_{BET} ranges from 10 m^2/g to over 900 m^2/g, with a mean of 150 m^2/g, indicating substantial differences in biochar structure. C% and O% content also show vast distributions, highlighting diversity among biochars. The distribution of data points among various feedstocks and synthesized biochars is visually represented in the form of the pi-charts, as illustrated in Fig. S4. The adsorption contact time, another influential feature, ranges from 0 to 4880 min, showing extensive investigation periods. Biochar loading is mostly below 0.1 g, and initial PO_4(III) concentration ranges from 1 to 100 mg/L. Initial solution pH is around 7.0, reflecting PO_4(III)'s amphoteric nature (Jaffari et al., 2023). Aqueous suspension volume is mainly below 0.1 L, with a solution temperature at 25 °C. Rotation speeds are concentrated at 160 rpm and 180 rpm. Over 90% of experiments were performed without anion. This comprehensive overview emphasizes the diverse and meticulous experimental conditions, contributing to a nuanced understanding of adsorption processes.

The Pearson correlation (PC) matrix illustrates the relationships among the numerical input features and their connection with the target output (q_e) (Fig. 4). Notably, S_{BET} values exhibit a positive correlation with pyrolysis time, heating rate and surface functional groups, indicating a substantial interdependence between these factors. Interestingly, only three input features—contact time, PO_4(III) concentration, and solution pH—demonstrate a slight positive correlation with q_e.

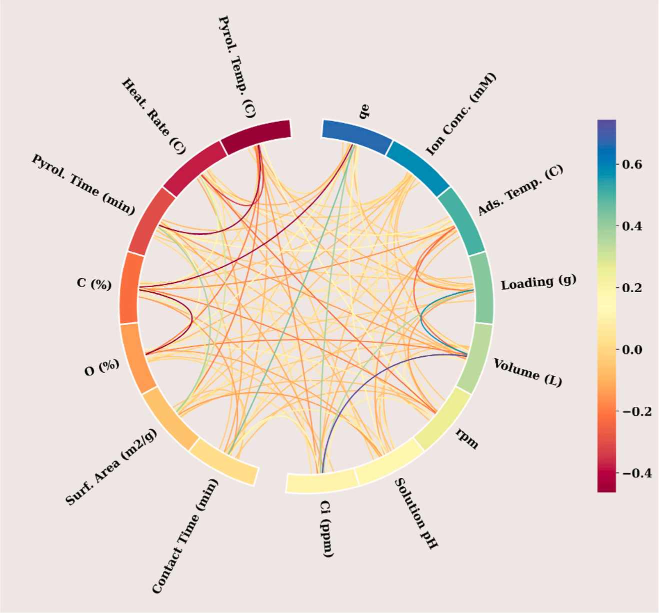


Fig. 4. Pearson correlation metrics.

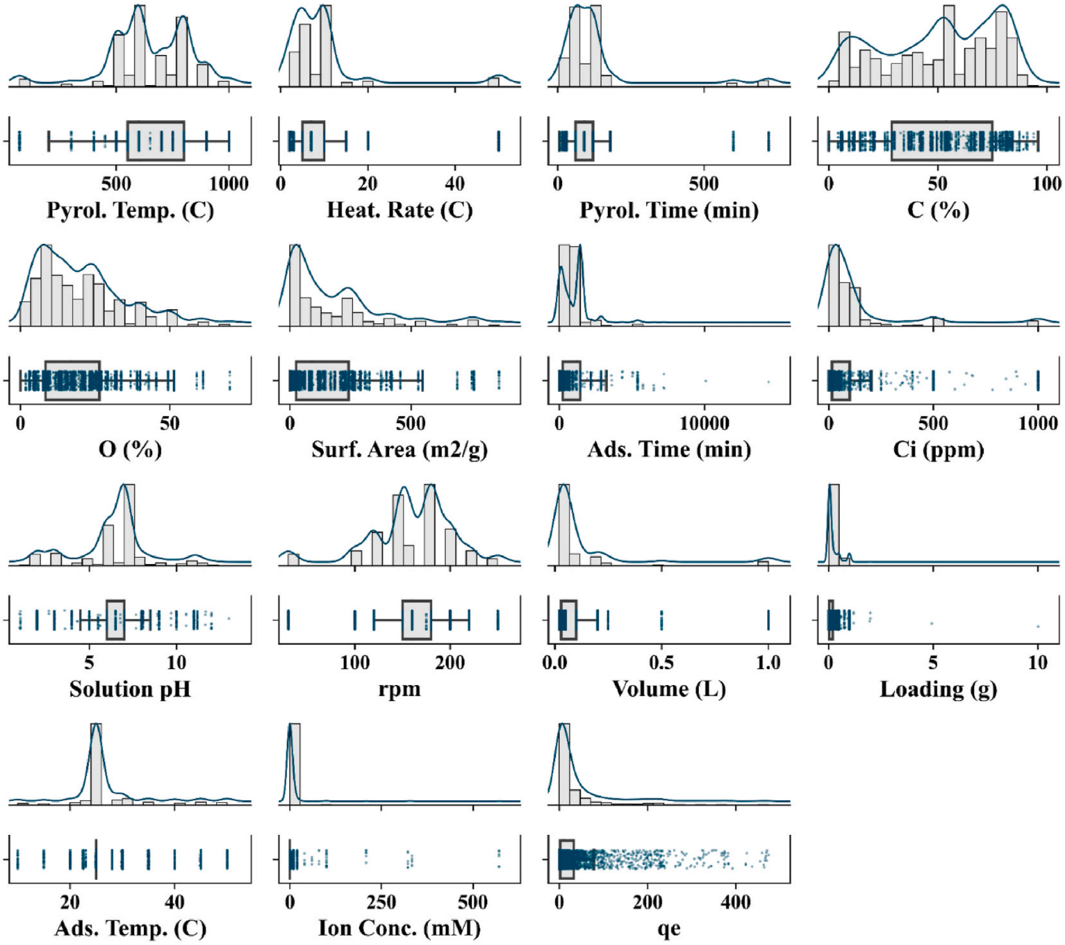


Fig. 3. Distribution of inputs and outputs for the biochar synthesis, characterizations and adsorption experimental conditions using histograms and boxplots.

Conversely, no notable correlation was observed between the residual input features and their correlation with q_e. The relatively weaker correlations among most of the input features suggest the importance of maintaining all of them during the model-building process (Hastie et al., 2009). Each feature is likely to contribute individually to the model's performance, emphasizing the holistic consideration of all input variables for a more comprehensive understanding and accurate prediction of the adsorption process (Hastie et al., 2009).

