Mirza *et al. Water Emerg Contam Nanoplastics* 2024;3:24
DOI: 10.20517/wecn.2024.70

Water Emerging Contaminants & Nanoplastics

Research Article

Open Access

Check for updates

Prediction of arsenic (III) adsorption from aqueous solution using non-neural network algorithms

Nazmul Hassan Mirza, Takeshi Fujino

Department of Environmental Science and Technology, Saitama University, Saitama 338-8570, Japan.

Correspondence to: Prof. Takeshi Fujino, Department of Environmental Science and Technology, Graduate School of Science and Engineering, Saitama University, 255 Shimo-Okubo, Sakura-Ku, Saitama 338-8570, Japan. E-mail: fujino@mail.saitama-u.ac.jp

How to cite this article: Mirza NH, Fujino T. Prediction of arsenic (III) adsorption from aqueous solution using non-neural network algorithms. *Water Emerg Contam Nanoplastics* 2024;3:24. https://dx.doi.org/10.20517/wecn.2024.70

Received: 24 Oct 2024 **First Decision:** 12 Nov 2024 **Revised:** 18 Nov 2024 **Accepted:** 22 Nov 2024 **Published:** 10 Dec 2024

Academic Editors: Zhijie Chen, Roberto Rosal **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

Abstract

Heavy metals such as arsenic can be effectively removed through adsorption. Through material property evaluation and adsorption parameter optimization, machine learning (ML) modeling provides an alternative to lengthy laboratory experimentation. In this work, adsorption data from an earlier study employing a waste-material composite were used. To create prediction models, four non-neural network algorithms - support vector machines (SVM), Gaussian process regression (GPR), linear regression, and ensemble approaches - were used and contrasted with neural network algorithms. Nine predictors were utilized, ranging from adsorbent composition alterations to experimental circumstances. Using principal component analysis (PCA) and feature selection, together with the F-test and minimum redundancy maximum relevance (MRMR) algorithms for feature reduction, optimization was accomplished. With an R-squared of 0.939, mean absolute error (MAE) of 5.778, and root mean squared error (RMSE) of 7.119 for training and an R-squared of 0.942, MAE of 5.450, and RMSE of 6.870 for testing, the optimized GPR method offered the best predictive performance. The best R-squared values found for other algorithms were: SVM (0.922), linear regression (0.925), and ensemble (0.927). The most important variables influencing adsorption efficiency were initial arsenic concentration, time, and the iron salt content. Local interpretable model-agnostic explanations (LIME), partial dependence plot (PDP), and Shapley additive explanations (SHAP) plots were used to explain these results. This work shows that, based on model-derived parameters, non-neural network algorithms may efficiently simulate and optimize arsenic adsorption tests, providing a trustworthy substitute for neural network techniques and markedly increasing adsorption efficiency.

Keywords: Adsorption, arsenic removal, machine learning, non-neural network regression, water quality

INTRODUCTION

The ecosystem and human health can suffer greatly when toxic heavy metals, even in trace amounts, are released into groundwater[1]. Most often, various industrial and agricultural activities, urbanization, and other factors release this type of hazardous element into the land, water, and atmosphere[2]. Arsenic (As) is one of the most well-known heavy metals for its toxicity. The current estimate of the affected population worldwide is approximately 150 million, which is certain to rise as more afflicted places are identified[3]. All natural waterways contain arsenic, which can originate from biological processes as well as natural processes occurring in the Earth's crust[4]. There are two primary types of inorganic arsenic: arsenite [As(III)] and arsenate [As(V)][5]. Arsenate is the main species present in naturally occurring surface water basins, while arsenite is the predominant form in groundwater[6]. While As(III) and As(V) are not equally dangerous, As(III) is approximately 60 times more poisonous[7]. Since As(III) frequently exists in neutral, undissociated forms, removal of it is quite challenging[8]. Arsenic poses a global health danger to humans. Long-term exposure to arsenic, mostly from tainted food and water, can cause serious illnesses such as lung, kidney, liver, and skin cancers[9]. The World Health Organization (WHO) has set a maximum allowable concentration of arsenic in drinking water at 10 μg/L[10].

Hence, it is still imperative to completely eradicate this toxic heavy metal. In addition to biological approaches, coagulation, precipitation, reverse osmosis, filtering, and adsorption are some of the ways that have been developed to remove arsenic[11,12]. Adsorption is arguably the most successful strategy among these[13]. With modified adsorbents such as chars, activated carbons, activated aluminium, thiol-functionalized chitin nanofibers, goethite-based adsorbent, zero-valent iron, synthetic siderite, titanium dioxide, and many more, adsorption is a commonly used physical separation method for the removal of hazardous heavy metals from groundwater[14]. For the removal of both organic and inorganic contaminants, adsorption is the method of choice in the water and wastewater sectors due to its ease of handling, low sludge production, affordability, and regeneration potential[15]. Adsorbents derived from solid waste are frequently employed in water and wastewater treatment procedures[16-18]. Moringa waste, a type of agricultural solid waste, can serve as a biosorbent to adsorb heavy metals[19]. Various parts of the Moringa oleifera plant, including seeds, leaves, and bark, have also been shown to possess the ability to remove heavy metals[20,21]. Metal salts, such as iron and aluminum, are frequently used to remove arsenic[22]. On the other hand, porous composite materials formed of metal salts combined with solid waste are employed to extract heavy metals. It is crucial to evaluate the many control elements affecting adsorption and overall efficiency in order to optimize composite adsorbents. Historically, mathematical models have been employed to optimize efficiency and evaluate the influence of parameters[23,24].

Currently, the water and wastewater treatment industry is increasingly employing machine learning (ML) to forecast treatment procedures, thanks to improvements in computational efficiency resulting from enhanced hardware and software[25-28]. Developing high-performance solid adsorbents typically involves intuition-based, factor-by-factor techniques, with response surface procedures serving as the primary means of experimental design[29]. The aqueous phase adsorption of heavy metals involves many factors, including the interaction between adsorbents and adsorbates and ambient variables[30,31]. A growing number of ML algorithms, including artificial neural networks (ANN), tree-based techniques, and support vector regression, are being used to simulate and examine the processes of heavy metal adsorption by solid adsorbents[32-34]. However, some research has been undertaken without considering appropriateness or context[35,36] and several papers omitted the findings of the present experiment and instead only incorporated a limited portion of data from previous publications[37,38]. Adsorption isotherms and batch studies are two common methods used to analyze the adsorption behavior of heavy metals and organic pollutants in the aqueous phase. However, both methods can be laborious and ineffective[39]. ML can be used to automatically

map complex systems and produce the mapping relationship between input and output variables[40]. The foundation of ML is sufficient and high-quality data, which effect model performance[41]. The ANN was first created to replicate the workings of the human brain using a rigorous data analysis technique. Consequently, numerous environmental studies pertaining to water pollution have been using an expanding number of different ANN methodologies. The ANN technique was effectively applied by Mandal *et al.* to model an As(III) removal process[42]. According to Sakizadeh, ANN have various shortcomings, especially in the environmental sciences, despite their effective application in many modeling studies[43]. One of the main constraints is the lack of sufficient data records. Training on a limited dataset may result in overfitting, a situation in which the model exhibits good generalization ability when it performs well on training data but badly on unseen data. In addition to ANN, the applications of other advanced machine learning algorithms (MLAs) such as support vector machine (SVM)[44,45] random forest (RF)[46], gradient boosted regression tree[47], Bayesian network[48], adaptive network-based fuzzy inference system[49], and Ensemble models[50] have been proven to be accurate and useful for different learning problems in multiple domains. A recent study investigated the adsorption of arsenic (III) using a porous filter media block (PFMB). The study employed classical isotherm and kinetic models alongside ANN modeling to analyze the adsorption process[51]. A PFMB made of biochar and moringa bark as bio-adsorbents and enhanced with sodium bicarbonate, ferric chloride, aluminum sulfate, and commercial gypsum was examined for its capacity to adsorb arsenic. Seven different PFMB samples, each prepared with different amounts of aluminum-iron salts and moringa-based bio-adsorbents (biomass and biochar), were used in batch studies. The experimental parameters encompassed different agitation rates, pH levels, duration, initial concentration, and quantity of adsorbent. This study will employ identical data sets to assess non-neural network approaches in forecasting arsenic adsorption and juxtapose them with the neural network model.

This study aims to predict the adsorption capacity of the PFMB using a total of four non-neural network machine-learning algorithms. The best non-neural network algorithms will be evaluated and compared with neural network results from the previous study. ANN were used in the previous study; although they are successful, they are computationally demanding and perform best with huge datasets. In order to overcome those constraints, this paper investigates non-neural network approaches for assessing adsorption trials. It draws attention to the effectiveness, relevance, and insights offered by non-neural models by contrasting neural with non-neural methods. The study provides a unique comparative examination of different approaches, with a focus on algorithm selection, model performance optimization, and finding critical experimental conditions. The efficacy of four non-neural network strategies - SVM, ensemble approaches, Gaussian process regression (GPR), and linear regression - in predicting adsorption capacity and identifying contributors in the PFMB will be evaluated. To provide a thorough explanation of the model results, the study will make use of Shapley additive explanations (SHAP), partial dependence plot (PDP) analysis, and local interpretable model-agnostic explanations (LIME).

METHODS

Data collection and preprocessing

A total of 507 data sets were gathered from batch experiments conducted in a previous study to facilitate ML analysis[51]. The dataset included nine predictive parameters, with arsenic uptake capacity serving as the response parameter [Table 1]. Modeling was performed using MATLAB R2023b, implementing four non-neural network algorithms: linear regression, SVM, ensemble methods, and GPR.

