scientific reports

Check for updates

OPEN

Machine learning approaches for predicting arsenic adsorption from water using porous metal–organic frameworks

Jafar Abdi[1]✉ & Golshan Mazloom[2]

Arsenic in drinking water is a serious threat for human health due to its toxic nature and therefore, its eliminating is highly necessary. In this study, the ability of different novel and robust machine learning (ML) approaches, including Light Gradient Boosting Machine (LightGBM), Extreme Gradient Boosting, Gradient Boosting Decision Tree, and Random Forest was implemented to predict the adsorptive removal of arsenate [As(V)] from wastewater over 13 different metal–organic frameworks (MOFs). A large experimental dataset was collected under various conditions. The adsorbent dosage, contact time, initial arsenic concentration, adsorbent surface area, temperature, solution pH, and the presence of anions were considered as input variables, and adsorptive removal of As(V) was selected as the output of the models. The developed models were evaluated using various statistical criteria. The obtained results indicated that the LightGBM model provided the most accurate and reliable response to predict As(V) adsorption by MOFs and possesses R^2, RMSE, STD, and AAPRE (%) of 0.9958, 2.0688, 0.0628, and 2.88, respectively. The expected trends of As(V) removal with increasing initial concentration, solution pH, temperature, and coexistence of anions were predicted reasonably by the LightGBM model. Sensitivity analysis revealed that the adsorption process adversely relates to the initial As(V) concentration and directly depends on the MOFs surface area and dosage. This study proves that ML approaches are capable to manage complicated problems with large datasets and can be affordable alternatives for expensive and time-consuming experimental wastewater treatment processes.

As a highly toxic material, arsenic is distributed all over environmental waters. Arsenic can be produced naturally through biological activity and earth crust. Also, it can be caused by human activity such as agriculture, mineral extraction, and discharge of industrial wastewater[1]. Inorganic arsenic mainly exists in two forms; arsenite [As(III)] and arsenate [As(V)]. Arsenate is the main species found in natural surface water bodies, while arsenite predominantly exists in the groundwater[2]. Both of them are highly toxic, but As(III) is approximately 60 times more toxic than As(V)[3]. Generally, As(III) holds neutral and un-dissociated forms, therefore its removal is very challenging[4] and it is required first to oxidize arsenite to arsenate for its effective removal. Arsenic is a global threat to human health. Long-term exposure to arsenic, mainly through contaminated water and food, can cause severe diseases such as kidney, liver, skin, and lung cancers[5]. World Health Organization (WHO) has set the maximum level of 10 µg/L for arsenic in drinking water[6]. Therefore, effective removal of this heavy metal is still a vital task. Different technologies have been developed for arsenic removal, including ion exchange, biological techniques, coagulation, precipitation, reverse osmosis, filtration, and adsorption[7,8]. Among these technologies, adsorption over the porous adsorbents is generally one of the most promising methods due to the high efficiency, cost affordable, and mild operating conditions[9]. Different porous adsorbents have been developed and studied for arsenic removal, such as zeolites, natural clay, carbonous materials, metal oxides, and metal–organic frameworks (MOFs)[10,11]. Recently, MOFs, as a new class of porous materials, have been attracted much interest in different fields[12–18]. MOFs consist of metal ions or clusters, which are coordinated through organic linkers. Owning to their characteristic features; i.e., large surface area, high porosity, adjustable pore size, high crystallinity, good thermal and chemical stability[19,20]. So far, many studies have shown the remarkable ability of different MOFs to

[1]Faculty of Chemical and Materials Engineering, Shahrood University of Technology, Shahrood, Iran. [2]Department of Chemical Engineering, Faculty of Engineering, University of Mazandaran, Babolsar, Iran. ✉email: Jafar.abdi@shahroodut.ac.ir

adsorb various contaminants, such as heavy metals[5,21,22]. For example, the performance of ZIF-8 was evaluated in the adsorptive removal of As(III) and As(V) by Jian et al.[23]. The authors have reported the maximum adsorption capacity of 49.49 and 60.03 mg/g for As(III) and As(V), respectively, at pH = 7 and room temperature. UiO-66 was successfully synthesized by Audu et al.[24] for adsorption of As(V) and As(III). They reported that faster and more efficient adsorption obtained as the pore sizes increased by reducing particle sizes. Li et al.[25] was compared the performance of acetate modulated MOF-76(Y) with pristine MOF-76(Y) in the adsorptive removal of As(V) in the alkaline solutions. It was shown that the acetate modulated sample exhibited a higher pore volume and smaller particle size, which indicated excellent performance in the adsorption of arsenate with a maximum adsorption capacity of 201.46 mg/g. The superior performance of pristine MIL-88A(Fe) and MIL-88A(Fe) decorated on cotton fiber was reported by Pang et al.[26] in the adsorption of As(III) and As(V).

Employing MOFs in the practical application requires many challenges and obstacles to be overcome. Evaluation of the MOFs' efficiency involves conducting the experiments, which are the most expensive and time-consuming steps. In addition, the results obtained for the removal of the contaminants in laboratory operations cannot be scaled up to real plants. While studying the effects of different operating variables is necessary for control and consequently optimization in large-scale processes. The development of mathematical models is the first attempt to study the various processes widely studied by different researchers[27–31]. However, providing mathematical models for complex processes is extensively CPU- and time-consuming requiring a lot of time and effort. Fortunately, other approaches based on machine learning (ML) have been developed for modeling and simulation of such complex processes. These robust alternative approaches can predict complicated processes without solving theoretical equations. Nowadays, ML methods have been utilized in various areas due to their excellent performance with acceptable accuracy and reliability[32–38].

In this study, we employ new approaches based on the ML for predicting As(V) adsorption from wastewater using MOFs at different operating conditions. The main novelty of the current work is the implementation of new innovative models that are efficient for managing extensive data collections. So far, the models based on the ML methods have not been previously employed for estimating As(V) removal over MOFs. Thus, a large dataset assigned to the As(V) adsorption by different MOFs adsorbents were collected at various conditions, including adsorbent dosage, arsenic concentration, contact time, temperature, solution pH, adsorbent surface area, and the presence of anions. Then, four powerful models Light Gradient Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost), Gradient Boosting Decision Tree (GBDT), and Random Forest (RF) were implemented for predicting As(V) adsorption efficiency.

Theory of the utilized model

Light gradient boosting machine (LightGBM). The LightGBM is developed according to the most basic concepts of gradient learning[39]. Comparing the LightGBM and the XGBoost throws light on the LightGBM's better efficiency and consumption of less memory. These advantages expedite the training phase of the model development[40]. Dividing eigenvalues, the LightGBM form 'k' different bins. Doing so, a histogram having a total width of 'k' can be constructed. As soon as the procedure mentioned above is completed, there would be no need for any ensemble with pre-sorted results. The resulting values could be stored in eight-bit memory space as an integer value. Therefore, the amount of memory needed for keeping the calculated values would be dipped drastically, and consequently, such an approach reduces the preciseness of the resulting model. The LightGBM also benefits from the Leaf-wise problem-solving method. It has been seen in investigations that the leaf-wise strategy is much more robust and expeditious than any other traditional strategy. The most salient and affecting factor in making the leaf-wise strategy more powerful and reliable than any level-wise approach is the fact that all the leaves of a specific layer will possibly be taken into consideration for calculations, which decreases memory allocation[41].

Extreme gradient boosting model (XGBoost). To find the minimum answer for a series of objective functions defined for an ensemble, classification and regression trees (CARTs) can be used expeditiously. Among all various forms of gradient boosting structures, the XGBoost method could be mentioned as a highly efficient tree-shaped approach. Typically, a CART model comprises three primary layers. Firstly, the main nodes could also be referred to as the root layer. Secondly, the interiors or internal nodes and what is located at the third layer could be named as the leaves or leaf nodes. A range of processes known as binary decision-making operations is responsible for dividing the root node and forming the internal nodes. These processes expeditiously develop the internal nodes from data sets made available in the root node. Finally, the classification operations will be completed in leaves of the modeling tree, resulting in the final classes. The robustness and the accuracy of every model could be improved by introducing various ensembles to the CARTs and developing them by assigning specified weight factors. The mentioned weight factor will determine how much an ensemble could affect the final result of the model[42].