3.2. Model development and testing

Three DL probabilistic models (PNN, BNN, and MC dropout) and two tree-based probabilistic models (NGBoost and XGBoostLSS) were developed to accurately predict the q_e values. To get the best prediction performance on the test data, the hyperparameters of all four models were first optimized using the Bayesian optimization algorithm. Table S2 contains a list of the hyperparameter names and optimized values used in the construction of these models. The convergence plots of hyperparameter optimization with error bars obtained from k-fold cross validation are depicted in Figs. S5–S9. A test dataset with novel biochar adsorbents was used to evaluate the predictive performance of five models that had their hyperparameters optimized. Fig. 5(a)–(e) presents the regression plots of predicted q_e values along with the corresponding experimental q_e values for the five selected models. A 70:30 data split is indicated by ridge plots around the x- and y-axes, with light-pink and light sky-blue bar plots representing the training and test data distributions, respectively. The perfect regression line is represented by the black solid lines ($y = x$). The presence of data points in this direction indicates that the predicted and experimental values are closely aligned. Upon assessing the performance of the models on the training sets, the NGBoost and XGBoostLSS exhibited a higher training R^2 of 0.97 and 0.94, respectively, followed by the PNN (0.82), BNN (0.72), and MC dropout (0.61) models. Similarly, the XGBoostLSS model exhibited a significantly lower RMSE of 15.1 mg/g compared to the other models. When predicting the test set, XGBoostLSS outperformed all other models, achieving the highest R^2 of 0.95, followed by NGBoost (0.91), PNN (0.82), BNN (0.72), and MC Dropout (0.59). Furthermore, NGBoost demonstrated a notably smaller RMSE of 20.4 mg/g on the test set, outperforming the PNN, BNN, and MC Dropout models (Table 1). Notably, in Fig. 5(a and b), data points align closely with the black solid line ($y = x$) for the XGBoostLSS and NGBoost model, while fitting lines for the other models deviate significantly from this line. The XGBoostLSS model exhibited a slightly higher R^2 value than the NGBoost model for the test dataset. However, the RMSE values of the NGBoost model were substantially lower than those of the XGBoostLSS model for both the training and test datasets. Therefore, it is imperative to conduct further statistical investigations into the performance of the models before determining the best model.

Fig. S10 provides a statistical summary of the prediction accuracy for each of the five probabilistic models used in this investigation. Using a Taylor plot, this figure compares the models' performances during training and validation (Taylor, 2001). When comparing the models' ability to match observations to predicted results, Taylor diagrams provide a concise statistical summary that takes variance, correlation, and root-mean-square difference into account (Taylor, 2001). All models have test correlation coefficients between 0.98 and 0.6, and standard deviations between 10 and 65. Notably, test values for NGBoost and XGBoostLSS are 0.96 and 0.94, respectively, while training correlation coefficients for both models are above 0.96. The findings suggest that the XGBoostLSS model stands out with the best test performance, boasting R^2 values of 0.89 each (Table 1).

The accuracy of these models was further scrutinized through the

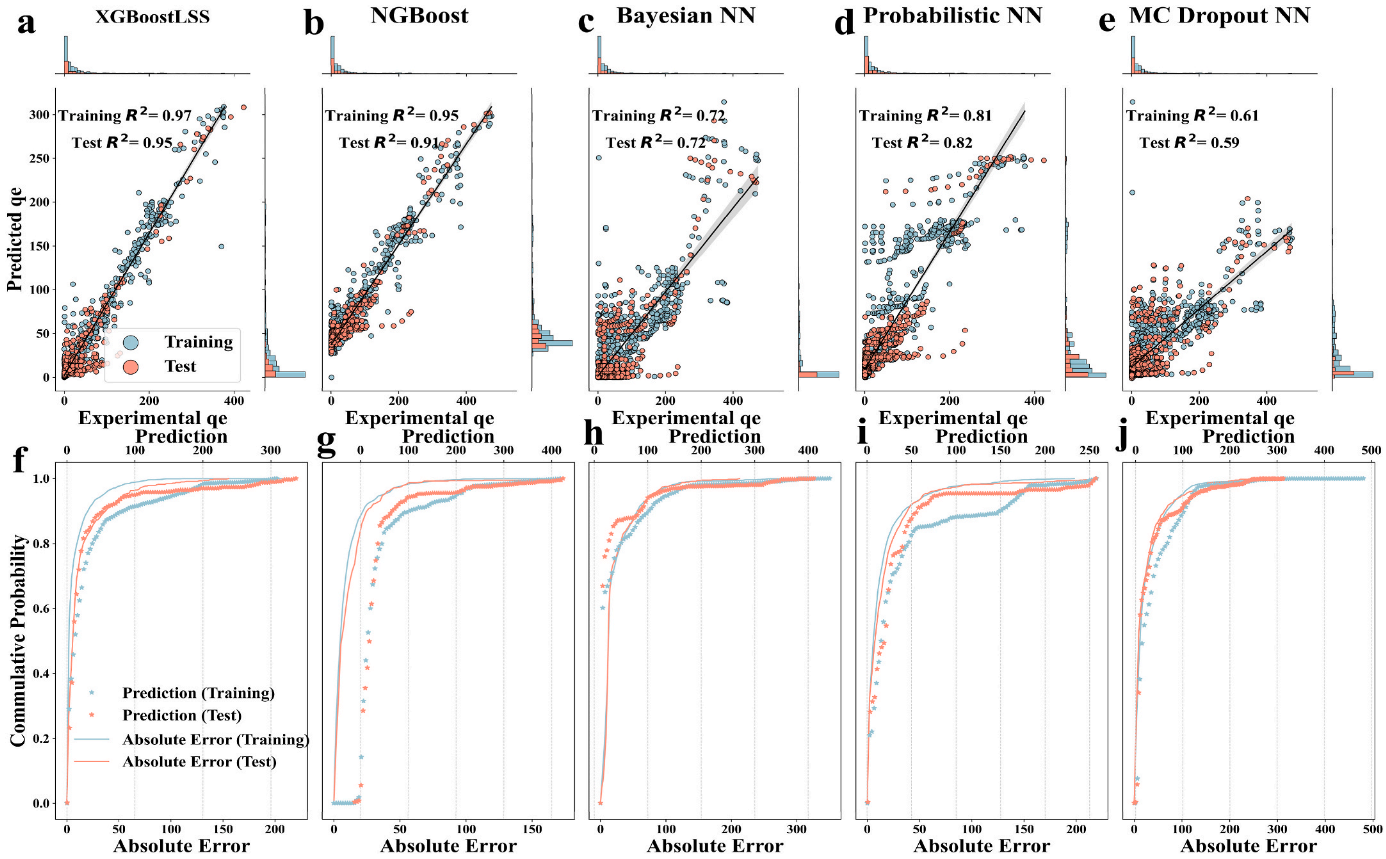


Fig. 5. Regression plots and data distributions of experimental vs predicted q_e for (a) XGBoostLSS, (b) NGBoost, (c) BNN, (d) PNN and (e) MC dropout NN, and Cumulative probability plot for (f) XGBoostLSS, (g) NGBoost, (h) BNN, (i) PNN and (j) MC dropout NN models.

Table 1
Statistical analysis (R^2 and RMSE) of NGBoost, PNN, BNN and MC dropout models using training and test dataset.

