ACS ES&T Engineering

pubs.acs.org/estengg

Article

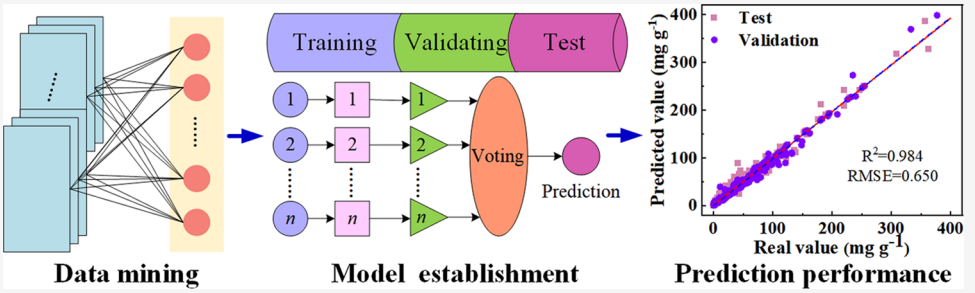
Machine Learning Prediction of Phosphate Adsorption on Six Different Metal-Containing Adsorbents

Yangyang Wu, Yingze Li, Ze Jiang, Ziyang Xu, Mingbao Yang, Jiahui Ding, and Changyong Zhang*

Cite This: https://doi.org/10.1021/acsestengg.3c00001

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Phosphate removal is a crucial objective in wastewater engineering to reduce harmful environmental impacts like eutrophication. Adsorption, a low-cost and efficient process for phosphate abatement, primarily relies on trapping phosphate on low-solubility solid surfaces. Metal-based materials, due to their abundance, low cost, environmental friendliness, and chemical stability, are considered the most promising phosphate adsorbents. However, the synthesis of appropriate adsorbents is complex and time-consuming. In addition, the diverse textural properties, the presence of various metals, and the selection of adsorption parameters make it challenging to the underlying mechanism of phosphate adsorption. In this study, we compiled a data set including 1800 data points mined from 128 peer-reviewed papers and adopted machine learning (ML) to systematically evaluate phosphate adsorption concerning textural properties, metal compositions, and adsorption parameters. We applied three different tree-based algorithms, including random forest (RF), decision trees (DTs), and extreme gradient boosting (XGBoost), to guide the design of adsorbents and predict the phosphate adsorption performances. Among the three algorithms, RF showed the best predictive performance with a high R^2 of 0.984 and a low root-mean-squared error (RMSE) of 0.650. Feature importance, based on the Shapley values, demonstrated the contributions of adsorbents' textural properties (e.g., surface area), adsorption parameters, and metal types in the order of precedence of phosphate adsorption, providing critical insights into guiding adsorbents design and synthesis for phosphate adsorption applications.

KEYWORDS: *machine learning, phosphate adsorption, tree-based algorithms, prediction*

1. INTRODUCTION

Phosphorus, a nonrenewable resource on the earth, is an indispensable element for all organisms.[1,2] Phosphorus, widely used in medicine, glass, ceramics, food, dyes, pesticides, metallurgy, and other aspects, has an important role, especially in agriculture for phosphate fertilizer production.[3] Nevertheless, when the total phosphorus concentration is greater than 0.02 mg L^{-1} in natural waterbodies, eutrophication will occur, leading to the uncontrollable growth of algae and breaking the ecological balance of aquatic systems.[4−7] Therefore, removing and recovering phosphorus from phosphorus-containing wastewater will not only eliminate water eutrophication but also alleviate global phosphorus shortage, representing great environmental and economic benefits.[8,9]

Phosphorus mainly exists in the form of inorganic phosphate in water and wastewater,[10] which can be removed by physical adsorption, chemical precipitation, ion exchange, and biological processes.[11−16] Compared with other methods, physical adsorption is an effective, reliable, and environmentally friendly phosphate removal process, which will be potentially applied in various situations, especially for decentralized wastewater treatment.[8,17−19] Regarding phosphorus adsorption, there are three interaction mechanisms including hydrogen bonding, shape complementarity (one of the mechanisms of phosphate adsorption, in which the selective removal of phosphate is achieved by designing specific cavities or shapes of molecularly imprinted polymers depending on the geometry of the phosphate anion), and inner-sphere complexation.[8,20,21]

Received: January 1, 2023
Revised: June 17, 2023
Accepted: June 28, 2023

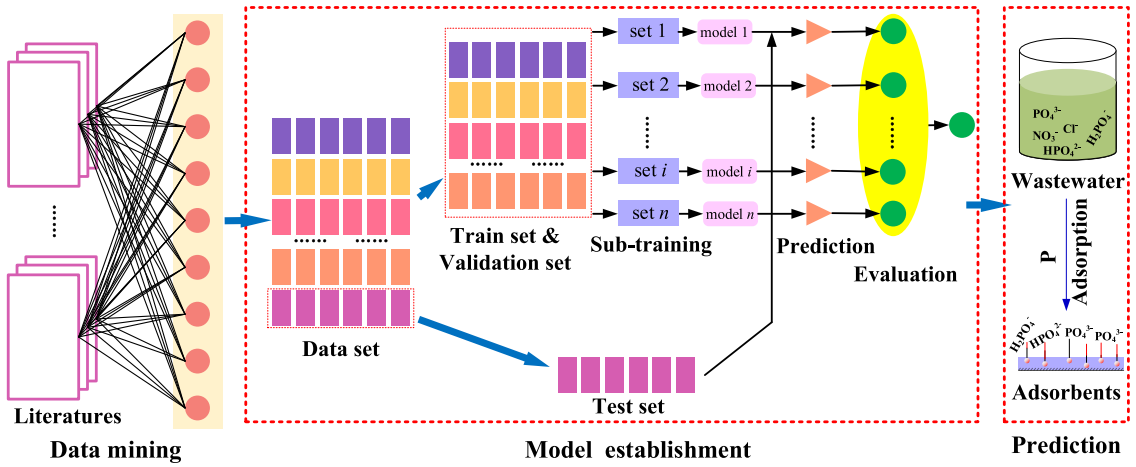


Figure 1. Schematic of the sequential approach followed in the current study, including three stages: (1) data mining: data on phosphate adsorption, including the adsorbent's characteristics and operating circumstances, were gathered from 128 peer-reviewed studies; (2) model establishment: this procedure comprises data splitting (training set, validation set, and test set) and model comparison (two metrics include RMSE and R^2); and (3) model prediction: the Sharpley value can assess the phosphate system's parameters and pinpoint crucial ones for phosphorus removal from wastewater that contains phosphorus, which helps fabricate phosphate adsorbents.

To date, various types of adsorbents have been developed for the removal of phosphate from wastewater. These include natural materials, metal oxides, and layered double hydroxides.[22,23] They are easy to get and operate, and they have high selectivity. More and more focus has been placed on the development of the phosphate adsorption capacity, selectivity by metal type, and microporous structure of metal-based adsorbents. Among them, metal-containing (e.g., lanthanum (La), aluminum (Al), iron (Fe), zirconium (Zr), calcium (Ca), and magnesium (Mg)) adsorbents are promising considering their environmental friendliness and high affinity for phosphate.[8] Apart from the types and amounts of the metals, the selection of appropriate base materials is also of great significance in fabricating extraordinary adsorbents. The phosphate adsorption isotherms of metal-based adsorbents at different pH values, temperatures, times, and rates of agitation have been widely studied for acquiring phosphate adsorption capacities to guide adsorbent design, optimization, and fabrication.[24] The final phosphate adsorbents are expected to have a high specific surface area, large pore volume, appropriate pore size, excellent hydrophilicity, and other features to achieve satisfying phosphate removal performance.[25]

Traditional adsorbents are made with a variety of characteristics in mind, including numerous hydrogen bonds, channels with regular shapes, and unique affinities. A variety of techniques have been applied to modify and activate adsorbents, including composite/doping, soaking, calcination, precipitation, acidification, and heat treatment. These conventional techniques have many drawbacks such as complex processes and high chemical and energy consumption. To comprehensively evaluate the phosphate adsorption processes, influencing factors such as the external environment, adsorption material, and adsorption solution must be considered. In contrast to a single evaluation metric, a data set with complex data and a large amount is required. Advancements in data science, such as machine learning (ML) and artificial intelligence (AI), have been applied to large data sets in many environment-related research areas to better interpret the complex relationships between system variables that affect system behavior and to provide new insights into their understanding and solution development.[26] In order to optimize the phosphate adsorption process and direct the choice and preparation of adsorption materials, we can simultaneously apply the ML approach to the adsorption of phosphate. Previous documents have reported the applications of ML in solving multiple environmental problems such as heavy metal removal,[27,28] micropollutant oxidation,[29,30] seawater desalination,[31,32] carbon dioxide adsorption,[33,34] and municipal solid-waste treatment.[35,36] A number of ML models such as bagging,[37] linear regression (LR),[38] neural networks (NNs),[39] and support vector machines (SVMs),[40] and tree-based ML models have been developed in previous studies. Particularly, decision tree-based algorithms, including gradient boosting decision tree (GBDT), decision tree (DT), extreme gradient boosting (XGBoost), categorical boosting (CatBoost), light gradient boosting machine (LightGBM), and random forest (RF), are a subcategory of supervised ML models.[29,41−47] DT is a tree structure that is like a binary tree or multitree. XGboost and LightGBM belong to GBDT (gradient boosting decision tree), and in these methods, a tree structure includes two separate steps. First, the appropriate structure for the tree must be found. Second, leaf values must be set as soon as the tree structure is finalized.[48] CatBoost, short for classification enhancement, is a state-of-the-art, open-source toolbox for gradient improvement that can handle the challenge of addressing the fundamentally distinct ideas of classification features. In comparison to deep learning models, RF in the scikit-learn package is a resilient ensemble model that can be used to make accurate predictions with a limited number of model parameters and is gaining more interest in the scientific and technical communities.[32] These tree-based algorithms have gained increasing popularity due to their ability to handle relatively small data sets (200−1000 data points) with more robust and faster hyperparameter tuning compared with widely used ANN and SVM models.[33] In the field of machine learning, the mainstream prediction algorithms include RF, DTs, SVM, NN, and others. We found that while NN can predict accurately based on training data, it often suffers from overfitting as it learns noise in the data. Additionally, the lack of gating in neural network algorithms leads to long running times, which reduces

https://doi.org/10.1021/acsestengg.3c00001
ACS EST Engg. XXXX, XXX, XXX−XXX

computational efficiency.[49] SVM algorithms are sensitive to parameter tuning, which has a significant impact on model convergence.[50] On the other hand, RF algorithms, which are based on multiple independent decision trees, can solve overfitting or underfitting problems. They are very robust and easy to understand and interpret and have good accuracy.[51] In comparison with XGBoost and DTs, RF can handle overfitting and has low sensitivity to noise owing to the ensemble training methods known as bootstrap aggregation.[32,52,53]

