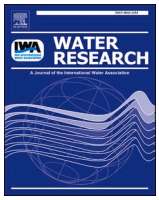
Water Research 271 (2025) 122815

Contents lists available at ScienceDirect

Water Research

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



Optimisation led energy-efficient arsenite and arsenate adsorption on various materials with machine learning

Jinsheng Huang [a,1], Waqar Muhammad Ashraf [b,1], Talha Ansar [c,1], Muhammad Mujtaba Abbas [c], Mehdi Tlija [d], Yingying Tang [a], Yunxue Guo [e], Wei Zhang [a,*]

[a] School of Environmental Science and Engineering, Guangzhou University, Guangzhou 510006, PR China
[b] The Sargent Centre for Process Systems Engineering, Department of Chemical Engineering, University College London, Torrington Place, London WC1E 7JE, UK
[c] Department of Mechanical Engineering, University of Engineering and Technology Lahore, New Campus, Kala Shah Kaku 39020, Pakistan
[d] Department of Industrial Engineering, College of Engineering, King Saud University, P.O. Box 800, Riyadh 11421, Saudi Arabia
[e] Key Laboratory of Tropical Marine Bio-resources and Ecology, South China Sea Institute of Oceanology, Chinese Academy of Sciences, No.1119, Haibin Road, Nansha District, Guangzhou 511458, China

ARTICLE INFO

Keywords:
Machine learning
Various materials
Process energy consumption
Web application
Sustainable development goals
Arsenic

ABSTRACT

The contamination of water by arsenic (As) poses a substantial environmental challenge with far-reaching influence on human health. Accurately predicting adsorption capacities of arsenite (As(III)) and arsenate (As(V)) on different materials is crucial for the remediation and reuse of contaminated water. Nonetheless, predicting the optimal As adsorption on various materials while considering process energy consumption continues to pose a persistent challenge. Literature data regarding the As adsorption on diverse materials were collected and employed to train machine learning models (ML), such as CatBoost, XGBoost, and LGBoost. These models were utilized to predict both As(III) and As(V) adsorption on a variety of materials using their reaction parameters, structural properties, and composition. The CatBoost model exhibited superior accuracy, achieving a coefficient of determination (R^2) of 0.99 and a root mean square error (RMSE) of 1.24 for As(III), and an R^2 of 0.99 and RMSE of 5.50 for As(V). The initial As(III) and As(V) concentrations were proved to be the primary factors influencing adsorption, accounting for 27.9 % and 26.6 % of the variance for As(III) and As(V) individually. The genetic optimization led optimisation process, considering the low energy consumption, determined maximum adsorption capacities of 291.66 mg/g for As(III) and 271.56 mg/g for As(V), using C-Layered Double Hydroxide with reduced graphene oxide and chitosan combined with rice straw biochar, respectively. To further facilitate the process design for different real-life applications, the trained ML models are embedded into a web-app that the user can use to estimate the As(III) and As(V) adsorption under different design conditions. The utilization of ML for the energy-efficient As(III) and As(V) adsorption is deemed essential for advancing the treatment of inorganic As in aquatic settings. This approach facilitates the identification of optimal adsorption conditions for As in various material-amended waters, while also enabling the timely detection of As-contaminated water.

Synopsis

We applied machine learning for prediction of arsenite and arsenate adsorption on various materials considering the process energy consumption.

1. Introduction

Arsenic (As) from naturally occurring or geogenic sources poses a pervasive global issue. It remains a significant groundwater contaminant, raising concerns due to its toxicity (Cui et al., 2020). Arsenite (As(III)) and arsenate (As(V)) are prevalent in aquatic environments (Sarkar et al., 2022). Meanwhile, As(III) is more toxic, reactive, and challenging to eliminate compared to As(V) from water, primarily due to its weaker adsorption affinity (Khan et al., 2020; Sarkar and Paul, 2020). Globally,

* Corresponding author.
E-mail address: zh_wei@gzhu.edu.cn (W. Zhang).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.watres.2024.122815
Received 27 August 2024; Received in revised form 17 October 2024; Accepted 18 November 2024
Available online 19 November 2024

elevated As concentrations in drinking water and groundwater are commonly observed, with As levels in polluted water and wastewater typically ranging 0.1–230 mg/L (Aftabtalab et al., 2022; Matschullat, 2000). Elevated As concentrations have been documented in regions like the Ganges River alluvial deposits and the Bengal Delta Plain, where As levels can exceed 2000 mg/L (Brickson, 2003). Globally, an estimated 220 million potentially exposed to As-contaminated groundwater in 2020, predominantly in Asia (94 %) (Michael, 2013; Podgorski and Berg, 2020). The advancement of techniques is imperative to ensure the provision of safe and potable water, particularly in heavily As-affected areas (Hou et al., 2023; Zhang et al., 2022b). Therefore, it is imperative to devise sustainable and economical techniques for eliminating As from drinking water and groundwater.

Adsorption is widely acknowledged as an environmentally friendly and cost-effective method alternatives such as for As adsorption in comparison to ion exchange, membrane filtration, and biological treatment (Gohr et al., 2022; Sakhiya et al., 2023). Various materials are employed for adsorption, including clay, silicon-based materials, metal-organic frameworks, carbon-based materials, and waste-derived materials. For instance, nano-scaled activated carbon, synthesized through iron and manganese oxide modification, has been utilized for As (V) removal (Gallios et al., 2017). Both natural and synthetic zeolites have been extensively researched for efficient As removal (Li et al., 2011). During the entire adsorption process, factors influencing the adsorption capacity of various materials for As include reaction parameters (such as adsorbent dosage and initial As concentration), structural properties (such as material type and surface area), and composition (such as pyrolysis temperature) (Zhang et al., 2023). Despite numerous studies, the majority have focused on laboratory experiments that systematically vary individual variables. Determining the respective contributions of these factors to adsorption efficiency poses challenges due to its time-consuming and intricate nature. Additional research is imperative to clarify the primary factors influencing As adsorption on various materials and to forecast the adsorption efficacy of different materials on As with low energy consumption and optimal efficiency. Addressing this limitation will significantly enhance the practical viability and efficacy of selecting As remediation materials, facilitating low consumption and high-efficiency adsorption of inorganic As in polluted water.

Machine Learning (ML) is an interdisciplinary field that harnesses extensive and intricate datasets to construct prediction models (Zhang et al., 2023). ML is applied across a wide array of research fields, including adsorption of organic compounds on biochar (Jaffari et al., 2023), CO_2 capture on engineered biochar (Yuan et al., 2024), heavy metals immobilization in contaminated soil (Palansooriya et al., 2022; Shi et al., 2023), distribution and contamination of As in groundwater (Podgorski & Berg, 2020), and the presence of As in wells, drinking water, and surface water (Ibrahim et al., 2022b; Lombard et al., 2021). The crux of ML theory involves the development and validation of algorithms that empower computers to autonomously learn. These algorithms analyze existing data structures automatically, deducing rules for making informed judgments and predictions about unfamiliar samples. This methodology aids in unveiling causal relationships between variables, uncovering intricate details that conventional analyses may overlook. As a result, ML emerges as a pivotal tool in addressing challenges linked to As contamination, especially in forecasting efficient adsorption and low-consumption of As across various scenarios, thus enhancing overall remediation outcomes. Consequently, leveraging ML technology to combat As contamination is indispensable for gaining deeper insights into the significance of each variable and streamlining the formulation of low-consumption, effective adsorption for recycling As-contaminated water.

Exploring As adsorption on various materials through ML technology remains limited in current literature (Liu et al., 2023; Zhang et al., 2023). In tackling this issue, we compiled an extensive dataset from diverse literature sources detailing As adsorption on various materials and elucidating the low consumption and high efficiency absorption influenced by multiple system variables. Given the dataset's high-dimensional complexity and the intricate interactions between input and target variables, a nonlinear function space was established to depict the As adsorption mechanism. To effectively capture the adsorption dynamics of As under diverse process conditions, a comprehensive dataset necessitates the integration of efficient function approximation algorithms capable of managing large data volumes and accurately predicting complex function spaces. In the context of such modelling endeavors, researchers commonly employ ML algorithms to effectively approximate the adsorption dynamics (Zhu et al., 2019a; 2019b). Employing an ML-based modeling algorithm, we estimated the adsorption behavior of various materials for As under different conditions. Our evaluation focused on the top three performing ML models - CatBoost, XGBoost, and LGBoost (Golde et al., 2019), showcasing robust predictive accuracy and generalization capabilities. These models effectively establish functional links between system variables, assign unbiased weights to pertinent variables, and yield precise predictions at both local and global scales.

This study aims to fill the knowledge gap by integrating research data on As adsorption on various materials under differing conditions. Through the compilation of datasets, ML models have been devised to forecast the cost-effective and efficient adsorption behavior of different materials for both As(III) and As(V) under diverse conditions, marking a significant advancement in this domain. Moreover, this study introduces a comprehensive framework to pinpoint the pivotal factors influencing As adsorption by different materials in water systems, thus enriching the comprehension of their role in As adsorption capacity, a facet lacking in current literature. The model considers the properties and forms of inorganic As in water on different materials, aiding in the prediction of inorganic As adsorption efficiency on varied materials with minimal consumption and heightened efficiency. This model diminishes the necessity for numerous experiments, streamlines parameter evaluation, and enhances the efficacy of environmental management. The application of these models demonstrates the potential to devise and refine As removal processes in polluted water. This research provides insightful understanding into the future advancement of environmentally friendly, resource-efficient, and highly effective As remediation in As-contaminated water.

2. Materials and methods

2.1. Data acquisition

Information on both As(III) and As(V) adsorption concerning the input and target variables was gathered from available literature sources. Relevant articles were searched in the Web of Science core collection using keywords "arsenic," "material," and "adsorption", leading to the retrieval of 93 articles published within the last decade. Relevant input variables associated with the target variable were meticulously identified. Information for the variables was extracted from each article, and all data points were amalgamated into a unified master file. For a thorough data collection process, specific data was immediately derived from tables, figures, and supplementary materials within the published articles. WebPlotDigitizer was used to extract data from graphs. Subsequently, the gathered research data underwent thorough analysis to identify any missing values, and methodologies for imputing missing data were subsequently applied to process missing values. Additional information on the missing data imputation process is provided below.