Gradient boosting decision tree (GBDT). In contrast with the Adaboost approach, the Gradient Boosting model utilizes the previously made residual errors of its precursor learners[43]. As in this method, a loss function is minimized during the model development procedure; it can be contemplated as a kind of decent gradient approach[44]. The currently presented study seeks to benefit from a combination of the Gradient Boosting approach and the decision trees, known as the gradient boosting decision tree (GBDT) method. Suppose an ensemble of experimentally obtained data with the form of $\{(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)\}$, the GBDT's steps could be presented as follows[41]:

Step A. Initialization of $f_0(x)$.

Step B. Iteration on tree learners from b = 1 to b = B.

B1. Calculation of negative gradient (Z_l).
B2. Setting $G_b(x)$ (regression tree) to the targets (Z_l, l = 1,…,N).
B3. Determination of the size of each decent gradient.
B4. Continuously update $f_b(x)$.
Step C. the output corresponding to each data point of (x, b) will be $f_B(x)$.

For developing a predictive model, a range of hyper-parameters had to be assigned, such as the number of decision tree learners, a subset of the ensemble for initial feeding to the learners, the upper limit of the allowable depth, the lowest number of leaves, number of features, and number of data points in the separated sample as the sample split[45].

Random forest (RF). The random forest modeling approach is fabricated from a combination of various decision trees. The random forest will train each tree simultaneously with other trees. However, it does not mean that trees have equal importance. In this predicting method, the algorithm is responsible for determining the superiority of every individual tree[46]. Additionally, to manage different features, the RF is enabled to select various features by implementing a built-in property of the RF classifier. This property will help the RF model to determine features without the elimination of some parameters and lowering the dimension of a complex problem[47]. Furthermore, a process known as bagging, which is the abbreviated version of bootstrap aggregating, is employed by the RF model to prevent similarity between trees in the forest and preserve their diversity. In the model development, the tree's population is typically determined for the model as an inputted integer. Afterward, data points will be discretized into various subgroups according to the number of required trees. As a method for randomized sampling, bagging will try to use approximately 30 percent of data points in the training phase of every individual subtree. The remaining 70 percent of data points must be considered out-of-bag (OOB) data points.

Models development

Data assembling. A large dataset consisting of 280 experimental dataset of As(V) elimination by various MOFs was collected using well-documented literature. The investigated MOFs include MIL-101(Fe), MIL-88A, MIL-100(Fe), UiO-66, UiO-66-NH_2, MIL-53(Fe), Co-MOF-74, Zn-MOF-74, MIL-88B, ZIF-8, AUBM-1, GUT-3, and MIL-125(Ti)[4,23,48–57]. The structures of the selected MOFs are schematically presented in Fig. 1.

The affecting parameters on adsorptive performance are the initial concentration of arsenic (mg/L), adsorbent dosage (g/L), contact time (min), solution pH, temperature (°C), the specific surface area of MOFs (m^2/g), and the presence of the anions. The statistical details of the dataset are listed in detail in Table 1. These parameters were considered as input data for the implemented models. At the same time, the removal percentage of As(V) was selected as the output of the models. Python, an open-source software, was used for modeling procedures. The training process of the models was performed using 85% of the data set called train subset. The performance of the models was investigated by 15% of the remaining dataset denoted test subset. Feature selection and classification using different algorithms for predicting the adsorption efficiency of As(V) by MOFs is presented in Fig. 2.

Detection of outliers. Outliers are usually existed, especially in large datasets. The accuracy and reliability of the models can be significantly influenced by the outliers. Therefore, all datasets should be refined. In this work, the Leverage method was employed for detecting and eliminating outliers. The Hat matrix of the Leverage method was calculated based on Eq. (1)[58,59]:

$$H = Y\left(Y^T Y\right)^{-1} Y^T \tag{1}$$

In Eq. (1), Y is a matrix with $m \times n$ dimensions, where m is the number of experimental data and n stands for the number of input variables. The Hat matrix diagonal elements are the Hat value of data. Outliers can be recognized by developing William's plot in which the normalized residuals are plotted versus the Hat values. The warning Leverage parameter (H^*) is calculated based on Eq. (2) and is also shown in William's plot[60]:

$$H^* = \frac{3(n+1)}{m} \tag{2}$$

Evaluation of the models quality. The quality and reliability of the developed models were assessed using different statistical techniques described as follows:

1. The average absolute relative deviation of the model results from the experimental values was calculated by the average absolute percent relative error (AAPRE) (Eq. (3)):

$$AAPRE(\%) = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{X(i)_{model} - X(i)_{exp}}{X(i)_{exp}}\right| \tag{3}$$