Model	R^2				RMSE			
	Train	Std	Test	Std	Train	Std	Test	Std
NGBoost	0.94	0.05	0.91	0.03	16.8	3.7	20.4	9.7
XGBoostLSS	0.97	0.08	0.95	0.05	15.1	4.73	24.9	11.7
PNN	0.82	0.3	0.82	0.37	26.7	14.5	31.0	15.4
BNN	0.72	0.27	0.72	0.35	32.7	17.2	35.6	19.4
MC dropout	0.61	0.45	0.59	0.48	45.2	20.5	42.9	24.6

examination of residual plots (Figs. S11(a)–(e)). A model is deemed more precise, when residual errors are randomly distributed without discernible patterns or trends (Jaffari et al., 2023). In accordance with these principles, the XGBoostLSS model's residual errors, as depicted in Fig. S11(a), exhibit a concentration around the zero line without any pattern, indicating well-distributed errors with maximum residuals. Contrastingly, the MC dropout model presents a trend and more scattered residual errors from the black base line, as evident in Fig. S11(e). This dispersion suggests a relatively lower prediction performance for the MC dropout model compared to the XGBoostLSS and other models. Residual analysis provides valuable insights into the robustness and accuracy of the models, with the XGBoostLSS model demonstrating more favorable characteristics in this regard.

Finally, to compare the effectiveness of the ML models, the cumulative probability of the q_e values against the absolute error was looked at. If a model's values remain parallel to the y-axis, it is thought to be more accurate (Bandler et al., 1994). As can be seen in Fig. 5(f–j), the three NN models have more than 25% of predicted data points and absolute errors greater than 25, whereas over 85% of the predicted data points and absolute errors for the XGBoostLSS model have an absolute error of less than 15. Regression plots, performance metrics, residual error plots, Taylor plots, and cumulative probability analysis all show that the XGBoostLSS model is the most exact and accurate of the models under consideration. As such, it is selected for additional post-processing analysis.

Overall, XGBoostLSS outperformed other models like NGBoost, BNN, PNN, and MC Dropout NN in predicting the adsorption capacity of PO_4(III) onto biochar-based materials, primarily due to its ability to model multiple distributional parameters and capture complex feature interactions. Overall, both tree-based models outperformed their NN counterparts because of their robustness against overfitting and effective handling of non-linear relationships which makes them particularly suitable for chemical processes characterized by stochastic variability. It has been extensively studied that for tabular datasets, treen-based algorithms can outperform their neural network counterparts (Abdi et al., 2022; Jaffari et al., 2023; Kim et al., 2024). In contrast, neural network-based models such as BNN and PNN often require extensive computational resources and are sensitive to training data adequacy and model assumptions, which can lead to overfitting or inadequate uncertainty quantification (Pombo et al., 2017). MC-Dropout NN, while useful for estimating prediction uncertainty, depends heavily on correct dropout rate settings, which can affect prediction reliability (Mae et al., 2021). Additionally, XGBoostLSS's capability to provide interpretable insights into feature importance through SHAP values offers valuable guidance on influencing factors in adsorption processes, further establishing its practical applicability and superiority in environmental modeling contexts.

3.3. Confidence interval

Fig. 6 analyzes confidence intervals for individual data points. A smaller area under the bell-shaped curve indicates greater confidence, while a larger area suggests higher error probability. Data point 764 shows the highest error potential, whereas most points display small areas, indicating strong model confidence (Dürr et al., 2020). Its confidence interval, derived from the target feature's probability distribution, details behavioral variability and outliers. Earlier ML models for wastewater treatment offered less accurate point estimates by not

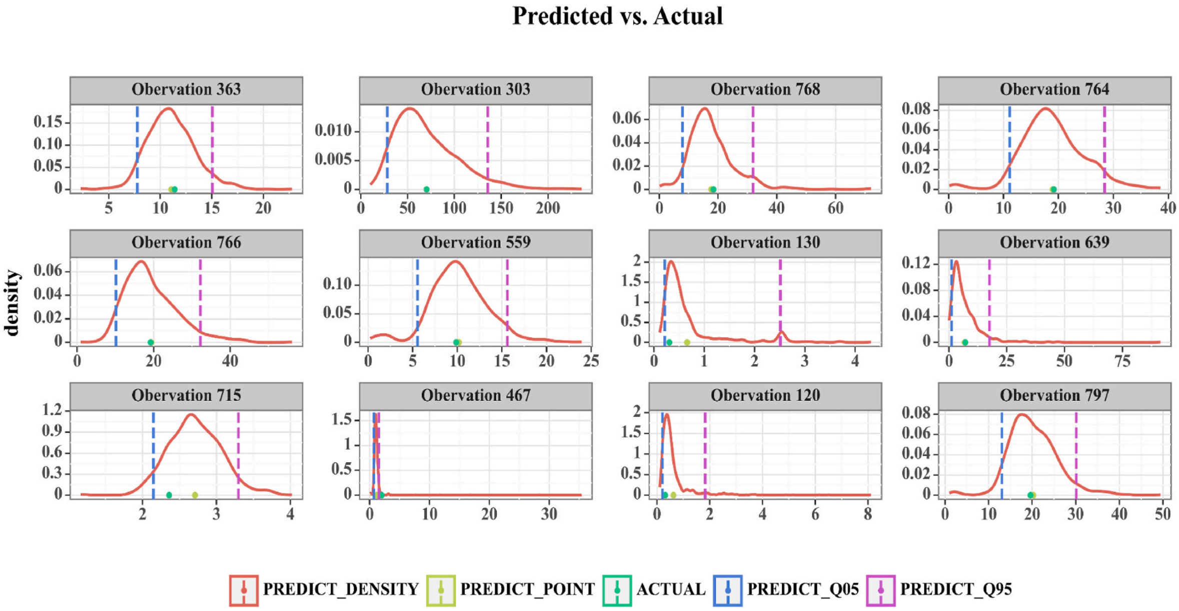


Fig. 6. Confidence intervals showing high and low areas under the bell-shaped curve. Herein, the observation number represent the number of data point in the test dataset.

considering uncertainty (Iftikhar et al., 2023; Kim et al., 2024). Confidence intervals thus enhance the model's reliability and informativeness.

Fig. S12 shows confidence intervals for both training and test sets for all five ML models. Furthermore, for further comparison of probabilistic performance of five models, standard deviations of R^2 and RMSE are given in Table 1. Fig. S12 and Table 1 indicates that both XGBoostLSS and NGBoost model performs better than rest of the models for confidence intervals. This is because the confidence intervals of these tree-based models are smaller than those of other models for both training and test sets. One reason for this could be that these models optimize the parameters of the predicted distribution itself, leading to tighter and more reliable confidence intervals (Duan et al., 2020; März, 2019). This characteristic makes XGBoostLSS particularly effective in scenarios where understanding the uncertainty associated with predictions is crucial, enhancing its performance over other models when it comes to detailed risk assessments and decision-making processes based on model output.