Table S1 and Table S2 offer datasets summary collected in this study, including the relevant reference publications. Altogether 682 and 923 data points were extracted for As(III) and As(V) adsorption, respectively, encompassing missing observations across different input variables. These missing data points underwent meticulous handling by employing rigorous techniques for data cleaning and imputation. To analyze the adsorption behavior of different materials and predict both As(III) and

As(V) adsorption capacity, eight influencing variables were investigated and classified into three primary clusters: (i) reaction parameters (adsorbent dosage, reaction temperature, initial As concentration, reaction time, and solution pH), (ii) structural properties (pore volume and BET surface area), and (iii) composition (pyrolysis temperature). These variables were selected according to their possible impact on the process of As adsorption and their correlation with the performances of the adsorbent materials.

After the completion of data collection, numerous values essential for model development are absent from the input features. For example, among the 682 and 923 data points gathered for modeling As(III) and As (V), there is a notable absence of information regarding BET surface area, pore volume, solution pH, reaction temperature, adsorbent dosage, and reaction time (Table S3). Specifically, the As(III) dataset is missing 108, 349, 42, 14, 68, and 53 data points for these parameters, while the As(V) dataset lacks 180, 497, 68, 35, 20, and 42 data points. Eliminating these missing data points from the dataset would significantly undermine the model's accuracy, and basic imputation methods like mean and mode are deemed inadequate due to the intricate and high-dimensional nature of the data.

In this study, a decision tree regression (DTR) model is utilized to impute missing feature values based on the complete dataset. Given that most features without missing values (categorical: M. type, BioC. Type, and B. type; continuous: Int. Total As conc. and P. temp) are categorical, the DTR model is anticipated to perform effectively due to the algorithm's interpretability (Rahman & Islam, 2011). The dataset without the missing values is divided into 80 % for training and 20 % for testing the DTR model to assess its predictive capabilities on the test dataset. Evaluation metrics for the DTR model on both training and test sets are detailed in Table S4. Subsequently, the DTR-based mapping function is employed to predict missing feature values.

2.2. Data visualization and pre-processing

Visualization of data is crucial in the development of ML models as it provides a graphical depiction of data distribution in both input and output variables. A key aspect of a dataset involves the dispersion of data within the ranges of variables. Therefore, to visualize the data trend profiles of the factors, box plots offer a succinct and efficient representation of data distribution, coupled with data statistics, enabling the exploration of data quality and diversity.

Heat maps are constructed in this research to illustrate the association of the variables related to As(III) and As(V) adsorption on various materials. In academic literature, the Pearson correlation coefficient (PCC) is widely used to measure the linear association between variable pairs, leveraging variable covariance and range invariance (Ishfaq et al., 2023; Tariq et al., 2024). Calculating the PCC among the variables defining the input-process-output system aids in identifying their linear interactions, providing data-driven insights into the system's operational characteristics. In the present investigation, the calculation of the PCC involves both the input and output variables in a dataset that includes As(III) and As(V) adsorption on various materials, with the objective of exploring the linear associations among the factors. The mathematical formula is presented below:

$$R_{xy} = \frac{\sum_i^N (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i^N (x_i - \overline{x})^2}\sqrt{\sum_i^N (y_i - \overline{y})^2}} \quad (1)$$

In this context, the PCC denoted as R_{xy} signifies the correlation between the input variable x and the target variable y. The PCC scale extends from –1 to 1. $R_{xy} = 1$ signifies a robust linear relationship between the variates, with the sign denoting positive or negative correlation. Conversely, $R_{xy} = 0$ indicates a lack of correlation between the variates.

2.3. Development and building of ML models

Three modeling algorithms—CatBoost, XGBoost, and LGBoost—are employed to predict levels of As(III) and As(V) adsorption by considering reaction parameters, as well as structural and compositional properties across various materials. These algorithms, recognized for their robustness in capturing interactive and nonlinear relationships among a multitude of input variables (Zhang et al., 2023), are adept at approximating complex function profiles within high-dimensional input spaces and large datasets (Yuan et al., 2024), while also demonstrating resilience against overfitting compared to multilayer perceptron models. Significantly, these algorithms exhibit excellent performance when applied to datasets comprising 200 to 1000 data points and input spaces with dimensions spanning from 5 to 15 (Roh et al., 2024). Therefore, these characteristics of the algorithms provide a competitive advantage in crafting process models for As adsorption on a range of materials within datasets that include diverse conditions of input variables.

Firstly, CatBoost is a gradient boosting algorithm tailored to proficiently manage categorical features in classification and regression tasks. CatBoost enhances prediction accuracy by utilizing ordered boosting and categorical feature encoding, thereby reducing overfitting and enhancing model generalization. It automatically handles categorical variables, mitigates prediction bias, and maintains robust and accurate model performance throughout training. Secondly, the XGBoost algorithm, as a scalable tree boosting system (Chen & Guestrin, 2016), integrates weak learners to create a strong learner. Its base learner, a classification and regression tree (CART), is continuously enhanced by greedily adding CARTs that optimize the model the most. This iterative process involves fitting residuals of previous CARTs through newly generated CARTs (Ye et al., 2023). Thirdly, LGBoost is an advanced variant of gradient boosted decision trees (GBDT) tailored to overcome the limitations of GBDT related to dataset magnitude and feature space complexity. In comparison to GBDT, LGBoost has demonstrated superior predictive and generalization abilities (Abdi & Mazloom, 2022; Wang et al., 2022). Through the utilization of exclusive feature bundling and gradient-boosting one-sided sampling techniques, the LGBoost model establishes efficient functional mappings between input and target factors, thereby optimizing the utilization of computational resources.

Achieving accurate predictions and robust generalization capabilities in ML modeling algorithms necessitates the meticulous selection of a diverse range of parameters. The parameter space is algorithm-specific, and the literature offers several methods for determining the optimal hyperparameter values. Frequently employed methods for hyperparameter tuning encompass random search, grid search, manual search, and Bayesian optimisation. Among these, Bayesian optimisation explores optimal hyperparameter combinations through sequential learning, converging to the best solution in a reasonable time frame with more efficient function evaluations compared to the complex and computationally expensive manual and grid search methods. Therefore, Bayesian optimization was utilized in this research to fine-tune the hyperparameters of the ML models. Overfitting poses a prevalent challenge in ML models when they fail to sufficiently capture the system's function space. To mitigate the concern, k-fold cross-validation is implemented during model training, effectively reducing overfitting.

2.4. Error metrics

The assessment of performance metrics is vital in gauging and contrasting the efficiency of the ML algorithm under scrutiny. These metrics include the coefficient of determination (R^2) and root mean square error (RMSE) (Ashraf & Dua, 2024a), and can be expressed mathematically:

$$R^2 = 1 - \frac{\sum_i^N (y_i - \widehat{y}_i)^2}{\sum_i^N (y_i - \overline{y}_i)^2} \quad (2)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \quad (3)$$

In this context, y_i denotes the true values of the target variable, while $\hat{y}_i$ signifies the predicted values of the target variable by the model; $\bar{y}_i$ signifies the average value for y_i; $i = 1, 2, 3, \ldots$; N corresponds to the total observed value. R^2 serves as a metric to evaluate the accuracy of predictions generated by the ML model. R^2 ranges from 0 to 1, indicting suboptimal predictive capability and ideal correspondence between the input and target variables, respectively. Meanwhile, in a given dataset, RMSE assesses the disparity between the observed responses and the model's predicted values.

2.5. Prediction intervals on the ML models

The inductive conformal prediction (ICP) is utilized to calculate prediction intervals for the model-based predictions, aiming to estimate the uncertainty bounds alongside the point-predictions. The ICP stands out as a robust method for generating prediction intervals, supported by theoretical assurances for deriving valid prediction intervals (Vovk et al., 1999). The ICP method uses data from the validation split to estimate non-conformity score for making the prediction intervals on the test dataset. The trained ML model predicts the validation dataset (x_{valid}) to generate predictions ($\hat{y}_{valid}$). The non-conformity function (ε) computes the deviation between the actual validation target values (y_{valid}) and the model-based predicted values ($\hat{y}_{valid}$). The 95th percentile of non-conformity scores, denoted as δ_{95}, associated with the validation dataset-based predictions is computed. δ_{95} is used to create a prediction interval for the test set predictions $\hat{y}_{test}$. The prediction intervals drawn by the ICP technique can be mathematically represented as:

$$\varepsilon = |y_{valid} - \hat{y}_{valid}| \quad (4)$$

$$\delta_{95} = {}_{(n(\varepsilon)*0.95)} \quad (5)$$

$$Prediction\ Interval = [\hat{y}_{test} - \delta_{95}\ ,\ \hat{y}_{test} + \delta_{95}] \quad (6)$$

The prediction interval, derived from the point prediction of the ML model for $\hat{y}_{test}$, signifies the range of uncertainty associated with the model's predictions on unseen test data. Narrow prediction intervals indicate high accuracy in the model's predictions, a crucial aspect for deploying ML models in production environments.

2.6. Feature importance analysis of ML model

The next critical step is to analyze the influence of input variables on the target variable to comprehend the predictive mechanism of the ML model. Various techniques have been documented to conduct feature importance analysis, encompassing SHapley Additive exPlanations (SHAP), the Monte Carlo method, one-factor-at-a-time analysis, and the partial derivative method (Ashraf & Dua, 2024b; Ashraf et al., 2021; Uddin et al., 2019). Among these, SHAP, which calculates SHapley values, emerges as the method that fulfills essential criteria such as efficiency, symmetry, dummy, and additivity. Notably, SHAP is recognized for delivering a distinct solution characterized by three critical attributes: local accuracy, consistency, and handling of missing data (Ibrahim et al., 2022a).

