Nuclear Science and Techniques (2025) 36:181
https://doi.org/10.1007/s41365-025-01759-4

# Prediction of radionuclide diffusion enabled by missing data imputation and ensemble machine learning

Jun-Lei Tian[1] · Jia-Xing Feng[1] · Jia-Cong Shen[1] · Lei Yao[1] · Jing-Yan Wang[1] · Tao Wu[1] · Yao-Lin Zhao[2]

Received: 3 December 2024 / Revised: 12 January 2025 / Accepted: 3 February 2025 / Published online: 17 July 2025

**Abstract**
Missing values in radionuclide diffusion datasets can undermine the predictive accuracy and robustness of the machine learning (ML) models. In this study, regression-based missing data imputation method using a light gradient boosting machine (LGBM) algorithm was employed to impute more than 60% of the missing data, establishing a radionuclide diffusion dataset containing 16 input features and 813 instances. The effective diffusion coefficient ($D_e$) was predicted using ten ML models. The predictive accuracy of the ensemble meta-models, namely LGBM-extreme gradient boosting (XGB) and LGBM-categorical boosting (CatB), surpassed that of the other ML models, with $R^2$ values of 0.94. The models were applied to predict the $D_e$ values of $EuEDTA^-$ and $HCrO_4^-$ in saturated compacted bentonites at compactions ranging from 1200 to 1800 kg/m$^3$, which were measured using a through-diffusion method. The generalization ability of the LGBM-XGB model surpassed that of LGB-CatB in predicting the $D_e$ of $HCrO_4^-$. Shapley additive explanations identified total porosity as the most significant influencing factor. Additionally, the partial dependence plot analysis technique yielded clearer results in the univariate correlation analysis. This study provides a regression imputation technique to refine radionuclide diffusion datasets, offering deeper insights into analyzing the diffusion mechanism of radionuclides and supporting the safety assessment of the geological disposal of high-level radioactive waste.

**Keywords** Machine learning · Radionuclide diffusion · Bentonite · Regression imputation · Missing data · Diffusion experiments

## 1 Introduction

Bentonite is often selected as an engineering barrier in a high-level radioactive waste (HLW) repositories due to its low hydraulic conductivity, which leads to a diffusion-controlled process for radionuclide transport [1–4]. The effective diffusion coefficient ($D_e$), a critical parameter in the safety assessment of repositories, describes the diffusion behavior of radionuclides in porous media [5–7]. Under complex disposal conditions, $D_e$ is affected by the properties of radionuclides, such as diffusing species and adsorption properties [8]; the characteristics of bentonite, such as compaction, pore structure, and physical and chemical properties [3, 9, 10]; and the porewater chemistry, such as pH and ionic strength [11–14]. Over the few decades, considerable attention has been devoted to determining the $D_e$ of radionuclides in compacted bentonite [1, 8, 15–17].

Predicting the $D_e$ of radionuclides is both challenging and crucial due to the nonlinear and complex interactions among radionuclides, porewater, and bentonite [2, 3]. Machine learning (ML) models are valuable tools for this task because they can manage complex and high-dimensional data. Various ML models, such as the light gradient boosting machine (LGBM), extreme gradient boosting (XGB), categorical boosting (CatB),

This work was partially supported by the National Natural Science Foundation of China (No. 12475340 and 12375350), Special Branch project of South Taihu Lake, and the Scientific Research Fund of Zhejiang Provincial Education Department (No. Y202456326).

✉ Tao Wu
twu@zjhu.edu.cn

✉ Yao-Lin Zhao
zhaoyaolin@mail.xjtu.edu.cn

[1] Huzhou Key Laboratory of Environmental Functional Materials and Pollution Control, Huzhou University, Huzhou 313000, China

[2] School of Nuclear Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China

support vector machine (SVM), random forest (RF), and artificial neural networks (ANN), have been applied to predict the $D_e$ of radionuclides in compacted bentonite [18–21]. Radionuclide diffusion datasets were compiled from experimental data published in the literatures and a radionuclide diffusion database established by the Japan Atomic Energy Agency (JAEA-DDB). These datasets included numerous input features ranged from 3 to 16 and the data size ranged from 293 instances to 956 instances [19–21]. It is worth mentioning that the JAEA-DDB collected over 5000 instances from radionuclide diffusion experiments spanning 1982 to 2009 [22]. However, the instances increased with decreasing input features, primarily due to the missing data, resulting in a potential impact on the accuracy and reliability of the ML model explanations.

The issues caused by missing data are a pervasive concern in databases [23, 24]. Missing data can lead to suboptimal outcomes, reduce predictive performance, and even result in misleading conclusions [25, 26]. For instance, the dry density and rock capacity factor have been reported as the two most influential factors in predicting the $D_e$ [20, 21]. In contrast, Wu et al. (2024) observed that the ion diffusion coefficient in water and dry density were the top-two contributors. This discrepancy can be attributed to an insufficient number of instances in the datasets used. Therefore, a comprehensive dataset is essential to provide a more reliable analysis of the diffusion mechanisms.

This study presents a novel, comprehensive radionuclide diffusion dataset with micro-mesoscopic features using ML models as regression imputation techniques. Firstly, the LGBM was employed as a regression-based missing data imputation method to impute over 60% of the missing data. Subsequently, ten ML models, including three ensemble ML algorithms (LGBM-CatB, LGBM-XGB, and LGBM-RF), four decision-tree algorithms (LGBM, CatB, XGB, and RF), support vector machine (SVM), and two neural networks (ANN and deep neural network (DNN)), were trained, optimized, and tested by fivefold cross-validation to predict $D_e$ values. Finally, through-diffusion experiments were conducted to measure the diffusion parameters of $EuEDTA^-$ and $HCrO_4^-$ in compacted bentonite, including $D_e$, rock capacity factor, accessible porosity, total porosity, and distribution coefficient, to evaluate the generalization of the trained ML models. The goal was to develop predictive models that exhibit high accuracy, strong robustness, and clear interpretability for radionuclide diffusion studies, which are crucial for the safety assessment of HLW repositories.

## 2 Materials and Methods

### 2.1 Material

Ba-bentonite was prepared by modifying Gaomiaozi (GMZ) bentonite with a $BaCl_2$ solution. The mass percentage of $BaCl_2$ in modified bentonite was 5%. The detailed procedures for this modification have been previously described [16]. Wyoming bentonite powder had the grain dry density of 2760 kg/m$^3$, montmorillonite content of 0.85, external surface area of 38 m$^2$/g, and cation exchange capacity of 78.7 meq/100 g [27, 28]. Ba-bentonite powder had the grain dry density of 2710 kg/m$^3$, montmorillonite content of 0.78, external surface area of 27.3 m$^2$/g, and cation exchange capacity of 58.7 meq/100 g [16].

All the solid chemicals were purchased from Aladdin. The pH values of the NaCl solution were adjusted to 5.0 ± 0.1 and 7.0 ± 0.1 for $EuEDTA^-$ and $HCrO_4^-$ diffusion experiments, respectively. A stock solution of $EuEDTA^-$ was prepared by dissolving a measured amount of $EuNO_3$. 6 $H_2O$ in 200 mL of a solution mixed with 0.6 mol/L NaCl and 0.01 mol/L EDTA. Similarly, a stock solution of $HCrO_4^-$ was prepared by dissolving a measured amount of $K_2Cr_2O_7$ in 200 mL of 0.5 mol/L NaCl solution. The initial concentrations of $HCrO_4^-$ and $EuEDTA^-$ were $1.8 \times 10^{-3}$ mol/L and $5.7 \times 10^{-4}$ mol/L, respectively, with corresponding pH values of 5.3 ± 0.1 and 6.8 ± 0.1. The uncertainty in the pH was determined based on the standard deviation derived from the five source solutions for $HCrO_4^-$ and $EuEDTA^-$. Excess EDTA ensured the complete complexation of Eu(III).

### 2.2 Through-diffusion method

A through-diffusion method was used to measure the diffusion parameters of $EuEDTA^-$ and $HCrO_4^-$ in compacted bentonites. The experiments were operated under ambient conditions, with pH 5.3 ± 0.1 and a temperature of 25 ± 3 °C for $EuEDTA^-$ diffusion, and pH 6.8 ± 0.1 and a temperature of 15 ± 3 °C for $HCrO_4^-$ diffusion. The bentonite powder was compacted into cylindrical blocks with dry densities in the range of 1200−1800 kg/m$^3$. The powder, with an initial water content of approximately 5%, was calculated to weigh between 7.8 and 11.4 g for the preparation of the bentonite blocks. During the weighing process and preparation of bentonite blocks in the experimental procedure, approximately 0.3 g of bentonite powder was lost. This loss represents the primary source of uncertainty in the compacted dry density. Table 1 summarizes the experimental conditions used in diffusion experiments. After the compacted bentonite blocks were mounted in the diffusion setups, they were saturated for five weeks with NaCl solution in the diffusion cells. The diffusion experiments lasted 90 days for $EuEDTA^-$ and 25 days for $HCrO_4^-$.

Cr and Eu concentrations were measured using an inductively coupled plasma optical emission spectrometer

**Table 1** Overview of the experimental condition for $EuEDTA^-$ and $HCrO_4^-$ diffusion experiments

| Experimental conditions | Detailed information | |
|---|---|---|
| Anion | $EuEDTA^-$ | $HCrO_4^-$ |
| Bentonite type | Ba-bent | Wyoming |
| Initial concentration (× $10^{-3}$ mol/L) | 0.57 ± 0.02 | 1.80 ± 0.10 |
| Ionic strength (mol/L) | 0.6 | 0.5 |
| Dry density (kg/$m^3$) | 1300−1700 | 1200−1800 |
| pH (−) | 5.3 ± 0.1 | 6.8 ± 0.1 |
| Temperature (°C) | 25 ± 3 | 15 ± 3 |
| Block dimension (cm) | $\Phi$2.54 × 1.3 | $\Phi$2.54 × 1.2 |
| Volume of source reservoir (mL) | 200 | |
| Volume of target reservoir (mL) | 10 | |

(Optima 7000DV, PerkinElmer, USA). Data processing was performed using fitting for diffusion parameters software to calculate diffusion parameters such as the $D_e$, rock capacity factor, distribution coefficient, total porosity, and accessible porosity. Further details regarding the experimental setup, operational steps, and data processing are available in previous studies [17, 29].