3.4. Feature importance analysis

SHAP feature importance analysis was employed to investigate the impact of each feature, and the results are shown in Fig. 7. Notably, adsorption experimental conditions emerged as the most crucial, constituting 43% of all features (Fig. 7(a)). Within the adsorption experimental conditions, the C_i of PO_4(III) exhibited the highest importance weight at +0.047, representing 31.73% of the overall importance of adsorption experimental conditions (Fig. S13). While the important weights of contact time (25.92%), rpm (16.81%), adsorbent loading (12.15%), solution pH (8.37%), adsorption temp (7.29%), solution volume (5.46%), ion concentration (2.56%) and the ion type (1.82%) were relatively lower than C_i of PO_4(III), yet they still contributed significantly. With an importance weight of 41.75%, the C (%) was the most significant property of the biochar and, all things considered, the second most influential feature for the adsorption of PO_4(III) from aqueous solution. Besides, the O and S_{BET} (m^2/g) also carried substantial importance weights of 35.66% and 22.58%, respectively in this class. The adsorbent synthesis conditions were the second most important class of input features collectively accounted for 29% of all features. Among these, various adsorbent materials emerged as the most influential factor with 25.18% importance. Apart from the adsorbent materials, heating rate (°C), pyrolysis temp (°C) and pyrolysis time (min) also contribute to the model prediction. These results suggested that all sixteen input features play a dominant role in achieving a higher adsorption performance.

To gain further insights into the effect of each feature on biochar adsorption capacity, the SHAP values of the features were further analyzed (Fig. 7(b)). Notably, features such as C_i of PO_4(III), C contents, O contents, contact time, various adsorbent materials, heating rate, rpm, S_{BET}, pyrolysis time, loading, feedstock and pyrolysis time (min) exhibited the most significant influence on the q_e of biochar (Fig. 7(b)). The increase of heating rate, pyrolysis time, and pyrolysis temperature suggested the creation of new surface active sites, thereby enhancing S_{BET} values and consequently improving q_e values (Zhao et al., 2023). Furthermore, the increase in O content indicated the enrichment of –OH surface functional groups, leading to enhanced q_e values (Palansooriya et al., 2022; Zhu et al., 2022b). In contrast, the rise in C content had two opposing effects on the adsorption capacity, suggesting that an excessively high C content might impede the adsorption of PO_4(III). Similarly, the increase in adsorption contact time exhibited a positive influence on the q_e values.

3.5. SHAP dependence of vital input features

The relationship between the nine most significant features and the q_e of biochar for PO_4(III) was examined using a one-dimensional SHAP dependence plot and SHAP feature importance analysis (Fig. 8). The *x*-axis displays experimental values, and light green columns show data distributions. SHAP values can be positive or negative, indicating positive or negative influences on the predicted target value, respectively. Higher SHAP values suggest higher predicted q_e values (Jaffari et al., 2024; Yao et al., 2023). By analyzing SHAP values, one can understand how changes in features affect the predicted q_e values.

The C_i of PO_4(III) emerged as the most critical input feature (Fig. 7),

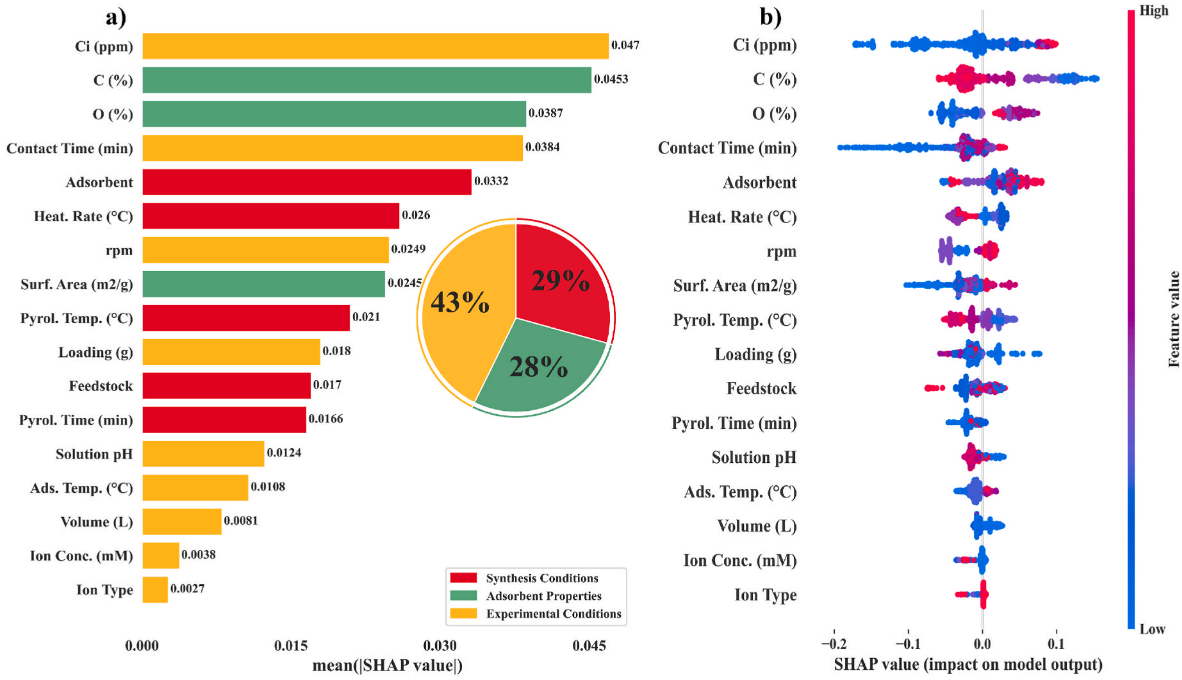


Fig. 7. SHAP feature importance analysis. (a) The input feature importance ranking according to mean|SHAP value| for q_e values. The inset shows the total percentage (%) of the categorical feature importance for the input features; (b) the global explanation of q_e values is provided by the positive and negative impacts of each input feature.