The first step in non-neural network modeling after data gathering is data preprocessing, which cleans and formats the data. Data preprocessing is an essential phase in the ML process that involves cleaning, transforming, and preparing raw data for training ML models[52]. This phase has a major impact on the final

Table 1. Data collection for the non-neural network analysis

Parameters	Ranges	Nos	Role
Initial concentration of arsenic [As(III)] in ppm (initial conc.)	0.25-4.0	507	Predictor
Amount of adsorbent used (g per 50 mL solution) (adsorbent)	0.25-1.0	507	Predictor
pH of the solution (pH)	2.5-11.3	507	Predictor
Shaking speed in revolutions per minute (rpm) (shaking)	100-200	507	Predictor
Contact time in minutes (time)	30-300	507	Predictor
$FeCl_3 \cdot 6H_2O$ used in PFMB preparation (g per 50 g of media) (Fe/Iron salt)	0.25-0.75	507	Predictor
$[Al_2(SO_4)_3]$ used in PFMB preparation (g per 50 g of media) (Al salt)	0.25-0.75	507	Predictor
Moringa bark biomass included in PFMB (g per 50 g of media) (M. Biomass)	0.25-0.75	507	Predictor
Moringa bark biochar included in PFMB (g per 50 g of media) (M. Biochar)	0.25-0.75	507	Predictor
Adsorption capacity for As(III) in micrograms per gram (μg/g) (arsenic uptake)		507	Response

PFMB: Porous filter media block.

models' efficacy and performance. Among the crucial phases in data preprocessing are data cleaning, data transformation, feature engineering, data normalization, scaling, normalizing, and encoding categorical information into numerical representations, which are common transformations[53,54]. Each predictor's significance was examined independently using the F-test technique, and the features were ranked based on the analysis of the p-values obtained from the F-test findings. In ML and data analysis, the minimum redundancy maximum relevance (MRMR) test is a feature selection technique that helps find a subset of characteristics with high relevance to the target variable while preserving low redundancy among the features chosen. The best feasible set of characteristics that can maximally and mutually diverge from one another and accurately represent the response variable was chosen using the MRMR algorithm. Principal component analysis (PCA) is a dimensionality reduction approach used in feature importance analysis that can be used to characterize the association between descriptors and output variables[55]. Information gain is used to demonstrate the significance of the descriptors[56]. Most of the information present in a big set of variables can be retained when converting it into a smaller set of variables using PCA[57]. A new variable that corresponds to a linear combination of the original variables is the resultant main component[58]. In regression learners, dimensionality reduction can produce regression models that lessen the risk of overfitting. Next, the data are divided into testing and training sets. The model is trained using algorithms such as ensemble techniques, GPR, SVM, and linear regression on the training data. Reliability is ensured using validation techniques such as cross-validation, and the model's performance is assessed on the testing set. Lastly, the model's predictions are interpreted and important characteristics influencing arsenic adsorption are identified using respective techniques [Figure 1].

Non-neural network algorithms

Adsorption behavior can be modeled by researchers using ML techniques, considering a range of parameters including arsenic concentrations, adsorbent properties, and environmental factors. Non-neural network ML algorithms have shown great promise in the field of arsenic adsorption modeling utilizing PFMBs[59]. Cost-effective and practical descriptors for assessing adsorbent properties and predicting adsorption efficiency encompass macro-level factors such as operational parameters (adsorbent dosage, adsorbate concentration, contact time, and stirring speed)[60] and synthesis conditions (composition and proportion of raw materials, synthesis duration, and pH)[61]. In model training, output variables that are easily accessible and have cost-effective experimental attributes, such as adsorption capacity, removal rate, and efficiency, are commonly used[62,63]. The main goal of regression analysis is to identify a linear, planar, or hyperplanar model that minimizes the gap between the observed and predicted values[64,65]. ML can be classified into four main types based on the nature of the datasets and their corresponding labels:

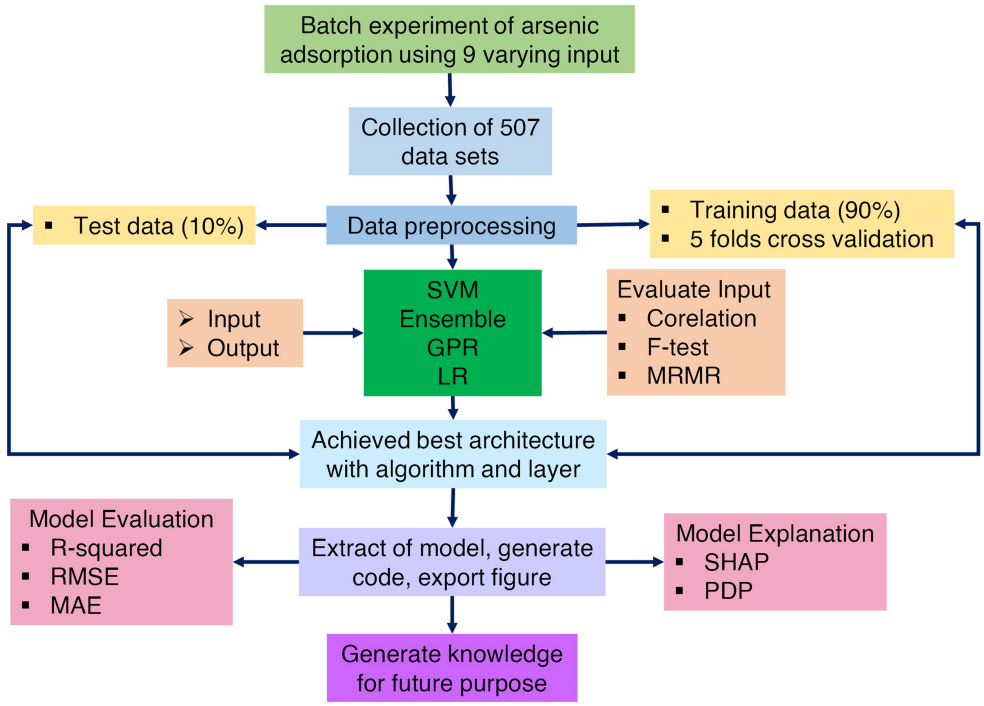

Figure 1. Schematic flow chart of ML toolbox. ML: Machine learning; SVM: support vector machines; GPR: Gaussian process regression; MRMR: minimum redundancy maximum relevance; RMSE: root mean squared error; MAE: mean absolute error; SHAP: Shapley additive explanations; PDP: partial dependence plot.

supervised, unsupervised, semi-supervised, and reinforcement learning[66]. The objective of this study is to utilize supervised ML to target the adsorption capacity. The techniques used for estimating adsorption capacity and identifying major contributing factors include ensemble methods, GPR, SVM, and linear regression. These techniques provide robust frameworks for these purposes [Table 2].

Linear regression is a simple and widely used statistical method for modeling the relationship between a dependent variable (target) and one or more independent variables (features) by fitting a linear equation to the observed data. A linear relationship between the independent variable X and the dependent variable y is assumed by the linear regression model. The model can be expressed as follows:

$$y = \beta 0 + \beta 1 \cdot x1 + \beta 2 \cdot x2 + \ldots + \beta n \cdot xn + \epsilon$$

When the error term is represented by ϵ, the coefficients for the independent variables are xi, the intercept term is $\beta 0$, and the coefficients for the independent variables are βi. According to linear regression, there must be a linear relationship between the variables, a normally distributed distribution of the errors, and a constant variance (homoscedasticity) of the errors. Among various available techniques, linear, interactions linear (IL), robust linear, and stepwise linear (SL) regression were evaluated in this study. The two best-performing models, interaction linear regression and stepwise linear regression, were chosen for further evaluation with varying input features, both with and without the application of PCA.

SVM regression concentrates on identifying a hyperplane that captures a given fraction of the data inside a margin around the predicted values[57], in contrast to classic regression models that seek to directly reduce the mean squared error. Finding the hyperplane that best fits the data while permitting some variation or error is the goal of SVM regression[67]. SVM maps the input space derived from the independent variables in

Table 2. Algorithms evaluated for each model

GPR	SVM	Linear regression	Ensembles of trees
Rational quadratic GPR	Linear SVM	Linear	Boosted trees
Squared exponential GPR	Quadratic SVM	Interactions linear	Bagged trees
Matern 5/2 GPR	Cubic SVM	Robust linear	Optimizable trees
Exponential GPR	Fine/medium/coarse SVM	Stepwise linear	
Optimizable GPR	Optimizable SVM		

GPR: Gaussian process regression; SVM: support vector machines.

a kernel function and focuses mostly on the border between classes[68]. The epsilon (ε), kernel function, and penalty parameter are all connected to the prediction accuracy of SVR. The parameter ε, which is utilized for training data fitting, governs the function of the ε-insensitive zone width. From available different presets, linear, quadratic, cubic, fine/medium/coarse and optimizable SVM were evaluated and the best one was selected as SVM. SVM, which operates based on statistical learning theory, is regarded as an exacting ML methodology[69]. SVM has been considered for regression analysis because of its great flexibility and a small number of tuning factors[70].

Ensemble methods, a supervised ML modeling technique for simulating the heavy metal adsorption process, are tree-based techniques[62]. It incorporates a variety of techniques, including decision trees, gradient boosting (GB), and RF, and can adapt to both classification and regression scenarios[71]. One of the ensemble learning methods, the boosting technique, is used to train the GB method[72]. By minimizing the prediction error for a regression problem, boosting is usually used to combine several weak prediction models into a final strong model with improved predictive performance. Bootstrap sampling is used in the RF algorithm to randomly choose the input variables and split the training data so that the trees develop independently of one another[73]. Cross-validation and other similar approaches are frequently used in hyperparameter tuning[74]. Among various Ensembles of trees, boosted trees, bagged trees, and optimizable trees were evaluated in this study before selecting the best one to train and model.

GPR is a non-parametric Bayesian regression technique that uses a distribution over functions to model the connection between input data and output values. When providing uncertainty estimates for forecasts, it is an effective way to capture intricate and non-linear interactions. A covariance (kernel) function and a mean function constitute a Gaussian Process. A key factor in GPR is the kernel function selection. It is crucial to select the right hyperparameters, such as those found in the kernel function. To increase the likelihood of the observed data, these hyperparameters are frequently optimized for GPR. Among various available techniques, rational quadratic, squared exponential, Matern 5/2, exponential, and optimizable GPR were tried and evaluated in this study.