2. The root mean square error (RMSE), which indicates the error dispersion, is calculated by Eq. (4):

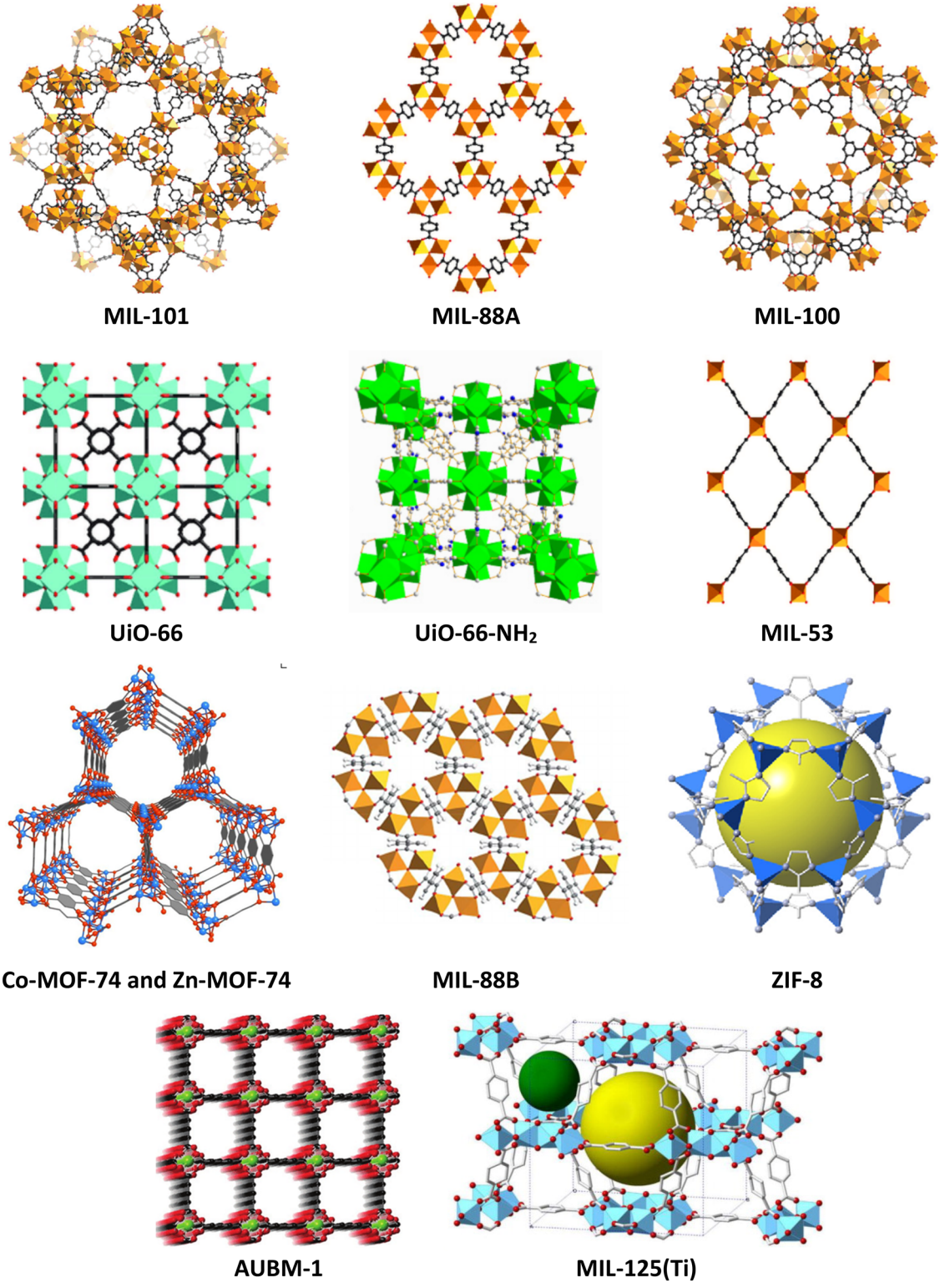

Figure 1. The framework structure of the investigated MOFs.

	Surface area (m^2/g)	Adsorbent dosage (g/L)	Arsenic concentration (mg/L)	Contact time (h)	Temperature (°C)	pH	Presence of anions	Removal efficiency (%)
Min	113.4	0.02	0.002	0.25	25	1.7	1	0.5
Max	1388	5	472.5	24	45	13	11	100

Table 1. Statistical details of the inputs and output parameters collected in this work.

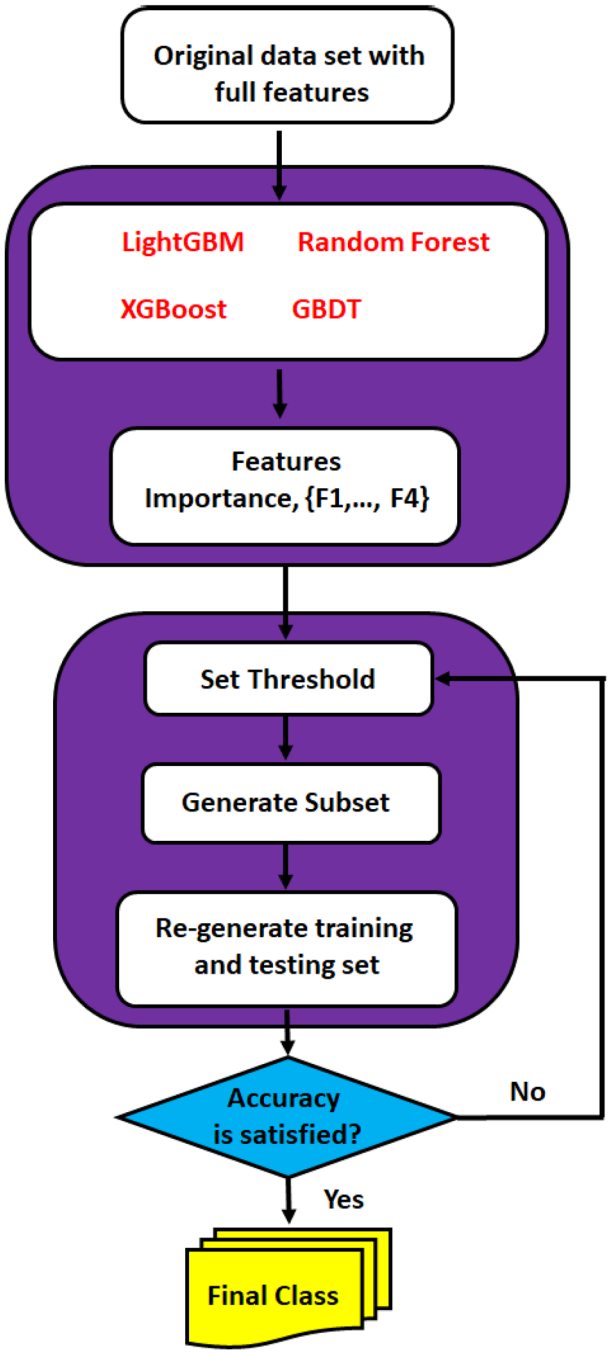

Figure 2. Feature selection and classification using different algorithm for predicting adsorption efficiency of arsenic by MOFs.

$$RMSE = \left(\frac{\sum_{i=1}^{n} (X(i)_{model} - X(i)_{exp})^2}{n} \right)^{1/2} \tag{4}$$

3. The dispersion of data is investigated by the standard deviation of errors (STD), which can be calculated using Eq. (5):

$$STD = \frac{1}{n} \sum_{i=1}^{n} \left(X(i)_{model} - \overline{X(i)}_{model} \right)^2)^{1/2} \tag{5}$$

4. The coefficient of determination (R^2) which assigns the accuracy of the predictions, can be calculated by Eq. (6). The R^2 value close to 1 determines that the estimation of experimental data is more accurate.