Given the aforementioned gaps, a data-driven ML model was developed to predict phosphate adsorption from wastewaters by considering both adsorbent design and operational conditions. We first mined previous documents for a data set of six metal hydroxide/oxide (i.e., lanthanum (La), aluminum (Al), iron (Fe), zirconium (Zr), calcium (Ca), and magnesium (Mg))-based adsorbents by collecting 1800 adsorption data points (SI Table S1). We then compared two different data spitting methods (i.e., point selection and group selection) to address the potential data leakage issue and selected the data that are the most appropriate as the training set to build a robust prediction model. We also investigated the performances of three different models, including RF, DT, and XGBoost, by comparing the root-mean-squared error (RMSE) and correlation of determination (R^2) (refer to the Section 2). Finally, to quantify the contribution of 12 descriptors to the overall adsorption capacity and identify the feature importance, Shapely values were calculated. These results help to analyze phosphate adsorption via metal hydroxide/oxide-based adsorbents, guide the adsorbent design and synthesis, and identify the influence of background ions on phosphate adsorption performances. A schematic of the sequential approach followed in this study is presented in Figure 1.

The practical application of the model guides the synthesis of the material according to the importance of the parameters of the adsorption system (e.g., if the specific surface area is the most prominent in the adsorption system, then the focus should be on increasing the specific surface area of the material when designing the material), which will facilitate the removal of phosphate from the water column. In view of the above, the model was developed by considering the material design and the adsorption process. The material design includes metal type, metal content, specific surface area, pore volume, and average pore size; the operating parameters of the adsorption process include temperature, pH, equilibrium concentration, and equilibrium adsorption volume. The model parameters were evaluated by averaging absolute Shapley values to derive the magnitude of their contribution to phosphate removal from water and thus guide phosphorus removal from the wastewater.

2. METHODOLOGY

2.1. Data Collection and Formatting. In the adsorption process of phosphate, it was important to distinguish between phosphate adsorption and deposition. Phosphate precipitation was the reaction of phosphate with cations such as Al^{3+}, Fe^{2+}/Fe^{3+}, and Ca^{2+} to form minerals such as aluminum phosphate, iron phosphate, and calcium phosphate, respectively. Precipitation was a slow process that permanently changed to metallic phosphate, and reversibility was very difficult.[54,55] However, the mechanism of phosphate adsorption involved hydrogen bond formation, shape complementarity, and internal complexation in three ways, making the process reversible.[11,56] So, the desorption of phosphate from the adsorbent material could be achieved by the addition of alkaline substances. We obtained data from the literature on phosphate adsorption by metal adsorbents, which was dominated by adsorption. Adsorption and surface precipitation were not well distinguished in the reviewed literature. However, the authors of these studies have used these experimental data to differentiate between the different modes of adsorption.

Data were collected according to the following steps: (i) searching literature through Web of Science and Google Scholar by using the topic "phosphate adsorption or phosphate removal" to find relevant studies from 2006 to 2022, (ii) manually checking the titles, abstracts, and keywords of these documents to narrow down the studies to phosphate adsorption by six metal hydroxide/oxide (i.e., La, Al, Fe, Zr, Ca, and Mg)-based adsorbents, (iii) carefully reading through these papers and selecting the papers that reported the key variables included in Table S1, which led to a final list of 128 papers (the full list of these publications is available in the SI), and (iv) thoroughly reading each paper and extracting phosphate adsorption performance data.

The following criteria and assumptions were incorporated during the data extraction process.

(1) Data set (14 descriptors) included metal types, surface area, pore size, pore volume, solution pH, system temperature, background, adsorption capacity, and adsorption equilibrium concentrations in peer-reviewed research papers.
(2) The isotherm fitting coefficient (R^2) needed to be higher than 0.9, with at least six experimental points in each adsorption isotherm.
(3) The experimental data were directly collected from the tables and texts or extracted from the figures by using PlotDigitizer software (http://plotdigitizer.sourceforge.net/) presented in the 128 papers. All data were carefully selected to avoid duplication during the data collection process.
(4) Descriptors can be classified into two types according to system features: (i) operational parameters including pH that can influence the form of phosphate in aqueous solution and temperature that will affect phosphate adsorption kinetics; (ii) the textural properties of the adsorbents such as the elemental compositions of metal contents (wt %), specific surface area (SA, $m^2\ g^{-1}$), total pore volume (TPV, $cm^3\ g^{-1}$), and average pore size (APS, nm) of the adsorbents.
(5) Note that the equilibrium phosphate adsorption capacity was the output of the black box system, whereas the descriptors were used as the input during the modeling process.

As mentioned above, we collected 1800 data points (SI Table S1) associated with six types of metal (i.e., La, Fe, Zr, Ca, Al, and Mg)-based adsorbents from 128 references.

2.2. Data Preprocessing. All data were transformed into consistent units for model training (details are given in SI Text S1). Note that TPV and/or APS data were sometimes absent in the literature. The missing TPV and/or APS data can be imputed through ML methods to avoid discarding records of these data. Specifically, multilinear regression (MLR), the least-squares method (LSM), and RF models can be employed to calculate and impute the missing values of TPV and APS (details are given in the Section 3). Note that some literature did not show the specific metal content (wt %), and we

https://doi.org/10.1021/acsestengg.3c00001
ACS EST Engg. XXXX, XXX, XXX−XXX

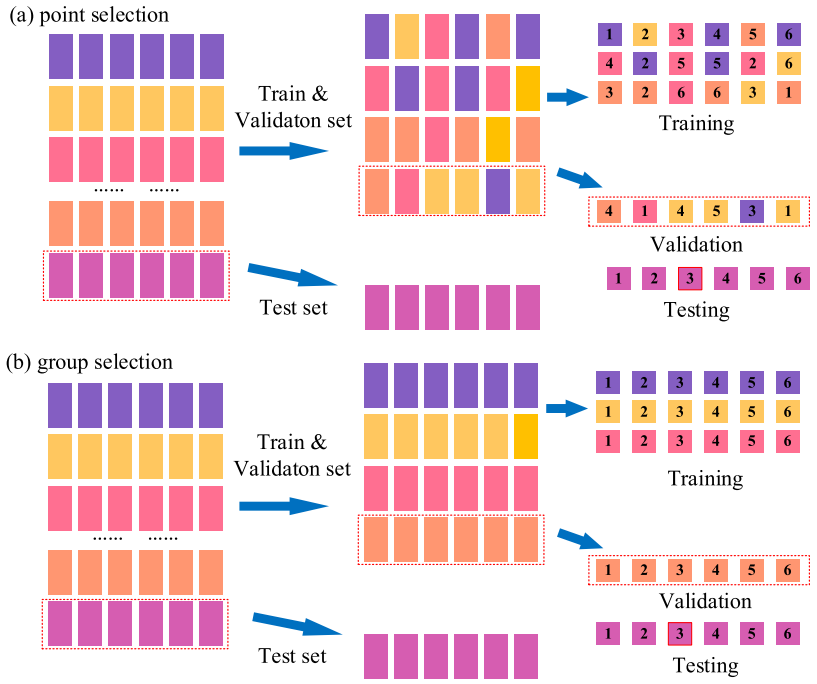


Figure 2. Two different data splitting methods (i.e., point selection and group selection). (a) Point selection: the test set is first filtered away from 15% of the data set. Then, the data in the training set and the validation set do not identify the experimental points (a particular isotherm) and are divided into 70 and 15% from the rest of the data set. (b) Group selection: the selection of the test set is the same as point selection, but the training set and the test set need to be divided into 70 and 15% from the rest of the data set according to groups of six experimental points on the adsorption isotherm.

ignored the detailed information but used "1" or "0" to represent the presence of metal or not during the model process.

The interrelationship between two parameters was described by the Pearson correlation coefficients (PCCs),[33,57,58] which are calculated as follows.

$$\rho = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - \overline{x})(y_j - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{j=1}^{n}(y_j - \overline{y})^2}} \quad (1)$$

where ρ is the value of PCC, $\overline{x}$ and $\overline{y}$ denote the means of input feature x and output variable y, respectively, and n denote the total amount of the sample. The value of ρ ranges from −1 to 1, where the two variables show no correlation when ρ is equal to 0. The closer the ρ to 1 after taking the absolute value, the stronger the correlation between the two variables.