Parameters optimization and model selection

Each of the four regression approaches, namely GPR, SVM, linear regression, and ensemble methods, offers a variety of pre-existing algorithms to choose from. The algorithms underwent rigorous testing and were evaluated based on their predictive accuracy. The algorithm with the highest effectiveness for each method was chosen, and feature selection techniques were subsequently utilized to improve the performance of the model. The feature selection approach entailed systematically lowering the least significant features to discover the ideal subset for each regression model, by assessing different combinations. Later on, PCA was utilized both with and without feature selection to assess its influence on model correctness. The integration of feature selection and PCA resulted in improved model accuracy. Ultimately, an optimization strategy was employed to further refine the model, utilizing the most effective combinations of PCA and feature selection

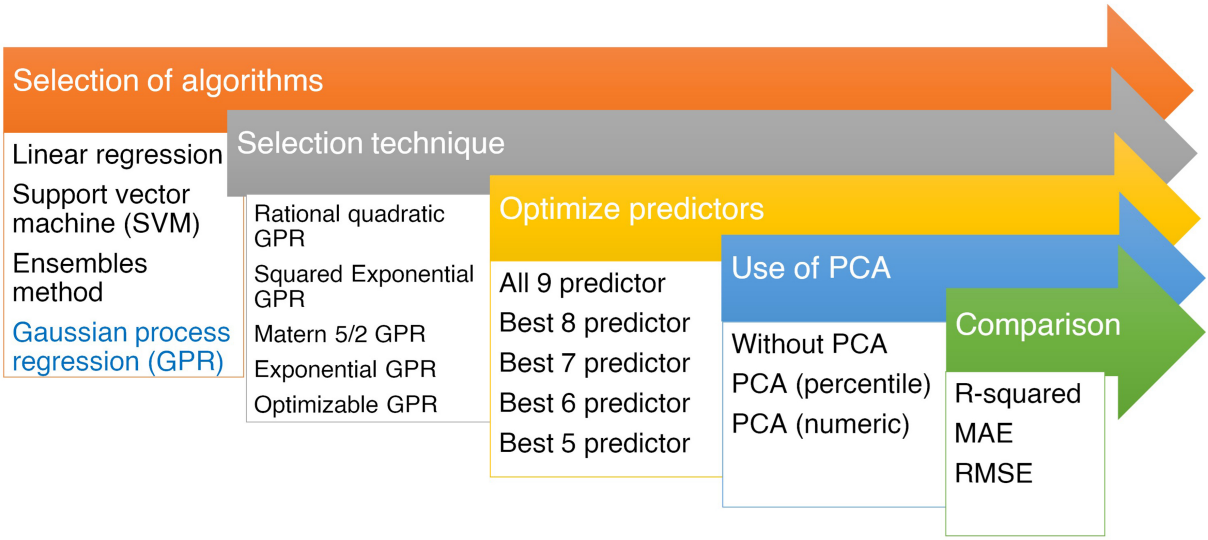

Figure 2. Evaluation and optimization of model (GPR). GPR: Gaussian process regression; PCA: principal component analysis; MAE: mean absolute error; RMSE: root mean squared error.

[Figure 2]. In order to achieve optimization, three techniques were taken into consideration: Bayesian optimization, grid search, and random search. Eventually, Bayesian optimization was chosen to fine-tune the models.

Evaluation of model and validation

In order to overcome the constraints and possible shortcomings of ML models utilized in arsenic adsorption research from aqueous solutions, evaluation and validation are essential. Inadequate or non-representative datasets, sensitivity to hyperparameters, or problems such as underfitting or overfitting can all pose problems for these models. To maximize the model's performance and adaptability for adsorption investigations under various experimental settings, rigorous validation guarantees the accuracy of predictions. In order to ensure the data's excellent quality, preprocessing procedures were employed. Both the trial-and-error methodology and hyperparameter tuning approaches, specifically Bayesian optimization in this case, were employed to fine-tune the hyperparameters. Regression models are assessed using several approaches, such as the coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE)[75]. Training ML models requires avoiding both overfitting and underfitting[76]. By adding a regularization loss function and expanding the training dataset, overfitting can be prevented[77]. On the other hand, lowering regularization parameters and raising the model's complexity - for example, by utilizing larger polynomial terms or more features - can lessen underfitting[78]. Three types of validation schemes are available: cross-validation, holdout validation, and resubstitution. Resubstitution is incapable of preventing overfitting, whereas holdout validation is more effective for large datasets. Cross-validation is an effective approach for evaluating a model's ability to generalize and handle variations[79]. Ten percent of the data in this study were allocated for testing, while the remaining 90% were used to train the model randomly using five-fold cross-validation[80]. After the training process, a random selection of 10% of the data was set aside to test the model and confirm its validation. ML models are often interpreted using the PDP, SHAP, and LIME. By visualizing the correlation between certain input properties and the anticipated result, PDP offers valuable insights into how modifications to particular variables impact the model's predictions. SHAP provides consistent and thorough explanations for the entire dataset by quantifying the contribution of each feature to the model's prediction using cooperative game theory. Rather than discussing the complete model, LIME concentrates on elucidating specific predictions. It achieves this by constructing a localized, simplified model approximation around the specific prediction being examined. When used in tandem,

these techniques improve comprehension of intricate ML models and make them more accessible and useful. R^2, RMSE and MAE equations are as follows: (1), (2), and (3). In these Equations, y represents the response value (sorption capacity), $\overline{y}$ represents the average response value, $\hat{y}$ is the model predicted value, and N is the observation number.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} \quad (1)$$

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

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}| \quad (3)$$

RESULTS AND DISCUSSION

Data collection and preprocessing

The initial data analysis involved calculating the mean, minimum, maximum, and standard deviation. Additionally, preprocessing processes were performed when required. No missing values were present. The histogram plot of all predictors is shown in Figure 3. In order to be suitable for regression analysis, data must undergo preprocessing techniques such as standardization, normalization, and modification. From the histogram plot, data transformation was done to make it normally distributed as much as possible. Figure 4 shows the correlation coefficient values of arsenic adsorption capacity with nine variables among them. There was no presence of collinearity among the nine variables examined in the study. The significance levels of the different factors were consistent, and the Pearson correlation coefficients revealed a similar level of linkage among these variables.

These plots did not include fixed parameters beyond those used in the plot. The strongest correlation between the adsorption capacity and the starting concentration amount was observed. Since there are more ions available for adsorption, a higher initial concentration of arsenic leads to higher adsorption. Furthermore, it might decrease at low adsorbent concentrations, which would leave fewer sites available for adsorption[81,82]. The correlation plot revealed that the amount of iron salt utilized in the production of PFMB, together with the duration of the batch experiment, were identified as the two most significant elements impacting the outcome. The plot of adsorbent quantity utilization demonstrates an initial increase with higher amounts, followed by a subsequent drop with further increases in quantity. Once a location becomes accessible for adsorption, it can significantly enhance the effectiveness of removal.

Oversaturation can lead to ineffective adsorption since more adsorption sites might go unused, which would reduce overall efficiency[83,84]. The observed decline in correlation was partly caused by variations in the relationship between the parameters, which were shown by the correlation plot. For example, 125 rpm was the shaking speed that resulted in the best adsorption efficiency, but both lower and higher speeds decreased adsorption[51]. As shaking speed increased, desorption increased and adsorption decreased at lower speeds. Additionally, as iron salts were more successful in adsorbing heavy metals, the adsorption capacity was negatively correlated with adsorbent quantities and the substitution of iron salts for aluminum salts[85]. Interestingly, despite the fact that biomass and charcoal were supposed to improve porosity, their presence had no discernible effect on adsorption. Research has shown that biosorbents made from Moringa

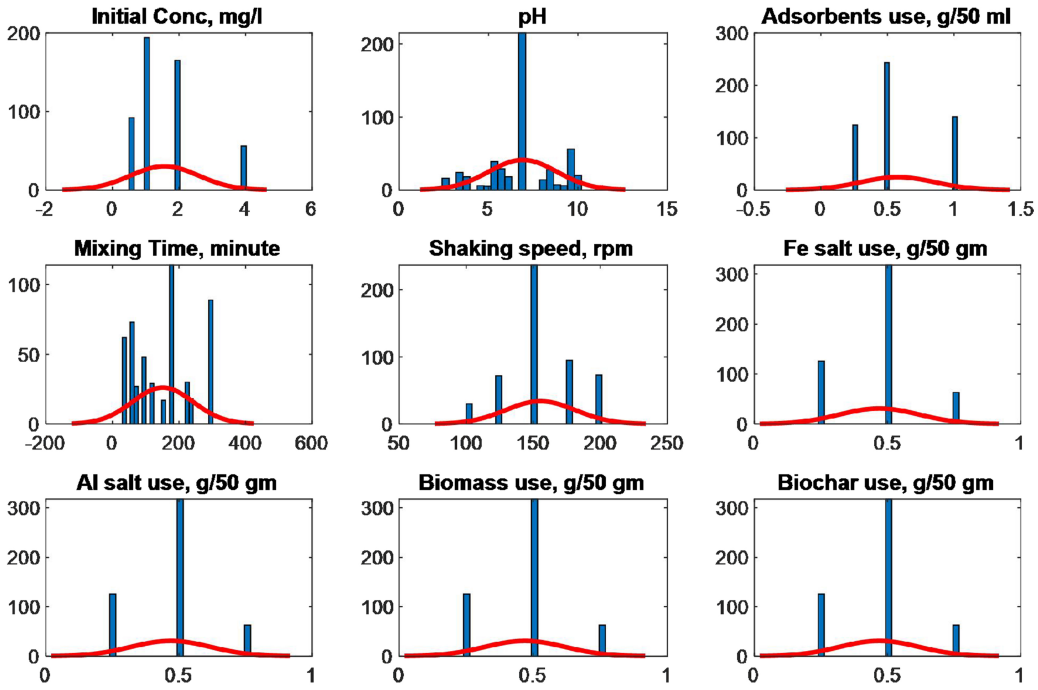

Figure 3. Histogram plot of 09 predictors (507 nos).

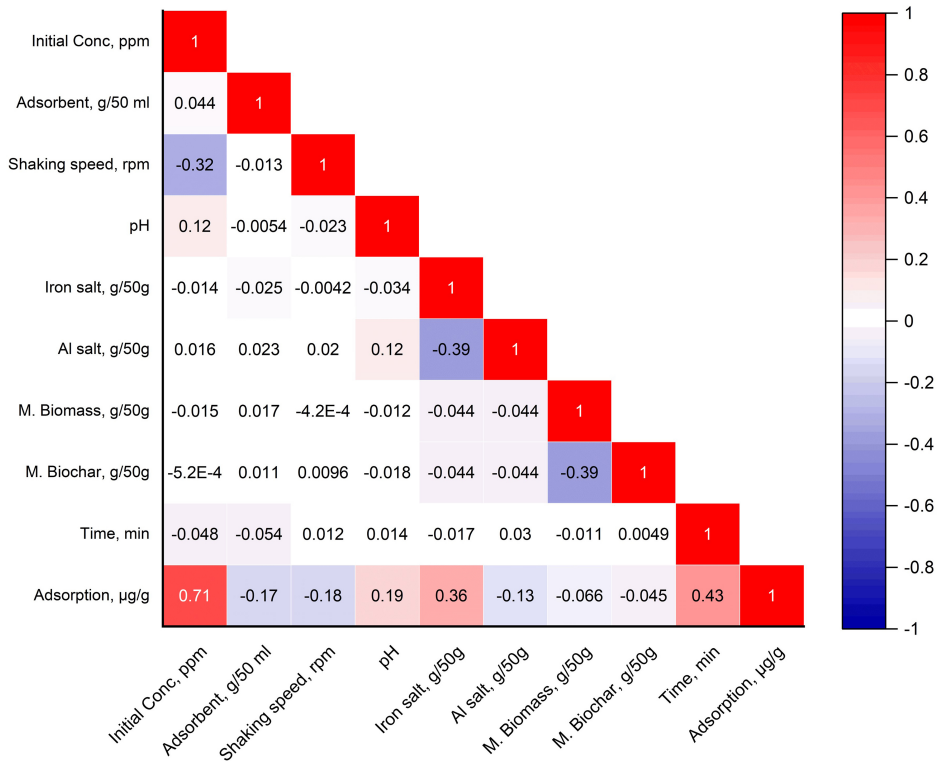

Figure 4. Correlation coefficient plot for predictors and response.