## 2.3 Data

### 2.3.1 Data compilation

The datasets were gathered from the JAEA-DDB and 16 published resources, covering the period from 1982 to 2024. The dataset comprised 16 input features and 324 experimental instances, including 304 instances obtained from Wu et al. (2024) and 20 experimental instances from three other studies [17, 20, 27]. Notably, the absence of pH values in 514 instances of the JAEA-DDB resulted in a significantly reduction in data size. To address this, regression imputation techniques using ML models were applied to predict the pH values based on a dataset of 324 instances, thereby expanding the dataset to 838 instances.

The dataset included 16 input features, which were categorized into three groups: (i) porewater properties, comprising the ionic strength ($I$), temperature ($T$), and pH; (ii) bentonite properties, including the montmorillonite content ($m$), external surface area ($A_{ext}$), dry density ($\rho_d$), grain density ($\rho_s$), total porosity ($\varepsilon_{tot}$), and montmorillonite stacking number ($n_c$); and (iii) radionuclide properties, encompassing the ion diffusion coefficient in water ($D_w$), molecular weight ($MW$), ion molar conductivity ($\lambda$), ionic radius ($r$), ionic charge ($z$), distribution coefficient ($K_d$), and rock capacity factor ($\alpha$).

### 2.3.2 Data preprocessing

The presence of outliers can reduce the predictive accuracy of ML models. To address this issue, the Mahalanobis distance (MD) method was employed to identify and remove outliers. The cutoff point ($d_i$) is given as:

$$d_i = \sqrt{(x-\mu)\cdot S^{-1}\cdot(x-\mu)}, \tag{1}$$

where $x$ represents the object vector, $\mu$ denotes the mean arithmetic vector, and $S$ is the covariance matrix of instances. The cutoff point was set to eight to ensure that the skewness of all input features was less than 10.

Three datasets were used to enhance the prediction of radionuclide diffusion. An overview of the features and instances of each dataset is summarized in Table 2. Dataset I included 15 input features, with pH as the output feature. To ensure the data quality and reduce noise, eight instances were removed using the MD method. This process yielded Dataset I, comprising 316 instances. The statistical details of Dataset I are presented in Table S1 of the supporting information. Datasets II and III comprised 16 input features, including the basic features (15 input features of Dataset I) and pH. The output feature for Datasets II and III was the $D_e$. Dataset III, comprising 813 instances, was obtained after removing 17 instances. It is noteworthy that these datasets comprised parameters at the micro-mesoscopic level. Specifically, the montmorillonite stacking number and ionic radius were classified as microscopic parameters, whereas the other parameters were considered as mesoscopic.

### 2.3.3 Imputation methods

Four decision-tree models, namely LGBM, CatB, XGB, and RF, were used as regression imputation methods to predict

**Table 2** Details of the features and instances of datasets

| Dataset | Input feature | Input number | Output feature | Instance number | Dataset Link |
|---|---|---|---|---|---|
| Dataset I | Basic features:<br>(i) Porewater: $I$, $T$.<br>(ii) Bentonite: $m$, $A_{ext}$, $\rho_d$, $\rho_s$, $\varepsilon_{tot}$, $n_c$.<br>(iii) Radionuclides: $D_w$, $r$, $z$, $\lambda$, $MW$, $K_d$, $\alpha$. | 15 | pH | 316 | https://doi.org/10.57760/sciencedb.j00186.00710 |
| Dataset II | Basic features and pH | 16 | $D_e$ | 316 | |
| Dataset III | Basic features and pH | 16 | $D_e$ | 813 | |

the pH values of Dataset I. LGBM exhibited superior predictive accuracy compared with the other models. This was consistent with the results of our previous study [21]. Dataset III was established by incorporating additional 514 instances with Dataset II using the LGBM for data imputation. Table S2 of the supporting information summarizes the statistical results of the input and output features for Dataset III.

## 2.4 Methodology

The $D_e$ values of radionuclides in compacted bentonite were predicted using ten ML models, including three ensemble ML algorithms (LGBM-CatB, LGBM-XGB, and LGBM-RF), four decision-tree algorithms (LGBM, CatB, XGB, and RF), SVM, and two neural networks (ANN and DNN). Ensemble ML models combine the strengths of multiple individual models to enhance overall predictive performance and stability, offering a promising solution to the challenges of bias and variance in individual models [30]. Since LGBM exhibited superior predictive performance compared with the other models, it was combined with CatB, XGB, and RF to predict the $D_e$ using a voting regressor method from the scikit-learn package [20, 31]. The voting regressor simultaneously applies multiple regression models to the same dataset, thereby optimizing the final output by synthesizing the prediction results of each model. During the training process, the system can adjust the weight distribution according to the performance of each model. The final prediction result $\hat{y}$ is calculated by:

$$\hat{y} = \sum_{i=1}^{n} y_i \omega_i, \tag{2}$$

where $y_i$ and $\omega_i$ represent the prediction result and the weight corresponding of the $i$-th model, respectively. This method optimized the weight ranges of the base learners within a model by initially pruning these ranges according to the gradient of the best base learner performance, thereby accelerating the model optimization [30]. The hyperparameters of ML models were tuned using the particle swarm optimization (PSO) algorithm. In this algorithm, potential solutions to an optimization problem are represented as a swarm of particles. Each particle $i$ possesses a position vector $x_i$ and a velocity vector $v_i$ within the search space. During the algorithmic evolution, iterative adjustments are performed on both the velocity and position of each particle. Specifically, the velocity of each particle is updated according to the individual's best-known position $p_i$ and the swarm's global best position $g_i$, as follows:

$$x_i^{k+1}(t+1) = x_i^k(t) + v_i^{k+1}(t+1), \tag{3}$$

$$\begin{aligned} v_i^{k+1}(t+1) = \omega v_i^k(t) + c_1 r_1 \left(p_i^k(t) - x_i^\kappa(t)\right) \\ + c_2 r_2 \left(g^k(t) - x_i^k(t)\right), \end{aligned} \tag{4}$$

where $\omega$ is inertia weight, which influences the particle's velocity based on its previous state. $c_1$ and $c_2$ represent the learning factor for individual and social adjustment, respectively. $r_1$ and $r_2$ denote random numbers uniformly distributed within [0, 1].

Figure 1 illustrates a workflow diagram for developing ML models to predict the $D_e$ values of radionuclides in various compacted bentonites. This study was organized into three parts: (i) Dataset augmentation: Missing pH values were predicted using decision-tree algorithms, thereby refining the radionuclide diffusion dataset. (ii) Model training and explanation: Ten ML models were employed to train prediction models with high predictive accuracy. The diffusion mechanism was analyzed using Spearman, Shapley additive explanations (SHAP), and partial dependence plots (PDP). (iii) Model application: The $D_e$ values of $\mathrm{EuEDTA}^-$ and $\mathrm{HCrO_4}^-$ in compacted bentonites were measured using a through-diffusion method, which was employed to evaluate the generalization capability of the best ML models.

## 2.5 Model development and evaluation

The datasets were randomly divided into a training set consisting of 80% of the instances and a test set containing the remaining 20%. Since data processing using logarithmic transformation and min–max normalization exhibited an insignificant impact on the predictive accuracy in predicting the $D_e$ of radionuclides in bentonite [19], logarithmic transformation was applied to the features, such as the ionic radius, ion diffusion coefficient in water, and $D_e$, owing to their significantly larger magnitudes compared to other features. A fivefold cross-validation method was used to reduce the risk of overfitting. Therefore, the 80% training data was further subdivided into a pretraining (80% of the training data) and a validation (20% of the remaining training data) datasets to pretrain the ML models and optimize the hyperparameters. The PSO technique was used to optimize the hyperparameters.

The predictive performance was evaluated by the coefficient of determination ($R^2$), and mean square error ($MSE$). These metrics are given as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(\log D_{\mathrm{e},i}^{\mathrm{exp}} - \log D_{\mathrm{e},i}^{\mathrm{pred}}\right)^2}{\sum_{i=1}^{N}\left(\log D_{\mathrm{e},i}^{\mathrm{exp}} - \log D_{\mathrm{e,ave}}^{\mathrm{exp}}\right)^2}, \tag{5}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(\log D_{\mathrm{e},i}^{\mathrm{exp}} - \log D_{\mathrm{e},i}^{\mathrm{pred}}\right)^2, \tag{6}$$

where $\log D_{\mathrm{e,i}}{}^{\mathrm{exp}}$ and $\log D_{\mathrm{e,ave}}{}^{\mathrm{exp}}$ are the experimental $D_{\mathrm{e}}$ and average experimental $D_{\mathrm{e}}$ measured from diffusion experiments, respectively. $\log D_{\mathrm{e,i}}{}^{\mathrm{pred}}$ is the predicted $D_{\mathrm{e}}$ using the ML models.

## 3 Results and discussion

### 3.1 Model development

#### 3.1.1 Regression imputation for predicting pH

Handling missing data is a crucial step affecting the quality and reliability of the data analysis. Various regression imputation techniques have been applied to impute missing data, such as ANNs, multivariate imputation by chained equations, k-nearest neighbors, time-series deep learning models, generative broad Bayesian imputation, principal component analysis imputation, and simple arithmetic averages. These methods have been applied to datasets with missing data percentages ranging from 0 to 80% [24, 26, 32–36]. Generally, three types of missing data mechanisms are recognized: missing completely at random, missing at random, and missing not at random [23]. Each mechanism presents different challenges and implications for imputation, highlighting the importance of identifying the underlying pattern of missingness before selecting an appropriate imputation strategy.