Finally, after completing data collection, data preprocessing for units' consistency, and filling in missing values, the data, including the input feature and output target, were normalized to obtain a uniform range of values. The following equation, eq 2, can achieve this goal.[33]

$$x_i^* = \frac{x_i - \overline{x}_i}{s_d} \quad (2)$$

where x_i represents the value of the input feature, x_i^* represents the standardized value of initial x_i, $\overline{x}_i$ is the mean value of x_i, and S_d indicates the standard deviation of x_i.

Data splitting is the first and probably the most critical step. During the ML model training process, it is of great significance to prevent data leakage by applying appropriate data splitting methods. In this study, two different data splitting methods (Figure 2) were employed and compared: (i) point selection (Figure 2a) that randomly divided data points from the originally collected data minus the test set into training and validation sets (note that the test set still only included the entire group); and (ii) group selection (Figure 2b) that selected the entire group as the training set, validation set, or test set. In brief, the entire group includes a series of experimental points and the characteristics of the adsorbents.

2.3. Modeling Methods and Hyperparameter Tuning.

Three tree-based ML algorithms (e.g., RF, DTs, and XGBoost) were evaluated and compared to predict phosphate adsorption on the adsorbents. Previous documents have proved that small data sets (e.g., 200−1000 data points) were suitable for the implementation of tree-based ML models;[40,57,58] thus, the collected data points (~1800) were enough for ML modeling. Hyperparameter tuning is an important process to find a set of hyperparameters to achieve optimal model performance. The optimal configuration of hyperparameter tuning was completed by Python 3.10 with a scikit-learn package (SI Text S2). Grid search, random search, and Bayesian optimization are the most common hyperparameter tuning methods reported in the literature. In the current study, the model hyperparameters were tuned with the grid search method, considering its reliability and easy implementation when tuning for a lower set of input features.

2.3.1. Error Metrics. RMSE and R^2 values were employed to evaluate the performance of the RF models with different evaluation metrics, XGBoost models with different kernels, and DT models with different hyperparameters. It is known that the lower the RMSE and the higher the R^2, the greater the model accuracy, as shown in eqs 3 and 4, respectively.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_1^n (y_i - \hat{y}_i)^2}{n}} \quad (3)$$

https://doi.org/10.1021/acsestengg.3c00001
ACS EST Engg. XXXX, XXX, XXX−XXX

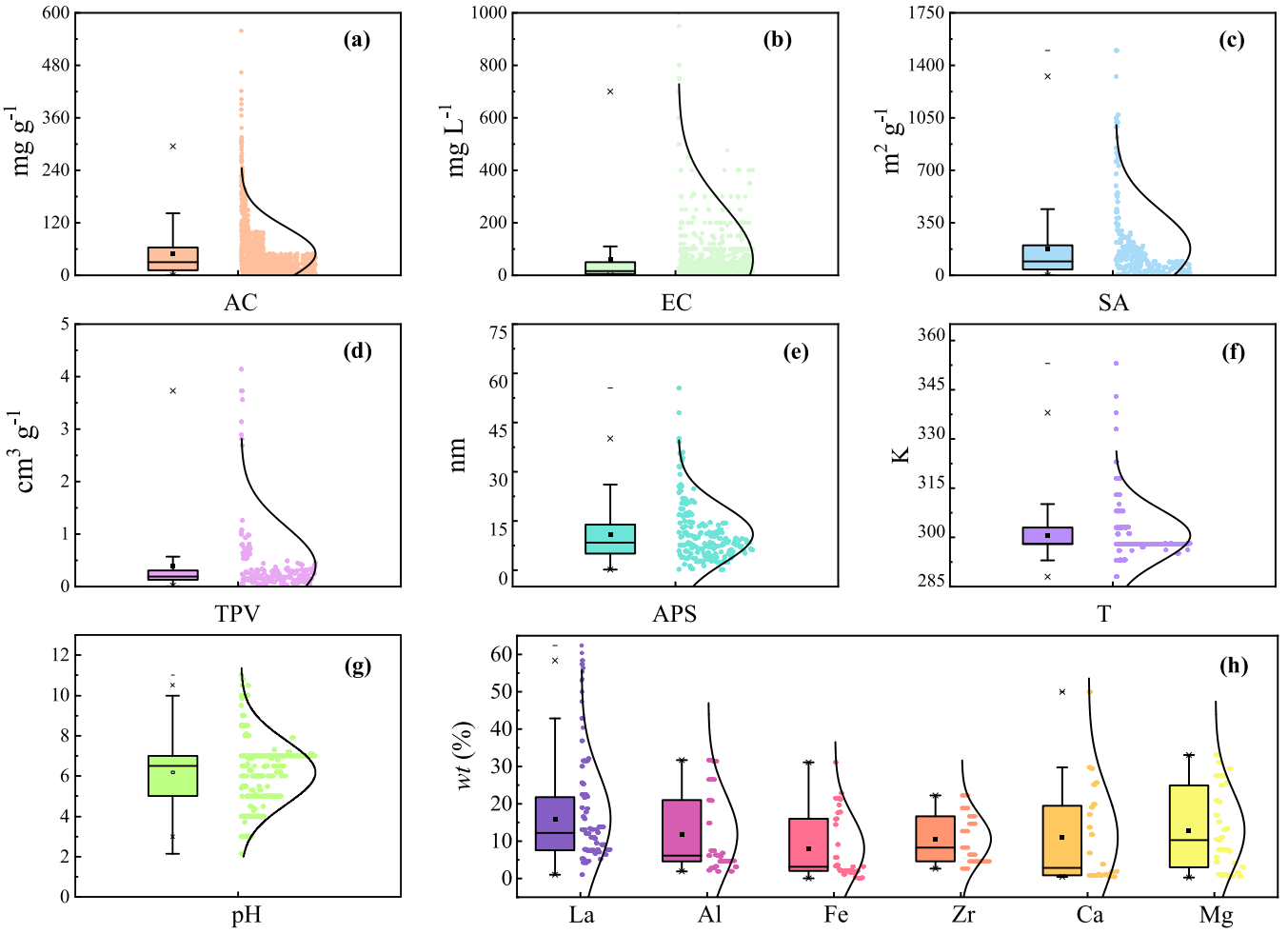


Figure 3. Box-normal plots representing the descriptive statistics of the data for each feature collected from peer-reviewed papers. The data were selected from 128 papers, which include both input and target features such as (a) adsorption capacity (AC, mg g^{-1}), (b) equilibrium concentration (EC, mg L^{-1}), (c) surface area (SA, m^2 g^{-1}), (d) total pore volume (TPV, cm^3 g^{-1}), (e) average pore size (APS, nm), (f) temperature (T, K), (g) pH, and (h) mass fraction of six metals: La (wt %), Al (wt %), Fe (wt %), Zr (wt %), Ca (wt %), and Mg (wt %).

$$R^2 = 1 - \frac{\sum_1^n (y_i - \hat{y}_i)^2}{\sum_1^n (y_i - \overline{y}_i)^2} \quad (4)$$

where y_i, $\hat{y}_i$, and $\overline{y}_i$ represent the actual, predicted, and mean values of the target feature, respectively, and n is the total number of objects in the validation/test set.

2.3.2. Feature Importance. To quantify the contribution of a specific input descriptor to the model is of great significance, as tree-based models are typically black boxes. Shapley additive explanation (SHAP), a package originating from Python (details are given in SI Text S3), can be employed to determine the Shapley values, which are generally used to quantify the contribution of each input feature to the overall adsorption performance and to evaluate whether the ML models violated any adsorption rules. Mean absolute Shapley (MAS) for an input feature is the mean of all of the Shapley values of the feature. Conceptually, the larger the MAS value, the more significant the descriptor in influencing the phosphate adsorption process.

3. RESULTS AND DISCUSSION

3.1. Descriptive Statistics on Phosphate Adsorption. Based on the output parameters, it was clear that the equilibrium adsorption capacity was dependent on the dosage of the adsorbent. However, there was a reciprocal restriction between the two, indicating that increasing the dosage of the adsorbent would lead to an increase in the equilibrium adsorption capacity up to a certain point, after which the capacity would start to decrease. This meant that there was an optimal dosage of adsorbent that could be used to achieve the maximum adsorption capacity. Furthermore, when evaluating the adsorption process, it is more scientifically pertinent to consider varying adsorbent capacities for the same dosage of the adsorbent instead of just the adsorbent dosage. To explicate the influence of different adsorbent dosages on the adsorption capacity under different conditions, it was necessary to determine the adsorption capacity of phosphate at the equilibrium phosphate concentration.