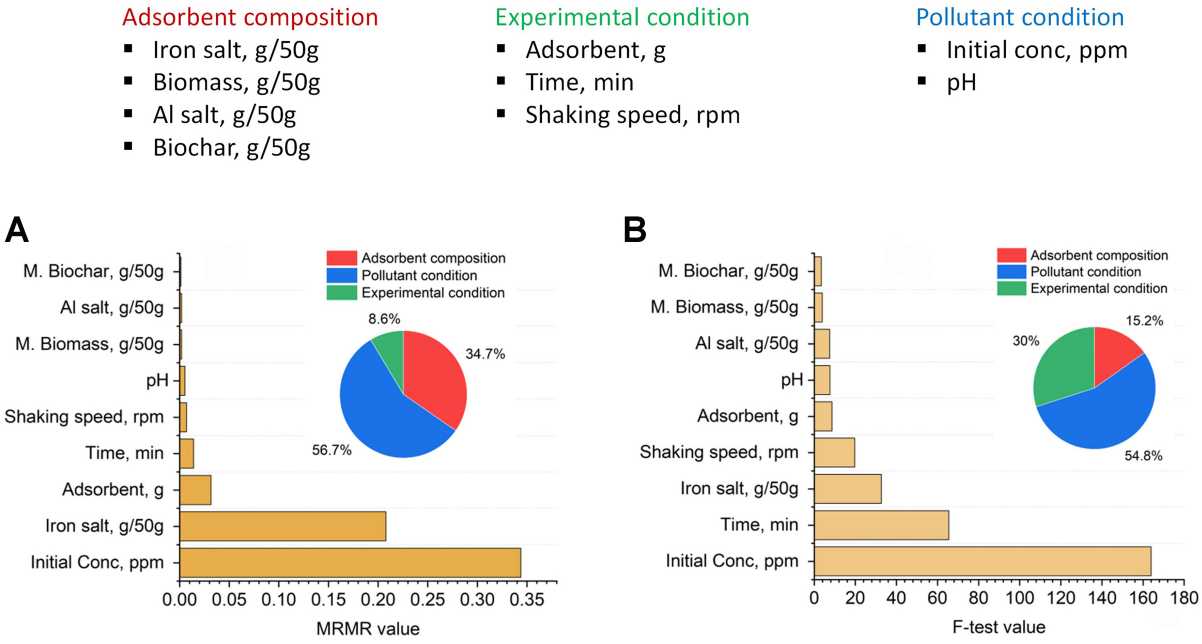

Figure 5. The analysis of feature importance using (A) MRMR and (B) F-test algorithms. MRMR: Minimum redundancy maximum relevance.

oleifera are effective in adsorption; nevertheless, compared to iron salts, metal salts may degrade composite performance[86]. Furthermore, interactions between the adsorbent and particular types of arsenic are significantly influenced by the point of zero charge (PZC). Higher pH values, usually between 5 and 10, improve adsorption, according to the link between pH and adsorption effectiveness[87]. Excellent As(III) adsorption capabilities were demonstrated by mesoporous iron oxide in the pH range of 5-9, with peak adsorption occurring at pH 8.0[88].

The analysis of feature importance using the MRMR and F-test algorithms is displayed in Figure 5. After that, the nine predictors were divided into three groups: composition of the adsorbent, condition of the pollutant, and experimental condition. The findings from both algorithms were quite similar, showing that the state of the pollutant had a greater influence on adsorption than any other factor, accounting for over 50% of the total. The adsorbent composition had the least significant impact on adsorption.

ML models

SL and IL were the two methods used in the linear regression model. Using the F-test and MRMR to assess each feature's relevance, the nine features were reduced to the top five. In addition, PCA was used. Based on the evaluation, it was determined that the IL model with six essential features produced the best results for training and validation when neither PCA nor the robust form were used. The three unused features were biochar, biomass, and Al salt. With a training duration of 7.7448 s, this model was able to predict 8,800 observations per second. Evaluation parameters for the most efficient model were as follows: RMSE 7.867, MSE 61.89, R-squared 0.92503, and MAE 6.4483 for training validation; and RMSE 8.6876, MSE 75.475, R-squared 0.9194, and MAE 7.4623 for testing [Figure 6].

The quadratic SVM was found to be the most effective model for our adsorption analysis during the training and testing phases of the SVM algorithm. The appropriate algorithms were chosen, and then feature selection, PCA, and optimization were carried out. The quadratic SVM, which employed an

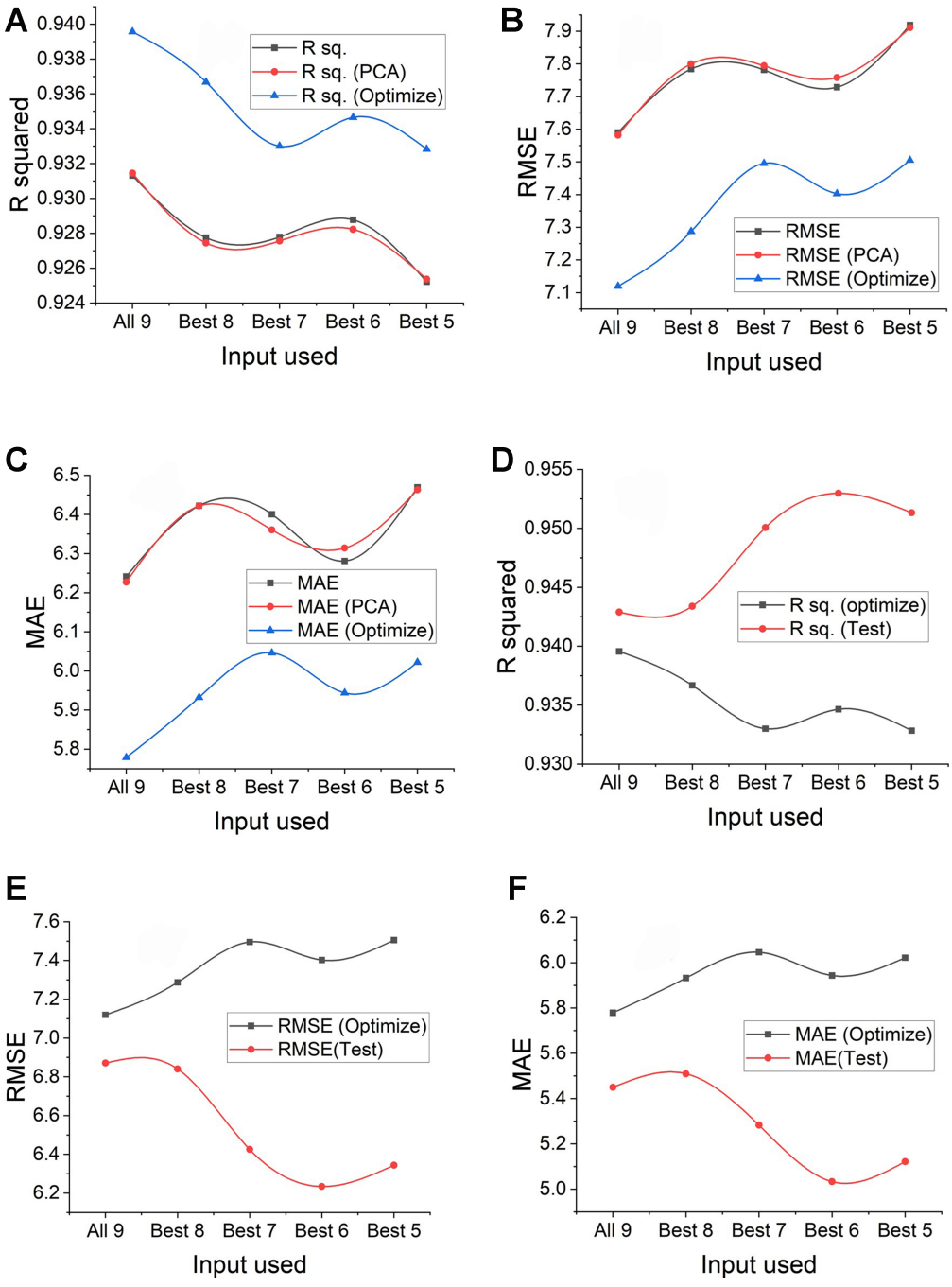

Figure 6. GPR, GPR with PCA, and GPR optimize algorithms were used with variating input features to (A) R squared; (B) RMSE; and (C) MAE. Model of GPR optimize algorithm was tested with (D) R squared; (E) RMSE; and (F) MAE. GPR: Gaussian process regression; PCA: principal component analysis; RMSE: root mean squared error; MAE: mean absolute error.

automatic epsilon, automatic box constant, automatic kernel size, and quadratic kernel function, was the best SVM model. PCA was used, without optimization, keeping nine numerical components with all nine features that were chosen. The model's prediction speed was about 4,200 observations per second, and its training time was 4.7034 s. We next compared these outcomes with those of other models. For training and validation, the most efficient SVM model showed RMSE of 8.0644, MSE of 65.034, R-squared of 0.92188,

and MAE of 6.5322; for testing, the results showed RMSE of 8.2772, MSE of 68.513, R-squared of 0.91671, and MAE of 6.7366 [Figure 6].

We used the bagged tree, boosted tree, and optimizable tree algorithms in the ensemble regression analysis. After applying feature selection, PCA was employed with every feature that had been chosen. The ensemble methods were then optimized, and the outcomes were contrasted to assess the performance of the regression. With an R-squared value of 0.92739, it was verified that the optimization ensembles including all 9 characteristics produced the best outcomes without the use of PCA. The strong predictive performance of this model is indicated by the close alignment of test results and training time validation findings. Table 3 lists all of the parameters that were used in the best ensemble model. An optimizable ensemble, the most efficient model, showed the following assessment metrics: RMSE was 7.7964, MSE was 60.784, R-squared was 0.92739, and MAE was 6.0745 for training and validation; RMSE was 6.85, MSE was 46.923, R-squared was 0.94389, and MAE was 5.414 for testing [Figure 6].