Employing a model-agnostic technique based on game theory principles, the computation of SHAP values for input variables involves constructing games that elucidate the impact of these variables on the target variable. By employing the SHAP technique, one can gain insights into sensitivity on both a local and global scale by choosing particular dataset arrays or examining the dataset as a whole. Consequently, through the computation of SHAP values for input variables, one can ascertain their importance and ranking. Understanding the impact of crucial input variables on the target variable is crucial, as it can guide laboratory experiments and enhance industrial processes efficiently.

2.7. Energy-efficient optimisation of As adsorption

Multi-objective optimisation involves the concurrent optimisation of multiple, often conflicting, objectives by traversing the feature sample space. In this research, we aim to maximize As(III) and As(V) adsorption on various material while minimizing the energy consumption of the process. Genetic algorithm (GA) is chosen for solving the optimisation problem due to its efficacy in handling multiple targets and extensive sample spaces. Genetic algorithms (GAs) represent a robust approach for multivariate optimisation, drawing inspiration from the principles of natural selection and genetics. These algorithms excel in addressing optimisation challenges characterized by numerous interdependent variables and nonlinear relationships. GA functions by generating a population of potential solutions and progressively refining them through processes like selection, crossover, and mutation. This iterative method facilitates the exploration of a vast search space to pinpoint optimal or near-optimal solutions. A primary strength of GAs lies in their capacity to evade local optima and simultaneously explore diverse regions within the search space, rendering them highly suitable for enhancing system responses. Within the realm of As(III) and As(V) adsorption on materials, GAs can be leveraged to optimize the compositional and structural attributes of materials, alongside reaction parameters, to maximize adsorption efficiency. The adaptability of GAs in managing both continuous and discrete variables offers a comprehensive approach to parameter optimisation.

2.8. Online model deployment

After completion of the ML model development, the subsequent phase involves deploying the model for real-life applications. In the scenario of ML models trained for As(III) and As(V) adsorption on diverse materials, a publicly accessible web application enables the research community to employ the model for estimating As(III) and As(V) adsorption across varying materials, material properties, and reaction conditions. This web application, hosting the ML model trained on data sourced from published studies, further enables exploration of various process design configurations and limitations to expedite the advancement of water treatment technologies.

3. Results and discussion

3.1. Data-descriptive analysis

Fig. 1a, b elucidate how the input variables relate to the target variables, focusing on As(III) and As(V). The majority of data points are concentrated within the interquartile range (IQR), specifically between the 25 % and 75 % percentiles of the dataset. A dual violin chart is employed to depict the distribution of As concentrations with respect to modified and unmodified biochar type showing the asymmetric distribution of data with respect to the condition of the biochar. A few observations distributed significantly far away from IQR are also observed on the box-plot based data-visualization; however, these observations are kept considering the information on the extreme process conditions and As adsorption. The data distribution patterns reveal a wide operational span for both input and target variables sourced from the literature data. These patterns outline the shared design and functional domain of these variables, aiding in the examination of As adsorption mechanisms on different materials. The expansive operational scope of these variables is beneficial for developing ML models capable of predicting As(III) and As(V) adsorption on a range of materials across varied input scenarios.

Another method for visualizing the interaction between different features is a ternary plot. This barycentric plot depicts three variables that collectively sum to a constant. The ternary plot showcases the ratios

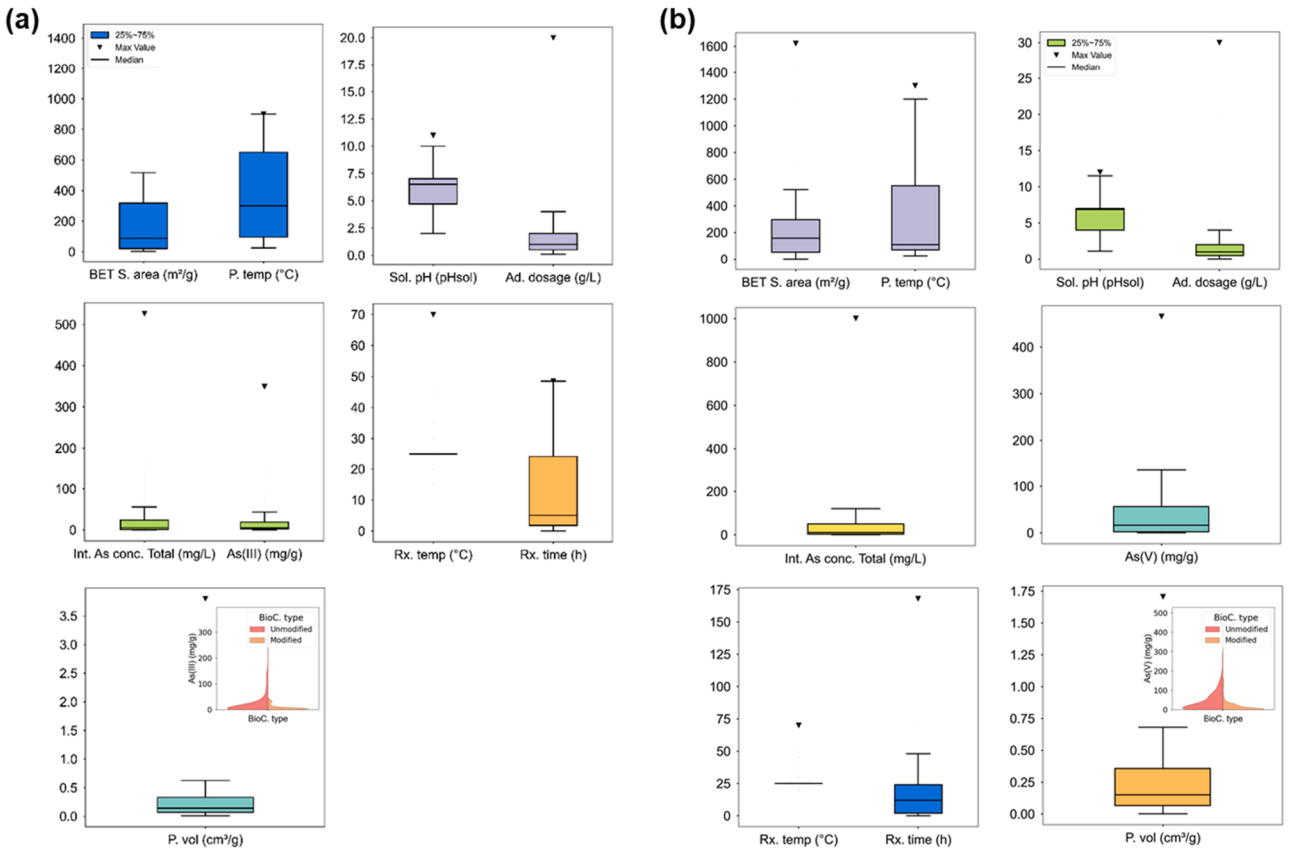


Fig. 1. Box plot based data-visualization of the input variables for As(III) (a) and As(V) (b). A good data-spread on the operating ranges of the input variables is observed corresponding to As(III) and As(V) adsorption.

of these variables within an equilateral triangle, standardizing the data on a 100 % basis. The proximity of a point to the triangle's corners signifies the relative impact of the corresponding feature. Ternary plots serve to illustrate the relationships between input variables (e.g., BET Surface Area) and their influence on As(III) and As(V) levels. The visualization illustrates the space of data distribution for input variables, target variables (particularly As(III) and As(V)), and the distribution of data (Fig. 2). Moreover, shapes and color maps are employed to elucidate the distribution of materials type (M. Type) and As(III) and As(V) concentrations, respectively. The ternary plots reveal that material types show distinct clustering patterns based on reaction temperature and BET surface area, with sorbents materials excelling in As(III) or As(V) sorption. Optimal As removal occurs within particular ranges of these variables while peak absorption is observed at lower BET surface areas and at higher reaction temperatures.

Fig. 3a, b displays heat maps according to the PCC, illustrating the relationships between the input variate and concentrations of As(III) and As(V) (represented by yellow and green colors, respectively). Most

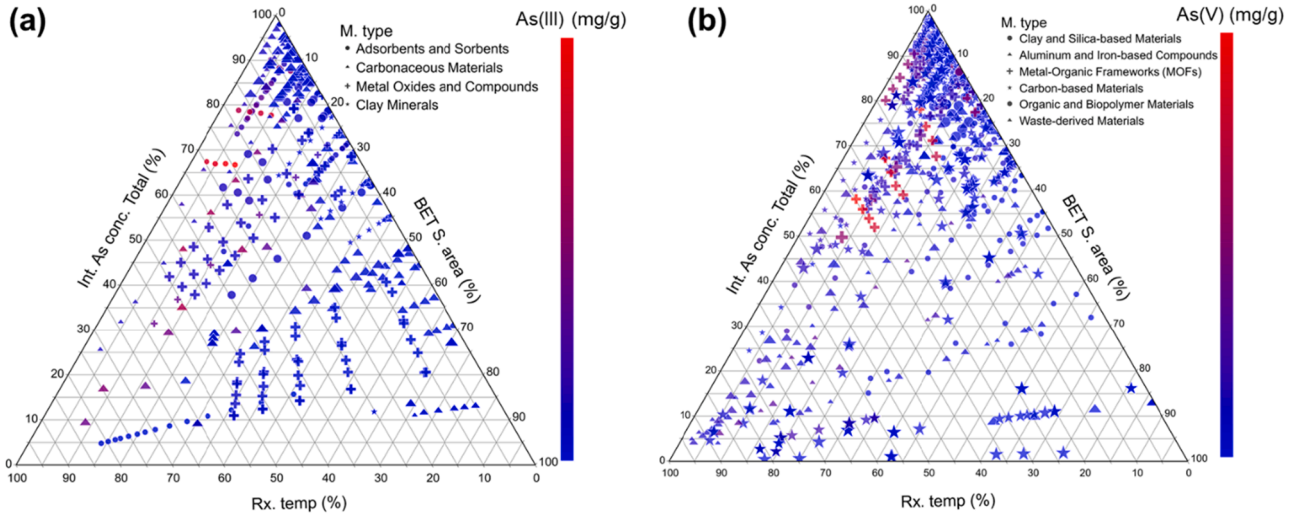


Fig. 2. Ternary plot-based data-visualization of the input variables for As(III) (a) and As(V) (b). A good data-spread on the operating ranges of the input variables is observed corresponding to As(III) and As(V) adsorption.