The JAEA-DDB database collected data from the literatures and reports covering 1982 to 2009. The instances have been derived from various diffusion experimental methods and numerous researchers. The absence of pH values in 514 instances within the JAEA-DDB database can be explained that these researches ignored the importance of pH values in their studies. In the JAEA-DDB database, missing data primarily resulted from ignoring or inadequately measuring the parameters that related to the radionuclide diffusion. The missing mechanism in the JAEA-DDB database was assumed to be missing completely at random, corresponding to a noncontinuous missingness. Based on the selected 16 input features, more than 60% of the dataset (514 instances) lacked pH values. Decision-tree models were employed to predict the missing pH values to augment the dataset and enhance the robustness of the ML models. Specifically, LGBM, CatB, XGB, and RF were employed to predict the pH values of Dataset I.

The predicted performances are summarized in Table 3. The LGBM exhibited superior robustness compared with the other models. For instance, the $R^2_{\mathrm{cv}}$ values for the test sets were ranked in descending order using a fivefold cross-validation as follows: LGBM> XGB> CatB > RF. The rank of $MSE_{\mathrm{cv}}$ values was the opposite of that of the $R^2_{\mathrm{cv}}$ values for the test datasets. Notably, LGBM achieved the highest performance metrics among all models, with an $MSE$ of 0.23 and $R^2$ of 0.92 for the test dataset. The hyperparameters of the optimal ML models are listed in Table S3 of the supporting information. Therefore, the missing pH values for 514 instances were predicted using the LGBM model, resulting in the establishment of Dataset III with 813 instances.

Figure 2 exhibits the data distribution and characteristics of the relationship between pH and each input

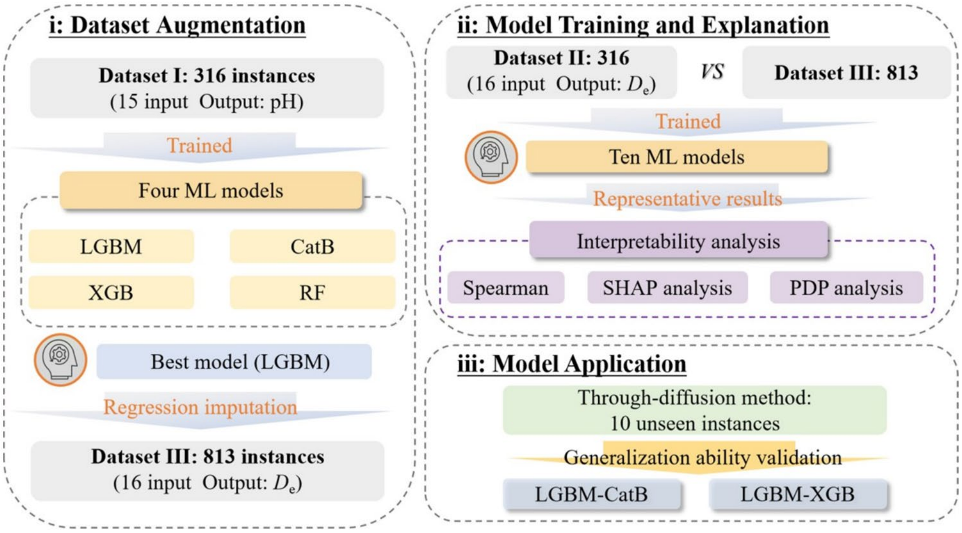

**Fig. 1** (Color online) Workflow diagram on building machine learning models for predicting the effective diffusion coefficient of radionuclides in various compacted bentonites

feature. Blue and orange represent the data distributions of Dataset I and the imputed 514 instances, respectively. It clearly demonstrates a nonlinear relationship between the pH and each input feature. The predicted pH values ranged from 5.0 to 9.0, exhibiting a Gaussian type distribution.

pH is an important porewater parameter that influences both the radionuclide species and surface charge of clay [37]. Figure 3 shows the pH dependence on the external surface area and ion molar conductivity, which are associated with the bentonite and radionuclide properties, respectively. Dataset I exhibits that the pH values ranged from 3.0 to 13.4. The predicted pH values are concentrated in the range from 5.0 to 9.0, suggesting a close adherence to a normal distribution of porewater for Dataset III.

#### 3.1.2 Model development for radionuclide diffusion

Ten ML models, namely LGBM-CatB, LGBM-XGB, LGBM-RF, LGBM, CatB, XGB, RF, ANN, DNN, and SVM, were used to predict the $D_e$ values of radionuclides in compacted bentonite. Figure 4 shows the performance metrics of the ML models for the test datasets of Dataset II and III using the optimal hyperparameters tuned with PSO techniques (Table S4 in the supporting information). The performance metrics were assessed using fivefold cross-validation. The red lines represent the smooth kernel curve of the distribution of performance metrics. The black lines within and outside the box plots denote the mean values and standard deviations of the performance metrics, respectively, with a lower standard deviation indicating strong robustness of the ML models. The detailed performance metrics for the training, validation, and test datasets are listed in Table S5 of the supporting information.

As the number of instances increased from 316 (Dataset II) to 813 (Dataset III), the performance metrics of all ML models improved significantly, as evidenced by the higher $R^2_{cv}$ values, lower $MSE_{cv}$, and reduced standard deviation. These findings indicate that expanding the dataset contributed to enhanced predictive performance and robustness of the ML models. It is noteworthy that the ensemble models were established by combining LGBM with other individual decision-tree models, primarily due to the relatively high training speed of the LGBM algorithm [38]. However, no significant difference is observed in the computational efficiencies of the ensemble and single models. The difference in running time was approximately five minutes. In the case of decision-tree algorithms, gradient boosting (GB) models (LGBM, CatB, and XGB) outperformed the RF models. The excellent predictive performance of GB models is consistent with previous findings in predicting the chloride diffusion coefficient in concrete [39]. In addition, the ensemble ML models (LGBM-CatB, LGBM-XGB, and LGBM-RF) and LGBM surpassed the other ML models, achieving an $R^2_{cv}$ above 0.90. This can be attributed to their ability to harness the strengths of various algorithms to thoroughly capture potentially complex patterns and errors within the data, thereby enhancing the prediction accuracy and robustness [30, 40]. For Dataset III, the $R^2_{cv}$ values of the ML models ranked in descending order as follows: LGBM-CatB ≈ LGBM-XGB > LGBM ≈ LGBM-RF > CatB ≈ XGB> ANN> DNN> RF > SVM. Notably, LGBM-CatB surpassed LGBM-XGB due to its lower standard deviation, indicating stronger robustness. SVM exhibited the lowest predictive performance based on Dataset III, with $R^2_{cv} = 0.75$ and $MSE_{cv} = 0.06$. Compared with ensemble models, SVM is a relatively simple model. The ensemble models are designed to capture more complex patterns and relationships in the data through a combination of multiple decision trees. This lack of complexity in the SVM limits its ability to generalize across different data instances in the dataset. Notably, some studies have reported test $R^2$ values below 0.80, such as an $R^2$ of 0.74 for predicting the retention rate of Cd in biochar [41] and an $R^2$ of 0.76 for predicting alcohol space-time yield [42]. Therefore, the prediction accuracy of SVM remained satisfactory, despite exhibiting a lower predictive performance than the other models.

Figure 5 shows the regression plots comparing the experimental and predicted $D_e$ values for the training (green triangle), validation (red circle), and test (purple square) datasets of Datasets II and III, using the LGBM-CatB, LGBM-XGB, LGBM, and LGBM-RF algorithms. These algorithms were selected owing to their excellent predictive accuracies. The plots reveal a close alignment between the experimental and predicted $D_e$ values with the slope line, underscoring the effective simulation capability of these ML models for predicting radionuclide diffusion processes. The performance metrics of the best-performing models are shown in Fig. 5. Notably, the ML models applied to the test dataset of Dataset III outperformed those applied to Dataset II. This disparity can be attributed to the augmentation of instances in Dataset III, which facilitates the models' ability to capture complex relationships within the data more effectively. For Dataset III, the ranking of models was as follows: LGBM-CatB ($R^2 = 0.94$) ≈ LGBM-XGB ($R^2 = 0.94$) > LGBM ($R^2 = 0.92$) ≈ LGBM-RF ($R^2 = 0.92$). These results indicate that both LGBM-CatB and LGBM-XGB exhibit high predictive accuracy.

### 3.2 Sensitivity analysis

#### 3.2.1 Spearman and Shapley additive explanation analyses

ML models can uncover predictive principles through analysis techniques that rank the importance of influencing

**Table 3** Mean performance metric values using five-fold cross-validation and the highest performance metrics for machine learning models to predict pH based on Dataset I

| Algorithms | Datasets | $R^2_{cv}$ | $MSE_{cv}$ | Best performance | |
|---|---|---|---|---|---|
| | | | | $R^2$ | $MSE$ |
| LGBM | Training | **0.99** | 0.01 | **0.99** | 0.01 |
| | Validation | **0.87** | 0.32 | **0.90** | 0.07 |
| | Test | **0.88** | 0.33 | **0.92** | 0.23 |
| XGB | Training | 0.98 | 0.05 | 0.98 | 0.06 |
| | Validation | 0.82 | 0.46 | 0.92 | 0.16 |
| | Test | 0.84 | 0.47 | 0.87 | 0.38 |
| CatB | Training | 0.99 | 0.01 | 0.99 | 0.01 |
| | Validation | 0.87 | 0.28 | 0.86 | 0.22 |
| | Test | 0.83 | 0.68 | 0.85 | 0.57 |
| RF | Training | 0.90 | 0.27 | 0.90 | 0.26 |
| | Validation | 0.77 | 0.61 | 0.79 | 0.67 |
| | Test | 0.77 | 0.38 | 0.80 | 0.32 |

factors in predictions, such as feature importance and SHAP analysis [19, 21, 43, 44]. Additionally, Spearman analysis, a nonparametric statistical method, assesses the monotonic relationship between two variables by correlating ranked data. These approaches provided valuable insights into the consistency and strength of the relationships within a dataset. It worthy notes that the reliability of these analytical techniques is intrinsically linked to the quality of the data used. Increasing the dataset size enhances the depth, broadness, and reliability of the ML models.