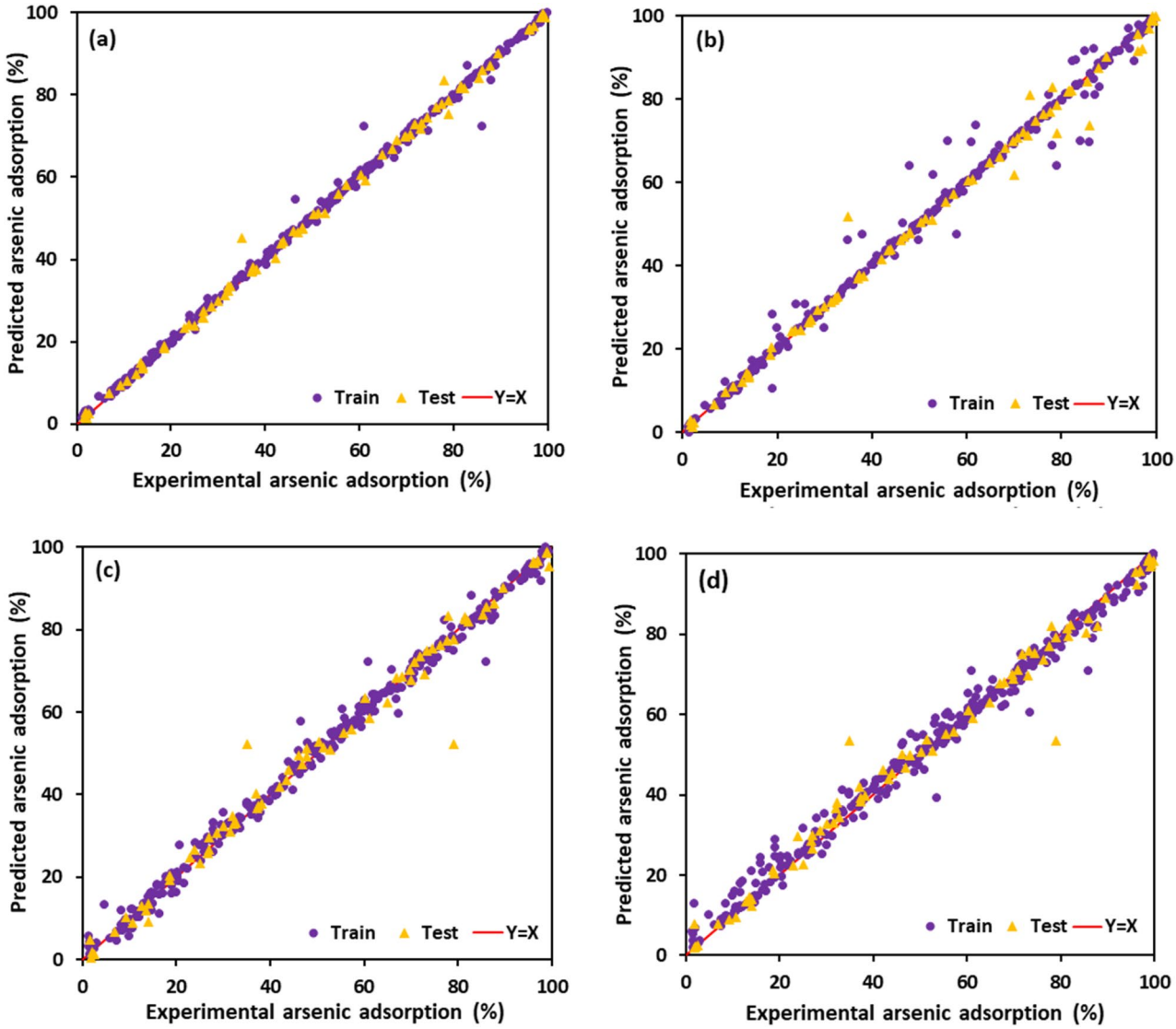

Figure 3. Cross plots of the proposed machine learning models in this study: (**a**) LightGBM, (**b**) XGBoost, (**c**) GBDT, and (**d**) RF.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(X(i)_{model} - X(i)_{exp}\right)^2}{\sum_{i=1}^{n} \left(X(i)_{model} - \overline{X(i)}_{exp}\right)^2} \tag{6}$$

Results and discussion

Validation of the developed models. The values of As(V) removal predicted by four developed models are illustrated in Fig. 3 regarding experimental data. When the predicted results are closer to the experimental data, the model with remarkable accuracy and reliability is achieved[61]. As shown in Fig. 3, all developed models revealed great accuracy where the data points scattered close enough to the line with a unit slope. It can be observed that the LightGBM model has excellently matched with experimental data amongst all the proposed models. The error distributions of all four developed models are depicted in Fig. 4. As shown, the errors fluctuated over zero line indicating that the models have been well developed with acceptable accuracy. However, the deviation of the LightGBM model from zero line was rarely notable compared with the others. The cumulative frequency of data versus AAPRE% is plotted in Fig. 5, visually indicating the model with higher accuracy. The model that is closest to the vertical axis is the most accurate. As can be seen, about 95% of data points can be predicted with AAPRE lower than 2% using the LightGBM and XGBoost. While, smaller than 40% of data points were predicted with AAPRE less than 2%, when RF and GBDT approaches were employed. Therefore, GBDT and RF models provided weak performance with less accuracy among the implemented models.

The performance of all developed models is also evaluated using additional statistical methods. Different statistical data attributed to the train, test and total data set of each model are listed in Table 2. With the highest total R^2 value of 0.9958 and the least RMSE value of 2.0688, the LightGBM model exhibited supreme performance.

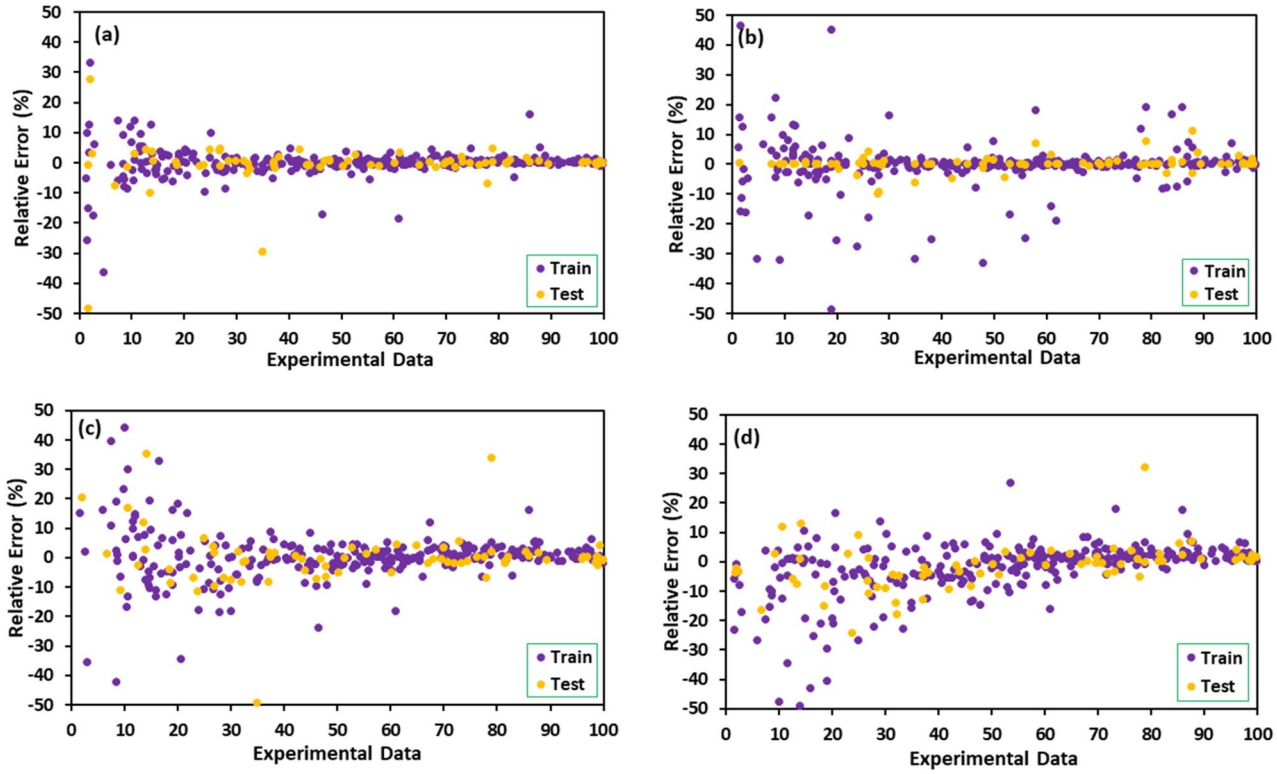

Figure 4. Error distribution plots of machine learning models for training and test sets: (**a**) LightGBM, (**b**) XGBoost, (**c**) GBDT, and (**d**) RF.

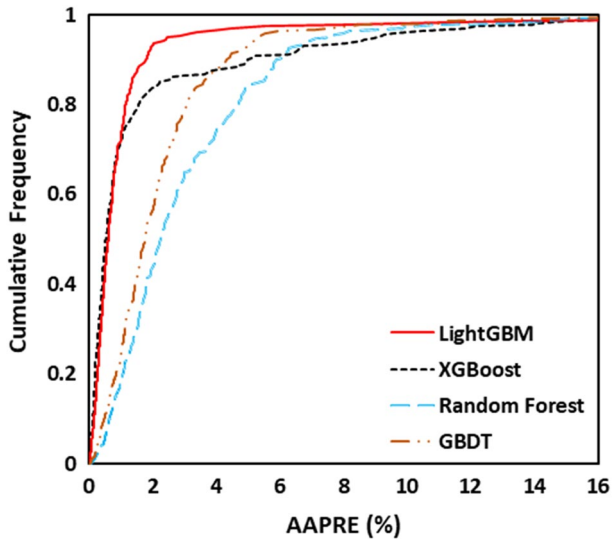

Figure 5. The cumulative frequency diagram of the proposed machine learning models.