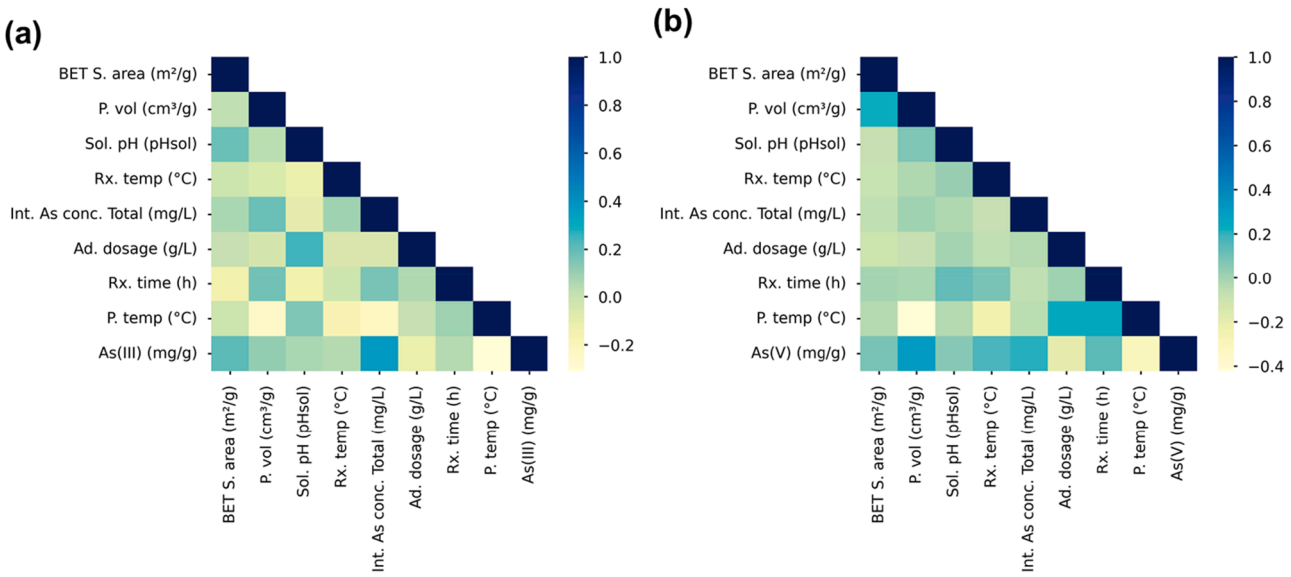


Fig. 3. PCC map constructed among the input variables as well as between the input and target variables for the dataset of As(III) (a) and As(V) (b). A few variables have large PCC values, while most of the variables have low PCC values indicating the presence of nonlinear relationships between the variables with respect to the compiled dataset.

features exhibit relatively weak linear relationships, with a maximum value of only 0.4 between BET Surface Area and As concentrations. Pore temperature (P. temp) and Pore volume (P. vol) show a weak negative correlation of –0.2. Nevertheless, the connections among the remaining variable pairs, encompassing input-input and input-target relationships within the As(III) and As(V) dataset, do not exhibit clear linearity. Low PCC values indicate a lack of direct linear connections, implying the existence of complex nonlinear relationships. This complexity can be effectively captured by ML models to delineate intricate patterns and connections within the dataset, thereby establishing an accurate functional mapping between the input and target variables within the system under investigation.

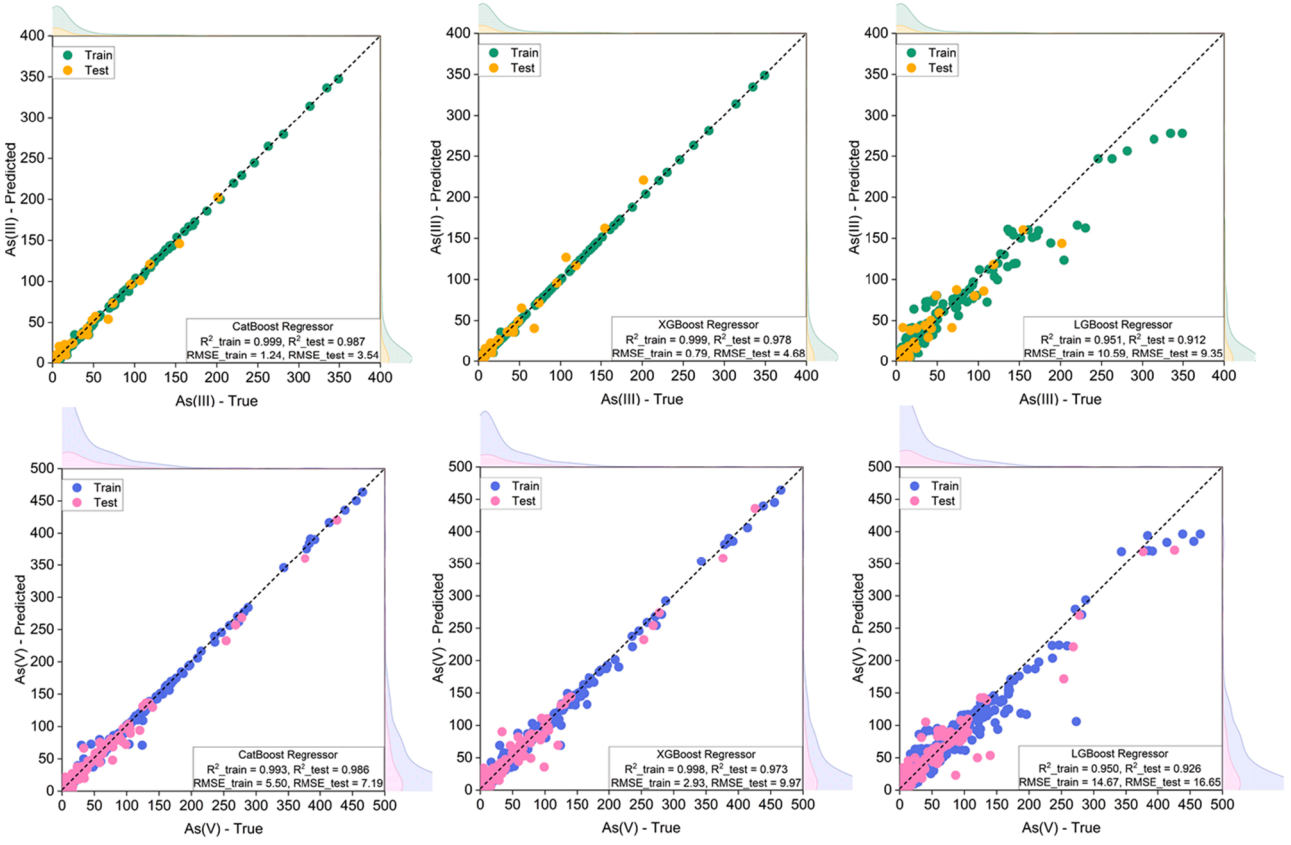


Fig. 4. The joint scatter plots constructed between actual and model predicted responses for As(III) and As(V) for the CatBoost, XGBoost, and LGBoost models. The trained models have comparable performance towards the prediction and training datasets.

3.2. CatBoost, XGBoost, and LGBoost models development

Utilizing the dataset outlined in the data collection, visualization, and processing section, three tree-based ML algorithms—CatBoost, XGBoost, and LGBoost—were employed to forecast the adsorption levels of As(III) and As(V) on diverse materials. A data split ratio of 0.70 for training, 0.15 for testing, and 0.15 for model validation was implemented. The functionality of gradient boosting algorithms was investigated to pinpoint the crucial hyperparameters influencing the modelling accuracy of XGBoost, LGBoost, and CatBoost algorithms (Roh et al., 2024; Suvarna et al., 2022). The hyperparameters for these models were optimized to enhance their predictive performance. For CatBoost, parameters such as loss function, learning rate, depth, number of iterations, L2 regularization, and max_bin were fine-tuned. XGBoost's hyperparameters were optimized for maximum depth, eta, sub-sample, the number of estimators, and colsample_bytree within the extensive parameter space. LGBoost's hyperparameters included maximum depth, learning rate, colsample_bytree, sub-sample, and number of estimators.

Displayed in Fig. 4 is a unified scatter plot showcasing the actual and forecasted results on both the training and testing datasets for As(III) through the application of three ML models: CatBoost, XGBoost, and LGBoost. In terms of the training dataset, the R^2 values, indicating model accuracy, are 0.99 for CatBoost, 0.99 for XGBoost, and 0.95 for LGBoost. On the testing dataset, the R^2 values for the models are 0.987, 0.978, and 0.912 for CatBoost, XGBoost, and LGBoost, respectively. Additionally, during the testing phase, CatBoost displays the lowest RMSE of 3.54, outperforming XGBoost (RMSE of 4.68) and LGBoost (RMSE of 9.35). Therefore, CatBoost demonstrates superior performance compared to XGBoost and LGBoost.

Similarly, Fig. 4 presents a combined scatter plot showing the observed and predicted responses for the adsorption of As(V) on different materials in the training and testing stages, utilizing CatBoost, XGBoost, and LGBoost. When evaluating the performance metrics of these models, they exhibit similar levels of performance. CatBoost performed exceptionally well in predicting the training dataset, attaining the highest R^2 value of 0.993 and a lower RMSE of 5.50, outperforming LGBoost, which had an RMSE of 14.67. Despite XGBoost showing a slightly better RMSE (2.93) during training, it displayed unsatisfactory performance in the experimental validation test. In the testing phase, CatBoost performed well with an R^2 of 0.986 and an RMSE of 7.19. Similarly, XGBoost showed results with an R^2 of 0.973 and an RMSE of 9.97, while LGBoost, though still performing adequately, had a slightly lower R^2 of 0.926 and a higher RMSE of 16.65. Consequently, the CatBoost model is selected for further analyses. While the models trained on the dataset displayed good modeling performance, it is essential to examine the uncertainty bounds associated with the model-based predictions. Within this investigation, CatBoost has demonstrated notably enhanced predictive efficacy in forecasting the As(III) and As(V) adsorption on diverse materials. As a result, prediction intervals are constructed for the CatBoost model-based predictions on the test dataset.