The GPR algorithm was used for regression analysis and modeling. Three presets were used: Matern 5/2 GPR, rational quadratic GPR, and squared exponential GPR. The Matern 5/2 preset yielded the best results out of all of them and was kept for additional examination. After that, feature selection was used to examine the top 8, 7, 6, and 5 features with and without PCA. To preserve numerical values and the 95% variance, PCA was used. Thereafter, an optimizable GPR model was used and contrasted with alternative GPR models. The optimizable GPR with all 9 characteristics and no PCA produced the best results [Figure 6]. Table 4 provides a thorough breakdown of the parameters used in the optimized GPR model. The optimized GPR model demonstrated the subsequent assessment measures, making it the most efficient model: In testing, RMSE was 6.8707, MSE was 47.207, R-squared was 0.94289, and MAE was 5.45; in training and validation, RMSE was 7.1194, MSE was 50.686, R-squared was 0.93956, and MAE was 5.7789.

Comparison of ML models

Upon evaluating the efficacy of distinct ML methods in forecasting arsenic (III) adsorption capacity through the utilization of PFMBs, significant insights become apparent. The optimal GPR model performed better than the others, with the highest R^2 values (0.94289 for the test and 0.93956 for validation) and the lowest RMSE (7.1194 for validation and 6.8707 for the test), demonstrating higher accuracy and dependability [Table 5]. For this approach, the test and training outcomes are likewise close, indicating the model's accuracy. Figure 7 displays the true reaction and forecast, demonstrating how the perfect prediction and true values line up. According to the optimization iteration, after about 30 iterations, the lowest MSE difference between the observed and model values was reached. With an R^2 of 0.94389 and a test RMSE of 6.85, the improved ensemble model performed admirably, closely trailing the GPR in terms of predictive power. Comparing the parameters of a neural network model with a non-neural network model (GPR) using similar data sets showed that the neural network performed better than the GPR model during the validation stage. However, in the testing phase, the non-neural network model exhibited superior efficiency. The allocation of data varied among the training, validation, and testing stages for each model. Within the neural network, 70% of the dataset was allocated for training purposes, while 15% was designated for validation and the other 15% was set aside for testing. Conversely, in the case of non-neural network models, 90% of the data were utilized for training the model, which involved validation using five-fold cross-validation. The remaining 10% was set aside specifically for testing purposes. This comparison demonstrates that non-neural network techniques can be equally effective as neural network algorithms when used in these experimental data sets. In comparison to GPR and the ensemble model, the SVM with PCA showed lower R^2 values (0.92188 for validation and 0.91671 for test) and somewhat higher error metrics (validation RMSE of 8.0644 and test RMSE of 8.2772), despite being effective. Using fewer features led to less accurate predictions, as evidenced by the interaction linear model with the best six features

Table 3. The parameters were used for the best Ensemble algorithms

ML model	Prediction speed	Training time	Preset	Method	Minimum leaf size	Nos of learners	Learning rate	Nos of predictors used	iterations	Training time limit
Ensemble	~600 obs/s	225.76 s	Optimize	Boosted trees	1	14	0.38311	9	30	False

ML: Machine learning.

Table 4. The parameters were used for the best GPR algorithms

ML model	Prediction speed	Training time	Signal standard deviation	Optimize numeric parameters	Basic function	Kernel function	Kernel scale	Sigma	iterations	Training time limit
GPR optimize	~1,300 obs/s	1,305.9 s	20.4671	Yes	Constant	Nonisotropic rational quadratic	0.12835	0.0020662	30	False

GPR: Gaussian process regression; ML: machine learning.

Table 5. Four ML algorithms with best model input conditions and results

Algorithms used	Feature used	RMSE (validation)	R squared (validation)	MAE (validation)	MAE (test)	RMSE (test)	R squared (test)
SVM (PCA)	All 9	8.0644	0.92188	6.5322	6.736	8.2772	0.91671
GPR (optimize)	All 9	7.1194	0.93956	5.7789	5.45	6.8707	0.94289
Ensemble (optimize)	All 9	7.7964	0.92739	6.0745	5.414	6.85	0.94389
Interaction linear	Best 6	7.867	0.92503	6.4483	7.4623	8.6876	0.9194
Neural network[51]	All 9	6.75	0.955			7.76	0.928

ML: Machine learning; RMSE: root mean squared error; MAE: mean absolute error; SVM: support vector machines; PCA: principal component analysis; GPR: Gaussian process regression.

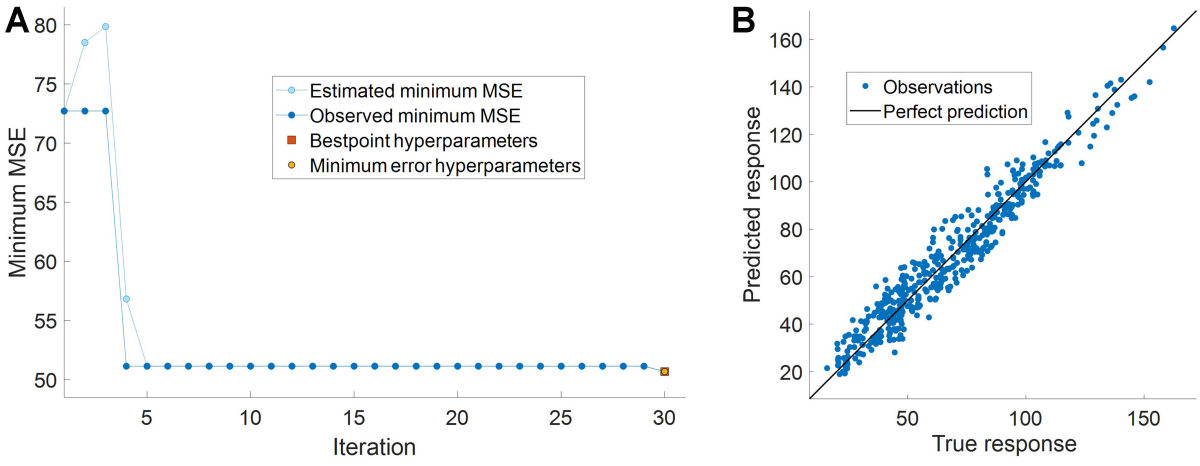

Figure 7. (A) Model plots of optimizable GPR minimum MSE plot, and (B) prediction *vs.* true response plot of optimized GPR. GPR: Gaussian process regression; MSE: mean squared error.

having the greatest error metrics (validation RMSE of 7.867 and test RMSE of 8.6876). The investigation, taken as a whole, emphasizes the value of ensemble approaches and feature optimization in reaching high

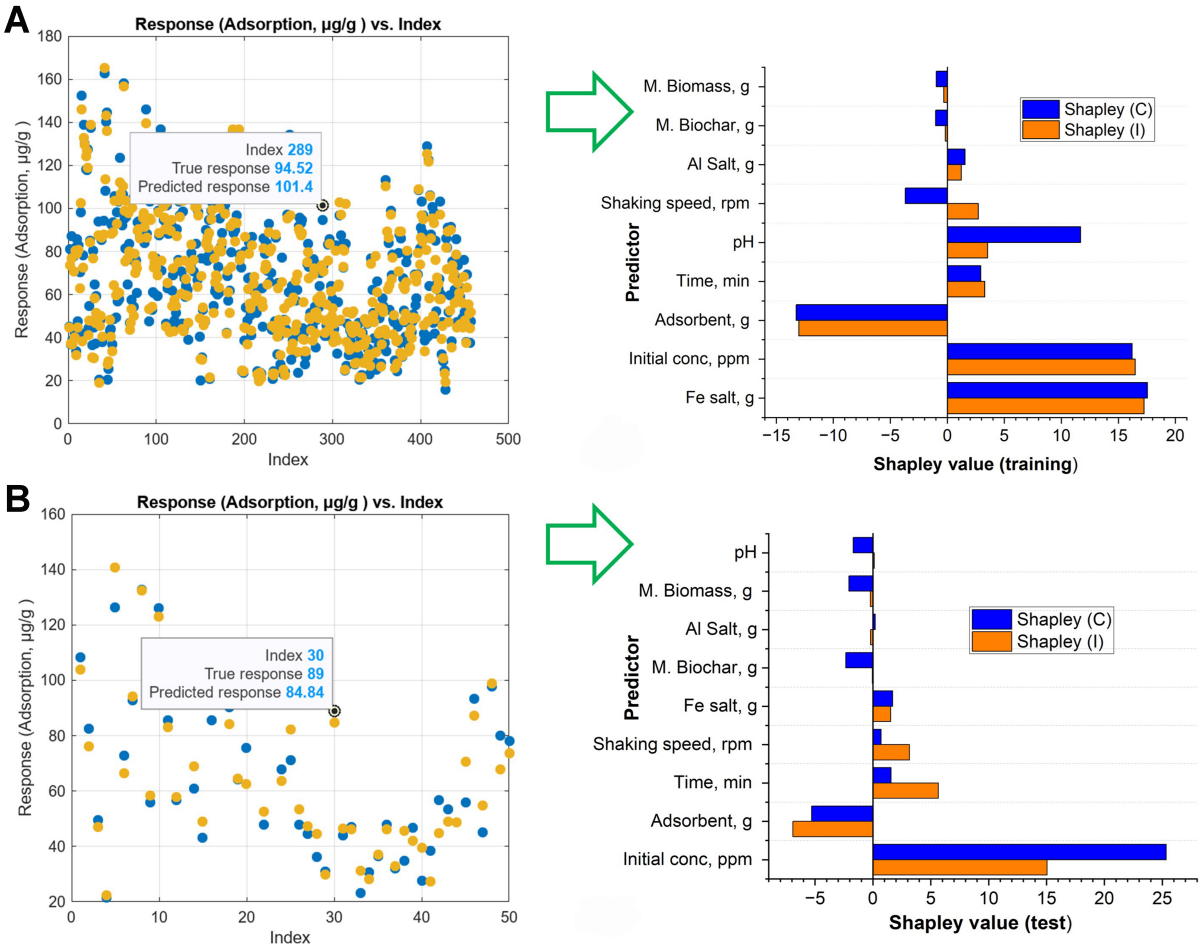

Figure 8. Shapley plot for training [(A) for index:289], and for test [(B) for index: 30] with response plot of the model optimized GPR. GPR: Gaussian process regression.

predicted accuracy for adsorption capacity.