In order to gain a deeper understanding of the adsorption process, the raw data was systematically described and analyzed. This involved identifying all of the input features and target variables and then looking at minimum, maximum, and average values using analytics. By doing this, it was possible to acquire a preliminary insight into the raw data, which set the foundation for further detailed analysis. Figure 3 presents a visual distribution of the input features and the target variables in the form of box-normal plots. The maximum and minimum phosphate adsorption capacities (ACs) were

https://doi.org/10.1021/acsestengg.3c00001
ACS EST Engg. XXXX, XXX, XXX−XXX

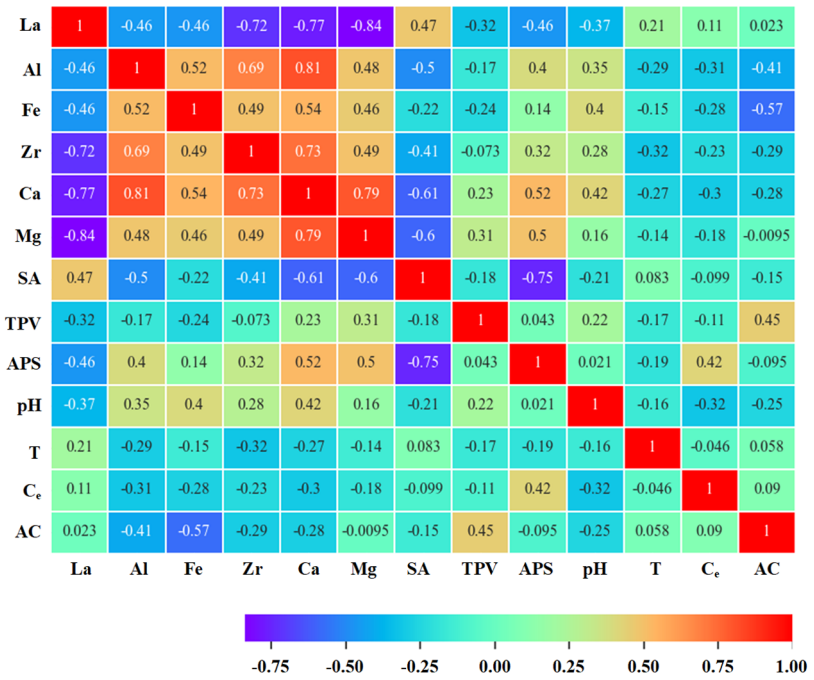


Figure 4. Pearson correlation matrix for all of the set of features included in this study. No significant correlation among the input variables was observed. La, Al, Fe, Zr, Ca, Mg, SA, TPV, APS, T, Ce, and AC denote lanthanum (wt %), aluminum (wt %), iron (wt %), zirconium (wt %), calcium (wt %), magnesium (wt %), specific surface area ($m^2\ g^{-1}$), total pore volume ($cm^3\ g^{-1}$), average pore size (nm), temperature (K), equilibrium concentration ($mg\ L^{-1}$), and adsorption capacities ($mg\ g^{-1}$), respectively.

558.78 and 0.05 $mg\ g^{-1}$, respectively, with an average value of 49.52 $mg\ g^{-1}$ (Figure 3a). As shown in Figure 3b, the values of equilibrium concentration (Ce) ranged from 0.1 to 800 $mg\ L^{-1}$, resulting in an average Ce value of 60.40 $mg\ L^{-1}$. At the same time, it is worth noting that a significant majority of the data set (specifically, 94% of observations) had equilibrium concentrations exceeding 10 $mg\ L^{-1}$. This observation implies that the model may have been optimized specifically for higher concentration levels and may not be as effective for lower values.[19] The properties of adsorbents can be described by analyzing the materials' SA, TPV, and APS, among which SA was reported in all of the literature. Previous documents also indicated that SA and TPV were significantly influenced by carbonization, activation treatment, and modification processes.[25,59] It can be noticed from Figure 3c−e that the values of SA, TPV, and APS fluctuated significantly, with mean values of 185.62 $m^2\ g^{-1}$, 0.40 $cm^3\ g^{-1}$, and 10.94 nm, respectively. To enhance the phosphate adsorption performances, metal-containing (e.g., La, Al, Fe, Zr, Ca, and Mg) adsorbents are developed, considering their environmental friendliness and high affinity for phosphate. The mass fraction of these metals loaded on the materials conformed to a Gaussian distribution (Figure 3h). In addition, the Shapley value was used to evaluate the contribution of every descriptor. Considering the lack of data about textual properties and the value of the exact metal weight fraction in the raw data set, applying a strong correlation algorithm could solve these problems and avoid shrinking the amount of the data (SI Text S4). Apart from the physicochemical characteristics of the adsorbents, operational parameters such as temperature and pH also play crucial roles in phosphate adsorption at liquid−solid interfaces.[60,61] The temperature will influence the random thermal motion of ions in the solution,[62] therefore resulting in the difference in phosphate adsorption kinetics. In addition, the solution pH is also closely related to factors including the net positive charge on adsorbents' surface and the form of aqueous phosphate species.[63] It can be observed from Figure 3f,g that most of the experiments were conducted in normal conditions (e.g., the temperature ranged from 298 to 302 K, and the pH ranged from 5 to 7).

Figure 4 illustrates the PCC values among various individual parameters. Most of these values reside within the range of −0.5 to 0.5, indicating a lack of significant correlation between the input variables. However, the top left corner of the figure highlights noteworthy correlations, specifically between La, Zr, Ca, Al, and Mg (PCC values: −0.84(La−Mg),0.79(Ca−Mg), and 0.81(Al−Ca)). This observation suggests that these parameters might be interdependent or interconnected, necessitating further investigation to elucidate their influence on the model output and optimize its performance. Three plausible reasons for this phenomenon are as follows: (i) One possible explanation for this phenomenon is that these two metal ions exhibit similar affinity during the adsorption of phosphate, resulting in their interactions causing concomitant changes in their respective contents and consequently displaying correlation.[64] (ii) Another conceivable reason is the synergistic effect of these metal ions during the phosphate adsorption process. Their mutual interaction enhances their interaction with phosphate, leading to concurrent alterations in their respective contents and manifesting a correlation.[65] (iii) Moreover, other factors might also impact the correlation between these metals. For instance, they may possess similar attributes, such as charge (La^{3+}, Al^{3+}, Zr^{2+}, Ca^{2+}, and Mg^{2+}) and equilibrium constant ($K_{sp}(LaPO_4)$ = 22.43, $K_{sp}(Mg_3(PO_4)_2)$ = 23.98), which contribute to a concomitant trend in phosphate adsorption.

3.2. Different Algorithms and Data Splitting Approaches. In this study, models based on three tree-based algorithms, RF, DT, and XGBoost, performed well in solving this problem. According to the RMSE and R^2 values, the model

https://doi.org/10.1021/acsestengg.3c00001
ACS EST Engg. XXXX, XXX, XXX−XXX

developed on DT and XGBoost can be accepted. However, the RF model achieves the best performance and thus is selected after comprehensively evaluating both R^2 and RMSE. The performances of the two data splitting methods (i.e., point selection and group selection) are presented in Figure 5.

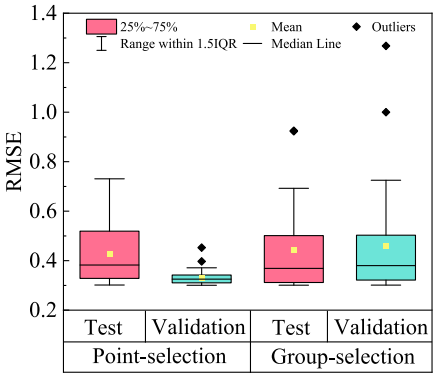


Figure 5. Different performances of data splitting approaches (e.g., point selection and group selection). It should be noticed that the collected data was split into a training set (70%), validation set (15%), and test set (15%), respectively.

A large difference in RMSE values between the two sets was observed in the point selection method, with the validation set showing a much lower RMSE value compared to the test set. In contrast, using the group selection approach resulted in comparable RMSE values for the validation and test sets. In addition, there was no obvious difference in RMSE values between the test set in point selection method and the test/validation set in the group selection method. These results indicate that the validation set in the group selection method can roughly reflect the prediction capability of the tree-based algorithms on the test set, confirming the robustness of this data splitting method.

The existence of the potential data leakage in the point selection approach could be attributed to the above results. The process of splitting randomly in the point selection approach needs to be given more attention because some data from a group would appear in the training set and the rest would remain in the validation set. That is, data leakage leads to overfitting, which to some extent, negatively influences the prediction performance of the model.[40,66] In this case, the training set has already covered some features (i.e., adsorbent properties and operational conditions) of the validation set in point selection, resulting in lower RMSE values of the validation set. However, the test set is independent of the training set; thus, the model might not perform well for the test set. For group selection, three data sets (i.e., training set, validation set, and test set) are independent, leading to better consistency in the prediction performance. In other words, this method can achieve similar RMSE values for the validation set and the test set. In contrast, the point selection method cannot satisfy the goal.[67] Therefore, the group selection method was selected in this study.

3.3. Cosine Similarity. Previous documents reported that when data distribution ranges in descriptor values are similar in the training set, validation set, and test set, a smaller training set can also achieve a relatively satisfactory prediction performance for the test/validation set compared to that using a much larger data set.[68] That is, rather than using all the collected data, we can select highly correlated data to build models and predict target values by using algorithms such as Cosine similarity,[69] Euclidean distance,[68] and City-block distance.[70] In addition to the problems that these algorithms can solve, more consideration should be paid to whether the algorithms conform to the physicochemical meaning of phosphate adsorption processes. To generalize the above finding, first, taking the cosine similarity, one of the different methods, for example, the method can reflect the physicochemical significance through the cosine value of the angle. Unlike cosine similarity, the distance seems to have no relation with the nature of the adsorption system. Second, these methods in RF are employed for comparison in selecting the training set. Then, through the calculation of the validation set and the test set of the data set as the criteria, Figure 6 shows

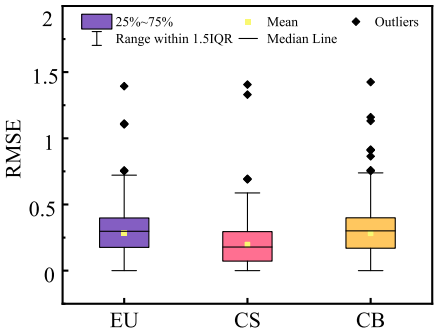


Figure 6. Comparison of the performance based on three methods selected: Euclidean distance (EU), cosine similarity (CS), and city-block distance (CB). The prediction target in each of the three methods was the same, and the difference was in the training sets.

that the cosine similarity method extensively yields better performance (the RMSE is 0.215) than the Euclidean approach (RMSE 0.381), which consecutively is better than the city-block rule (RMSE 0.436).