Fig. 5 depicts the prediction intervals generated for CatBoost models for both As(III) and As(V) utilizing the inductive conformal prediction technique. The actual values of the test dataset for As(III) and As(V) are plotted to evaluate the credibility of the established prediction intervals concerning the point predictions produced by the CatBoost models. It is evident that the actual values of the test dataset lie within the prediction intervals created for both As(III) and As(V), confirming the reliable uncertainty assessment of the trained model. Therefore, validated prediction intervals that quantify the uncertainty bounds along with the point predictions by the trained CatBoost models for As(III) and As(V) serve as valuable tools for assessing modelling accuracy in the local context. Additionally, the tighter confidence intervals established for As (III) and As(V) demonstrate the robust predictive performance of the trained CatBoost models, reflecting confidence in the point predictions generated by these models.

3.3. The significance of the input variables on adsorption of As(III) and As(V)

The ML model developed with the existing data serves as a portrayal of the system under examination. Conducting SHAP analysis is essential to delve into the system's underlying physics and pinpoint the input variables that wield substantial influence on the process. This necessitates a well-trained model with strong predictive capabilities. To achieve this, a feature importance analysis based on SHAP analysis is carried out. Considering the exceptional predictive accuracy of the CatBoost model in forecasting As(III) and As(V) adsorption on various materials using

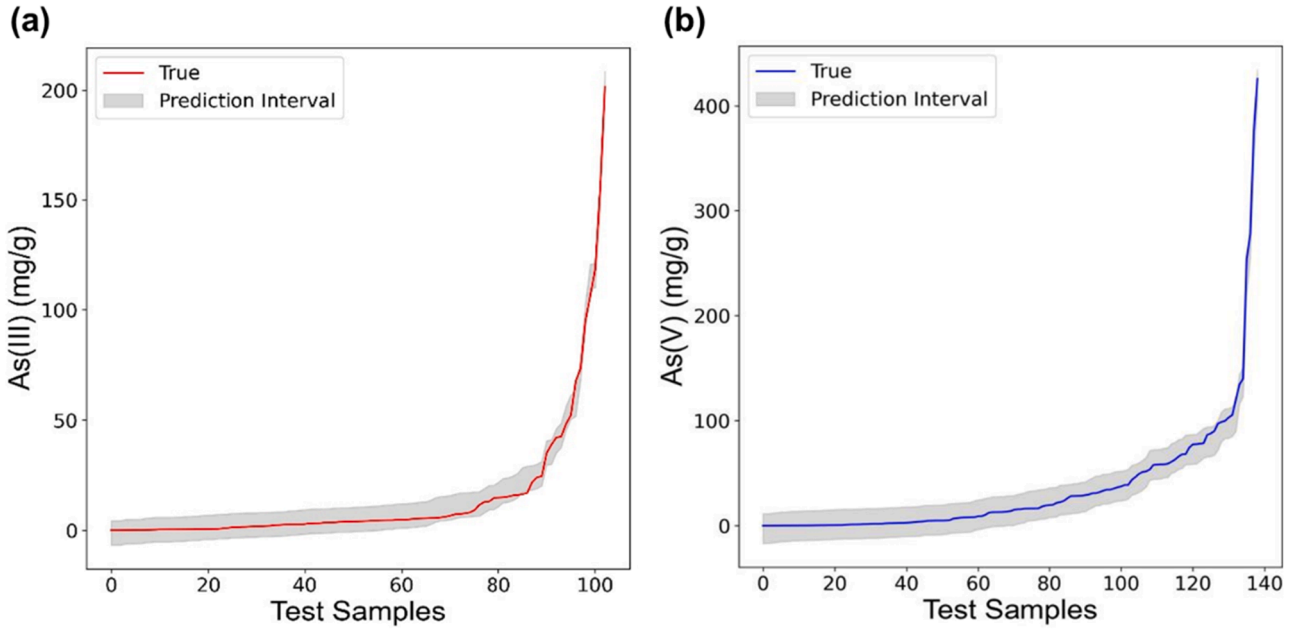


Fig. 5. The model uncertainty line plot constructed between 95 % confidence region, and model predicted responses for As(III) (a) and As(V) (b) for the CatBoost. The trained models have shown good calibration and least variability.

input parameters, the developed models are now integrated into the SHAP analytical framework, assisting in identifying the critical input variables of the process. Fig. 6 presents the SHAP analysis, unveiling the significance of input variables in the As(III) and As(V) adsorption on diverse materials. The SHAP values elucidate the individual impact of each input variable on the target outcome. The initial As(III) and As(V) levels exert the greatest influence on their adsorption across materials. For As(III) adsorption on biochar, particle temperature and BET surface area follow as the second and third most critical factors. Similarly, adsorbent dosage and particle temperature are highlighted as the second and third most crucial variables affecting As(V) adsorption on diverse materials (Fig. 6). Importantly, these findings align with the current understanding of the mechanisms involved in the absorption of As by different materials.

The As(III) and As(V) levels at the outset significantly influences the adsorption process onto biochar. Prior study has emphasized the critical role of the initial metal concentration in adsorption behavior (Ali et al., 2022; Jaffari et al., 2024; Mirza & Fujino, 2024; Zhang et al., 2023; Zhou et al., 2017). The adsorption capacity of aqueous As(III) using moringa bark was found to increase with rising initial concentrations (Mirza and Fujino, 2024). The adsorption of As on biochar aligns with a pseudo-second-order kinetics model, where elevated initial As concentrations prompt rapid binding on the biochar surface, leading to heightened occupancy of active sites with increasing initial As levels (Zhang et al., 2022a). Usually, the absorption equilibrium of As on biochar follows either the Freundlich or Langmuir isotherm models. Initially, there is an increase in absorption as the As solution concentrations rise, culminating in a saturation threshold. Given that absorption on biochar primarily occurs at the monolayer level, higher initial As concentrations bolster absorption efficiency (Khan et al., 2020). Consequently, the initial As(III) and As(V) levels are recognized as the predominant factors impacting the adsorption process.

The P. temp emerges as the second most crucial input variables for As (III) adsorption and the third for As(V), affecting adsorption on different materials. As the particle temperature increases, the carbon content (C) in biochar rises, resulting in a greater proportion of recalcitrant carbon in biochar produced at higher temperatures. This results in an expanded surface area, thereby enhancing heavy metals immobilization (Han et al., 2020; Igalavithana et al., 2019). Additionally, the Ad. dosage ranks second for As(V) and fifth for As(III) as an input variable that influences the adsorption of As on various materials. The ideal adsorption of As(III) and As(V) occurs at 1 g/L and 2 g/L of the adsorbent, respectively, highlighting the requirement for different dosages of modified biochar depending on the levels of As(III) and As(V) in the water (Zhang et al., 2022a). The BET S. area is recognized as the third most critical input parameter that impacts As(III) adsorption on biochar. Increased S. area facilitates the infiltration of As into the biochar pores, creating additional adsorption sites on the surface to bind As ions effectively (Wongrod et al., 2018). The increased S. area facilitates greater interaction with As ions. Upon reaching saturation of As adsorption on the material surface, the structural characteristics aid in the transportation of As towards the internal areas, consequently influencing the reaction equilibrium (Joshi et al., 2023; Peng et al., 2022). Consequently, the initial As(III) and As(V) concentrations, P. temp, Ad. dosage, and BET S. area all play pivotal roles in controlling the As(III) and As(V) adsorption on diverse materials.

3.4. Energy-efficient adsorption of As under various initial As levels

The trained ML models encapsulate the fundamental patterns and mechanisms of As adsorption on diverse materials. These proficient models are utilized to devise an objective function aimed at optimizing As adsorption while reducing process energy consumption. The product of *Rx. temp* and *Rx. time* accounts for the process energy consumption and heat exchange in the workbench, taking into account the constant mass flow rate of the heating medium. Thus, the multi-objective function formulated for the energy-efficient As(III) and As(V) adsorption on different materials is expressed as follows:

$$f(x) = -\ normalized(As) + (normalized(Rx.\ temp) \times normalized(Rx.\ time)$$

$$h(x) = 0$$

$$x^L < x < x^U$$

here, negative sign indicates that As adsorption on biochar is to be maximized and vice versa. $h(x)$ is the equality constraint representing the trained ML model. Whereas, x represents the set of input variables; x^L and x^U are the bounds on input variables that restricts the search space for the estimation of the solution. To further enhance the quality of the solution estimated by the solver, we have scaled the two terms in the objective function into [0,1] so that solver may converge to efficient solution.