Model explanations

In this investigation, the contributions of individual variables to the adsorption capacity predictions were interpreted using SHAP. The SHAP values suggest that the initial concentration of arsenic and the time of contact have the most significant beneficial impact on adsorption efficiency. This highlights the critical roles that these factors play in the process. Furthermore, the interactions between features were emphasized using SHAP value distributions. For example, the combined impacts of temperature and pH had a greater influence on adsorption capacity than each variable alone. These observations are consistent with theoretical interpretations of adsorption dynamics, confirming the resilience of the model and improving our understanding of the underlying mechanisms. SHAP analysis was carried out utilizing both conditional and interventional approaches for a particular training data point (index 289), where the true value was 94.52 and the predicted value was 101.4, and a test data point (index 30), with a true value of 89 and a predicted value of 84.84 [Figure 8]. The use of iron salt and the starting concentration of arsenic were shown to have the greatest effects in both situations, whereas the amount of adsorbent had the opposite impact. The dependability of the model's explanations is supported by the consistency of these outcomes across approaches. These results were visually verified by the SHAP plots, which displayed the various contributions to the model's predictions and how well they matched the actual values[89,90]. PCA was used to

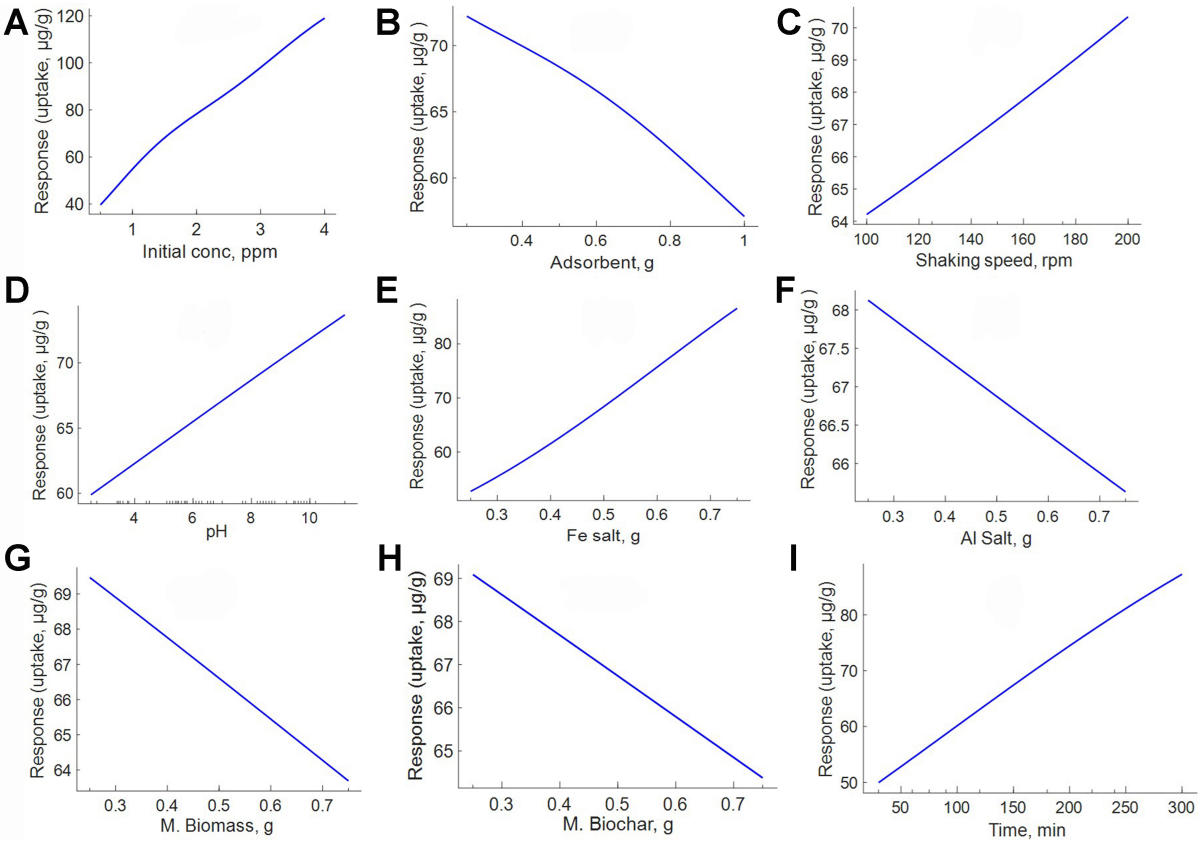

Figure 9. PDP of the optimized GPR model. PDP: Partial dependence plot; GPR: Gaussian process regression.

reduce components to improve accuracy and simplify the model. PCA was applied in both available ways: by specifying explained variance and the number of components. For this study, all regression analyses showed better results when specifying the number of components than the explained variance. The figure shows the number of components required to explain 95% of the variance in the data.

The optimized GPR model was used to illustrate the association between nine important factors and the anticipated adsorption capacity of arsenic (III) onto PFMBs using PDPs [Figure 9]. The PDPs have detected some noteworthy tendencies, which are as follows: A direct correlation was observed between the sorption efficiency and the starting arsenic concentration, as the adsorption capacity increased rapidly from 40 to 120 µg/g with an increase in the initial arsenic concentration ranging from 0.5 to 4 ppm. However, the sorption efficiency decreased from 75 to 55 µg/g when the quantity of adsorbent was raised from 0.25 to 1 g per 50 mL. The adsorption capacity exhibited a slight increase from 60 to 70 µg/g and 60 to 75 µg/g, respectively, with varying shaking speeds (100 to 200 rpm) and pH levels (3 to 12). The pH range specified denotes the optimal range for achieving maximum efficiency. The addition of iron salt (used as 0.25 to 0.75 g per 50 g) enhanced the adsorption capacity from 50 to 90 µg/g. However, the addition of aluminum salt somewhat decreased it from 68 to 65 µg/g. The addition of 0.25 to 0.75 g of biochar and biomass reduced the adsorption capacity from 70 to 60 µg/g. Capacity rose from 50 to 90 µg/g when the adsorption duration was extended from 30 to 300 min. These results draw attention to the intricate interactions between variables and offer suggestions for improving adsorbent formulations and testing setups[91,92].

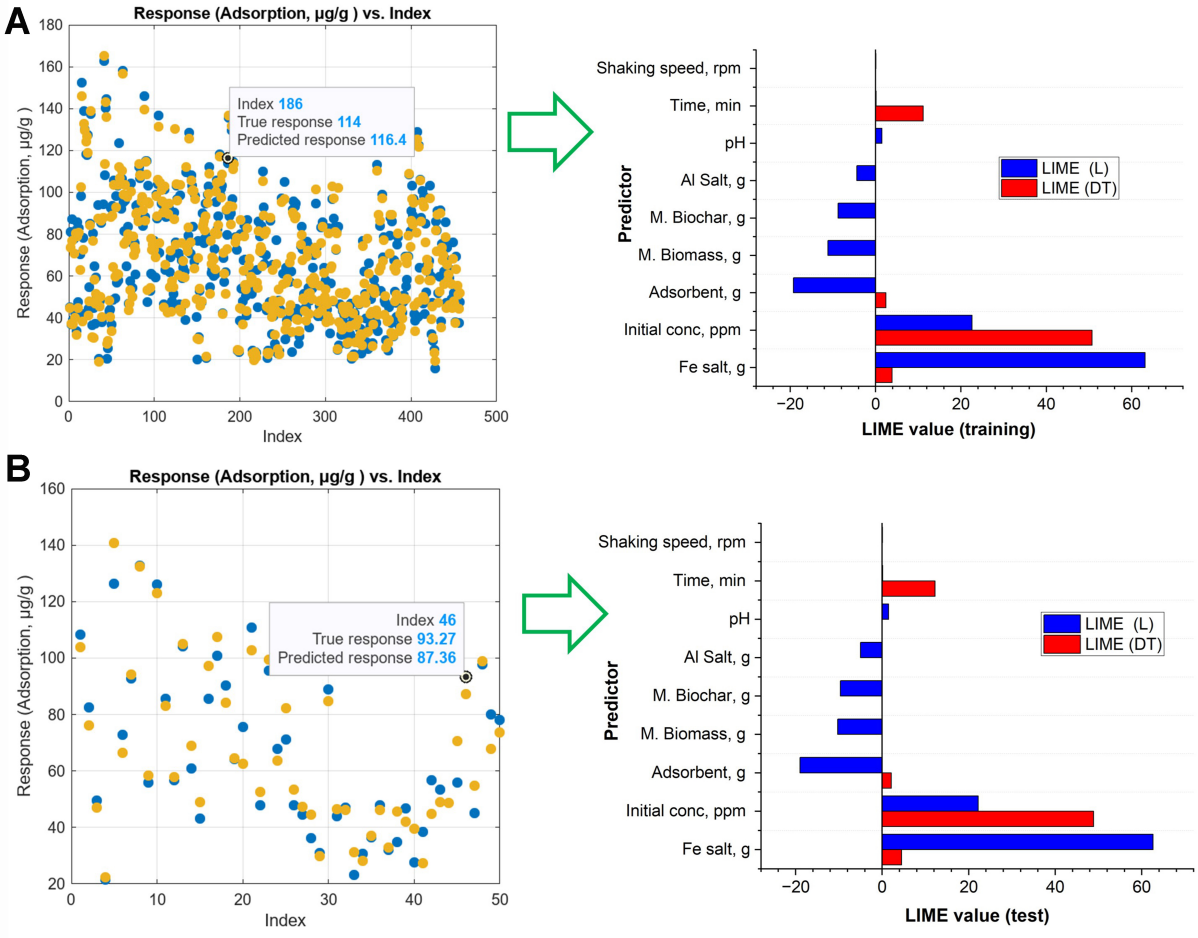

Figure 10. LIME plots for training [(A) index:186], and for test [(B) index: 46] with response plot of the model optimized GPR using linear and decision tree model. LIME: Local interpretable model-agnostic explanations; GPR: Gaussian process regression.

Here, optimized GPR predictions on adsorption capacity were analyzed using artificial predictor data with 5,000 samples, and LIME was used to provide an explanation. At index 46, where the experimental adsorption capacity was 93.27 and the model anticipated 87.36, we looked at the test data while maintaining a focus on local data locality with 1,500 neighbors and a kernel width of 0.75. Indicating that the linear model offered a closer approach to the model's prediction, the LIME explanation values were 86.56 for the linear model and 63.89 for the tree-based model. As can be seen in Figure 10, the LIME explanations for the training data at index 186, with an experimental value of 114 and a model prediction of 116.4, were 112.04 for the linear model (with a kernel width of 0.95) and 113.38 for the tree-based model. This implies that although the tree-based model can also give plausible explanations depending on the situation, the linear model constantly provides closer approximations.