3.4. Model Performance. Compared with tree-based models, specifically RF, DT, and XGBoost, the RF model achieved high accuracy in predicting the adsorption of phosphate (Figure 7). All models are developed based on the same data samples, among which hyperparameters are tuned through the grid search. According to the features of mining data, the content of each metal was not clear in terms of the adsorbent. For devising and evaluating tree-based models in the context of prediction, the input of the metal type needs to be considered in the aforementioned method (details are given in the Section 2.2). Therefore, on the one hand, it can make full use of the limited amount of data to avoid data set waste; on the other hand, for the prediction of the adsorption effect, we analyze and qualitatively compare the adaptability of various metals to the algorithm (Figure 7a–c). At the same time, using the existing metal content of 654 data points to establish three tree-based models, the RF is more conducive to the practical application of analytical guidance (Figure 7d–f).

Figure 7d–f shows that the actual content of the metal makes the predictions of the model slightly broader compared to the data set of the existing metal content. From Figure 7a–c, the R^2 values of tree-based ML models using the master data set via the group method are as follows: XGBoost: validation R^2 = 0.986, test R^2 = 0.962; RF: validation R^2 = 0.990, test R^2 = 0.984; and DT: validation R^2 = 0.984, test R^2 = 0.980. It

https://doi.org/10.1021/acsestengg.3c00001
ACS EST Engg. XXXX, XXX, XXX−XXX

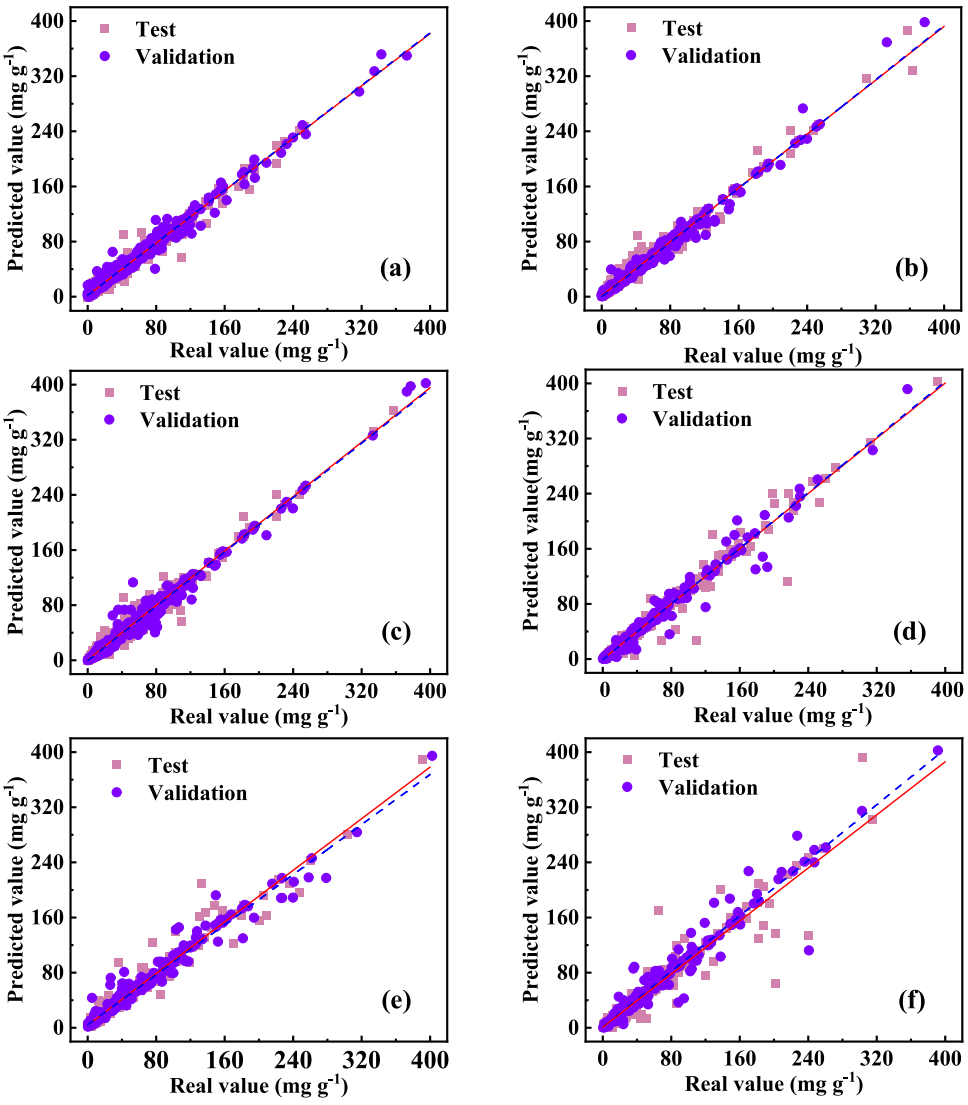


Figure 7. Performance of three tree-based models for the adsorption of phosphate. The prediction based on the cosine similarity method is acquired for each group of input data. (a–c) Different performances between the prediction and observation in the entire data set using XGBoost, RF, and DT, respectively. (d–f) Results calculated through the existing metal content of 654 data points using XGBoost, RF, and DT, respectively. The blue dot line represents the fitting line of the validation test, and the red solid line represents the fitting line of the test set.

suggests that RF performs the best among the three models with the highest R^2 in test samples. Furthermore, the RF model for the content of metals specifically also presents robustness and reliability with the smallest RMSE of 0.650 (Table 1). In addition, the effect of the doping of different metals on the material is given in the following section.

3.5. Feature Importance Analysis. To determine the most influential effects on phosphate adsorption, the Shapley value and MAS were used to evaluate the contribution of each descriptor to equilibrium adsorption (Figures 8 and 9, and more details are in SI Figure S2). In this study, although the RF model has the best performance, specific performance needs to be conducted on the different sets mentioned above in terms of practical application. According to the currently collected data, the three strategies for processing data sets are conducive to establishing models, predicting classification, and acquiring material properties. The Shapley value originates from game theory to solve the distribution in cooperation.[71] The Shapley value can be quantified to express the donations that are not equivalent for all factors in this study (Figures 8a, 9a, and S2a). The positive or negative Shapley values, calculated by each descriptor in a certain system, represent the positive or negative effects on the equilibrium adsorption capacity, respectively, and vice versa. Furthermore, the effect of all descriptors on the adsorption was assessed through MAS

Table 1. Summary of the Performance Based on Different Algorithms by Comparison between Semiqualitative and Quantitative Metals

algorithms	validation R^2	test R^2	RMSE
XGBoost[a]	0.986	0.962	0.756
RF[a]	0.990	0.984	0.650
DT[a]	0.984	0.980	0.778
XGBoost[b]	0.973	0.966	0.699
RF[b]	0.968	0.956	0.695
DT[b]	0.952	0.910	0.796

[a]The data range of original data sets, among which all of metal content were treated as 0 and 1. [b]The data range of 654 data points included the existing metal content.

https://doi.org/10.1021/acsestengg.3c00001
ACS EST Engg. XXXX, XXX, XXX−XXX

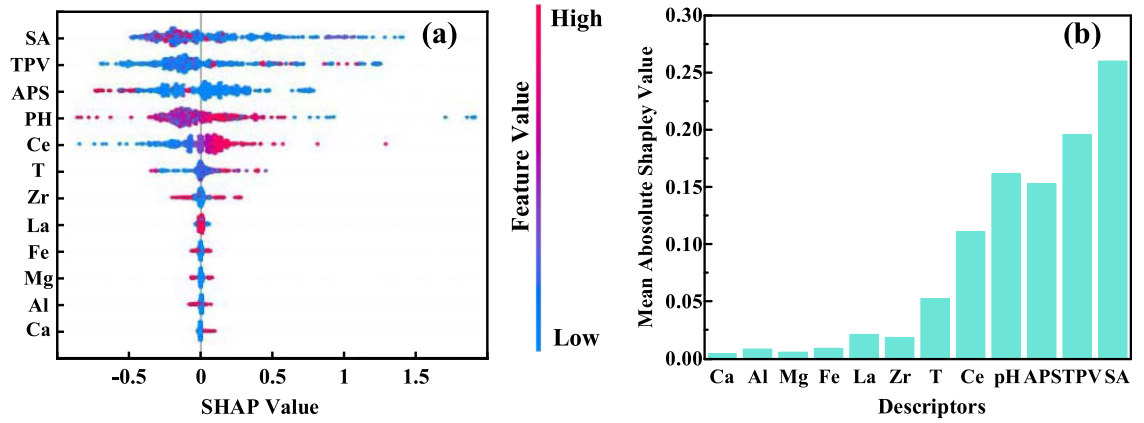


Figure 8. (a) Shapley values of all descriptors in the adsorption of phosphate by distinguishing the existence of metals as 0 and 1 with all data points and (b) mean absolute Shapley (MAS) values for all of the descriptors in the selected model. La, Al, Fe, Zr, Ca, and Mg denote lanthanum (wt %), aluminum (wt %), iron (wt %), zirconium (wt %), calcium (wt %), and magnesium (wt %), respectively. SA, TPV, APS, T, and Ce denote the surface area (m^2 g^{-1}), total pore volume (cm^3 g^{-1}), average pore size (nm), temperature (K), and equilibrium concentration (mg L^{-1}), respectively.