Genetic algorithm is employed for solving the multi-objective optimization problem, which excels in optimizing both categorical and regression-type features within the in-sample space (Ishfaq et al., 2024). Furthermore, GA solver can approximate the global solution for the specified optimization problem efficiently in terms of computing resources and time, providing flexibility and high solution quality for the

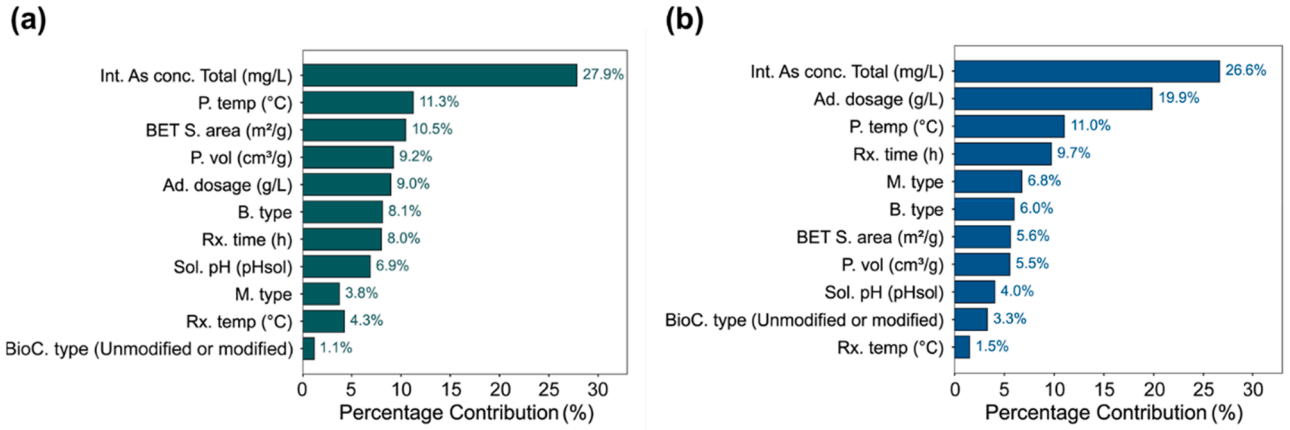


Fig. 6. SHAP analysis-based listing of significant input variables for the adsorption of As(III) (a) and As(V) (b) on biochar. Initial As concentration has the most significant impact on its adsorption on the biochar. P. temp & Ad. dosage and BET S. area & P. temp are the second and third-most significant input variables impacting As adsorption on biochar.

given problem (Deb et al., 2002). Additional insights into the operational principles the GA solver are available in earlier study (Deb et al., 2002). In addressing the optimization problem, the GA solver employed specific parameter configurations: a cap of 500 iterations, a 0.1 mutation likelihood, an elitism ratio of 0.01, a 0.5 crossover probability, a parent segment of 0.3, and utilized the uniform crossover technique. These parameter values remain consistent across the GA solver for determining the optimal operational conditions for both As(III) and As(V), except for the population size, set at 200 for As(III) and 400 for As(V). For identify energy-efficient operational conditions for As(III) and As(V) adsorption on different materials under varying loading rates of initial As concentration, 10 different values were selected for the initial As concentration representing incremental 10th percentile values within the initial As concentration range. For each specific initial As value, the optimization problem is iteratively solved 10 times to ascertain a robust and efficient solution, considering modelling inaccuracies and the solver's capabilities in estimating the solution. The progression of objective value enhancement across iterations from the 10 iterations of solving the objective function is illustrated in Fig. 7 for both As(III) and As(V). The iterative search for solutions appears robust across different As loading rates for As(III), as indicated by tighter prediction intervals compared to those of As(V). Furthermore, the maximum As adsorption achieved through multiple rounds of optimization problem solving is estimated while considering energy implications.

In this study, the solver achieved the highest concentration value of 291.66 ± 7 mg/g for As(III) when the initial As concentration of was 153.40 mg/L. This was accomplished by employing C-Layered Double Hydroxide material combined with reduced graphene oxide. For As(V), a peak concentration of 271.56 ± 22.5 mg/g was observed at an initial concentration of 100 mg/L, using Chitosan components in conjunction with rice straw biochar (Fig. 7). Therefore, the maximum adsorption of As under different initial As concentrations considering the minimal energy consumption allows to design the process on the optimisation led selection of the material that yields the optimal performance for As adsorption under the process conditions.

3.5. Public deployment of ML models

After completing the lifecycle of a ML project, the subsequent crucial phase involves deployment, especially when catering to the research community for the development of water treatment systems. Web deployment stands out as a user-friendly and easily accessible method, facilitating broad access. Utilizing Streamlit serves as the framework for both the frontend and backend of the web application. Github acts as the hosting platform for the application on the cloud, while Streamlit cloud services are employed to host the machine, providing an expandable domain for the website. The Streamlit-based interface, supported by trained ML models, is user-friendly and enables users to input data on

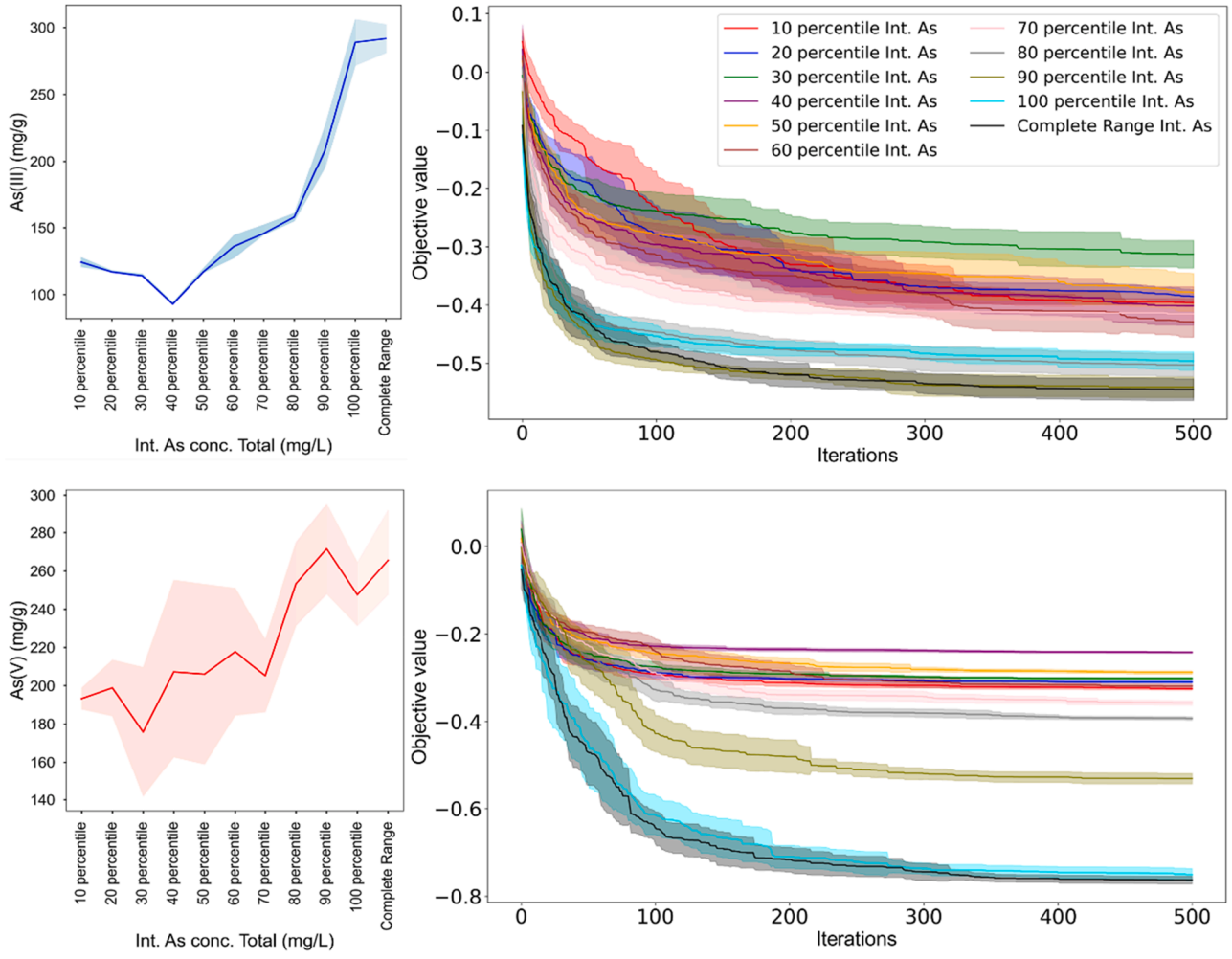


Fig. 7. Optimisation results of Genetic Algorithm of input variables for the adsorption of As(III) and As(V) on biochar. Initial As concentration has the most significant impact. Different value constraints for different scenarios. Optimisation results of Genetic Algorithm of input variables for the adsorption of As(III) and As(V) on biochar. As(III) and As(V) concentrate against different constraints values of Int. As conc. Highest As(III) achieved at Int. As conc of 153.40 mg/L and As(V) at Int. As conc of 100 mg/L.

the process conditions and material type into tabs to estimate the adsorption of As(III) and As(V). Streamlit is well-regarded in the ML community for its streamlined deployment of ML models, effective query management, and seamless interaction with ML models, thereby enhancing user experience for specific inquiries.

Given the continuous updates to Python libraries, there are no anticipated compatibility concerns. The existing setup, designed for straightforward and low-profile hosting, accommodates up to 100 simultaneous visitors, surpassing the anticipated peak of 10 visitors. The system utilizes containerization and launches a separate instance of the code for each user accessing the application. Our commitment involves the ongoing operation of the application and the regular updating of the model with new data, employing transfer learning techniques to optimize ML applications for both research and real-world scenarios. Therefore, individuals with minimal prior experience in ML-based modeling can engage with the developed web application to predict the adsorption of As based on process conditions and material types. This capability can significantly aid in the advancement of water treatment technologies.

Fig. 8 visually represents the web application. Below are the links for accessing these web applications. As(III) concentration prediction: https://as-iii-concentration-prediction.streamlit.app/; As(V) concentration prediction: https://as-v-concentration-prediction.streamlit.app/. Therefore, by leveraging our model and the developed Streamlit interface, we can input different influencing factors and materials to ascertain the As(III) and As(V) adsorption efficiency. To illustrate the practical application of the developed web-app for predicting the adsorption of As on biochar, three laboratory experiments were conducted and the experimental adsorption of As was compared with the web-based predicted values. The biochar material was modified, with ZrO_2 specified for As(III) and As(V) in the experiments. The experimental conditions are detailed in the supplementary data provided under "Table S5 Application Data". Fig. 9 presents a comparison between the experimental and web-app based predicted values of As(III) and As(V) for the three experiments conducted under varying conditions. A significant agreement is observed between the experimental and web-app based predicted values for As(III) and As(V) across the three experimental conditions. Future work will involve data augmentation and model updates using literature data to further improve the capabilities and applicability of the web-app for predicting As adsorption on different materials. The proposed methodology and workflow have the potential to reduce the experimental workload and resource consumption, offering valuable insights and applications for the remediation of As(III) and As(V) in contaminated-water environments.