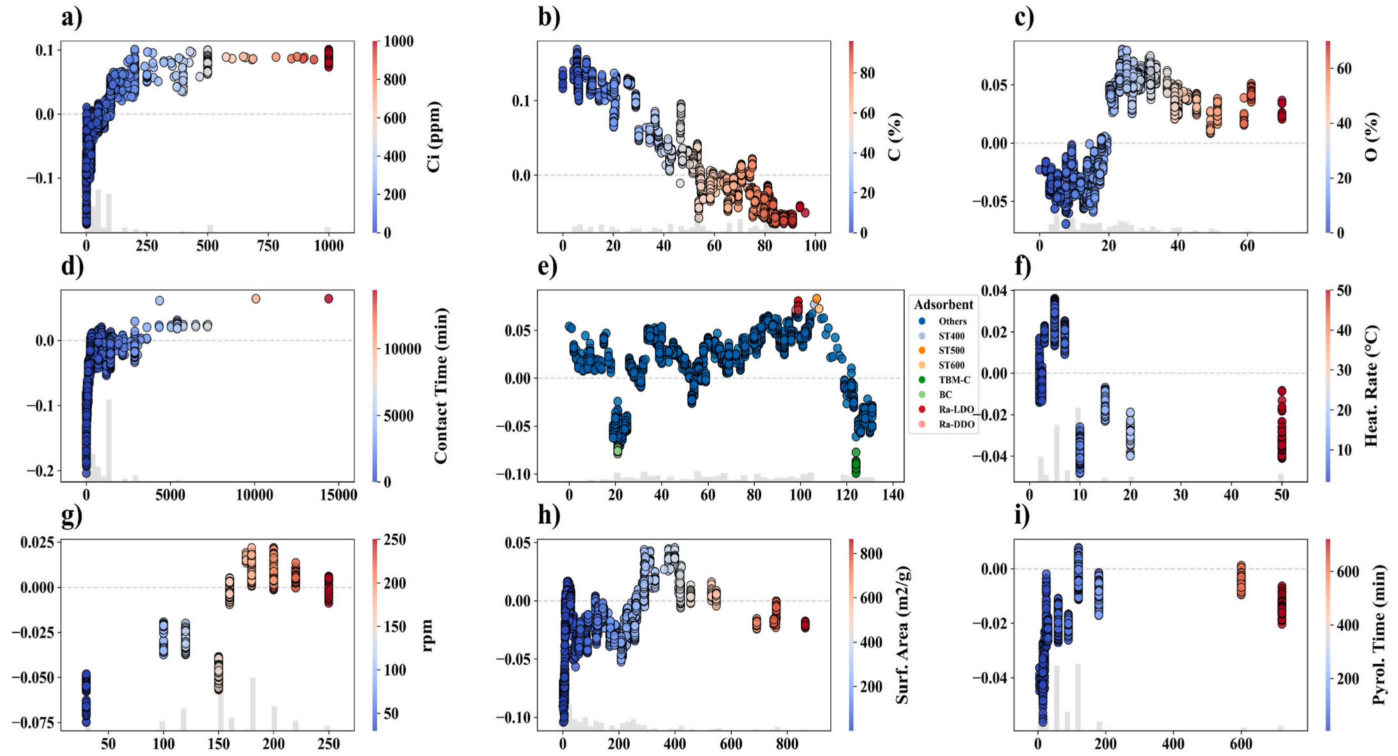


Fig. 8. SHAP feature dependency plots of (a) initial concentration, (b) C (%), (c) O (%), (d) Contact time (min), (e) Various adsorbents, (f) Heating rate (^{O}C/min), (g) rpm, (h) S_{BET} (m^2/g) and (i) Pyrolysis Time (min). The x-axis indicates value of feature and y-axis its indicate SHAP value.

significantly contributing to higher PO_4(III) adsorption. Fig. 8(a) shows a rapid increase in SHAP dependency values as C_i increases from 0 to 700 ppm, followed by a slower enhancement up to 1000 ppm. This trend is likely due to the availability of more active sites for adsorption as PO_4(III) concentration rises from 0 to 700 ppm, leading to a sharp increase in SHAP values. As the concentration continues to increase, adsorption sites may start to saturate, causing a slower rise in SHAP values (Xie et al., 2024). Additionally, competition among PO_4(III) ions for adsorption sites could further contribute to the diminished increase in SHAP dependency values as saturation becomes more pronounced.

After C_i of PO_4(III), the C (%) was the second most important input for the adsorption process. As shown in Fig. 8(b), the SHAP dependency value decreases significantly with increasing C% values. A higher C% could alter the surface properties of biochar materials, influencing their interaction with PO_4(III) ions and potentially leading to a decline in SHAP dependency values (Eduah et al., 2020). Moreover, an increased C % might introduce additional adsorption sites or alternative sites with lower affinity for PO_4(III) ions, likely due to trade-offs between C% content and other critical parameters, such as oxygen-based functional groups (Li et al., 2016). This shift may reduce the relative importance of C% as a key factor in the adsorption process, as reflected by the decreased SHAP dependency values.

O (%) was the third most influential input feature, which controls the predicted q_e value on the biochar surface. Fig. 8(c) illustrates how the SHAP dependency values increased as the O (%) value increased, reaching a maximum of 35%. Positive PO_4(III) ions can be electrostatically attracted to the excess negatively charged oxygen-based surface functional groups (−O, –OH) on the surface of biochar (Lam et al., 2021; Sadegh et al., 2023). However, if the C (%) is increased beyond 35%, the SHAP dependence value decreases, indicating that the q_e value is negatively impacted. The reason for the decreasing trend in q_e prediction could be attributed to the excess O-based surface functional groups of biochar, which could lead to the formation of a water film barrier that impeded the availability of interior active sites (Wan et al., 2019).

The model identified contact time (min) as the fourth most influential input feature, as shown in Fig. 8(d). The SHAP value increased rapidly during the first 720 min, indicating quick adsorption, then slowed and reached equilibrium around 1440 min. This early increase is due to higher S_{BET} and more active sites on the biochar, facilitating immediate PO_4(III) ion capture (Jaffari et al., 2023). Over time, access to vacant sites decreased as they became saturated, leading to equilibrium. When PO_4(III) ion concentration balances with the interface, adsorption equilibrium is achieved (Iftikhar et al., 2023). These findings suggest 1440 min as the ideal contact time to reach equilibrium.

The type of adsorbent was the fifth most influential feature in this study. The SHAP dependency plot of all 132 adsorbent materials is shown in Fig. S14. Fig. 8(e) shows the SHAP dependence plot of the three most effective and three least effective biochar adsorbents towards the PO_4(III) removal from wastewater. The rape (Ra) combined with layered double oxide (Ra-LDO) biochar displayed the largest partial dependence of all the applied biochar, indicating that it has the best q_e for the removal of PO_4(III). The interaction between the enriched surface functional groups, higher S_{BET}, pore volume, and PO_4(III) ions may be the cause of Ra-LDO's higher q_e (Chien et al., 2022; Zhang et al., 2019). Conversely, chitosan modified (TBM-C) biochar that was pyrolyzed at 500 °C had the lowest PDP value. Reduced porosity and S_{BET} could be the result of highly crystalline structures forming at higher pyrolysis temperatures, which could explain the lower PDP value (Cui et al., 2016). The adsorption performance decline may also be attributed to NaOH treatment since alkali-treated biochar can generate new surface functional groups like hydroxyl (OH) and carboxyl (COOH). There may be fewer active adsorption sites available overall as a result of these groups' competition with other functional groups for adsorption sites on the biochar surface (Yang et al., 2022). In a similar vein, the other alkaline-treated biochars considered in this investigation, including Alkali-SCG, PB600, and NaOH-SCW, also displayed comparatively lower PDP values.

Fig. 8(f) displays the SHAP dependency plot of heating rate (^{O}C) and its impact on the q_e value. The SHAP values show that lower heating rates have a positive impact on the q_e value. The slower heating rates could lead to the better exfoliation of the biochar material, resulting in a higher S_{BET} with increased accessibility. Besides, the pyrolysis at lower heating rates can potentially create additional active sites on the surface of the biochar. These extra active sites may offer more opportunities for the adsorption of PO_4(III) ions, contributing to an increased q_e values (Li et al., 2016). Conversely, the SHAP plot indicates that higher heating rates have a negative impact on the q_e value. It can be attributed to the reason that faster heating rates may lead to more rapid and intense thermal decomposition of the biochar. This could result in the loss of functional groups or the creation of less porous and less reactive biochar structures, diminishing its q_e values (Kumar et al., 2022).