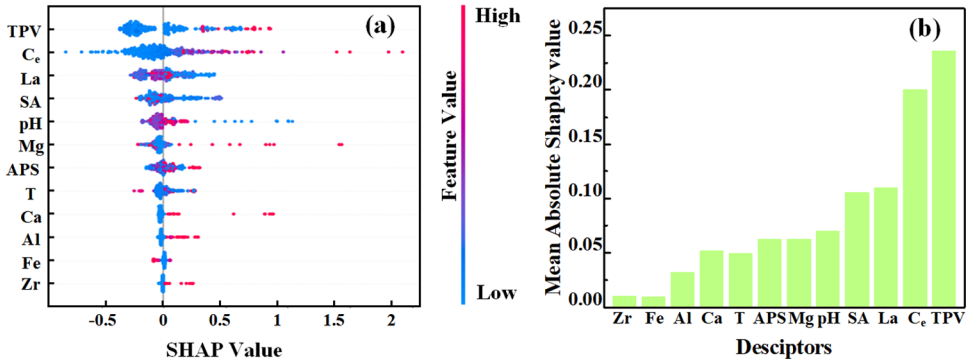


Figure 9. (a) Shapley values of all descriptors in the adsorption of phosphate by adding the definite content of the metal with partial data sets and (b) mean absolute Shapley (MAS) values for all of the descriptors in the selected model. La, Al, Fe, Zr, Ca, and Mg denote lanthanum (wt %), aluminum (wt %), iron (wt %), zirconium (wt %), calcium (wt %), and magnesium (wt %), respectively. SA, TPV, APS, T, and Ce denote the surface area (m^2 g^{-1}), total pore volume (cm^3 g^{-1}), average pore size (nm), temperature (K), and equilibrium concentration (mg L^{-1}), respectively.

values, which is the averaged value from all of the Shapley values. As a result, the larger the MAS value of the descriptor, the greater its effect on phosphate adsorption.

For data sets that treated metals as 0 and 1, SA performs the most influential with the largest value of MAS of 0.26, given in Figure 8b. Regardless of the type of metal, the total pore volume and average pore size of the adsorbent contribute more to the adsorption, followed by the properties of the solution (Figure 9). It should be noted that the specific content of the metal used to calculate the MAS values shows substantial differences compared to the simplified method mentioned above. From the perspective of adsorbent preparation, this is very meaningful. That is, as shown in Figure 9b, the lanthanide in these metals shows a great influence on the adsorption of phosphate, followed by calcium, magnesium, and aluminum. Instead, total pore volume has become the most influential factor in adsorption, and it is also well understood that loading metals depend on the porosity of the material. The equilibrium concentration of phosphate in solution also has a great influence on phosphate adsorption. This analysis is based on the interpretation of the RF model built on Shapley as well as MAS values. We found that the importance of the involved features follows the order of TPV > Ce > La > SA > pH > Mg > APS > Ca > T > Al > Fe > Zr in the specific percentages of different metals in the adsorbent (Figure 9). As a result, we conclude that the total pore volume and the equilibrium concentration are the most influential factors for phosphate adsorption in terms of the physicochemical meaning. These results have been supported by previous studies, and some examples can be found in their related literature works.[72−74]

To further study the influence of background ions in the actual environment on phosphate adsorption, adsorption removal experiments of phosphate solution containing other ions (i.e., chloride ions (Cl^-), nitrate ions (NO_3^-), and cadmium ions (Cd^{2+})) were selected in the literature. However, there are few studies on isotherms for phosphate and other ions in solution, so the data of phosphate adsorption with background ions cannot well reflect the influence of other ions on phosphate adsorption (details are given in the Section 3.6). According to the MAS value obtained from the existing data analysis, the background ions have little interference with the phosphate by other ions, which may be attributed to the selectivity of the adsorbent material (SI Figure S2b).

3.6. Limitations of the Built Models. In the process of model development, the tree-based model performs better than other algorithms in terms of the RMSE and R^2 metrics. There are three main drawbacks that limit the practical applications: (i) metal hydroxide/oxide materials, applied to phosphate adsorption, need more precise descriptors for a given material; the preparation and characterization of

https://doi.org/10.1021/acsestengg.3c00001
ACS EST Engg. XXXX, XXX, XXX−XXX

materials while ignoring the importance of applications will result in the lack of the specific composition of material quantification. (ii) For phosphate in natural lakes along with actual wastewater, there are many interference factors (i.e., ionic species). In contrast, most studies only focus on the effect of phosphate removal in a simulated environment, ignoring its application in real wastewater. (iii) The concentration level of a variable can indeed impact the effectiveness of a model, and it may be necessary to adjust the model's settings to ensure that it can accommodate different ranges of data. Failing to do so could potentially affect the model's overall accuracy and usefulness. After reviewing the literature, we found that such high capacities are often reported at equilibrium concentrations much greater than 10 mg L^{-1}, which may represent unrealistic conditions affecting the model's performance when applied to more realistic, lower concentration levels in effluent polishing scenarios. (iv) Many studies on phosphate adsorption mainly concentrate on the adsorption process and the materials and factors influencing it, often neglecting the desorption and regeneration aspects. This dearth of comprehensive information on desorption and regeneration hampers the accumulation of sufficient data for applying machine learning techniques to this facet of adsorbent research. Nonetheless, this study highlights the necessary targets and objectives for future phosphate adsorption experiments. Incorporating descriptors such as the presence of background ions, the composition of the adsorbent, and adsorbent recyclability in experiments will contribute to refining the model's construction and understanding. Further investigation into the analysis process mechanisms will enhance the future applicability assessment of adsorbents in practical scenarios.

4. ENVIRONMENTAL IMPLICATIONS

In this study, we employed a tree-based algorithm to build robust, tuneable, and chemically meaningful models for predicting phosphate adsorption in aqueous solutions, utilizing 1800 experimental data points from 128 studies. By employing a group selection approach during model training and the cosine similarity method for the original data, the improvement in model performance was found to be reliant on data preprocessing techniques. Although the accuracy of the ML modeling process is contingent on data volume, the cosine similarity method allows for better predictions based on a limited number of experimental data sets.

With the interpretation of the RF model built on Shapley and MAS values, we concluded that the total pore volume and the equilibrium concentration are the most influential for phosphate adsorption in terms of physicochemical significance. Concurrently, lanthanide metal-loaded materials exhibited superior performance compared to other metals. Owing to the insights from this research, the next endeavors include the following: (i) To improve the applicability of the adsorbent in practical scenarios, it is crucial to conduct experiments that not only focus on phosphate adsorption but also consider other ions commonly present in wastewater, such as organic matter, nitrate ions, and chloride ions. Through obtaining high-quality characterization data from these experiments, the model can be better constructed and understood, leading to more effective use of the adsorbent; (ii) to develop an instructive synthesis approach to achieve the maximum phosphate adsorption, which will benefit from the rational design of adsorbents and adsorption parameters; (iii) given the different environments in which phosphates are located, consideration based on the multiangle, deep-level, and all-round preparation of suitable materials is applied in appropriate scenarios; and (iv) the open database source that is convenient for other researchers need to add new experimental data points for optimizing the model in operational applications.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsestengg.3c00001.

More detailed explanation of the methods and training process in this study, the figures mentioned in the main text, source data compiled for this research, and additional figures and tables to support the training process (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Changyong Zhang − *CAS Key Laboratory of Urban Pollutant Conversion, Department of Environmental Science and Engineering, University of Science and Technology of China, Hefei 230026, P. R. China;* orcid.org/0000-0002-9288-627X; Email: changyongzhang@ustc.edu.cn

Authors

Yangyang Wu − *CAS Key Laboratory of Urban Pollutant Conversion, Department of Environmental Science and Engineering, University of Science and Technology of China, Hefei 230026, P. R. China*

Yingze Li − *Guangdong Provincial Key Laboratory for Green Chemical Product Technology, School of Chemistry and Chemical Engineering, South China University of Technology, Guangzhou 510640, P. R. China*

Ze Jiang − *School of Civil and Environmental Engineering, University of New South Wales, Sydney, New South Wales 2052, Australia;* orcid.org/0000-0002-3472-0829

Ziyang Xu − *CAS Key Laboratory of Urban Pollutant Conversion, Department of Environmental Science and Engineering, University of Science and Technology of China, Hefei 230026, P. R. China*

Mingbao Yang − *State Key Laboratory of Public Big Data, College of Computer Science and Technology, Guizhou University, Guiyang 550025, P. R. China*

Jiahui Ding − *CAS Key Laboratory of Urban Pollutant Conversion, Department of Environmental Science and Engineering, University of Science and Technology of China, Hefei 230026, P. R. China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsestengg.3c00001

Notes

The authors declare no competing financial interest.

The authors present the ML pipeline in this study in the form of a Python document hosted on GitHub (https://github.com/Tomas-Wu-SCUT/Phosphate-adsorption-by-six-different-metal-containing-adsorbents). Two clean Excel datasheets with the essential data are also attached to this repository, which can be directly imported into the Python compiler (such as PyCharm, Spider, and others). The pipeline is generic and cross-deployable in nature and can be used as a template for the analysis of phosphate adsorbents.

https://doi.org/10.1021/acsestengg.3c00001
ACS EST Engg. XXXX, XXX, XXX−XXX

■ ACKNOWLEDGMENTS

The authors acknowledge the financial support received from the Hundred-Talent Program of the Chinese Academy of Sciences (KY2400000016).