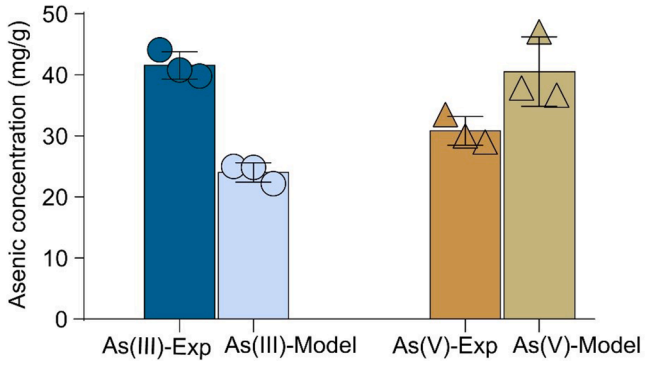


Fig. 9. Real-world application of deployed web-app predicting adsorption of As (III) and As(V) on modified biochar.

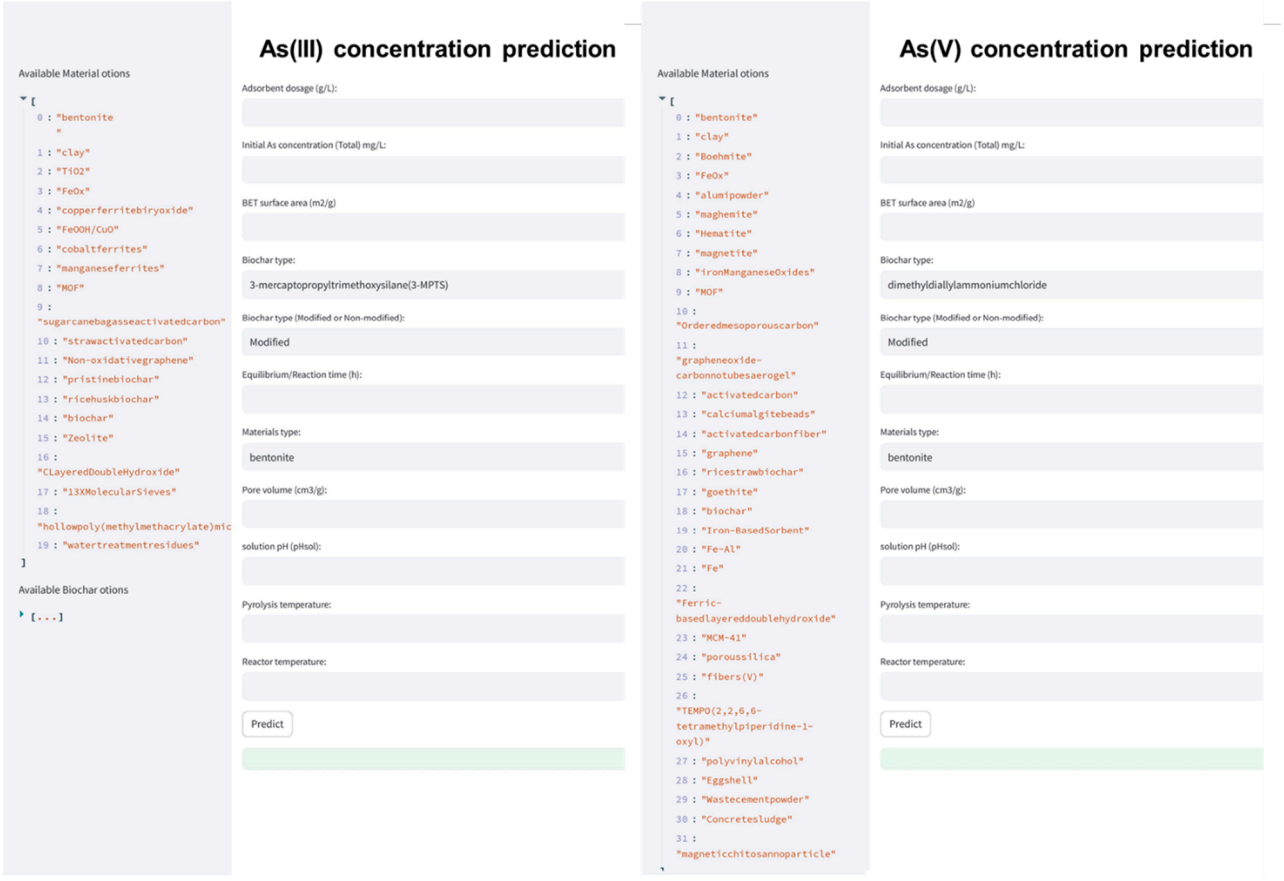


Fig. 8. Web application for predicting adsorption of As(III) and As(V) on biochar.

4. Conclusion

Arsenic contamination in water is a global concern due to its detrimental impact on public health. The utilization of various materials for the removal of both As(III) and As(V) from polluted water has emerged as a promising and cost-effective technique with numerous benefits, including safeguarding public health, environmental preservation, carbon sequestration, and economic advantages. The ML models have the potential to enhance the water treatment system for As removal by leveraging adsorption on diverse materials, an aspect that remains largely unexplored in the existing literature. ML models, trained on extensive literature data, undergo rigorous procedures for training and evaluation to accurately predict the adsorption behavior of As(III) and As(V) under different operational conditions. Noteworthy is the exceptional performance of CatBoost models, surpassing a 98 % accuracy in forecasting the As(III) and As(V) adsorption using diverse materials in aqueous settings. SHAP analysis revealed that the initial As(III) and As (V) levels wield the greatest influence on their adsorption onto varied materials. Through the utilization of a genetic algorithm to optimize operational parameters and select suitable materials, the models predicted the maximum As(III) and As(V) adsorption at varying initial As concentrations. Web-based tools have been developed to predict the adsorption behavior of As(III) and As(V) on a range of materials, aiding in the design of water treatment processes while considering multiple process and material-related factors. The amalgamation of ML and optimization techniques has significantly bolstered our ability to tackle the critical challenge of As contamination in water sources, thereby advancing public health and environmental sustainability.

Associated content

Author Information.

CRediT authorship contribution statement

Jinsheng Huang: Writing – original draft, Methodology, Investigation, Data curation. **Waqar Muhammad Ashraf:** Writing – review & editing, Visualization, Software, Resources, Methodology. **Talha Ansar:** Visualization, Methodology. **Muhammad Mujtaba Abbas:** Visualization, Validation, Software. **Mehdi Tlija:** Visualization, Validation, Software. **Yingying Tang:** Investigation, Data curation. **Yunxue Guo:** Writing – review & editing, Conceptualization. **Wei Zhang:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Funding acquisition, 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.

Acknowledgments

This work was supported by the Outstanding Youth Project of Guangdong Natural Science Foundation (2022B1515020030). The authors extend their appreciation to King Saud University for funding this work through Researchers Supporting Project (RSPD2025R685), King Saud University, Riyadh, Saudi Arabia.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.watres.2024.122815.

Data availability

Data will be made available on request.

References

Abdi, J., Mazloom, G., 2022. Machine learning approaches for predicting arsenic adsorption from water using porous metal-organic frameworks. Sci. Rep. 12, 16458.
Aftabtalab, A., Rinklebe, J., Shaheen, S.M., Niazi, N.K., Moreno-Jimenez, E., Schaller, J., Knorr, K.H., 2022. Review on the interactions of arsenic, iron (oxy)(hydr)oxides, and dissolved organic matter in soils, sediments, and groundwater in a ternary system. Chemosphere 286, 131790.
Ali, H., Ahmed, S., Hsini, A., Kizito, S., Naciri, Y., Djellabi, R., Abid, M., Raza, W., Hassan, N., Rehman, M.S.U., Khan, A.J., Khan, M., Ul Haq, M.Z., Aboagye, D., Irshad, M.K., Hassan, M., Hayat, A., Wu, B., Qadeer, A., Ajmal, Z., 2022. Efficiency of a novel nitrogen-doped FeO impregnated biochar (N/FeO@BC) for arsenic (III and V) removal from aqueous solution: insight into mechanistic understanding and reusability potential. Arab. J. Chem. 15 (11), 104209.
Ashraf, W.M., Dua, V., 2024a. Data Information integrated Neural Network (DINN) algorithm for modelling and interpretation performance analysis for energy systems. Energy Ai 16, 100363.
Ashraf, W.M., Dua, V., 2024b. Partial derivative-based dynamic sensitivity analysis expression for non-linear auto regressive with exogenous (NARX) model-case studies on distillation columns and model's interpretation investigation. Chem. Eng. J. Adv. 18, 100605.
Ashraf, W.M., Uddin, G.M., Arafat, S.M., Krzywanski, J., Wang, X.N., 2021. Strategic-level performance enhancement of a 660 MWe supercritical power plant and emissions reduction by AI approach. Energy Convers. Manag. 250, 114913.
Brickson, B.E., 2003. Field kits fail to provide accurate measure of arsenic in groundwater. Environ. Sci. Technol. 37 (1), 35a–38a.
Chen, T.Q., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. Kdd'16:. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794.
Cui, D., Zhang, P., Li, H.P., Zhang, Z.X., Song, Y., Yang, Z.G., 2020. The dynamic effects of different inorganic arsenic species in crucian carp (*Carassius auratus*) liver during chronic dietborne exposure: bioaccumulation, biotransformation and oxidative stress. Sci. Total Environ. 727, 138737.
Deb, K., Pratap, A., Agarwal, S., Meyarivan, T., 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. 6, 182–197.
Gallios, G.P., Tolkou, A.K., Katsoyiannis, I.A., Stefusova, K., Vaclavikova, M., Deliyanni, E.A., 2017. Adsorption of arsenate by nano scaled activated carbon modified by iron and manganese oxides. Sustainability 9, 1684. -Basel.
Gohr, M.S., Abd-Elhamid, A.I., El-Shanshory, A.A., Soliman, H.M.A., 2022. Adsorption of cationic dyes onto chemically modified activated carbon: kinetics and thermodynamic study. J. Mol. Liq. 346, 118227.
Golde, C.E., Rothrock, M.J., Mishra, A., 2019. Comparison between random forest and gradient boosting machine methods for predicting spp. prevalence in the environment of pastured poultry farms. Food Res. Int. 122, 47–55.
Han, L.F., Sun, K., Yang, Y., Xia, X.H., Li, F.B., Yang, Z.F., Xing, B.S., 2020. Biochar's stability and effect on the content, composition and turnover of soil organic carbon. Geoderma 364, 114184.
Hou, D.Y., Al-Tabbaa, A., O'Connor, D., Hu, Q., Zhu, Y.G., Wang, L.W., Kirkwood, N., Ok, Y.S., Tsang, D.C.W., Bolan, N.S., Rinklebe, J., 2023. Sustainable remediation and redevelopment of brownfield sites. Nat. Rev. Earth Environ. 4 (4), 271–286.
Ibrahim, B., Ewusi, A., Ahenkorah, I., 2022a. 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. -Sui.
Ibrahim, B., Ewusi, A., Ahenkorah, I., Ziggah, Y.Y., 2022b. Modelling of arsenic concentration in multiple water sources: a comparison of different machine learning methods. Groundw. Sustain. Dev. 17, 100745.
Igalavithana, A.D., Kwon, E.E., Vithanage, M., Rinklebe, J., Moon, D.H., Meers, E., Tsang, D.C.W., Ok, Y.S., 2019. Soil lead immobilization by biochars in short-term laboratory incubation studies. Environ. Int. 127, 190–198.
Ishfaq, K., Asad, M., Ashraf, W.M., Sana, M., Anwar, S., Zhang, W., Dua, V., 2024. Towards artificial intelligence empowered performance enhancement of EDM process using nano-graphene mixed bio-dielectric supporting the carbon neutrality and sustainable development. J. Clean. Prod. 457, 142482.
Ishfaq, K., Sana, M., Ashraf, W.M., 2023. Artificial intelligence-built analysis framework for the manufacturing sector: performance optimization of wire electric discharge machining system. Int. J. Adv. Manuf. Technol. 128 (11–12), 5025–5039.
Jaffari, Z.H., Abbas, A., Kim, C.M., Shin, J., Kwak, J., Son, C., Lee, Y.G., Kim, S., Chon, K., Cho, K.H., 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., 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.
Joshi, M., Bhatt, D., Srivastava, A., 2023. Enhanced adsorption efficiency through biochar modification: a comprehensive review. Ind. Eng. Chem. Res. 62 (35), 13748–13761.
Khan, Z.H., Gao, M.L., Qiu, W.W., Qaswar, M., Islam, M.S., Song, Z.G., 2020. The sorbed mechanisms of engineering magnetic biochar composites on arsenic in aqueous solution. Environ. Sci. Pollut. Res. 27 (33), 41361–41371.