Spearman correlation and SHAP analysis techniques were employed to analyze the correlation and importance of the input features, presenting intuitively global interpretations

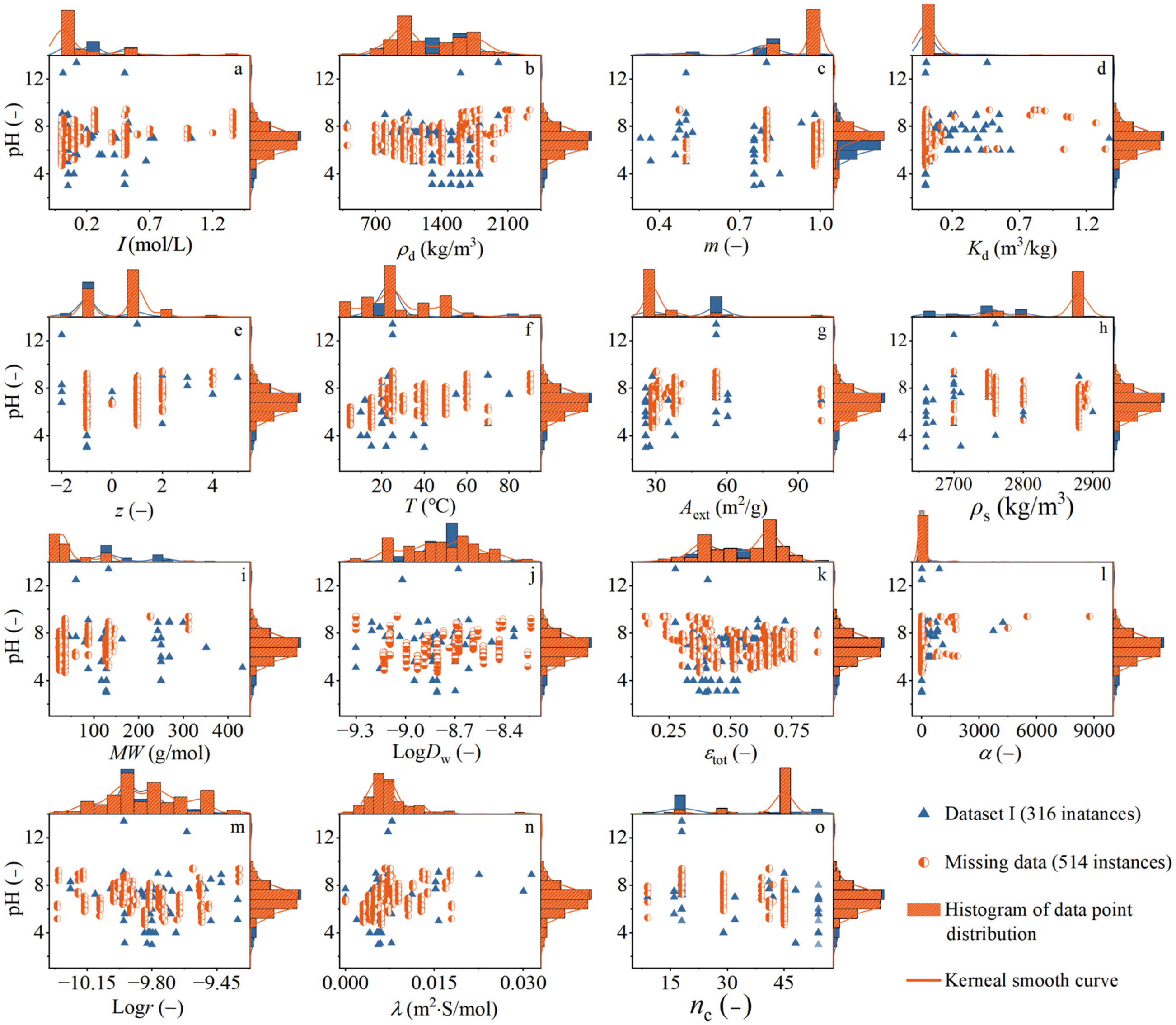

**Fig. 2** (Color online) Data distribution of features and the relationship between pH and each input feature

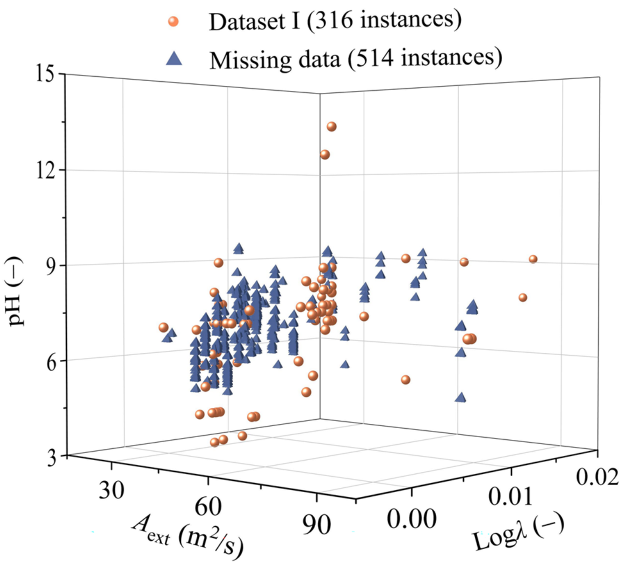

**Fig. 3** (Color online) Analyzing the dependency of pH on the external surface area and ion molar conductivity

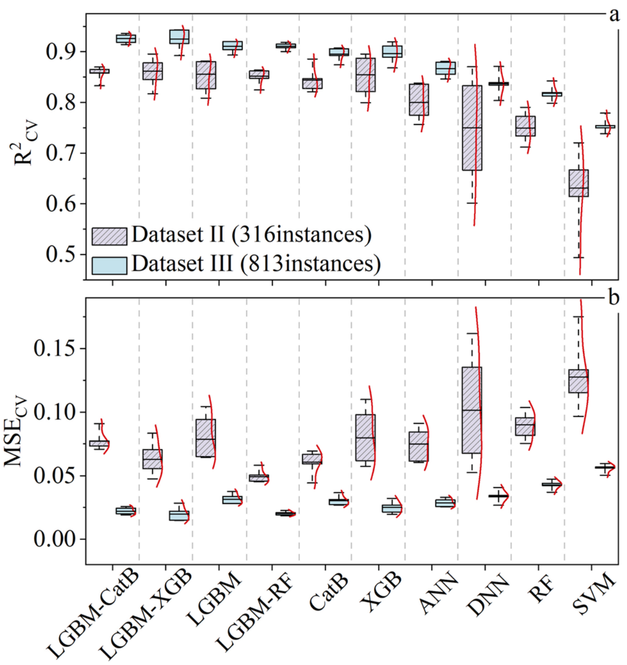

**Fig. 4** (Color online) Mean performance metric values using fivefold cross-validation for machine learning models in the test datasets of Dataset II and III

of the ML models (Fig. 6). The features were ranked from left to right according to their correlation and contribution to the prediction. The Spearman correlation analysis revealed that the most influential factor among the 16 input features was the ion diffusion coefficient in water for Dataset II, and the total porosity for Dataset III . This feature exhibited a positive correlation with $D_e$ (Fig. 6a, b). This is consistent with the previous findings [19] and Archie's law [31, 45].

In the case of Dataset II, the SHAP analysis revealed that the most important input features varied across different ML models: the compacted dry density for LGBM-CatB, ionic radius for LGBM-XGB, and ion diffusion coefficient in water for LGBM (Fig. 6c, e, g). Notably, only the SHAP results for LGBM were consistent with the Spearman correlation analysis. This discrepancy can be attributed to differences in the feature importance assessment and prediction mechanisms inherent to each ML algorithm. As the number of instances increased from 316 (Dataset II) to 813 (Dataset III), both Spearman and SHAP analyses identified the total porosity as the primary contributor, which is consistent with Archie's law [31, 45]. The total porosity for radionuclide diffusion in compacted bentonite blocks is expressed as a percentage of the total interconnected pore space within the blocks. A higher total porosity implies greater availability of transport pathways. These findings suggest that larger datasets may reduce the discrepancies between ML models in terms of feature importance assessment and prediction mechanisms.

### 3.2.2 Partial dependence plots

The dependence of $D_e$ on the 16 input features has been discussed in our previous study [19]. However, some relationships may remain unclear due to the limited size of the dataset. To address this, PDP analysis was performed to visually represent the univariate correlations and examine the influence of the size of the dataset on these relationships (Fig. 7). The histograms and lines correspond to the data distribution and correlation with each input feature and the PDP. Generally, a more concentrated data distribution generally leads to more accurate analytical results. These findings indicate that Dataset III, which was larger than Dataset II, exhibited more continuous PDP curves, suggesting a more stable and clear relationship between the features and $D_e$.