CONCLUSIONS

Arsenic removal is sometimes difficult and expensive; however, using inexpensive adsorbents is a viable way to produce safe water. An efficient method for estimating adsorbent capacity without requiring a lot of trial-and-error laboratory testing is ML modeling. In this work, we employed non-neural network algorithms as a feasible substitute for conventional neural network techniques in modeling arsenic adsorption studies. The data used in this work came from earlier batch tests on the adsorption of arsenic using a PFMB made of

biomass, aluminum salt, gypsum-based iron salt, and moringa bark biochar. Predictive models were created using four non-neural network algorithms: ensemble techniques, linear regression, GPR, and SVM. Nine predictors were considered, including variations in the adsorbent composition, such as the proportions of iron salt, aluminum salt, moringa biochar, and moringa biomass, as well as experimental circumstances such as pH, shaking speed, initial concentration, adsorbent amount, and time. Every algorithm underwent optimization via feature selection and PCA application. F-test and MRMR algorithms were utilized for feature reduction. Using R-squared, MAE, and RMSE values for comparative study, it was shown that the optimized GPR algorithm performed the best in terms of prediction. The most important variables in adsorption efficiency were the initial arsenic concentration, time, and iron salt content in the adsorbent. This was further clarified using SHAP, LIME, and PDP plots, which also demonstrated the effectiveness of the most efficient model. This paper shows that adsorption tests can be efficiently modeled by non-neural network techniques, and the performance of adsorbents may be optimized by using parameters that are generated from the model. By adding more controllable parameters or predictors, such as more specific physicochemical characteristics and experimental settings, the model's effectiveness can be further increased. For the effective removal of heavy metals or other contaminants, this methodology provides a useful way to optimize composite adsorbents in removal technologies. In order to optimize the efficacy of the adsorbent composition and potentially expand its use in a variety of environmental cleanup initiatives, future research should investigate the integration of these extra components.

DECLARATIONS

Authors' contributions

Conceived and designed collaboratively: Mirza NH, Fujino T
Experimental data collection: Mirza NH
Computational model development and simulations: Mirza NH
Data analysis, interpretation of results, and manuscript writing: Mirza NH, Fujino T
Both authors approved the final version of the manuscript before submission.

Availability of data and materials

The datasets used in the study are available from the corresponding author upon reasonable request.

Financial support and sponsorship

None.

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

REFERENCES

1. Ren X, Chen C, Nagatsu M, Wang X. Carbon nanotubes as adsorbents in environmental pollution management: a review. *Chem Eng J* 2011;170:395-410. DOI

2. Chung JY, Yu SD, Hong YS. Environmental source of arsenic exposure. *J Prev Med Public Health* 2014;47:253-7. DOI PubMed PMC
3. Ravenscroft P, Brammer H, Richards K. Arsenic pollution: a global synthesis. Oxford, UK: Wiley-Blackwell, 2009. DOI
4. Smedley P, Kinniburgh D. A review of the source, behaviour and distribution of arsenic in natural waters. *Appl Geochem* 2002;17:517-68. DOI
5. Chutia P, Kato S, Kojima T, Satokawa S. Arsenic adsorption from aqueous solution on synthetic zeolites. *J Hazard Mater* 2009;162:440-7. DOI PubMed
6. Song W, Zhang M, Liang J, Han G. Removal of As(V) from wastewater by chemically modified biomass. *J Mol Liq* 2015;206:262-7. DOI
7. Sigdel A, Park J, Kwak H, Park P. Arsenic removal from aqueous solutions by adsorption onto hydrous iron oxide-impregnated alginate beads. *J Ind Eng Chem* 2016;35:277-86. DOI
8. Sun J, Zhang X, Zhang A, Liao C. Preparation of Fe-Co based MOF-74 and its effective adsorption of arsenic from aqueous solution. *J Environ Sci* 2019;80:197-207. DOI
9. Wang C, Luan J, Wu C. Metal-organic frameworks for aquatic arsenic removal. *Water Res* 2019;158:370-82. DOI PubMed
10. Holm TR. Effects of CO_3^{2-}/bicarbonate, Si, and PO_4^{3-} on Arsenic sorption to HFO. *J AWWA* 2002;94:174-81. DOI
11. Bissen M, Frimmel FH. Arsenic - a review. Part II: oxidation of Arsenic and its removal in water treatment. *Acta hydrochim hydrobiol* 2003;31:97-107. DOI
12. Choong TS, Chuah T, Robiah Y, Gregory Koay F, Azni I. Arsenic toxicity, health hazards and removal techniques from water: an overview. *Desalination* 2007;217:139-66. DOI
13. Norberto J, Zoroufchi Benis K, Mcphedran KN, Soltan J. Microwave activated and iron engineered biochar for arsenic adsorption: life cycle assessment and cost analysis. *J Environ Chem Eng* 2023;11:109904. DOI
14. Masue Y, Loeppert RH, Kramer TA. Arsenate and arsenite adsorption and desorption behavior on coprecipitated aluminum:iron hydroxides. *Environ Sci Technol* 2007;41:837-42. DOI PubMed
15. Mohan D, Pittman CU Jr. Arsenic removal from water/wastewater using adsorbents - a critical review. *J Hazard Mater* 2007;142:1-53. DOI PubMed
16. Lee B, Kim Y, Lee H, Yi J. Synthesis of functionalized porous silicas via templating method as heavy metal ion adsorbents: the introduction of surface hydrophilicity onto the surface of adsorbents. *Micropor Mesopor Mat* 2001;50:77-90. DOI
17. Ahmaruzzaman M. Industrial wastes as low-cost potential adsorbents for the treatment of wastewater laden with heavy metals. *Adv Colloid Interface Sci* 2011;166:36-59. DOI PubMed
18. Zhang L, Zeng Y, Cheng Z. Removal of heavy metal ions using chitosan and modified chitosan: a review. *J Mol Liq* 2016;214:175-91. DOI
19. Tizhe B, Osemeahon S, Nkafamiya I, Shagal M. Biosorption of metal ions from aqueous solution by immobilized moringa oleifera bark. *IRJPAC* 2015;5:238-44. DOI
20. Barua S, Rahman IMM, Nazimuddin M, Hasegawa H. Evaluation of Moringa oleifera carbon for the As(III) removal from contaminated groundwater. *Int J Innov Appl Stud* 2014;8:1390-9. Available from: https://www.cabidigitallibrary.org/doi/full/10.5555/20153089415. [Last accessed on 7 Dec 2024]
21. Mirza NH, Fujino T. Gypsum-based porous media filter with solid waste: a novel composite for removing arsenic (III) from aqueous solution. *Int J Eng* 2025;38:735-43. DOI
22. Ouyang D, Zhuo Y, Hu L, Zeng Q, Hu Y, He Z. Research on the adsorption behavior of heavy metal ions by porous material prepared with silicate tailings. *Minerals* 2019;9:291. DOI
23. Mazloom G, Farhadi F, Khorasheh F. Kinetic modeling of pyrolysis of scrap tires. *J Anal Appl Pyrol* 2009;84:157-64. DOI
24. Khraibet SA, Mazloom G, Banisharif F. Comparative study of different two-phase models for the propane oxidative dehydrogenation in a bubbling fluidized bed containing the VO_x/γ-Al_2O_3 catalyst. *Ind Eng Chem Res* 2021;60:9729-38. DOI
25. Jaffari ZH, Abbas A, Kim CM, et al. Transformer-based deep learning models for adsorption capacity prediction of heavy metal ions toward biochar-based adsorbents. *J Hazard Mater* 2024;462:132773. DOI
26. Yaseen ZM. An insight into machine learning models era in simulating soil, water bodies and adsorption heavy metals: Review, challenges and solutions. *Chemosphere* 2021;277:130126. DOI PubMed
27. Xiong T, Cui J, Hou Z, et al. Prediction of arsenic adsorption onto metal organic frameworks and adsorption mechanisms interpretation by machine learning. *J Environ Manage* 2023;347:119065. DOI
28. Liu R, Zuo L, Zhang P, Zhao J, Tao D. A deep learning neural network approach for predicting the factors influencing heavy-metal adsorption by clay minerals. *Clay Miner* 2022;57:70-6. DOI
29. Shaheen SM, Niazi NK, Hassan NE, et al. Wood-based biochar for the removal of potentially toxic elements in water and wastewater: a critical review. *Int Mater Rev* 2019;64:216-47. DOI
30. Mongioví C, Crini G, Gabrion X, et al. Revealing the adsorption mechanism of copper on hemp-based materials through EDX, nano-CT, XPS, FTIR, Raman, and XANES characterization techniques. *Chem Eng J Adv* 2022;10:100282. DOI
31. Mallik AK, Moktadir MA, Rahman MA, Shahruzzaman M, Rahman MM. Progress in surface-modified silicas for Cr(VI) adsorption: a review. *J Hazard Mater* 2022;423:127041. DOI PubMed
32. Ismail UM, Onaizi SA, Vohra MS. Aqueous Pb(II) removal using ZIF-60: adsorption studies, response surface methodology and machine learning predictions. *Nanomaterials* 2023;13:1402. DOI PubMed PMC