Li, Z.H., Jean, J.S., Jiang, W.T., Chang, P.H., Chen, C.J., Liao, L.B., 2011. Removal of arsenic from water using Fe-exchanged natural zeolite. J. Hazard. Mater. 187 (1–3), 318–323.
Liu, J.X., Xu, Z.L., Zhang, W.J., 2023. Unraveling the role of Fe in As(III & V) removal by biochar via machine learning exploration. Sep. Purif. Technol. 311, 123245.
Lombard, M.A., Bryan, M.S., Jones, D.K., Bulka, C., Bradley, P.M., Backer, L.C., Focazio, M.J., Silverman, D.T., Toccalino, P., Argos, M., Gribble, M.O., Ayotte, J.D., 2021. Machine learning models of arsenic in private wells throughout the conterminous United States as a tool for exposure assessment in human health studies. Environ. Sci. Technol. 55 (8), 5012–5023.
Matschullat, J., 2000. Arsenic in the geosphere-a review. Sci. Total Environ. 249 (1–3), 297–312.
Michael, H.A., 2013. An Arsenic Forecast for China. Science 341 (6148), 852–853 (1979).
Mirza, N.H., Fujino, T., 2024. Aqueous arsenic (III) removal using a novel solid waste based porous filter media block: traditional and machine learning (ML) approaches. Desalin. Water Treat. 319, 100536.
Palansooriya, K.N., Li, J., Dissanayake, P.D., Suvarna, M., Li, L.Y., Yuan, X.Z., Sarkar, B., Tsang, D.C.W., Rinklebe, J., Wang, X.N., Ok, Y.S., 2022. Prediction of soil heavy metal immobilization by biochar using machine learning. Environ. Sci. Technol. 56 (7), 4187–4198.
Peng, Y.R., Azeem, M., Li, R.H., Xing, L.B., Li, Y.M., Zhang, Y.C., Guo, Z.Q., Wang, Q., Ngo, H.H., Qu, G.Z., Zhang, Z.Q., 2022. Zirconium hydroxide nanoparticle encapsulated magnetic biochar composite derived from rice residue: application for As(III) and As(V) polluted water purification. J. Hazard. Mater. 423, 127081.
Podgorski, J., Berg, M., 2020. Global threat of arsenic in groundwater. Science 368 (6493), 845–850 (1979).
Rahman, M.G., Islam, M.Z., 2011. A Decision Tree-Based Missing Value Imputation Technique For Data Pre-Processing. Australian Computer Society Inc, pp. 41–50.
Roh, J., Park, H., Kwon, H., Joo, C., Moon, I., Cho, H., Ro, I., Kim, J., 2024. Interpretable machine learning framework for catalyst performance prediction and validation with dry reforming of methane. Appl. Catal. B Environ. 343, 123454.
Sakhiya, A.K., Kaushal, P., Vijay, V.K., 2023. Potential of rice straw derived activated biochar to remove arsenic and manganese from groundwater: a cleaner approach in the Indo-Gangetic Plain. Appl. Surf. Sci. Adv. 17, 100443.
Sarkar, A., Paul, B., 2020. Evaluation of the performance of zirconia-multiwalled carbon nanotube nanoheterostructures in adsorbing As(III) from potable water from the perspective of physical chemistry and chemical physics with a special emphasis on approximate site energy distribution. Chemosphere 242, 125234.
Sarkar, A., Paul, B., Darbha, G.K., 2022. The groundwater arsenic contamination in the Bengal Basin-A review in brief. Chemosphere 299, 134369.
Shi, L., Li, J., Palansooriya, K.N., Chen, Y.H., Hou, D.Y., Meers, E., Tsang, D.C.W., Wang, X.N., Ok, Y.S, 2023. Modeling phytoremediation of heavy metal contaminated soils through machine learning. J. Hazard. Mater. 441, 129904.
Suvarna, M., Araújo, T.P., Pérez-Ramírez, J., 2022. A generalized machine learning framework to predict the space-time yield of methanol from thermocatalytic CO_2 hydrogenation. Appl. Catal. B Environ. 315, 121530.
Tariq, R., Mohammed, A., Alshibani, A., Ramírez-Montoya, M.S., 2024. Complex artificial intelligence models for energy sustainability in educational buildings. Sci. Rep. 14 (1), 15020.
Uddin, G.M., Arafat, S.M., Kazim, A.H., Farhan, M., Niazi, S.G., Hayat, N., Zeid, I., Kamarthi, S., 2019. Artificial intelligence-based Monte-Carlo numerical simulation of aerodynamics of tire grooves using computational fluid dynamics. AI EDAM 33 (3), 302–316.
Vovk V., Gammerman A., Saunders C. 1999. Machine-learning applications of algorithmic randomness. Machine Learning, Proceedings, 444–453.
Wang, D.N., Li, L., Zhao, D., 2022. Corporate finance risk prediction based on LightGBM. Inf. Sci. 602, 259–268 (Ny).
Wongrod, S., Simon, S., van Hullebusch, E.D., Lens, P.N.L., Guibaud, G., 2018. Changes of sewage sludge digestate-derived biochar properties after chemical treatments and influence on As(III and V) and Cd(II) sorption. Int. Biodeterior. Biodegradation 135, 96–102.
Ye, M., Zhu, L., Li, X.J., Ke, Y.H., Huang, Y., Chen, B.B., Yu, H.L., Li, H., Feng, H., 2023. Estimation of the soil arsenic concentration using a geographically weighted XGBoost model based on hyperspectral data. Sci. Total Environ. 858, 159798.
Yuan, X.Z., Suvarna, M., Lim, J.Y., Pérez-Ramírez, J., Wang, X.N., Ok, Y.S., 2024. Active Learning-Based Guided Synthesis of Engineered Biochar for CO_2 Capture. Environ. Sci. Technol. 58, 6628–6636.
Zhang, J.C., Huang, L.P., Ye, Z.J., Zhao, Q.Y., Li, Y.J., Wu, Y., Zhang, W., Zhang, H.G., 2022a. Removal of arsenite and arsenate from contaminated water using Fe-ZrO-modified biochar. J. Environ. Chem. Eng. 10 (6), 108765.
Zhang, W., Ashraf, W.M., Senadheera, S.S., Alessi, D.S., Tack, F.M.G., Ok, Y.S., 2023. Machine learning based prediction and experimental validation of arsenite and arsenate sorption on biochars. Sci. Total Environ. 904, 166678.
Zhang, W., Cho, Y., Vithanage, M., Shaheen, S.M., Rinklebe, J., Alessi, D.S., Hou, C.H., Hashimoto, Y., Withana, P.A., Ok, Y.S., 2022b. Arsenic removal from water and soils using pristine and modified biochars. Biochar. 4, 55.
Zhou, N., Chen, H.G., Xi, J.T., Yao, D.H., Zhou, Z., Tian, Y., Lu, X.Y., 2017. Biochars with excellent Pb(II) adsorption property produced from fresh and dehydrated banana peels via hydrothermal carbonization. Bioresour. Technol. 232, 204–210.
Zhu, X., Li, Y., Wang, X., 2019a. Machine learning prediction of biochar yield and carbon contents in biochar based on biomass characteristics and pyrolysis conditions. Bioresour. Technol. 288, 121527.
Zhu, X., Wang, X., Ok, Y.S., 2019b. The application of machine learning methods for prediction of metal sorption onto biochars. J. Hazard. Mater. 378, 120727.