Figure 7a, b shows that both the rock capacity factor and distribution coefficient exhibit a clear positive correlation with the prediction for Dataset III. This finding is consistent with studies on radionuclides diffusion in crystalline rocks [46] and sodium montmorillonite [47]. Consistently, Fig. 7d illustrates the positive impact of ionic charge, where cations exhibit a higher $D_e$ than neutral species, and anions display lower $D_e$ values. This is consistent with previous studies, which attributed the differences in diffusion mechanisms to electrostatic interactions between the radionuclide species and charged bentonite surfaces [3]. Specifically, cation diffusion is controlled by surface diffusion effects, whereas anions diffusion is driven by anionic exclusion effects [47, 48].

pH values in the range from 6 to 9 negatively influence the prediction for Dataset III, whereas a peak was observed at approximately pH 8 for Dataset II (Fig. 7c). The negative effect of Dataset III might be more convincing because of its larger data size. Figure 7e shows a positive impact on the prediction when ion molar conductivity exceeded 0.01 $m^2$ S/mol for Dataset III. However, the relationships among the external surface area, montmorillonite stacking number, grain density, and ionic strength remained unclear for both Datasets II and III (Fig. 7f–i). This lack of clarity can be attributed to data dispersion, despite the larger dataset size.

The case of remaining input features, such as the total porosity, ion diffusion coefficient in water, and temperature, exhibited positive impacts on the prediction, whereas the dry density, montmorillonite content, ionic radius, and molecular weight showed negative impacts (Fig. 7j–p). The positive influences of the total porosity and ion diffusion coefficient in water could be explained by Archie's law [16, 45], whereas the positive impact of temperature followed the Arrhenius equations [49–51]. The detailed explanations are provided in our previous studies [19, 21]. It is worth mentioning that a negative influence of ionic radius was observed at Log$r < -9.6$ (2.5 Å). This positive relationship can be attributed to the limited data for species with ionic radius above 2.5 Å. Overall, the univariate correlation results visualized using the PDP technique align with the diffusion laws observed in the experiments and diffusion mechanisms derived from the numerical models. This consistency underscores the reliability of the interpretation capabilities of the ML models.

## 3.3 Diffusion experiments and model application

Anionic radionuclides with long half-life are important for the safety evaluation of HLW repositories because of their high diffusivities. A through-diffusion method was employed to measure the diffusion parameters of $EuEDTA^-$ and $HCrO_4^-$ in compacted bentonites at compacted dry densities ranged from 1200 to 1800 kg/$m^3$. Their $D_e$ values were predicted using LGBM-CatB and LGBM-XGB to test the generalization ability.

### 3.3.1 Determination of the diffusion parameters using diffusion experiments

Figure 8 shows the breakthrough curves of $EuEDTA^-$ and the species distribution of EuEDTA complexes. $A_{cum}$ denotes the accumulated mass of $EuEDTA^-$ and $HCrO_4^-$ that penetrated a 1.2 cm thick bentonite block to reach the sample reservoirs. The data show that the accumulated mass

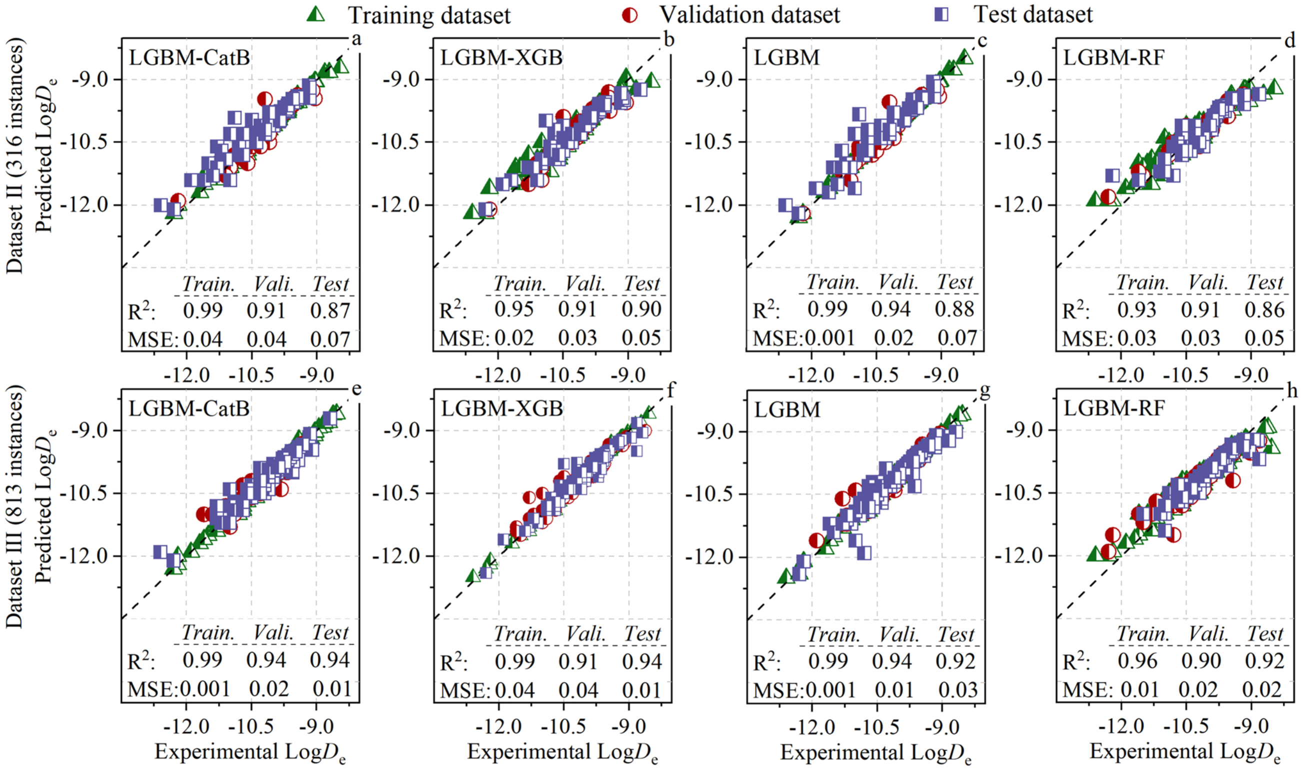

**Fig. 5** (Color online) Regression plots of experimental versus predicted effective diffusion coefficients based on Datasets II and III: **a**, **e** LGBM-CatB, **b**, **f** LGBM-XGB, **c**, **g** LGBM, and **d**, **h** LGBM-RF

increased with decreasing dry density, which is consistent with the general understanding that lower dry density facilitates radionuclide diffusion through porous media [3, 5]. The pH was maintained at 5.3 ± 0.1 during the Eu(III) diffusion experiments. Simulations using Vision MINTEQ indicated that Eu(III) exists as a mixture of species, including $Eu^{3+}$, EuHEDTA(aq), $EuEDTA^-$, and $EuCl^{2+}$, in 0.6 mol/L NaCl solution (Fig. 8c). $EuEDTA^-$ was the main species at pH above 2.0. It indicates that this study measured the diffusion parameters of $EuEDTA^-$ in compacted Ba-bentonite.

Table 4 summarizes the diffusion parameters of $HCrO_4^-$ and $EuEDTA^-$, including $D_e$, rock capacity factor, accessible porosity, total porosity, and distribution coefficient. Both $D_e$ and distribution coefficient are important parameters in the safety assessment of repositories, whereas the other parameters play a crucial role in elucidating the diffusion mechanism. The error in the compacted dry density measurement was primarily attributed to a loss of approximately 0.3 g during the preparation of bentonite blocks. Both $HCrO_4^-$ and $EuEDTA^-$ are monovalent anions that cannot access the interlayer pores of compacted bentonite [17, 21]. The rock capacity factor of $HCrO_4^-$ was lower than the total porosity, indicating that the accessible porosity was equal to the rock capacity factor. This suggests that the predominant diffusion path of $HCrO_4^-$ was through the free pores of compacted bentonite. In contrast, $EuEDTA^-$ exhibited an adsorptive behavior similar to that of simulated trivalent actinide complexes, such as $AmEDTA^-$ and $CmEDTA^-$, with the rock capacity factor being higher than the total porosity. The distribution coefficient, $K_d$, of $EuEDTA^-$ was calculated as follows:

$$K_d = \frac{\alpha - \varepsilon_{acc}}{\rho_d}, \tag{7}$$

where the accessible porosity, $\varepsilon_{acc}$, was obtained using the $I^-$ diffusion experiments [19].

All diffusion parameters decreased with increasing dry density for both $EuEDTA^-$ and $HCrO_4^-$. The distribution coefficient of $EuEDTA^-$ ranged from 4.2× $10^{-4}$ $m^3$/kg to 6.7 × $10^{-4}$ $m^3$/kg, which is lower than the range reported for $EuEDTA^-$ in hard rock clay (1.3 × $10^{-3}$–3.2 × $10^{-3}$ $m^3$/kg) [52] and for $CeEDTA^-$ in compacted Zhisin bentonite (0.8 × $10^{-3}$–1.2 × $10^{-3}$ $m^3$/kg) [17] . The distribution coefficient of $EuEDTA^-$ was lower than that of $Eu^{3+}$, indicating that EDTA facilitated the diffusion of Eu(III), thereby reducing the retardation capacity of the bentonite barrier [52, 53]. This observation is consistent with the diffusion behavior of $CeEDTA^-$ and $CoEDTA^{2-}$ [17, 19, 31].