■ REFERENCES

(1) Liu, X.; Yuan, Z.; Liu, X.; Zhang, Y.; Hua, H.; Jiang, S. Historic Trends and Future Prospects of Waste Generation and Recycling in China's Phosphorus Cycle. *Environ. Sci. Technol.* **2020**, *54*, 5131−5139.
(2) Zhang, C.; Wang, M.; Xiao, W.; Ma, J.; Sun, J.; Mo, H.; Waite, T. D. Phosphate selective recovery by magnetic iron oxide impregnated carbon flow-electrode capacitive deionization (FCDI). *Water Res* **2021**, *189*, No. 116653.
(3) Carvalho, A.; Wang, M.; Zhu, X.; Rodin, A. S.; Su, H.; Castro Neto, A. H. Phosphorene: from theory to applications. *Nat. Rev. Mater.* **2016**, *1*, 1−16.
(4) Seviour, R. J.; Mino, T.; Onuki, M. The microbiology of biological phosphorus removal in activated sludge systems. *Fems Microbiol. Rev.* **2003**, *27*, 99−127.
(5) Mayer, B. K.; Baker, L. A.; Boyer, T. H.; Drechsel, P.; Gifford, M.; Hanjra, M. A.; Parameswaran, P.; Stoltzfus, J.; Westerhoff, P.; Rittmann, B. E. Total Value of Phosphorus Recovery. *Environ. Sci. Technol.* **2016**, *50*, 6606−6620.
(6) Lin, S. S.; Shen, S. L.; Zhou, A.; Lyu, H. M. Assessment and management of lake eutrophication: A case study in Lake Erhai, China. *Sci. Total Environ.* **2021**, *751*, No. 141618.
(7) Wang, Z.; Wang, C.; Jiang, H.; Liu, H. Higher dissolved oxygen levels promote downward migration of phosphorus in the sediment profile: Implications for lake restoration. *Chemosphere* **2022**, *301*, No. 134705.
(8) Wu, B.; Wan, J.; Zhang, Y.; Pan, B.; Lo, I. M. C. Selective Phosphate Removal from Water and Wastewater using Sorption: Process Fundamentals and Removal Mechanisms. *Environ. Sci. Technol.* **2020**, *54*, 50−66.
(9) Wang, C.; Jiang, H.-L. Chemicals used forin situimmobilization to reduce the internal phosphorus loading from lake sediments for eutrophication control. *Crit. Rev. Environ. Sci. Technol.* **2016**, *46*, 947−997.
(10) Jia, Y.; Sun, S.; Wang, S.; Yan, X.; Qian, J.; Pan, B. Phosphorus in water: A review on the speciation analysis and species specific removal strategies. *Crit. Rev. Environ. Sci. Technol.* **2023**, *53*, 435−456.
(11) Peng, L.; Dai, H.; Wu, Y.; Peng, Y.; Lu, X. A comprehensive review of phosphorus recovery from wastewater by crystallization processes. *Chemosphere* **2018**, *197*, 768−781.
(12) Ye, Y.; Ngo, H. H.; Guo, W.; Liu, Y.; Li, J.; Liu, Y.; Zhang, X.; Jia, H. Insight into chemical phosphate recovery from municipal wastewater. *Sci. Total Environ.* **2017**, *576*, 159−171.
(13) Liu, Y.; Fan, Q.; Wang, S.; Liu, Y.; Zhou, A.; Fan, L. Adsorptive removal of fluoride from aqueous solutions using Al-humic acid-La aerogel composites. *Chem. Eng. J.* **2016**, *306*, 174−185.
(14) Mangal, M. N.; Salinas-Rodriguez, S. G.; Dusseldorp, J.; Kemperman, A. J. B.; Schippers, J. C.; Kennedy, M. D.; van der Meer, W. G. J. Effectiveness of antiscalants in preventing calcium phosphate scaling in reverse osmosis applications. *J. Membr. Sci.* **2021**, *623*, No. 119090.
(15) Nazarian, R.; Desch, R. J.; Thiel, S. W. Kinetics and equilibrium adsorption of phosphate on lanthanum oxide supported on activated carbon. *Colloid Surf. A-Physicochem. Eng. Asp.* **2021**, *624*, No. 126813.
(16) Zhang, Y.; Tang, Q.; Sun, Y.; Yao, C.; Yang, Z.; Yang, W. Improved utilization of active sites for phosphorus adsorption in FeOOH/anion exchanger nanocomposites via a glycol-solvothermal synthesis strategy. *J. Environ. Sci.* **2022**, *111*, 313−323.
(17) Loganathan, P.; Vigneswaran, S.; Kandasamy, J.; Bolan, N. S. Removal and Recovery of Phosphate From Water Using Sorption. *Crit. Rev. Env. Sci. Technol.* **2014**, *44*, 847−907.
(18) Othman, A.; Dumitrescu, E.; Andreescu, D.; Andreescu, S. Nanoporous Sorbents for the Removal and Recovery of Phosphorus from Eutrophic Waters: Sustainability Challenges and Solutions. *ACS Sustainable Chem. Eng.* **2018**, *6*, 12542−12561.
(19) Kumar, P. S.; Korving, L.; van Loosdrecht, M. C. M.; Witkamp, G. J. Adsorption as a technology to achieve ultra-low concentrations of phosphate: Research gaps and economic analysis. *Water Res. X* **2019**, *4*, No. 100029.
(20) Ewen, S. L.; G S, J. H. Molecularly imprinted polymers using anions as templates. *Recognit. Anions* **2008**, 207−248.
(21) Vasapollo, G.; Sole, R. D.; Mergola, L.; Lazzoi, M. R.; Scardino, A.; Scorrano, S.; Mele, G. Molecularly imprinted polymers: present and future prospective. *Int. J. Mol. Sci.* **2011**, *12*, 5908−5945.
(22) Ahmed, S.; Ashiq, M. N.; Li, D.; Tang, P.; Leroux, F.; Feng, Y. Recent Progress on Adsorption Materials for Phosphate Removal. *Recent Pat. Nanotechnol.* **2019**, *13*, 3−16.
(23) He, Q.; Zhao, H.; Teng, Z.; Wang, Y.; Li, M.; Hoffmann, M. R. Phosphate removal and recovery by lanthanum-based adsorbents: A review for current advances. *Chemosphere* **2022**, *303*, No. 134987.
(24) Elkhlifi, Z.; Sellaoui, L.; Zhao, M.; Ifthikar, J.; Jawad, A.; Shahib, I. I.; Sijilmassi, B.; Lahori, A. H.; Selvasembian, R.; Meili, L.; Gendy, E. A.; Chen, Z. Lanthanum hydroxide engineered sewage sludge biochar for efficient phosphate elimination: Mechanism interpretation using physical modelling. *Sci. Total Environ.* **2022**, *803*, No. 149888.
(25) Loganathan, P.; Vigneswaran, S.; Kandasamy, J.; Bolan, N. S. Removal and Recovery of Phosphate From Water Using Sorption. *Crit. Rev. Environ. Sci. Technol.* **2014**, *44*, 847−907.
(26) Lowry, G. V.; Boehm, A. B.; Brooks, B. W.; Gago-Ferrero, P.; Jiang, G.; Jones, G. D.; Liu, Q.; Ren, Z. J.; Wang, S.; Zimmerman, J. Data Science for Advancing Environmental Science, Engineering, and Technology: Upcoming Special and Virtual Issues in ES&T and ES&T Letters. *Environ. Sci. Technol. Lett.* **2022**, *9*, 581−582.
(27) Palansooriya, K. N.; Li, J.; Dissanayake, P. D.; Suvarna, M.; Li, L.; Yuan, X.; Sarkar, B.; Tsang, D. C. W.; Rinklebe, J.; Wang, X.; Ok, Y. S. Prediction of Soil Heavy Metal Immobilization by Biochar Using Machine Learning. *Environ. Sci. Technol.* **2022**, *56*, 4187−4198.
(28) Yang, H.; Huang, K.; Zhang, K.; Weng, Q.; Zhang, H.; Wang, F. Predicting Heavy Metal Adsorption on Soil with Machine Learning and Mapping Global Distribution of Soil Adsorption Capacities. *Environ. Sci. Technol.* **2021**, *55*, 14316−14328.
(29) Cha, D.; Park, S.; Kim, M. S.; Kim, T.; Hong, S. W.; Cho, K. H.; Lee, C. Prediction of Oxidant Exposures and Micropollutant Abatement during Ozonation Using a Machine Learning Method. *Environ. Sci. Technol.* **2021**, *55*, 709−718.
(30) Baek, S. S.; Choi, Y.; Jeon, J.; Pyo, J.; Park, J.; Cho, K. H. Replacing the internal standard to estimate micropollutants using deep and machine learning. *Water Res.* **2021**, *188*, No. 116535.
(31) Priya, P.; Nguyen, T. C.; Saxena, A.; Aluru, N. R. Machine Learning Assisted Screening of Two-Dimensional Materials for Water Desalination. *ACS Nano* **2022**, *16*, 1929−1939.
(32) Bonny, T.; Kashkash, M.; Ahmed, F. An efficient deep reinforcement machine learning-based control reverse osmosis system for water desalination. *Desalination* **2022**, *522*, No. 115443.
(33) Yuan, X.; Suvarna, M.; Low, S.; Dissanayake, P. D.; Lee, K. B.; Li, J.; Wang, X.; Ok, Y. S. Applied Machine Learning for Prediction of CO2 Adsorption on Biomass Waste-Derived Porous Carbons. *Environ. Sci. Technol.* **2021**, *55*, 11925−11936.
(34) Burns, T. D.; Pai, K. N.; Subraveti, S. G.; Collins, S. P.; Krykunov, M.; Rajendran, A.; Woo, T. K. Prediction of MOF Performance in Vacuum Swing Adsorption Systems for Postcombustion CO2 Capture Based on Integrated Molecular Simulations, Process Optimizations, and Machine Learning Models. *Environ. Sci. Technol.* **2020**, *54*, 4536−4544.
(35) Zhu, X.; Yang, G. Study on HHV prediction of municipal solid wastes: A machine learning approach. *Int. J. Energy Res.* **2021**, *46*, 3663−3673.
(36) Liang, R.; Chen, C.; Kumar, A.; Tao, J.; Kang, Y.; Han, D.; Jiang, X.; Tang, P.; Yan, B.; Chen, G. State-of-the-art applications of machine learning in the life cycle of solid waste management. *Front. Environ. Sci. Eng.* **2022**, *17*, No. 44.