Also, based on the AAPRE% and STD, the LightGBM model presented the best performance. In the LightGBM model, the values of AAPRE and STD for train, test and total dataset were 2.11%, 0.0554, and 2.43%, and 0.0825, 2.88%, and 0.0628, respectively, which were the least obtained values among the developed models. In addition, the implemented models were selected before, and their accuracy in the training and testing stages and over the whole of the database was monitored using four statistical matrices. It is hard to most accurate one through visual inspection. Therefore, the ranking analysis is employed for doing so[62]. Figure 6 provides the results of model ranking in each stage based on the average values of the four statistical criteria reported in Table 2. The LightGBM model in the learning step is the best; nevertheless, it shows the second-ranking in the testing stage, and XGBoost depicts the best performance. On the other hand, the LightGBM with the first ranks over the whole database is

Statistical criteria		R^2	RMSE	STD	AAPRE (%)
LightGBM	Train	0.9983	1.4013	0.0554	2.11
	Test	0.9852	3.6872	0.0825	2.43
	Total	0.9958	2.0688	0.0628	2.88
XGBoost	Train	0.9931	2.6873	0.0668	2.66
	Test	0.9832	3.7892	0.1007	3.52
	Total	0.9879	2.8081	0.0854	3.19
GBDT	Train	0.9922	2.3791	0.2731	8.55
	Test	0.9826	4.1125	0.2452	9.09
	Total	0.9812	2.9137	0.2698	8.72
RF	Train	0.9804	2.0364	0.4561	9.59
	Test	0.9762	4.4207	0.4233	10.72
	Total	0.9799	3.3845	0.4507	10.13

Table 2. Calculated statistical criteria for the developed models.

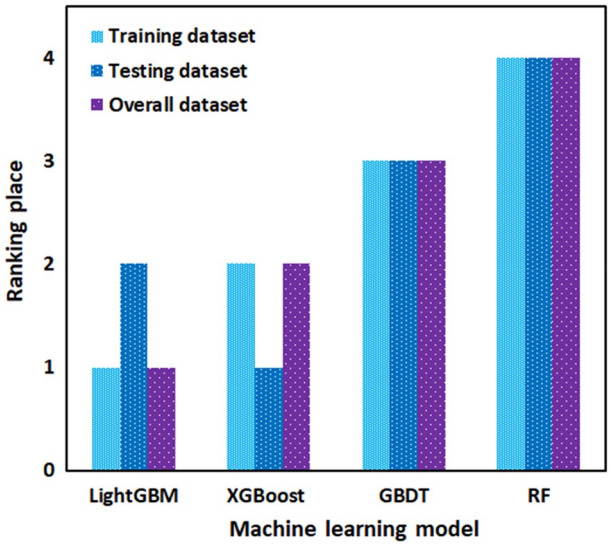

Figure 6. Comparison between AARPE of the developed models.

the best model for predicting arsenic removal using MOFs. According to the results, the developed predictive models can be summarily ranked in terms of their accuracy as follows: LightGBM > XGBoost > GBDT > RF.

Trend analysis of the LightGBM. The adsorptive removal of As(V) is highly influenced by operating conditions. In this section, the capability of the most accurate model, the LightGBM, was examined for predicting the trends of different parameters on the arsenic adsorption using various MOFs. In Fig. 7, the obtained results of the developed LightGBM model are compared with experimental data. The affecting parameters examined under different conditions were temperature, As(V) concentration, solution pH, and the coexistence of anions. As seen in Fig. 7, for all operating conditions, the LightGBM model exhibited excellent performance, predicting experimental data with high accuracy. Figure 7a illustrates the maximum adsorption capacity of As(V) obtained from different MOFs in comparison with predicted values by the LightGBM model. MOFs with different structural and morphological properties depicted various adsorption capacities, but interestingly, the performance of MOFs could be predicted with excellent accuracy using the developed model. Comparison between all mentioned MOFs, the Zn-MOF-74 had a maximum capacity of 328 mg/g because of its high surface area (604 m^2/g), nearly twelve times of GUT-3 (209 m^2/g) with the adsorption capacity of 29 mg/g. It has also been found that the amount of arsenic adsorption on the MOF structure is directly attributed to surface properties (e.g., charge, functional group, morphology, etc.) besides the surface area.

Another important factor influencing the adsorption capacity of adsorbents is the initial concentration of pollutants. As shown in Fig. 7b, the effects of As(V) initial concentration on the adsorption capacity of UiO-66-NH_2 at different temperatures[51] can be precisely predicted using the developed LightGBM approach. It can be seen that the adsorption capacity of the adsorbent increased with increasing the initial concentration of As(V) due to the more available metal ions present in the solution. Moreover, the enhancement of temperature from 25 to 45 °C resulted in a steady decrease in the adsorptive removal efficiency of As(V) over UiO-66-NH_2. This

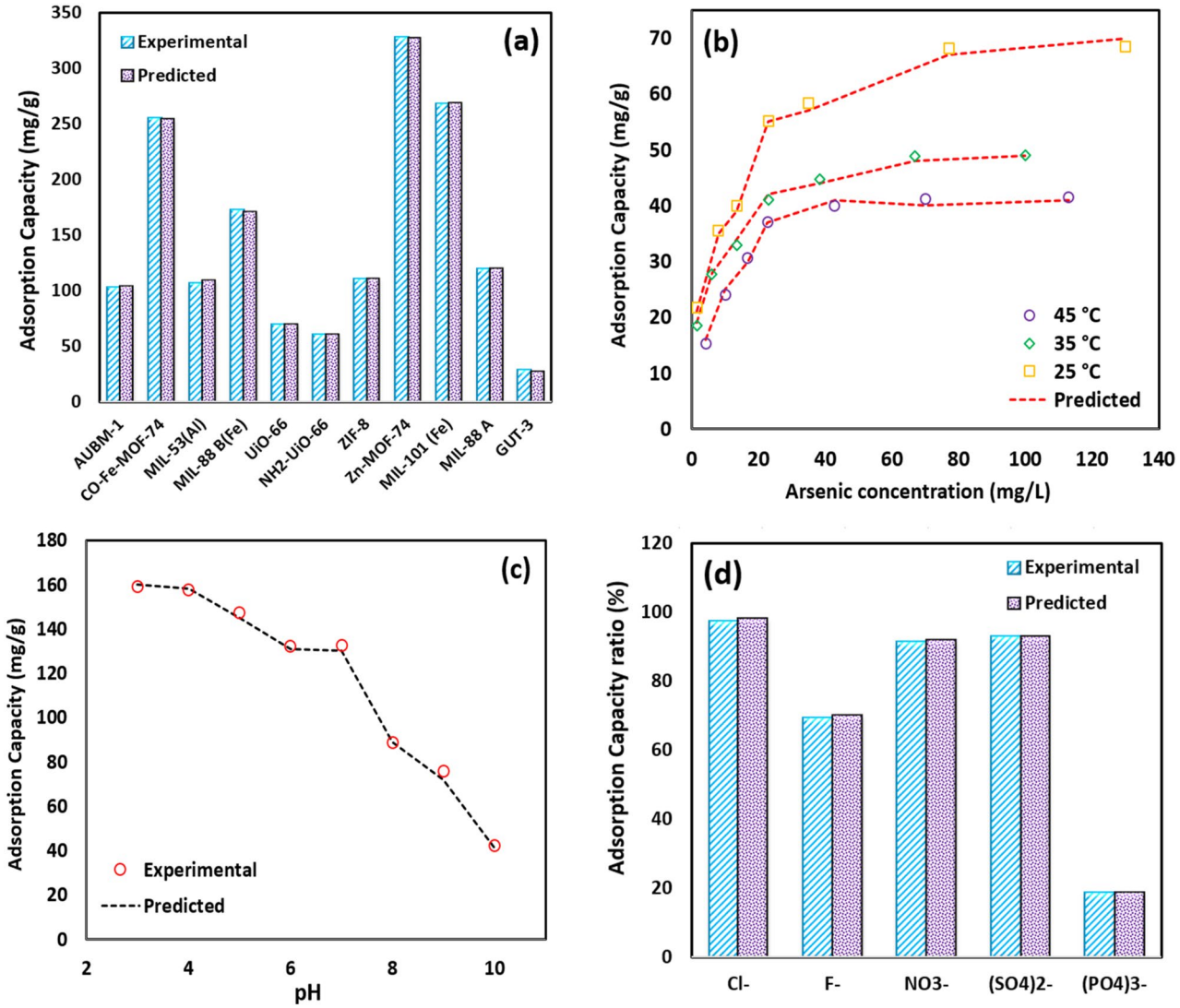

Figure 7. (**a**) The performance of different MOFs in the adsorption of As(V), (**b**) The effect of temperature on the adsorption capacity of UiO-66-NH_2 in different As(V) concentration (pH = 9.2), (**c**) The effect of solution pH on arsenic removal by Fe-Co-MOF-74 (Initial arsenic concentration = 100 mg/L, adsorbent dose = 0.5 g/L, temperature = 25 °C), and (**d**) The effect of competition anions on the adsorption capacity of ZIF-8 (pH = 7).

observation confirms that As(V) adsorption is exothermic, therefore increasing temperature probably weakens the binding forces formed between the pollutants and the active sites of the adsorbent[63].