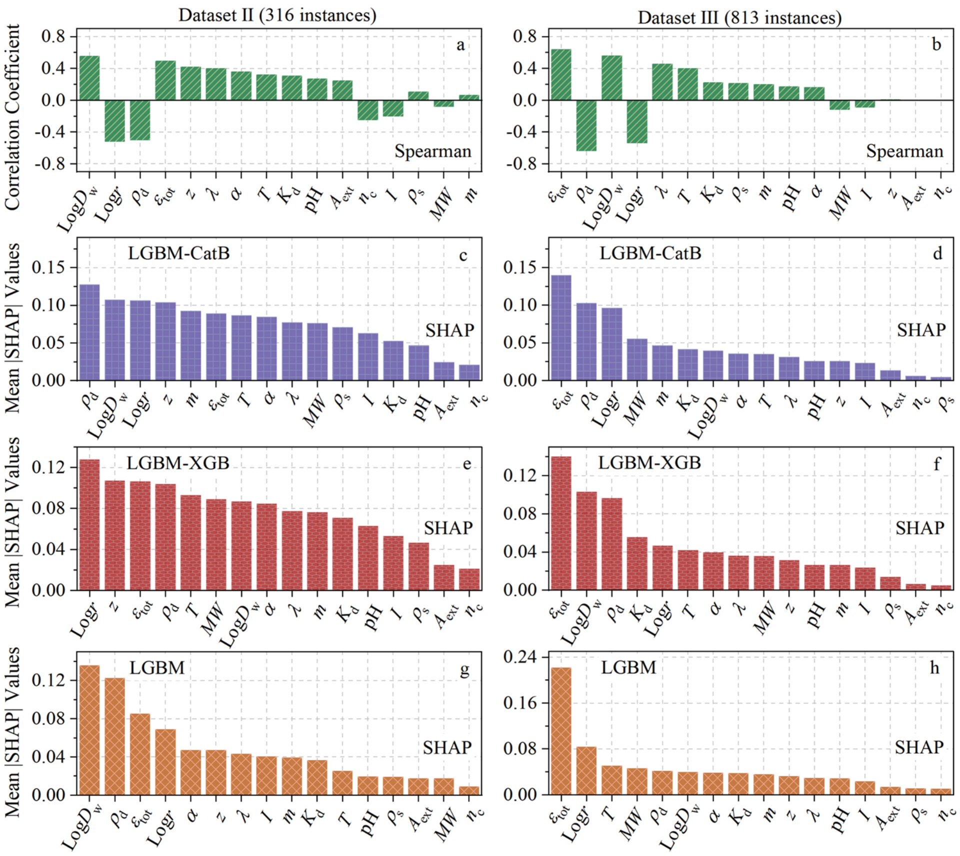

**Fig. 6** (Color online) **a**, **b** Spearman correlation analysis and global interpretations of ML models based on Dataset II and III: **c**, **d** LGBM-CatB, **e**, **f** LGBM-XGB, and **g**, **h** LGBM

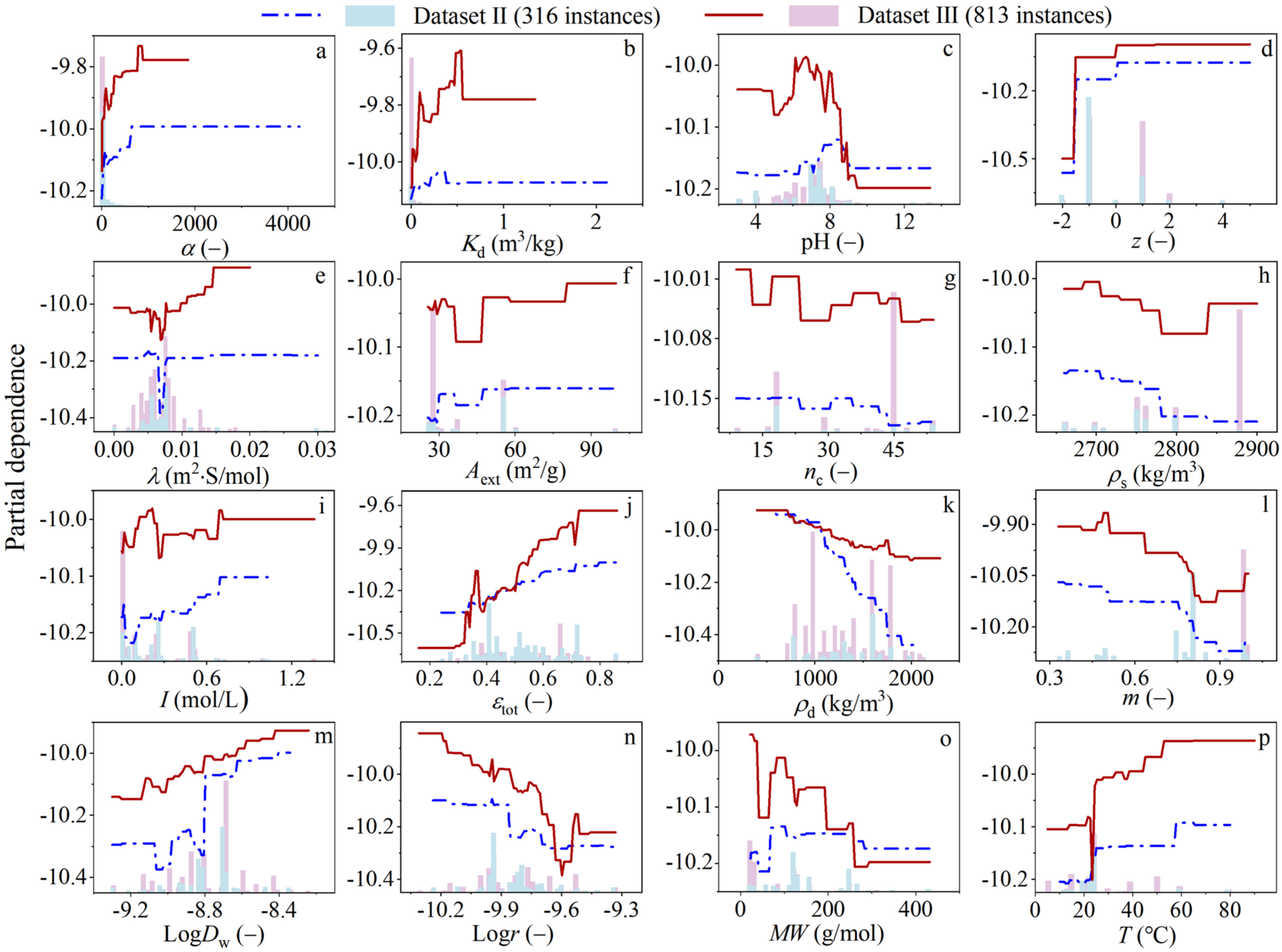

**Fig. 7** (Color online) Partial dependence plot for **a** the rock capacity factor, **b** distribution coefficient, **c** pH, **d** ionic charge, **e** ion molar conductivity, **f** external surface area, **g** montmorillonite stacking number, **h** grain density, **i** ionic strength, **j** total porosity, **k** dry density, **l** montmorillonite content, **m** ion diffusion coefficient in water, **n** ionic radius, **o** molecular weight, and **p** temperature

### 3.3.2 Model application

The LGBM-CatB and LGBM-XGB models were employed to predict the $D_e$ of $HCrO_4^-$ in compacted Wyoming bentonite and EuEDTA$^-$ in compacted Ba-bentonite, which were compared with published diffusion experimental results for $HCrO_4^-$ and the simulated actinides CeEDTA$^-$ and CoEDTA$^{2-}$ [17, 19, 21]. Additionally, both models were used to predict the $D_e$ of radionuclide cation $^{137}Cs^+$ and neutral species HTO [8, 54, 55] (Fig. 9). It shows that $D_e/D_w$ decreased with increasing compacted dry density, which is consistent with the result of previous studies [3, 5, 45]. In this study, the $D_w$ value for metal-EDTA complexes was assumed to be $5.0 \times 10^{-10}$ m$^2$/s [56]. The $D_e$ of EuEDTA$^-$ was observed to be higher than those of CeEDTA$^-$ [17]. and CoEDTA$^{2-}$ [19]. The LGBM-CatB and LGBM-XGB models successfully predict $D_e$, as evidenced by the good agreement with the experimental $D_e$ values (Fig. 9a).

Figure 9b shows that the $D_e$ of $HCrO_4^-$ in compacted Wyoming bentonite is lower than that in Anji bentonite [19] and GMZ bentonite [21], likely due to the higher montmorillonite content. LGBM-CatB slightly underestimated $D_e$ for $HCrO_4^-$ in Wyoming bentonite, with the predicted $D_e$ values being 25%–47% lower than the experimental $D_e$. Although this discrepancy is less pronounced than the predictions for $HCrO_4^-$ in GMZ and Anji bentonites using LGBM and PSO-LGBM, the difference was reported to be 9%–27% [19, 21]. This performance is significantly superior to that predicted using Archie's law, according to which the predictive $D_e$ values were 1.0–1.5 orders of magnitude higher than the experimental results [45].

Figure 9c shows that the predicted $D_e$ values of $^{137}Cs^+$ are consistent with the experimental results at a compacted density of 1400 kg/m$^3$. However, a significant underestimation was observed at a compacted density of 800 kg/m$^3$, with a difference of approximately four times. This can be explained by the limited number of experimental data points available for this density in the dataset, which comprised only 58 instances, accounting for approximately 7% of the total dataset. It indicates that additional diffusion experiments for $^{137}Cs^+$ should be conducted at a compact density of approximately 800 kg/m$^3$ to facilitate the identification of diffusion patterns using ML models. Figure 9d illustrates that both the LGBM-CatB and LGBM-XGB models accurately predict the $D_e$ of HTO. Under similar experimental conditions, the $D_e$ in Wyoming bentonite (red squares) was higher than that in FEBEX bentonite (blue pentagrams), primarily because of the lower montmorillonite content, with $m = 0.85$ for Wyoming bentonite and $m = 0.92$ for FEBEX bentonite [54, 55].