The SHAP feature importance analysis suggests that rpm is the seventh most important feature (Fig. 7). In Fig. 8(g), insights into the SHAP dependence plot of rpm and its impact on adsorption performance are presented. Remarkably, as the rpm increased from 25 to 200, the SHAP dependence value transitioned from negative to positive, reaching its peak at 200 rpm. This highest SHAP dependence value at 200 rpm could be attributed to the efficient mixing of biochar adsorbent material in the solution, which improves the contact between PO_4(III) ions and biochar, ultimately leading to enhanced adsorption performance. Upon further increasing the agitation speed beyond 200 rpm, the SHAP dependence plot again exhibited a declining trend, indicating adverse effects on adsorption performance. This could be credited to the fact that excessive agitation could reduce the available time for PO_4(III) ions to be adsorbed on the surface of biochar.

S_{BET} was identified as the seventh important input feature, influencing the prediction of the q_e of biochar materials. In Fig. 8(h), an increase in the SHAP dependency values was observed with an increase in S_{BET} values. This trend suggests that a surplus of accessible adsorption sites can enhance the interaction between PO_4(III) ions and the biochar surface, resulting in higher q_e values. The noteworthy impact of S_{BET} on the adsorption of PO_4(III) ions on the biochar surface aligns with the trends observed in the PC metrics (Fig. 4). However, a slight decline in SHAP dependency value is noted beyond a certain point. This decline implies that the excessive pursuit of a higher S_{BET} should be avoided during the synthesis of biochar materials for the removal of PO_4(III) ions from wastewater (Zhu et al., 2022a). Finding the optimal balance in the specific surface area is crucial, as excessively high values may not necessarily translate to improve adsorption performance and even can be counterproductive.

The pyrolysis time (min) emerged as the eighth most influential feature in this study. The SHAP dependence analysis conducted on pyrolysis time (min) revealed intriguing insights into its influence on adsorption behavior (Fig. 8(i)). The findings indicate a notable improvement in the SHAP dependence plot with increasing pyrolysis time, particularly from 0 min to 120 min. Notably, the highest SHAP dependence values were recorded at 120 min, suggesting an optimal duration for the pyrolysis of feedstock materials. However, beyond this optimal point, the SHAP dependence values exhibited a decline, albeit to a lesser extent, indicating diminishing returns or a plateau effect on adsorption performance with further increases in pyrolysis time. These results underscore the importance of optimizing pyrolysis time to maximize the active sites and pores on the surface of biochar materials.

3.6. Graphical user interface design

Generally, some basic knowledge of programming is necessary to utilize ML modeling. Learning these skills and writing code on computer terminal can be time-consuming and may not be suitable for widespread application of these trained models. To address this issue and make the developed models more accessible to the wider scientific community, a user-friendly tool with a graphical user interface (GUI) was developed that runs online using streamlit engine. This tool is based on all the five probabilistic ML models developed in this study. With this online GUI tool, predicting aqueous adsorption on biochar-based adsorbents becomes straightforward. The ML-based GUI application features a single-screen interface and provides functionalities such as importing experimental data and exporting images and results. The GUI for the 17 selected input parameters was implemented to streamline adsorption predictions for PO_4(III) removal, thereby making the utilization of trained ML models much more convenient and user-friendly, as shown in Fig. S15.

The online web-based GUI is hosted at https://phosphateefficiencypredictionwithai.streamlit.app/and can be used by the users from around the globe. The first step in using GUI is importing the data using a csv or excel sheet consisting of 17 features as columns. The user can also manually impute the values of all 17 input features which include adsorbent, Feedstock, Pyrolysis temperature, Heating rate, Pyrolysis time, Carbon (%), O (%), surface area, adsorption time, Initial concentration, solution pH, rpm, volume, loading, adsorption temperature, ion concentration and ion type. After data import, the user can retrain the ML model on the input data or get the prediction results including point prediction, upper limit and lower limit. and its source code is available via an open source readthedocs link (https://po4-removal-ml.readthedocs.io/).

3.7. Applications and drawbacks of the current study

The XGBoostLSS model developed in this work serves two key purposes. Firstly, it enhances the understanding of how biochar-based adsorbents remove PO_4(III) ions. The adsorption process involves various mechanisms influenced by factors such as adsorbent synthesis conditions (e.g., pyrolysis temperature and time), biochar characteristics (e.g., surface functional groups), and experimental conditions (e.g., initial concentration, contact time, solution pH). Unlike hypothesis models that struggle with uncertainty due to limited measurements, the XGBoostLSS model offers a thorough understanding of these processes. Additionally, it allows for accurate predictions of biochar's adsorption performance without conducting experiments. This is particularly beneficial for wastewater treatment, where biochar materials are used. Environmental engineers can utilize these probabilistic models and their GUI to predict or visualize the q_e value of biochar, saving time and resources. The model's predictive capabilities also assist in fine-tuning experimental conditions to optimize adsorption processes, suggesting that biochar composition can be enhanced by adjusting pyrolysis time and temperature to improve adsorption capacity.

Despite the strong performance of the developed probabilistic ML models, several limitations need further investigation. First, the dataset could be expanded to include different surface functional groups, as these can significantly influence the q_e values, particularly when biochar is modified. The relationship between the ML algorithm and dataset requires iterative optimization, meaning that identifying the optimal model involves trial and error. This process should be carefully carried out on the adsorption dataset before deep training and fine-tuning the final model. Additionally, while several probabilistic ML models were compared, exploring simpler tree-based models could offer valuable insights and potentially improve model performance. Data dimension reduction is also strongly recommended to enhance prediction accuracy and reduce computational resource requirements. By reducing the dimensionality of the dataset, we can improve the model's efficiency and simplify its application without sacrificing performance. From a practical perspective, the impact of highly toxic competitor inorganics (e.g., heavy metal ions, NO_3) and organics (e.g., humic substances, proteins, saccharides) on the biochar's active sites with PO_4(III) should be considered. These factors are crucial in real-world wastewater treatment scenarios, where their presence may influence the adsorption process. Addressing these challenges, including dataset biases and model scalability, will lead to more accurate predictions and improve the applicability of biochar-based adsorbents in complex, real-world environmental applications.

4. Conclusion

This study demonstrated the application of five probabilistic ML models—XGBoostLSS, NGBoost, Bayesian NN, probabilistic NN, and MC dropout NN—for predicting the q_e of PO_4(III) onto biochar-based adsorbents. Using a dataset of 2959 data points, XGBoostLSS achieved the highest predictive accuracy, with the greatest test R^2 and lowest RMSE values. SHAP analysis identified adsorption experimental conditions (43%), adsorbent synthesis conditions (29%), and adsorbent characteristics (28%) as the most influential factors. These findings provide valuable insights into the key parameters affecting PO_4(III) removal and highlight the potential of ML for optimizing adsorbent performance in wastewater treatment. The study underscores the practical applicability of these models in addressing environmental challenges while acknowledging their inherent limitations, paving the way for further advancements in sustainable engineering solutions.