The solution pH is a very critical parameter affecting the adsorptive removal of different pollutants. The pH can influence the surface potential of the adsorbent as well as the type of the pollutant species in the solution[64]. The effect of different solution pH on the As(V) removal over Fe-Co-based MOF-74[4] are compared with the estimated results by the LightGBM model in Fig. 7c. As depicted, the adsorption capacity indicated a steady decreasing trend with increasing pH from 3 to 10. Based on the obtained results, in the examined pH range, the adsorbent maintained a positive surface charge, while, As(V) existed in the form of negatively charged species[4]. Thus, the electrostatic interaction between As(V) and Fe-Co based MOF-74 can explain the adsorption process. The decline in the adsorption capacity with increasing the solution pH can be assigned to the decreasing adsorbent surface potential[4]. As seen in Fig. 7c, the LightGBM model was a reliable technique providing accurate predictions for adsorption capacity in the whole examined pH.

The effect of the coexistence of various anions such as NO_3^-, PO_4^{3-}, Cl^-, SO_4^{2-}, F^- on the adsorption capacity of As(V) using ZIF-8[65] was compared with the predicted results of the LightGBM model in Fig. 7d. It can be seen that PO_4^{3-} exhibited an intense inhibitory impact on the adsorption process. This may be due to the similar structure of PO_4^{3-} with AsO_4^{3-} and their competition for adsorption over active sites of MOFs. The presence of F^- also revealed the adverse effects on the As(V) removal. While the negative effects of other ions were negligible. As it is obvious in Fig. 7d, the implemented LightGBM approach was strongly capable for predicting the impact of coexistence of anions on the adsorption of As(V).

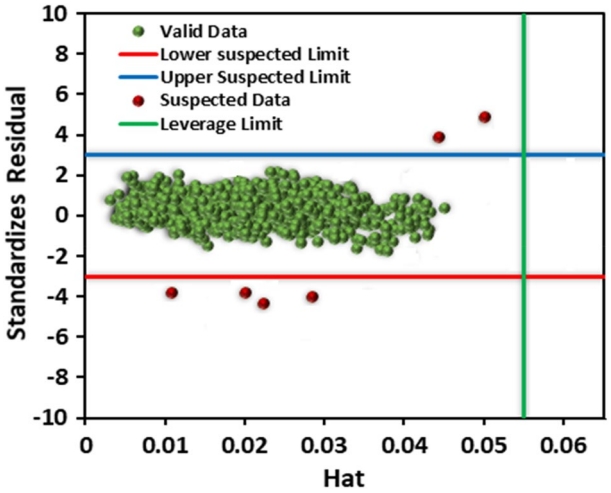

Figure 8. Outlier detection using William's plot for the LightGBM model.

The applicability domain of the LightGBM model. To evaluate the applicability area of the LightGBM model with the best performance, William's plot was plotted in Fig. 8. In the Leverage approach, the data in the area determined by standard residuals of + 3 to − 3 on the y-axis and 0 to H* on the x-axis are only valid. Based on Fig. 8, only six data points were placed out of the valid domain of William's plot. Accordingly, these points were considered experimentally uncertain data. Since the number of doubtful data was small compared to the entire data set, it can be concluded that both collected experimental data and the LightGBM model were statistically valid and trustworthy.

The sensitivity analysis. The sensitivity analysis was performed to evaluate the magnitude of the impacts of all input parameters on the As(V) removal predicted by the LightGBM model[66]. The value of the relevancy factor (r) determines the extent of each input parameter's effect on the As(V) adsorption[67]. The factor "r" can be a negative or positive value. A positive value for unique input data confirms that the output variable directly interacts with that input data. Whereas the negative value reveals the inverse interaction between output and input variable[68]. In addition, the greater the absolute value of "r" for a particular input parameter, the more significant the impact of that variable on the model output[69]. The relevancy factor is computed by Eq. (7):

$$r(I_i, \omega) = \frac{\sum_{j=1}^{N} \left(I_{ij} - \overline{I_i}\right)\left(\omega_j - \overline{\omega}\right)}{\left(\sum_{j=1}^{N} \left(I_{ij} - \overline{I_j}\right)^2 \sum_{j=1}^{n} \left(\omega_j - \overline{\omega}\right)^2\right)^{0.5}} \tag{7}$$

where ω_j and $\overline{\omega}$ are the jth and the mean value of the predicted As(V) removal, respectively. I_{ij} and $\overline{I_i}$ represent the ith and the mean value of the ith input variable, respectively. N is the total number of data.

The calculated relevancy factor of all input parameters on the As(V) adsorption predicted by the LightGBM model is plotted in Fig. 9. As mentioned above, the input parameters were MOFs surface area, adsorbent dosage, arsenic concentration, contact time, temperature, solution pH, and presence of anions. As illustrated in Fig. 9, the adsorbent dosage and surface area of the MOFs exhibited the most positive impacts on As(V) adsorption. This confirms that any increase in the adsorbent content, as well as the adsorbent specific surface area would result in increasing the amount of As(V) removal. On the other hand, the initial arsenic concentration inversely influenced the adsorption process. Other parameters such as the presence of anions and pH had negligible effects on the model output.

Conclusion

In this study, the potential of four different ML approaches, LightGBM, XGBoost, GBDT, and RF, were investigated to estimate As(V) adsorption from wastewater. An experimental dataset of As(V) removal using 13 different MOFs was selected with various operating conditions. Validation of the proposed models was performed using statistical methods. The LightGBM model with the least AAPRE value of 2.88% and the least STD value of 0.0628 was the most trustworthy model. Based on the cumulative frequency diagram of the LightGBM model, about 95% of data points can be estimated with AAPRE lower than 2%. In addition, the Leverage approach proved that most of the data points of the LightGBM model were scattered within the valid domain of William's plot. Moreover, the effects of different operating parameters such as initial arsenic concentration, temperature, solution pH, and the presence of anions can be predicted accurately on the As(V) removal. This study confirms ML approaches that are cost affordable and straightforward can be effectively employed for wastewater treatment.

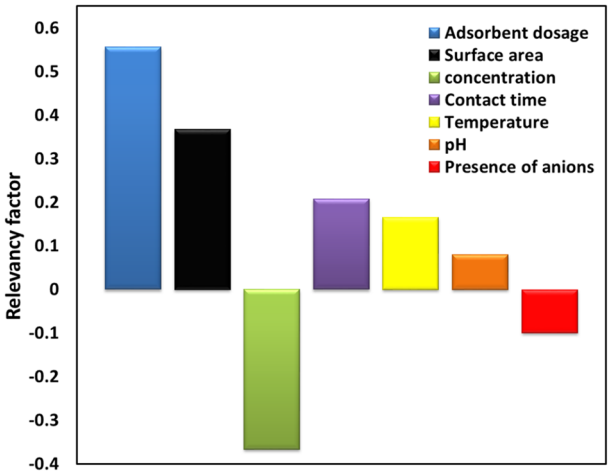

Figure 9. Sensitivity analysis for the developed LightGBM model on different input variables.