33. Abdi J, Mazloom G. Machine learning approaches for predicting arsenic adsorption from water using porous metal-organic frameworks. *Sci Rep* 2022;12:16458. DOI PubMed PMC
34. Aftab RA, Zaidi S, Danish M, Ansari KB, Danish M. Novel machine learning (ML) models for predicting the performance of multi-metal binding green adsorbent for the removal of Cd (II), Cu (II), Pb (II) and Zn (II) ions. *Environ Adv* 2022;9:100256. DOI
35. Zhu X, Wan Z, Tsang DC, et al. Machine learning for the selection of carbon-based materials for tetracycline and sulfamethoxazole adsorption. *Chem Eng J* 2021;406:126782. DOI
36. Yin G, Jameel Ibrahim Alazzawi F, Mironov S, et al. Machine learning method for simulation of adsorption separation: comparisons of model's performance in predicting equilibrium concentrations. *Arab J Chem* 2022;15:103612. DOI
37. Dashti A, Raji M, Riasat Harami H, Zhou JL, Asghari M. Biochar performance evaluation for heavy metals removal from industrial wastewater based on machine learning: application for environmental protection. *Sep Purif Technol* 2023;312:123399. DOI
38. Almalawi A, Khan AI, Alqurashi F, Abushark YB, Alam MM, Qaiyum S. Modeling of remora optimization with deep learning enabled heavy metal sorption efficiency prediction onto biochar. *Chemosphere* 2022;303:135065. DOI PubMed
39. Elbana TA, Magdi Selim H, Akrami N, Newman A, Shaheen SM, Rinklebe J. Freundlich sorption parameters for cadmium, copper, nickel, lead, and zinc for different soils: influence of kinetics. *Geoderma* 2018;324:80-8. DOI
40. Ducamp M, Coudert F. Prediction of thermal properties of zeolites through machine learning. *J Phys Chem C* 2022;126:1651-60. DOI
41. Ma S, Liu Z. Machine learning for atomic simulation and activity prediction in heterogeneous catalysis: current status and future. *ACS Catal* 2020;10:13213-26. DOI
42. Mandal S, Mahapatra S, Sahu M, Patel R. Artificial neural network modelling of As(III) removal from water by novel hybrid material. *Process Saf Environ* 2015;93:249-64. DOI
43. Sakizadeh M. Artificial intelligence for the prediction of water quality index in groundwater systems. *Model Earth Syst Environ* 2016;2:63. DOI
44. Guo Y, Bartlett PL, Shawe-Taylor J, Williamson RC. Covering numbers for support vector machines. *IEEE Trans Inform Theory* 2002;48:239-50. DOI
45. Durbha SS, King RL, Younan NH. Support vector machines regression for retrieval of leaf area index from multiangle imaging spectroradiometer. *Remote Sens Environ* 2007;107:348-61. DOI
46. Čeh M, Kilibarda M, Lisec A, Bajat B. Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. *IJGI* 2018;7:168. DOI
47. Wei L, Yuan Z, Zhong Y, Yang L, Hu X, Zhang Y. An improved gradient boosting regression tree estimation model for soil heavy metal (arsenic) pollution monitoring using hyperspectral remote sensing. *Appl Sci* 2019;9:1943. DOI
48. Cha Y, Kim YM, Choi JW, Sthiannopkao S, Cho KH. Bayesian modeling approach for characterizing groundwater arsenic contamination in the Mekong River basin. *Chemosphere* 2016;143:50-6. DOI
49. Jang JSR. ANFIS: adaptive-network-based fuzzy inference system. *IEEE Trans Syst Man Cybern* 1993;23:665-85. DOI
50. Abdullahi J, Rufai I, Rimtip NN, Orhon D, Aslanova F, Elkiran G. A novel approach for precipitation modeling using artificial intelligence-based ensemble models. *Desalin Water Treat* 2024;317:100188. DOI
51. Mirza NH, Fujino T. Aqueous arsenic (III) removal using a novel solid waste based porous filter media block: traditional and machine learning (ML) approaches. *Desalin Water Treat* 2024;319:100536. DOI
52. Zhu JJ, Yang M, Ren ZJ. Machine learning in environmental research: common pitfalls and best practices. *Environ Sci Technol* 2023;57:17671-89. DOI PubMed
53. Zhao S, Li J, Chen C, Yan B, Tao J, Chen G. Interpretable machine learning for predicting and evaluating hydrogen production via supercritical water gasification of biomass. *J Clean Prod* 2021;316:128244. DOI
54. Li J, Li L, Tong YW, Wang X. Understanding and optimizing the gasification of biomass waste with machine learning. *Green Chem Eng* 2023;4:123-33. DOI
55. Alemneh ST, Emire SA, Hitzmann B, Zettel V. Comparative study of chemical composition, pasting, thermal and functional properties of Teff (Eragrostis tef) flours grown in Ethiopia and South Africa. *Int J Food Prop* 2022;25:144-58. DOI
56. Smith A, Keane A, Dumesic JA, Huber GW, Zavala VM. A machine learning framework for the analysis and prediction of catalytic activity from experimental data. *Appl Catal B Environ* 2020;263:118257. DOI
57. Li J, Zhang L, Li C, et al. Data-driven based in-depth interpretation and inverse design of anaerobic digestion for CH_4-rich biogas production. *ACS EST Eng* 2022;2:642-52. DOI
58. Jamei M, Karbasi M, Alawi OA, et al. Earth skin temperature long-term prediction using novel extended Kalman filter integrated with artificial intelligence models and information gain feature selection. *Sustain Comput Infor* 2022;35:100721. DOI
59. Breiman L. Random forests. *Mach Learn* 2001;45:5-32. DOI
60. Zhu X, Wang X, Ok YS. The application of machine learning methods for prediction of metal sorption onto biochars. *J Hazard Mater* 2019;378:120727. DOI
61. Pathy A, Meher S, P B. Predicting algal biochar yield using eXtreme Gradient Boosting (XGB) algorithm of machine learning methods. *Algal Res* 2020;50:102006. DOI
62. Kooh MRR, Thotagamuge R, Chou Chau Y, Mahadi AH, Lim CM. Machine learning approaches to predict adsorption capacity of Azolla pinnata in the removal of methylene blue. *J Taiwan Inst Chem Eng* 2022;132:104134. DOI
63. Ghaedi M, Daneshfar A, Ahmadi A, Momeni M. Artificial neural network-genetic algorithm based optimization for the adsorption of phenol red (PR) onto gold and titanium dioxide nanoparticles loaded on activated carbon. *J Ind Eng Chem* 2015;21:587-98. DOI

64. Zhang S, Yuan Y, Liu C, et al. Modeling and optimization of porous aerogel adsorbent for removal of cadmium from crab viscera homogenate using response surface method and artificial neural network. *LWT* 2021;150:111990. DOI
65. Jiang H, Yang Z, Li Z. Non-parallel hyperplanes ordinal regression machine. *Knowl Based Syst* 2021;216:106593. DOI
66. Emmert-Streib F, Dehmer M. Taxonomy of machine learning paradigms: a data-centric perspective. *WIREs Data Min Knowl* 2022;12:e1470. DOI
67. Ye H, Ma X, Yang T, Hou Y. Comparative study on the performance prediction of fuel cell using support vector machine with different kernel functions. In: Proceedings of China SAE Congress 2018: Selected Papers; Singapore. 2020. pp. 337-51. DOI
68. Lee SH, Li J, Wang X, Yang K. Online-learning-aided optimization and interpretation of sugar production from oil palm mesocarp fibers with analytics for industrial applications. *Resour Conserv Recy* 2022;180:106206. DOI
69. Zhang L, Chao B, Zhang X. Modeling and optimization of microbial lipid fermentation from cellulosic ethanol wastewater by Rhodotorula glutinis based on the support vector machine. *Bioresour Technol* 2020;301:122781. DOI
70. Da T, Ren H, He W, Gong S, Chen T. Prediction of uranium adsorption capacity on biochar by machine learning methods. *J Environ Chem Eng* 2022;10:108449. DOI
71. Li J, Zhang W, Liu T, et al. Machine learning aided bio-oil production with high energy recovery and low nitrogen content from hydrothermal liquefaction of biomass with experiment verification. *Chem Eng J* 2021;425:130649. DOI
72. Li J, Suvarna M, Pan L, Zhao Y, Wang X. A hybrid data-driven and mechanistic modelling approach for hydrothermal gasification. *Appl Energy* 2021;304:117674. DOI
73. Cai J, Xu K, Zhu Y, Hu F, Li L. Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest. *Appl Energy* 2020;262:114566. DOI
74. Dehghanian N, Ghaedi M, Ansari A, et al. A random forest approach for predicting the removal of Congo red from aqueous solutions by adsorption onto tin sulfide nanoparticles loaded on activated carbon. *Desalin Water Treat* 2016;57:9272-85. DOI
75. He Z, Wang Q, Rao P, et al. WS_2 significantly enhances the degradation of sulfachloropyridazine by Fe(III)/persulfate. *Sci Total Environ* 2022;850:157987. DOI PubMed
76. Haq MA, Rahim Khan MA. DNNBoT: deep neural network-based botnet detection and classification. *Comput Mater Con* 2022;71:1729-50. DOI
77. Koeshidayatullah A. Optimizing image-based deep learning for energy geoscience via an effortless end-to-end approach. *J Pet Sci Eng* 2022;215:110681. DOI
78. Fan Y, Yang W. A backpropagation learning algorithm with graph regularization for feedforward neural networks. *Inform Sci* 2022;607:263-77. DOI
79. Yaqoob U, Younis MI. Chemical gas sensors: recent developments, challenges, and the potential of machine learning - a review. *Sensors* 2021;21:2877. DOI
80. Yuan X, Li J, Lim JY, et al. Machine learning for heavy metal removal from water: recent advances and challenges. *ACS EST Water* 2024;4:820-36. DOI
81. Alam MA, Shaikh WA, Alam MO, et al. Adsorption of As (III) and As (V) from aqueous solution by modified Cassia fistula (golden shower) biochar. *Appl Water Sci* 2018;8:839. DOI
82. Nguyen TH, Pham TH, Nguyen Thi HT, et al. Synthesis of iron-modified biochar derived from rice straw and its application to arsenic removal. *J Chem* 2019;2019:1-8. DOI
83. Yang S, Wu Y, Aierken A, et al. Mono/competitive adsorption of Arsenic(III) and Nickel(II) using modified green tea waste. *J Taiwan Inst Chem Eng* 2016;60:213-21. DOI
84. Imran M, Iqbal MM, Iqbal J, et al. Synthesis, characterization and application of novel MnO and CuO impregnated biochar composites to sequester arsenic (As) from water: modeling, thermodynamics and reusability. *J Hazard Mater* 2021;401:123338. DOI
85. Giles DE, Mohapatra M, Issa TB, Anand S, Singh P. Iron and aluminium based adsorption strategies for removing arsenic from water. *J Environ Manage* 2011;92:3011-22. DOI
86. Sumathi T, Alagumuthu G. Adsorption studies for arsenic removal using activated *Moringa oleifera*. *Int J Chem Eng* 2014;2014:1-6. DOI
87. Bae J, Kim S, Kim KS, Hwang H, Choi H. Adsorptive removal of arsenic by mesoporous iron oxide in aquatic systems. *Water* 2020;12:3147. DOI
88. Mudzielwana R, Gitari WM, Ndungu P. Removal of As(III) from synthetic groundwater using Fe-Mn bimetal modified kaolin clay: adsorption kinetics, isotherm and thermodynamics studies. *Environ Process* 2019;6:1005-18. DOI
89. Islam SR, Eberle W, Bundy S, Ghafoor SK. Infusing domain knowledge in AI-based "black box" models for better explainability with application in bankruptcy prediction. arXiv. [Preprint.] May 30, 2019 [accessed 2024 Dec 7]. Available from: https://doi.org/10.48550/arXiv.1905.11474.
90. Li J, Pan L, Suvarna M, Tong YW, Wang X. Fuel properties of hydrochar and pyrochar: prediction and exploration with machine learning. *Appl Energy* 2020;269:115166. DOI
91. Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. *J Comput Graphical Stat* 2015;24:44-65. DOI
92. Yuan X, Suvarna M, Low S, et al. Applied machine learning for prediction of CO_2 adsorption on biomass waste-derived porous carbons. *Environ Sci Technol* 2021;55:11925-36. DOI PubMed