CRediT authorship contribution statement

Sara Iftikhar: Writing – original draft, Data curation. **Rehan Ishtiaq:** Writing – original draft, Methodology, Data curation. **Nallain Zahra:** Formal analysis, Data curation. **Fazila Ruba:** Investigation, Data curation. **Sze-Mun Lam:** Visualization, Investigation, Formal analysis. **Ather Abbas:** Writing – review & editing, Supervision, Methodology. **Zeeshan Haider Jaffari:** Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Conceptualization.

Declaration of competing interest

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

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemosphere.2024.144031.

Data availability

All the figures and tables presented in this study are completely reproducible. The python code is available at https://github.com/Sara-Iftikhar/po4_removal_ml and the Jupyter notebooks reproducing the presented findings are hosted at readthedocs (www.readthedocs.io) and sphinx-gallery applications, which can be found at https://po4-removal-ml.readthedocs.io/.

References

Abdi, J., Hadipoor, M., Hadavimoghaddam, F., Hemmati-Sarapardeh, A., 2022. Estimation of tetracycline antibiotic photodegradation from wastewater by heterogeneous metal-organic frameworks photocatalysts. Chemosphere 287, 132135.

Bandler, J.W., Biernacki, R.M., Cai, Q., Chen, S.H., 1994. A novel approach to statistical modeling using cumulative probability distribution fitting. In: 1994 IEEE MTT-S International Microwave Symposium Digest (Cat. No. 94CH3389-4), pp. 385–388.

Charnock, T., Perreault-Levasseur, L., Lanusse, F., 2022. Bayesian neural networks. In: Artificial Intelligence for High Energy Physics. World Scientific, pp. 663–713.

Chien, S.-W.C., Ng, D.-Q., Kumar, D., Lam, S.-M., Jaffari, Z.H., 2022. Investigating the effects of various synthesis routes on morphological, optical, photoelectrochemical and photocatalytic properties of single-phase perovskite $BiFeO_3$. J. Phys. Chem. Solid. 160, 110342.

Cui, X., Dai, X., Khan, K.Y., Li, T., Yang, X., He, Z., 2016. Removal of phosphate from aqueous solution using magnesium-alginate/chitosan modified biochar microspheres derived from Thalia dealbata. Bioresour. Technol. 218, 1123–1132.

Da, T.-X., Ren, H.-K., He, W.-K., Gong, S.-Y., Chen, T., 2022. Prediction of uranium adsorption capacity on biochar by machine learning methods. J. Environ. Chem. Eng. 10, 108449.

Davison, E.K., Neville, J.C., Sperry, J., 2023. Phosphorus sustainability: a case for phytic acid as a biorenewable platform. Green Chem. 25, 5390–5403.

Duan, T., Anand, A., Ding, D.Y., Thai, K.K., Basu, S., Ng, A., Schuler, A., 2020. Ngboost: Natural gradient boosting for probabilistic prediction. In: International Conference on Machine Learning, pp. 2690–2700.

Dürr, O., Sick, B., Murina, E., 2020. Probabilistic Deep Learning: with python, Keras and Tensorflow Probability. Manning Publications, New York.

Eduah, J.O., Nartey, E.K., Abekoe, M.K., Henriksen, S.W., Andersen, M.N., 2020. Mechanism of orthophosphate (PO_4-P) adsorption onto different biochars. Environ. Technol. Innov. 17, 100572.

Foroutan, R., Mohammadi, R., Razeghi, J., Ahmadi, M., Ramavandi, B., 2024. Amendment of Sargassum oligocystum bio-char with $MnFe_2O_4$ and lanthanum MOF obtained from PET waste for fluoride removal: a comparative study. Environ. Res. 251, 118641.

Foroutan, R., Peighambardoust, S.J., Amarzadeh, M., Korri, A.K., Peighambardoust, N.S., Ahmad, A., Ramavandi, B., 2022. Nickel ions abatement from aqueous solutions and shipbuilding industry wastewater using ZIF-8-chicken beak hydroxyapatite. J. Mol. Liq. 356, 119003.

Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: representing model uncertainty in deep learning. In: International Conference on Machine Learning, pp. 1050–1059.

Gao, D., Li, B., Huang, X., Liu, X., Li, R., Ye, Z., Wu, X., Huang, Y., Wang, G., 2023. A review of the migration mechanism of antibiotics during struvite recovery from wastewater. Chem. Eng. J. 466, 142983.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Am. Stat. Assoc. 102, 359–378.

Grossmann, L., 2024. Sustainable media feedstocks for cellular agriculture. Biotechnol. Adv. 73, 108367.

Guth, S., Mojahed, A., Sapsis, T.P., 2024. Quality measures for the evaluation of machine learning architectures on the quantification of epistemic and aleatoric uncertainties in complex dynamical systems. Comput. Methods Appl. Mech. Eng. 420, 116760.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning; Data mining, Inference and Prediction, 2nd edition. Springer, New York.

Hertel, V., Chow, C., Wani, O., Wieland, M., Martinis, S., 2023. Probabilistic SAR-based water segmentation with adapted Bayesian convolutional neural network. Remote Sens. Environ. 285, 113388.

Huang, J., Zhang, J., Li, X., Qiao, Y., Zhang, R., Kumar, G.S., 2023. Investigating the effects of ensemble and weight optimization approaches on neural networks' performance to estimate the dynamic modulus of asphalt concrete. Road Mater. Pavement Des. 24, 1939–1959.

Iftikhar, S., Zahra, N., Rubab, F., Sumra, R.A., Khan, M.B., Abbas, A., Jaffari, Z.H., 2023. Artificial neural networks for insights into adsorption capacity of industrial dyes using carbon-based materials. Sep. Purif. Technol. 326, 124891.

Ishtiaq, R., Zahra, N., Iftikhar, S., Rubab, F., Sultan, K., Abbas, A., Lam, S.-M., Jaffari, Z. H., Park, K.Y., Others, 2024. Adsorption of Cr(VI) ions onto fluorine-free niobium carbide (MXene) and machine learning prediction with high precision. J. Environ. Chem. Eng. 12, 112238.

Jaffari, Z.H., Abbas, A., Kim, C., Shin, J., Kwak, J., Son, C., Lee, Y., Kim, S., Chon, K., Hwa, K., 2024. Transformer-based deep learning models for adsorption capacity prediction of heavy metal ions toward biochar-based adsorbents. J. Hazard Mater. 462, 132773.

Jaffari, Z.H., Abuabdou, S.M.A., Ng, D.Q., Bashir, M.J.K., 2021. Insight into two-dimensional MXenes for environmental applications: recent progress, challenges, and prospects. FlatChem 28, 100256.

Jaffari, Z.H., Jeong, H., Shin, J., Kwak, J., Son, C., Lee, Y.-G., Kim, S., Chon, K., Cho, K. H., 2023. Machine-learning-based prediction and optimization of emerging contaminants' adsorption capacity on biochar materials. Chem. Eng. J. 466, 143073.

Kim, C.-M., Jaffari, Z.H., Abbas, A., Chowdhury, M.F., Cho, K.H., 2024. Machine learning analysis to interpret the effect of the photocatalytic reaction rate constant (k) of semiconductor-based photocatalysts on dye removal. J. Hazard Mater. 465, 132995.