Data availability

The collected experimental databank has been added to the manuscript (please see Supplementary Information: Dataset).

Received: 10 May 2022; Accepted: 19 September 2022
Published online: 30 September 2022

References

1. Smedley, P. L. & Kinniburgh, D. G. A review of the source, behaviour and distribution of arsenic in natural waters. *Appl. Geochem.* **17**, 517–568 (2002).
2. Song, W., Zhang, M., Liang, J. & Han, G. Removal of As (V) from wastewater by chemically modified biomass. *J. Mol. Liq.* **206**, 262–267 (2015).
3. Sigdel, A., Park, J., Kwak, H. & Park, P.-K. Arsenic removal from aqueous solutions by adsorption onto hydrous iron oxide-impregnated alginate beads. *J. Ind. Eng. Chem.* **35**, 277–286 (2016).
4. 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.* **80**, 197–207 (2019).
5. Wang, C., Luan, J. & Wu, C. Metal-organic frameworks for aquatic arsenic removal. *Water Res.* **158**, 370–382 (2019).
6. Holm, T. R. Effects of CO32–/bicarbonate, Si, and PO43–on Arsenic sorption to HFO. *J. Am. Water Works Assoc.* **94**, 174–181 (2002).
7. Choong, T. S., Chuah, T., Robiah, Y., Koay, F. G. & Azni, I. Arsenic toxicity, health hazards and removal techniques from water: An overview. *Desalination* **217**, 139–166 (2007).
8. Bissen, M. & Frimmel, F. H. Arsenic—a review. Part II: Oxidation of arsenic and its removal in water treatment. *Acta Hydrochim. Hydrobiol.* **31**, 97–107 (2003).
9. Mohan, D. & Pittman, C. U. Jr. Arsenic removal from water/wastewater using adsorbents—a critical review. *J. Hazard. Mater.* **142**, 1–53 (2007).
10 Gupta, A. D., Rene, E. R., Giri, B. S., Pandey, A. & Singh, H. Adsorptive and photocatalytic properties of metal oxides towards arsenic remediation from water: A review. *J. Environ. Chem. Eng.* **9**, 106376 (2021).
11. Gupta, K., Joshi, P., Gusain, R. & Khatri, O. P. Recent advances in adsorptive removal of heavy metal and metalloid ions by metal oxide-based nanomaterials. *Coord. Chem. Rev.* **445**, 214100 (2021).
12 Tahir, M. A., Arshad, N. & Akram, M. Recent advances in metal organic framework (MOF) as electrode material for super capacitor: A mini review. *J. Energy Storage* **47**, 103530 (2021).
13. Oladoye, P. O., Adegboyega, S. A. & Giwa, A.-R.A. Remediation potentials of composite metal-organic frameworks (MOFs) for dyes as water contaminants: A comprehensive review of recent literatures. *Environ. Nanotechnol. Monit. Manag.* **16**, 100568 (2021).
14. Al-Rowaili, F. N. *et al.* A review for metal-organic frameworks (MOFs) utilization in capture and conversion of carbon dioxide into valuable products. *J. CO2 Util.* **53**, 101715 (2021).
15. Duan, C. *et al.* Recent advances in the synthesis of nanoscale hierarchically porous metal–organic frameworks. *Nano Mater. Sci.* https://doi.org/10.1016/j.nanoms.2021.12.003 (2022).
16. Khataee, A. *et al.* State-of-the-art progress of metal-organic framework-based electrochemical and optical sensing platforms for determination of bisphenol A as an endocrine disruptor. *Environ. Res.* **212**, 113536. https://doi.org/10.1016/j.envres.2022.113536 (2022).
17. Abdi, J., Izadi, M. & Bozorg, M. Improvement of anti-corrosion performance of an epoxy coating using hybrid UiO-66-NH2/carbon nanotubes nanocomposite. *Sci. Rep.* **12**, 10660. https://doi.org/10.1038/s41598-022-14854-y (2022).
18. Song, Y., Xie, W., Shao, M. & Duan, X. Integrated electrocatalysts derived from metal organic frameworks for gas-involved reactions. *Nano Mater. Sci.* https://doi.org/10.1016/j.nanoms.2022.01.003 (2022).
19. Abdi, J., Sisi, A. J., Hadipoor, M. & Khataee, A. State of the art on the ultrasonic-assisted removal of environmental pollutants using metal-organic frameworks. *J. Hazard. Mater.* **424**, 127558. https://doi.org/10.1016/j.jhazmat.2021.127558 (2022).
20. Shahmirzaee, M. *et al.* Metal-organic frameworks as advanced sorbents for oil/water separation. *J. Mol. Liq.* **363**, 119900. https://doi.org/10.1016/j.molliq.2022.119900 (2022).
21. Abdi, J., Banisharif, F. & Khataee, A. Amine-functionalized Zr-MOF/CNTs nanocomposite as an efficient and reusable photocatalyst for removing organic contaminants. *J. Mol. Liq.* **334**, 116129. https://doi.org/10.1016/j.molliq.2021.116129 (2021).
22. Kobielska, P. A., Howarth, A. J., Farha, O. K. & Nayak, S. Metal–organic frameworks for heavy metal removal from water. *Coord. Chem. Rev.* **358**, 92–107. https://doi.org/10.1016/j.ccr.2017.12.010 (2018).