https://doi.org/10.1021/acsestengg.3c00001
ACS EST Engg. XXXX, XXX, XXX−XXX

(37) Prasad, A. M.; Iverson, L. R.; Liaw, A. Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems* **2006**, *9*, 181−199.
(38) Timofei S, K. L.; Suzuki, T.; et al. Multiple Linear Regression (MLR) and Neural Network (NN) calculations of some disazo dye adsorption on cellulose. *Dyes Pigment.* **1997**, *34*, 181−193.
(39) Rumelhart, D. E.; Geoffrey, E. H.; Williams, R. J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533−536.
(40) Li, J.; Pan, L.; Suvarna, M.; Tong, Y. W.; Wang, X. Fuel properties of hydrochar and pyrochar: Prediction and exploration with machine learning. *Appl. Energy* **2020**, *269*, No. 115166.
(41) Belgiu, M.; Drăguţ, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS-J. Photogramm. Remote Sens.* **2016**, *114*, 24−31.
(42) Boulesteix, A.-L.; Janitza, S.; Kruppa, J.; König, I. R. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdiscip. Rev.: Data Mining Knowl. Discov.* **2012**, *2*, 493−507.
(43) 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.* **2021**, *11*, No. 24468.
(44) Chen, K.; Chen, H.; Zhou, C.; Huang, Y.; Qi, X.; Shen, R.; Liu, F.; Zuo, M.; Zou, X.; Wang, J.; Zhang, Y.; Chen, D.; Chen, X.; Deng, Y.; Ren, H. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water Res.* **2020**, *171*, No. 115454.
(45) Hu, Y.; Du, W.; Yang, C.; Wang, Y.; Huang, T.; Xu, X.; Li, W. Source identification and prediction of nitrogen and phosphorus pollution of Lake Taihu by an ensemble machine learning technique. *Front. Environ. Sci. Eng.* **2022**, *17*, No. 55.
(46) Yang, D.; Wang, L.; Yuan, P.; An, Q.; Su, B.; Yu, M.; Chen, T.; Hu, K.; Zhang, L.; Lu, Y.; Du, G. Cocrystal virtual screening based on the XGBoost machine learning model. *Chin. Chem. Lett.* **2023**, *34*, No. 107964.
(47) Zeng, T.; Liang, Y.; Dai, Q.; Tian, J.; Chen, J.; Lei, B.; Yang, Z.; Cai, Z. Application of machine learning algorithms to screen potential biomarkers under cadmium exposure based on human urine metabolic profiles. *Chin. Chem. Lett.* **2022**, *33*, 5184−5188.
(48) Dorogush, A. V.; Ershov, V.; Gulin, A.CatBoost: gradient boosting with categorical features support *ArXiv* 2018, 1810, p 11363.
(49) Tu, J. V. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J Clin Epidemiol* **1996**, *49*, 1225−1231.
(50) Anguita, D.; Ghio, A.; Greco, N.; Oneto, L.; Ridella, S. *Model Selection for Support Vector Machines: Advantages and Disadvantages of the Machine Learning Theory*; IJCNN, 2010; pp 1−8.
(51) Sheykhmousa, M.; Mahdianpari, M.; Ghanbari, H.; Mohammadimanesh, F.; Ghamisi, P.; Homayouni, S. Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review. *IEEE J. Sel. Top Appl. Earth Obs. Remote Sens.* **2020**, *13*, 6308−6325.
(52) Rodriguez-Galiano, V. F.; Ghimire, B.; Rogan, J.; Chica-Olmo, M.; Rigol-Sanchez, J. P. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS-J. Photogramm. Remote Sens.* **2012**, *67*, 93−104.
(53) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.
(54) Prasad R, C. D.Phosphorus basics Understanding phosphorus forms and their cycling in the soil *Alabama Coop. Ext. Syst.* 2019, pp 1−4.
(55) Ye, Y.; Ngo, H. H.; Guo, W.; Liu, Y.; Li, J.; Liu, Y.; Zhang, X.; Jia, H. Insight into chemical phosphate recovery from municipal wastewater. *Sci. Total Environ.* **2017**, *576*, 159−171.
(56) Wu, B.; Wan, J.; Zhang, Y.; Pan, B.; Lo, I. M. C. Selective Phosphate Removal from Water and Wastewater using Sorption: Process Fundamentals and Removal Mechanisms. *Environ. Sci. Technol.* **2020**, *54*, 50−66.
(57) Li, J.; Zhu, X.; Li, Y.; Tong, Y. W.; Ok, Y. S.; Wang, X. Multi-task prediction and optimization of hydrochar properties from high-moisture municipal solid waste Application of machine learning on waste-to-resource. *J. Cleaner Prod.* **2021**, *278*, No. 123928.
(58) Zhu, X.; Wang, X.; Ok, Y. S. The application of machine learning methods for prediction of metal sorption onto biochars. *J. Hazard. Mater.* **2019**, *378*, No. 120727.
(59) Huang, Y.; Lee, X.; Grattieri, M.; Yuan, M.; Cai, R.; Macazo, F. C.; Minteer, S. D. Modified biochar for phosphate adsorption in environmentally relevant conditions. *Chem. Eng. J.* **2020**, *380*, No. 122375.
(60) Liu, J.; Wan, L.; Zhang, L.; Zhou, Q. Effect of pH, ionic strength, and temperature on the phosphate adsorption onto lanthanum-doped activated carbon fiber. *J. Colloid Interface Sci.* **2011**, *364*, 490−496.
(61) Mezenner, N. Y.; Bensmaili, A. Kinetics and thermodynamic study of phosphate adsorption on iron hydroxide-eggshell waste. *Chem. Eng. J.* **2009**, *147*, 87−96.
(62) Kılıç, M.; Kırbıyık, Ç.; Çepelioğullar, Ö.; Pütün, A. E. Adsorption of heavy metal ions from aqueous solutions by bio-char, a by-product of pyrolysis. *Appl. Surf. Sci.* **2013**, *283*, 856−862.
(63) He, J.; Xu, Y.; Wang, W.; Hu, B.; Wang, Z.; Yang, X.; Wang, Y.; Yang, L. Ce(III) nanocomposites by partial thermal decomposition of Ce-MOF for effective phosphate adsorption in a wide pH range. *Chem. Eng. J.* **2020**, *379*, No. 122431.
(64) Adak, M. K.; Sen, A.; Mukherjee, A.; Sen, S.; Dhak, D. Removal of fluoride from drinking water using highly efficient nano-adsorbent, Al(III)-Fe(III)-La(III) trimetallic oxide prepared by chemical route. *J. Alloys Compd.* **2017**, *719*, 460−469.
(65) Yu, J.; Xiang, C.; Zhang, G.; Wang, H.; Ji, Q.; Qu, J. Activation of Lattice Oxygen in LaFe (Oxy)hydroxides for Efficient Phosphorus Removal. *Environ. Sci. Technol.* **2019**, *53*, 9073−9080.
(66) Porumb, M.; Iadanza, E.; Massaro, S.; Pecchia, L. A convolutional neural network approach to detect congestive heart failure. *Biomed. Signal Process. Control* **2020**, *55*, No. 101597.
(67) Saeb, S.; Lonini, L.; Jayaraman, A.; Mohr, D. C.; Kording, K. P.Voodoo machine learning for clinical predictions *BioRxiv* 2016059774.
(68) Zhang, K.; Zhong, S.; Zhang, H. Predicting Aqueous Adsorption of Organic Compounds onto Biochars, Carbon Nanotubes, Granular Activated Carbons, and Resins with Machine Learning. *Environ. Sci. Technol.* **2020**, *54*, 7008−7018.
(69) Xia, P.; Zhang, L.; Li, F. Learning similarity with cosine similarity ensemble. *Inf. Sci.* **2015**, *307*, 39−52.
(70) Rizo Rodríguez, S. I.; Tenório de Carvalho, F. d. A. Clustering interval-valued data with adaptive Euclidean and City-Block distances. *Expert Syst. Appl.* **2022**, *198*, No. 116774.
(71) Winter, E. Chapter 53 The Shapley Value. In *Handbook of Game Theory with Economic Applications*; Elsevier, 2002; Vol. 3, pp 2025−2054.
(72) Li, S.; Ma, X.; Ma, Z.; Dong, X.; Wei, Z.; Liu, X.; zhu, L. Mg/Al-layered double hydroxide modified biochar for simultaneous removal phosphate and nitrate from aqueous solution. *Environ. Technol. Innov.* **2021**, *23*, No. 101771.
(73) Wang, Z.; Huang, Z.; Zheng, B.; Wu, D.; Zheng, S. Efficient removal of phosphate and ammonium from water by mesoporous tobermorite prepared from fly ash. *J. Environ. Chem. Eng.* **2022**, *10*, No. 107400.
(74) Kuhn, D. C.; Cabral, L. L.; Pereira, I. C.; Gonçalves, A. J.; Maciel, G. M.; Haminiuk, C. W. I.; Nagalli, A.; Passig, F. H.; Carvalho, K. Q. d. Development of aerated concrete waste/white cement composite for phosphate adsorption from aqueous solutions: Characterization and modeling studies. *Chem. Eng. Process.* **2023**, *184*, No. 109284.

https://doi.org/10.1021/acsestengg.3c00001
ACS EST Engg. XXXX, XXX, XXX−XXX