Notably, the experimental diffusion data from this study, as well as from the $^{137}Cs^+$ [8] and HTO [54, 55] diffusions, were not included in the test datasets, highlighting the strong generalization ability of both LGBM-CatB and LGBM-XGB models. The generalization ability of LGBM-XGB was superior to that of LGBM-CatB, indicating that model selection plays a crucial role in accurately predicting radionuclide diffusion in complex geological environments. Given that HLW repositories have been designed to operate for over 10,000 years, the prediction of radionuclide diffusion in bentonite barriers must consider the complex coupling effects among radionuclides, porewater, and bentonite under intrinsic disposal conditions. Current diffusion datasets remain insufficient for safety assessments of bentonite barriers owing to limitations in data size and dimensionality. Therefore, additional diffusion experiments should be conducted to enhance the dimensionality and scale of the datasets.

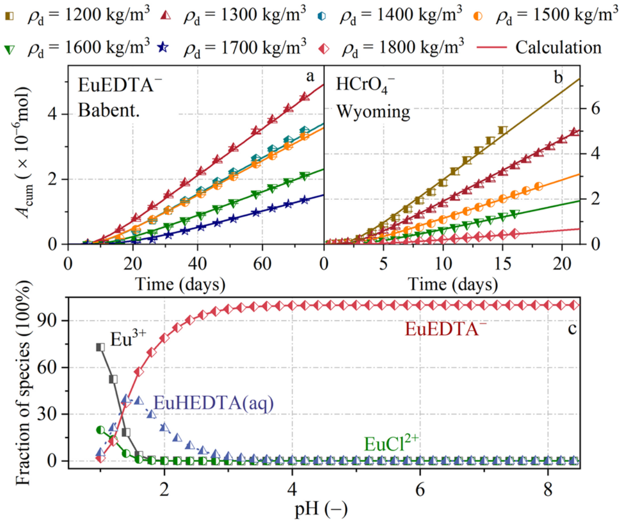

**Fig. 8** (Color online) Relationship between the accumulated mass ($A_{cum}$) and time for **a** EuEDTA$^-$ and **b** HCrO$_4^-$ in saturated compacted bentonites. **c** Species distribution of Eu(III)-EDTA system in aqueous solution

## 4 Conclusion

A radionuclide diffusion dataset comprising 16 input features and 813 instances was developed using regression imputation machine learning (ML) methods. Ten ML algorithms were employed to predict the effective diffusion coefficient ($D_e$) of radionuclides in compacted bentonite. The light gradient boosting machine (LGBM)-extreme gradient boosting (XGB) and LGBM-categorical boosting (CatB)

**Table 4** Overview of diffusion parameters of EuEDTA$^-$ and HCrO$_4^-$ in compacted bentonite

| $\rho_d$ (kg/m$^3$) | $m_{bent}$ (g) | $D_e$ (× $10^{-11}$ m$^2$/s) | $D_a$ (× $10^{-11}$ m$^2$/s) | $\alpha$ (−) | $\varepsilon_{acc}$ (−) | $\varepsilon_{tot}$ (−) | $K_d$ (× $10^{-4}$ m$^3$/kg) |
|---|---|---|---|---|---|---|---|
| EuEDTA$^-$ in Ba-bentonite | | | | | | | |
| 1300 ± 45 | 8.7 ± 0.3 | 3.6 ± 0.4 | 3.0 ± 0.3 | 1.2 ± 0.1 | 0.33 ± 0.01# | 0.52 | 6.7 ± 0.6 |
| 1400 ± 45 | 9.3 ± 0.3 | 2.8 ± 0.3 | 2.6 ± 0.2 | 1.1 ± 0.1 | 0.31 ± 0.01# | 0.48 | 5.6 ± 0.6 |
| 1500 ± 46 | 9.8 ± 0.3 | 2.6 ± 0.3 | 2.7 ± 0.2 | 1.0 ± 0.1 | 0.30 ± 0.01# | 0.45 | 4.7 ± 0.5 |
| 1600 ± 46 | 10.5 ± 0.3 | 1.8 ± 0.2 | 1.9 ± 0.1 | 1.0 ± 0.1 | 0.26 ± 0.01# | 0.41 | 4.3 ± 0.5 |
| 1700 ± 47 | 11.2 ± 0.3 | 1.3 ± 0.1 | 1.5 ± 0.1 | 0.9 ± 0.1 | 0.19 ± 0.01# | 0.37 | 4.2 ± 0.3 |
| HCrO$_4^-$in Wyoming bentonite | | | | | | | |
| 1200 ± 46 | 7.8 ± 0.3 | 6.2 ± 0.6 | 11.9 ± 0.5 | 0.52 ± 0.04 | 0.52 ± 0.04 | 0.57 | – |
| 1300 ± 52 | 7.7 ± 0.3 | 3.9 ± 0.3 | 8.1 ± 0.3 | 0.48 ± 0.04 | 0.48 ± 0.04 | 0.53 | – |
| 1500 ± 45 | 10.0 ± 0.3 | 2.7 ± 0.2 | 10.2 ± 0.2 | 0.26 ± 0.02 | 0.26 ± 0.02 | 0.46 | – |
| 1600 ± 47 | 10.2 ± 0.3 | 1.8 ± 0.1 | 7.7 ± 0.2 | 0.23 ± 0.02 | 0.23 ± 0.02 | 0.42 | – |
| 1800 ± 47 | 11.4 ± 0.3 | 0.7 ± 0.1 | 5.7 ± 0.1 | 0.12 ± 0.01 | 0.12 ± 0.01 | 0.35 | – |

#Data from [19]

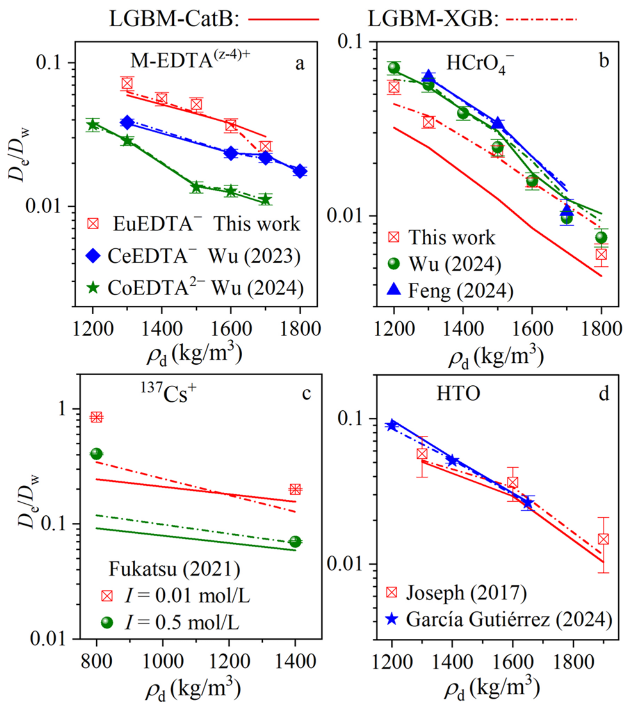

**Fig. 9** (Color online) Generalization ability validation of LGBM-CatB and LGBM-XGB: **a** M-EDTA$^{(z-4)+}$ diffusion, **b** HCrO$_4^-$ diffusion, **c** $^{137}$Cs$^+$ diffusion, and **d** HTO diffusion

algorithms surpassed the other ML models, achieving $R^2$ values of 0.94 based on the imputed dataset. This improvement indicates that the imputed dataset enabled the ML models to achieve high predictive performance and strong robustness.

The generalizability of the LGBM-CatB and LGBM-XGB models was evaluated by applying them to predict the $D_e$ values of EuEDTA$^-$ in compacted Ba-bentonite and HCrO$_4^-$ in compacted Wyoming bentonite. Both models exhibited excellent predictive accuracy for EuEDTA$^-$, whereas LGBM-CatB slightly underestimated $D_e$ for HCrO$_4^-$ in Wyoming bentonite, with predicted $D_e$ values 25%–47% lower than the experimental $D_e$. This indicates that the generalization ability of LGBM-XGB surpassed that of LGBM-CatB.

It has been widely accepted that the quality and quantity of datasets play a crucial role in the predictive performance of ML models. However, a significant number of experimental diffusion results were excluded from the diffusion datasets due to incomplete or missing data. To address this limitation, additional experiments are necessary to comprehensively characterize the properties of porewater and bentonite. These experiments should include but are not limited to mineral composition, elemental, and particle size analyses.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s41365-025-01759-4.