23. Jian, M., Liu, B., Zhang, G., Liu, R. & Zhang, X. Adsorptive removal of arsenic from aqueous solution by zeolitic imidazolate framework-8 (ZIF-8) nanoparticles. *Colloids Surf. A* **465**, 67–76 (2015).
24. Audu, C. O. *et al.* The dual capture of As V and As III by UiO-66 and analogues. *Chem. Sci.* **7**, 6492–6498 (2016).
25. Li, Z. *et al.* Efficient capture of arsenate from alkaline smelting wastewater by acetate modulated yttrium based metal-organic frameworks. *Chem. Eng. J.* **397**, 125292 (2020).
26. Pang, D. *et al.* Superior removal of inorganic and organic arsenic pollutants from water with MIL-88A (Fe) decorated on cotton fibers. *Chemosphere* **254**, 126829 (2020).
27 Mazloom, G. & Alavi, S. M. Kinetic study of selective propane oxidation to acrylic acid over Mo1V0. 3Te0. 23Nb0. 12Ox using the genetic algorithm. *React. Kinet. Mech. Catal.* **110**, 387–403 (2013).
28. Mazloom, G., Farhadi, F. & Khorasheh, F. Kinetic modeling of pyrolysis of scrap tires. *J. Anal. Appl. Pyrol.* **84**, 157–164 (2009).
29. Khraibet, S. A., 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-Al2O3 catalyst. *Ind. Eng. Chem. Res.* **60**, 9729–9738 (2021).
30. Mazloom, G. A modified three-phase multistage fluid bed model by considering axial dispersion in bubble side. *Part. Sci. Technol.* **34**, 648–657 (2016).
31 Mazloom, G. & Alavi, S. M. Partial oxidation of propane over Mo1V0. 3Te0. 23Nb0. 12Ox. catalyst in a fluidized bed reactor. *Part. Sci. Technol.* **33**, 204–212 (2015).
32. Fan, M., Hu, J., Cao, R., Ruan, W. & Wei, X. A review on experimental design for pollutants removal in water treatment with the aid of artificial intelligence. *Chemosphere* **200**, 330–343 (2018).
33. Zhang, H. *et al.* Machine learning for novel thermal-materials discovery: early successes, opportunities, and challenges. *arXiv preprint* arXiv:1901.05801 *(2019)*.
34. Al Aani, S., Bonny, T., Hasan, S. W. & Hilal, N. Can machine language and artificial intelligence revolutionize process automation for water treatment and desalination?. *Desalination* **458**, 84–96 (2019).
35. Wang, Y. *et al.* A new machine learning algorithm to optimize a reduced mechanism of 2-butanone and the comparison with other algorithms. *ES Mater. Manuf.* **6**, 28–37 (2019).
36. Joshi, S. C. Knowledge based data boosting exposition on CNT-engineered carbon composites for machine learning. *Adv. Compos. Hybrid Mater.* **3**, 354–364 (2020).
37 Wu, L., Xiao, Y., Ghosh, M., Zhou, Q. & Hao, Q. Machine learning prediction for bandgaps of inorganic materials. *ES Mater. Manuf.* https://doi.org/10.30919/esmm5f756 (2020).
38. Chen, C. *et al.* Recent advances in solar energy full spectrum conversion and utilization. *ES Energy Environ.* **11**, 3–18 (2021).
39. Ke, G. *et al.* Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **30**, 3146–3154 (2017).
40. Yang, X., Dindoruk, B. & Lu, L. A comparative analysis of bubble point pressure prediction using advanced machine learning algorithms and classical correlations. *J. Petrol. Sci. Eng.* **185**, 106598 (2020).
41. Zhou, B. *et al.* Pressure of different gases injected into large-scale coal matrix: Analysis of time–space dependence and prediction using light gradient boosting machine. *Fuel* **279**, 118448. https://doi.org/10.1016/j.fuel.2020.118448 (2020).
42. Chen, T. & Guestrin, C. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794.
43. Hastie, T., Friedman, J. & Tibshirani, R. Unsupervised Learning. In: *The Elements of Statistical Learning. Springer Series in Statistics*. (Springer, New York, NY, 2001). https://doi.org/10.1007/978-0-387-21606-5_14.
44. Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367–378 (2002).
45. Amar, M. N., Shateri, M., Hemmati-Sarapardeh, A. & Alamatsaz, A. Modeling oil-brine interfacial tension at high pressure and high salinity conditions. *J. Petrol. Sci. Eng.* **183**, 106413 (2019).
46. Wu, Y. & Misra, S. Intelligent image segmentation for organic-rich shales using random forest, wavelet transform, and hessian matrix. *IEEE Geosci. Remote Sens. Lett.* **17**, 1144–1147 (2019).
47. Shaikhina, T. *et al.* Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. *Biomed. Signal Process. Control* **52**, 456–462 (2019).
48. Li, Z., Liu, X., Jin, W., Hu, Q. & Zhao, Y. Adsorption behavior of arsenicals on MIL-101 (Fe): The role of arsenic chemical structures. *J. Colloid Interface Sci.* **554**, 692–704 (2019).
49. Wu, H. *et al.* Arsenic removal from water by metal-organic framework MIL-88A microrods. *Environ. Sci. Pollut. Res.* **25**, 27196–27202 (2018).
50. Cai, J., Wang, X., Zhou, Y., Jiang, L. & Wang, C. Selective adsorption of arsenate and the reversible structure transformation of the mesoporous metal–organic framework MIL-100 (Fe). *Phys. Chem. Chem. Phys.* **18**, 10864–10867 (2016).
51. He, X. *et al.* Exceptional adsorption of arsenic by zirconium metal-organic frameworks: Engineering exploration and mechanism insight. *J. Colloid Interface Sci.* **539**, 223–234 (2019).
52. Vu, T. A. *et al.* Arsenic removal from aqueous solutions by adsorption using novel MIL-53 (Fe) as a highly efficient adsorbent. *RSC Adv.* **5**, 5261–5268 (2015).
53. Yu, W. *et al.* Metal-organic framework (MOF) showing both ultrahigh As (V) and As (III) removal from aqueous solution. *J. Solid State Chem.* **269**, 264–270 (2019).
54. Hou, S. *et al.* Green synthesis and evaluation of an iron-based metal–organic framework MIL-88B for efficient decontamination of arsenate from water. *Dalton Trans.* **47**, 2222–2231 (2018).
55. Atallah, H., Mahmoud, M. E., Jelle, A., Lough, A. & Hmadeh, M. A highly stable indium based metal organic framework for efficient arsenic removal from water. *Dalton Trans.* **47**, 799–806 (2018).
56. Zheng, X. *et al.* Efficient removal of As (V) from simulated arsenic-contaminated wastewater via a novel metal–organic framework material: Synthesis, structure, and response surface methodology. *Appl. Organomet. Chem.* **34**, e5584 (2020).
57. Liu, Z. *et al.* Synthesis of uniform-sized and microporous MIL-125 (Ti) to boost arsenic removal by chemical adsorption. *Polyhedron* **196**, 114980 (2021).
58 Abdi, J., Hadipoor, M., Hadavimoghaddam, F. & Hemmati-Sarapardeh, A. Estimation of tetracycline antibiotic photodegradation from wastewater by heterogeneous metal-organic frameworks photocatalysts. *Chemosphere* **287**, 132135 (2021).
59. Abdi, J. *et al.* Assessment of competitive dye removal using a reliable method. *J. Environ. Chem. Eng.* **2**, 1672–1683 (2014).
60. Rousseeuw, P. J. & Leroy, A. M. *Robust Regression and Outlier Detection* Vol. 589 (John Wiley & Sons, 2005).
61. Abdi, J., Hadavimoghaddam, F., Hadipoor, M. & Hemmati-Sarapardeh, A. Modeling of CO2 adsorption capacity by porous metal organic frameworks using advanced decision tree-based models. *Sci. Rep.* **11**, 24468. https://doi.org/10.1038/s41598-021-04168-w (2021).
62. Jiang, Y., Zhang, G., Wang, J. & Vaferi, B. Hydrogen solubility in aromatic/cyclic compounds: Prediction by different machine learning techniques. *Int. J. Hydrogen Energy* **46**, 23591–23602 (2021).
63. Al-Ghouti, M. A. & Al-Absi, R. S. Mechanistic understanding of the adsorption and thermodynamic aspects of cationic methylene blue dye onto cellulosic olive stones biomass from wastewater. *Sci. Rep.* **10**, 1–18 (2020).
64. Ye, S. *et al.* Facile assembled biochar-based nanocomposite with improved graphitization for efficient photocatalytic activity driven by visible light. *Appl. Catal. B* **250**, 78–88 (2019).
65. Li, J. *et al.* Zeolitic imidazolate framework-8 with high efficiency in trace arsenate adsorption and removal from water. *J. Phys. Chem. C* **118**, 27382–27387 (2014).

66. Mousavi, S. P. *et al.* Viscosity of ionic liquids: Application of the Eyring's theory and a committee machine intelligent system. *Molecules* **26**, 156 (2021).
67. Mousavi, S.-P. *et al.* Modeling surface tension of ionic liquids by chemical structure-intelligence based models. *J. Mol. Liq.* **342**, 116961 (2021).
68. Hajirezaie, S., Wu, X. & Peters, C. A. Scale formation in porous media and its impact on reservoir performance during water flooding. *J. Nat. Gas Sci. Eng.* **39**, 188–202 (2017).
69. Hosseinzadeh, M. & Hemmati-Sarapardeh, A. Toward a predictive model for estimating viscosity of ternary mixtures containing ionic liquids. *J. Mol. Liq.* **200**, 340–348 (2014).

Acknowledgements
The authors are thankful to Shahrood University of Technology for the support.

Author contributions
J.A. Writing-Review & Editing, Data curation, Methodology, Validation, Supervision. G.M. Writing-Original Draft.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-20762-y.

Correspondence and requests for materials should be addressed to J.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.