Kirichenko, L., Lavrynenko, R., 2023. Probabilistic machine learning methods for fractional brownian motion time series forecasting. Fractal Fract 7, 517.

Krstajic, D., Buturovic, L.J., Leahy, D.E., Thomas, S., 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. J. Cheminf. 6, 1–15.

Kumar, A., Bhattacharya, T., Shaikh, W.A., Chakraborty, S., Sarkar, D., Biswas, J.K., 2022. Biochar modification methods for augmenting sorption of contaminants. Curr. Pollut. Rep. 8, 519–555.

Lam, S.-M., Jaffari, Z.H., Sin, J.-C., Zeng, H., Lin, H., Li, H., Mohamed, A.R., Ng, D.-Q., 2021. Surface decorated coral-like magnetic $BiFeO_3$ with Au nanoparticles for effective sunlight photodegradation of 2,4-D and *E. coli* inactivation. J. Mol. Liq. 326, 115372.

Li, M., Liu, J., Xu, Y., Qian, G., 2016. Phosphate adsorption on metal oxides and metal hydroxides: a comparative review. Environ. Rev. 24, 319–332.

Mae, Y., Kumagai, W., Kanamori, T., 2021. Uncertainty propagation for dropout-based Bayesian neural networks. Neural Network. 144, 394–406.

März, A., 2022. Multi-target XGBoostLSS regression. arXiv Prepr. arXiv2210, 06831.

März, A., 2019. XGBoostLSS–An extension of XGBoost to probabilistic forecasting. https://doi.org/10.48550/arXiv.1907.03178. https://arxiv.org/abs/1907.03178.

Nasab, E.A., Nasseh, N., Damavandi, S., Amarzadeh, M., Ghahrchi, M., Hoseinkhani, A., Alver, A., Khan, N.A., Farhadi, A., Danaee, I., 2023. Efficient purification of aqueous solutions contaminated with sulfadiazine by coupling electro-Fenton/ultrasound process: optimization, DFT calculation, and innovative study of human health risk assessment. Environ. Sci. Pollut. Res. 30, 84200–84218.

Palansooriya, K.N., Li, J., Dissanayake, P.D., Suvarna, M., Li, L., Yuan, X., Sarkar, B., Tsang, D.C.W., Rinklebe, J., Wang, X., others, 2022. Prediction of soil heavy metal immobilization by biochar using machine learning. Environ. Sci. Technol. 56, 4187–4198.

Pauletto, P.S., Dotto, G.L., Salau, N.P.G., 2020. Optimal artificial neural network design for simultaneous modeling of multicomponent adsorption. J. Mol. Liq. 320, 114418.

Pombo, N., Garcia, N., Bousson, K., 2017. Classification techniques on computerized systems to predict and/or to detect Apnea: a systematic review. Comput. Methods Progr. Biomed. 140, 265–274.

Qu, J., Meng, Q., Peng, W., Shi, J., Dong, Z., Li, Z., Hu, Q., Zhang, G., Wang, L., Ma, S., others, 2023. Application of functionalized biochar for adsorption of organic pollutants from environmental media: synthesis strategies, removal mechanisms and outlook. J. Clean. Prod. 423, 138690.

Rodriguez, P., Bautista, M.A., Gonzalez, J., Escalera, S., 2018. Beyond one-hot encoding: lower dimensional target embedding. Image Vis Comput. 75, 21–31.

Sadegh, F., Sadegh, N., Wongniramaikul, W., Apiratikul, R., Choodum, A., 2023. Adsorption of volatile organic compounds on biochar: a review. Process Saf. Environ. Protect. 182, 559–578.

Scala, F., Ceschini, A., Panella, M., Gerace, D., 2023. A general approach to dropout in quantum neural networks. Adv. Quantum Technol., 2300220

Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res. Atmos. 106, 7183–7192.

Turan, N.G., Mesci, B., Ozgonenel, O., 2011. The use of artificial neural networks (ANN) for modeling of adsorption of Cu (II) from industrial leachate by pumice. Chem. Eng. J. 171, 1091–1097.

Wan, D., Wu, L., Liu, Y., Chen, J., Zhao, H., Xiao, S., 2019. Enhanced adsorption of aqueous tetracycline hydrochloride on renewable porous clay-carbon adsorbent derived from spent bleaching earth via pyrolysis. Langmuir 35, 3925–3936.

Xie, Y., Abaee, M., Navazeni, R., Shamshiri, V., Frontistis, Z., Amarzadeh, M., 2024. Engineering S-scheme heterojunction MgO/WO_3-integrated Graphene photocatalyst for robust detoxification of tetracycline: mechanistic insight and actual matrix remediation. Surface. Interfac. 104591.

Yang, Y., Piao, Y., Wang, R., Su, Y., Qiu, J., Liu, N., 2022. Mechanism of biochar functional groups in the catalytic reduction of tetrachloroethylene by sulfides. Environ. Pollut. 300, 118921.

Yao, Y., Qiu, Y., Cui, Y., Wei, M., Bai, B., 2023. Insights to surfactant huff-puff design in carbonate reservoirs based on machine learning modeling. Chem. Eng. J. 451, 138022.

Zhang, Y., Wang, X., Hu, Z., Xiao, Q., Wu, Y., 2025. Capturing and recovering phosphorus in water via composite material: research progress, future directions, and challenges. Sep. Purif. Technol. 353, 128453.

Zhang, Z., Yan, L., Yu, H., Yan, T., Li, X., 2019. Adsorption of phosphate from aqueous solution by vegetable biochar/layered double oxides: fast removal and mechanistic studies. Bioresour. Technol. 284, 65–71.

Zhao, F., Tang, L., Jiang, H., Mao, Y., Song, W., Chen, H., 2023. Prediction of heavy metals adsorption by hydrochars and identification of critical factors using machine learning algorithms. Bioresour. Technol. 383, 129223.

Zheng, Y., Wan, Y., Zhang, Y., Huang, J., Yang, Y., Tsang, D.C.W., Wang, H., Chen, H., Gao, B., 2023. Recovery of phosphorus from wastewater: a review based on current phosphorous removal technologies. Crit. Rev. Environ. Sci. Technol. 53, 1148–1172.

Zhu, X., He, M., Sun, Y., Xu, Z., Wan, Z., Hou, D., Alessi, D.S., Tsang, D.C.W., 2022a. 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.

Zhu, X., Wang, X., Ok, Y.S., 2019. The application of machine learning methods for prediction of metal sorption onto biochars. J. Hazard Mater. 378, 120727.

Zhu, X., Xu, Z., You, S., Komárek, M., Alessi, D.S., Yuan, X., Palansooriya, K.N., Ok, Y.S., Tsang, D.C.W., 2022b. Machine learning exploration of the direct and indirect roles of Fe impregnation on Cr(VI) removal by engineered biochar. Chem. Eng. J. 428, 131967.

Ziel, F., 2022. M5 competition uncertainty: overdispersion, distributional forecasting, GAMLSS, and beyond. Int. J. Forecast. 38, 1546–1554.