**Author contributions** All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by Jun-Lei Tian, Jia-Xing Feng, and Yao-Lin Zhao. The first draft of the manuscript was written by Tao Wu, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

**Data availability** The data that support the findings of this study are openly available in Science Data Bank at https://cstr.cn/31253.11.sciencedb.j00186.00710 and https://doi.org/10.57760/sciencedb.j00186.00710.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

1. L. Baborová, E. Viglašová, D. Vopálka, Cesium transport in Czech compacted bentonite: planar source and through diffusion methods evaluated considering non-linearity of sorption isotherm. Appl. Clay Sci. **245**, 107150 (2023). https://doi.org/10.1016/j.clay.2023.107150
2. L. Cui, W. Ye, Q. Wang et al., A model for describing advective and diffusive gas transport through initially saturated bentonite with consideration of temperature. Eng. Geol. **323**, 107215 (2023). https://doi.org/10.1016/j.enggeo.2023.107215
3. P. Krejci, T. Gimmi, L.R. Van Loon et al., Relevance of diffuse-layer, Stern-layer and interlayers for diffusion in clays: a new model and its application to Na, Sr, and Cs data in bentonite. Appl. Clay Sci. **244**, 107086 (2023). https://doi.org/10.1016/j.clay.2023.107086
4. R. Zuo, Z. Xu, X. Wang et al., Adsorption characteristics of strontium by bentonite colloids acting on claystone of candidate high-level radioactive waste geological disposal sites. Environ. Res. **213**, 113633 (2022). https://doi.org/10.1016/j.envres.2022.113633
5. M. García Gutiérrez, J. Cormenzana, T. Missana et al., Diffusion coefficients and accessible porosity for HTO and $^{36}$Cl in compacted FEBEX bentonite. Appl. Clay Sci. **26**, 65–73 (2004). https://doi.org/10.1016/j.clay.2003.09.012
6. H. Lyu, Z. Xu, J. Zhong et al., Machine learning-driven prediction of phosphorus adsorption capacity of biochar: insights for adsorbent design and process optimization. J. Environ. Manag. **369**, 122405 (2024). https://doi.org/10.1016/j.jenvman.2024.122405
7. Y. Yang, S.V. Churakov, R.A. Patel et al., Pore-scale modeling of water and ion diffusion in partially saturated clays. Water Resour. Res. **60**, e2023WR035595 (2024). https://doi.org/10.1029/2023WR035595
8. Y. Fukatsu, K. Yotsuji, T. Ohkubo et al., Diffusion of tritiated water, $^{137}$Cs$^+$, and $^{125}$I$^-$ in compacted Ca-montmorillonite: Experimental and modeling approaches. Appl. Clay Sci. **211**, 106176 (2021). https://doi.org/10.1016/j.clay.2021.106176
9. A. Asaad, F. Hubert, E. Ferrage et al., Role of interlayer porosity and particle organization in the diffusion of water in swelling clays. Appl. Clay Sci. **207**, 106089 (2021). https://doi.org/10.1016/j.clay.2021.106089
10. C. Wigger, L.R. Van Loon, Importance of interlayer equivalent pores for anion diffusion in clay-rich sedimentary rocks.

Environ. Sci. Technol. **51**, 1998–2006 (2017). https://doi.org/10.1021/acs.est.6b03781

11. A. Muurinen, O. Karnland, J. Lehikoinen, Ion concentration caused by an external solution into the porewater of compacted bentonite. Phys. Chem. Earth **29**, 119–127 (2004). https://doi.org/10.1016/j.pce.2003.11.004
12. P. Wersin, M. Kiczka, K. Koskinen, Porewater chemistry in compacted bentonite: application to the engineered buffer barrier at the Olkiluoto site. Appl. Geochem. **74**, 165–175 (2016). https://doi.org/10.1016/j.apgeochem.2016.09.010
13. P. Wersin, M. Mazurek, T. Gimmi, Porewater chemistry of Opalinus clay revisited: findings from 25 years of data collection at the Mont Terri Rock Laboratory. Appl. Geochem. **138**, 105234 (2022). https://doi.org/10.1016/j.apgeochem.2022.105234
14. C. Wigger, L.R. Van Loon, Effect of the pore water composition on the diffusive anion transport in argillaceous, low permeability sedimentary rocks. J. Contam. Hydrol. **213**, 40–48 (2018). https://doi.org/10.1016/j.jconhyd.2018.05.001
15. I.C. Bourg, A.C. Bourg, G. Sposito, Modeling diffusion and adsorption in compacted bentonite: a critical review. J. Contam. Hydrol. **61**, 293–302 (2003). https://doi.org/10.1016/S0169-7722(02)00128-6
16. T. Wu, Z. Feng, Z. Geng et al., Restriction of Re(VII) and Se(IV) diffusion by barite precipitation in compacted bentonite. Appl. Clay Sci. **232**, 106803 (2023). https://doi.org/10.1016/j.clay.2022.106803
17. T. Wu, Y. Hong, D. Shao et al., Experimental and modeling study of the diffusion path of Ce(III)-EDTA in compacted bentonite. Chem. Geol. **636**, 121639 (2023). https://doi.org/10.1016/j.chemgeo.2023.121639
18. Z. Feng, Z. Gao, Y. Wang et al., Application of machine learning to study the effective diffusion coefficient of Re(VII) in compacted bentonite. Appl. Clay Sci. **243**, 107076 (2023). https://doi.org/10.1016/j.clay.2023.107076
19. T. Wu, J. Tian, X. Shi et al., Predicting anion diffusion in bentonite using hybrid machine learning model and correlation of physical quantities. Sci. Total Environ. **946**, 174363 (2024). https://doi.org/10.1016/j.scitotenv.2024.174363
20. X. Shi, J. Tian, J. Shen et al., Application of machine learning in predicting the apparent diffusion coefficient of Se(IV) in compacted bentonite. J. Radioanal. Nucl. Chem. **333**, 5811–5821 (2024). https://doi.org/10.1007/s10967-024-09637-w
21. Z. Feng, J. Tian, T. Wu et al., Unveiling the Re, Cr, and I diffusion in saturated compacted bentonite using machine-learning methods. Nucl. Sci. Tech. **35**, 93 (2024). https://doi.org/10.1007/s41365-024-01456-8
22. Y. Tochigi, Y. Tachi, Development of diffusion database of buffer materials and rocks-expansion and application method of foreign buffer materials. JAEA-Data/Code 2009–029. (2010). Japan Atomic Energy Agency
23. H.N. Haliduola, F. Bretz, U. Mansmann, Missing data imputation using utility-based regression and sampling approaches. Comput. Meth. Prog. Biol. **226**, 107172 (2022). https://doi.org/10.1016/j.cmpb.2022.107172
24. W.S. Loh, L. Ling, R.J. Chin et al., A comparative analysis of missing data imputation techniques on sedimentation data. Ain Shams Eng. J. **15**, 102717 (2024). https://doi.org/10.1016/j.asej.2024.102717
25. Y. Kim, S.M. Yi, J. Heo et al., Is replacing missing values of PM2.5 constituents with estimates using machine learning better for source apportionment than exclusion or median replacement? Environ. Pollut. **354**, 124165 (2024). https://doi.org/10.1016/j.envpol.2024.124165
26. M. Pastorini, R. Rodríguez, L. Etcheverry et al., Enhancing environmental data imputation: a physically-constrained machine learning framework. Sci. Total Environ. **926**, 171773 (2024). https://doi.org/10.1016/j.scitotenv.2024.171773
27. Z. Feng, J. Tian, X. Shi et al., Analyzing porosity of compacted bentonite via through diffusion method. J. Radioanal. Nucl. Chem. **333**, 1185–1193 (2024). https://doi.org/10.1007/s10967-024-09368-y
28. A. Idiart, M. Pkala, Models for diffusion in compacted bentonite. SKB TR-16-06 (2016). Swedish Nuclear Fuel and Waste Management Company
29. T. Wu, Y. Yang, Z. Wang et al., Anion diffusion in compacted clays by pore-scale simulation and experiments. Water Resour. Res. **56**, e2019WR027037 (2020). https://doi.org/10.1029/2019WR027037
30. N. Hou, Y. Tong, M. Zhou et al., New Strategies for constructing and analyzing semiconductor photosynthetic biohybrid systems based on ensemble machine learning models: Visualizing complex mechanisms and yield prediction. Bioresour. Technol. **412**, 131404 (2024). https://doi.org/10.1016/j.biortech.2024.131404
31. Z. Feng, J. Feng, J. Tian et al., Predicting the diffusion of $CeEDTA^-$ and $CoEDTA^{2-}$ in bentonite using decision tree hybridized with particle swarm optimization algorithms. Appl. Clay Sci. **262**, 107596 (2024). https://doi.org/10.1016/j.clay.2024.107596
32. S.C. Kuok, K.V. Yuen, T. Dodwell et al., Generative broad Bayesian (GBB) imputer for missing data imputation with uncertainty quantification. Knowl. Based Syst. **301**, 112272 (2024). https://doi.org/10.1016/j.knosys.2024.112272
33. M.J. Kim, Y. Cho, Imputation of missing values in well log data using k-nearest neighbor collaborative filtering. Comput. Geosci. **193**, 105712 (2024). https://doi.org/10.1016/j.cageo.2024.105712
34. C. Carpenter, Machine learning aids imputation of missing petrophysical data in Iraqi reservoir. J. Pet. Technol. **76**, 58–61 (2024). https://doi.org/10.2118/0824-0058-JPT
35. H.B. Abdulkhaleq, K.A. Khalil, W.J. Al Mudhafar et al., Advanced machine learning for missing petrophysical property imputation applied to improve the characterization of carbonate reservoirs. Geoemgry Sci. Eng. **238**, 212900 (2024). https://doi.org/10.1016/j.geoen.2024.212900
36. G. Antariksa, R. Muammar, A. Nugraha et al., Deep sequence model-based approach to well log data imputation and petrophysical analysis: A case study on the West Natuna Basin. Indonesia. J. Appl. Geophy. **218**, 105213 (2023). https://doi.org/10.1016/j.jappgeo.2023.105213
37. J. Yang, Z. Zhang, Z. Chen et al., Co-transport of U(VI) and gibbsite colloid in saturated granite particle column: role of pH, U(VI) concentration and humic acid. Sci. Total Environ. **688**, 450–461 (2019). https://doi.org/10.1016/j.scitotenv.2019.05.395
38. Z. Gao, Y. Wang, H. Lü et al., Machine learning the nuclear mass. Nucl. Sci. Tech. **32**, 109 (2021). https://doi.org/10.1007/s41365-021-00956-1
39. V.Q. Tran, Machine learning approach for investigating chloride diffusion coefficient of concrete containing supplementary cementitious materials. Constr. Build. Mater. **328**, 127103 (2022). https://doi.org/10.1016/j.conbuildmat.2022.127103
40. L.J. Yang, J.Y. Peng, F. Qiu et al., Classification of superconducting radio-frequency cavity faults of CAFE2 using machine learning. Nucl. Sci. Tech. **36**, 104 (2025). https://doi.org/10.1007/s41365-025-01685-5
41. J. Li, L. Pan, Z. Li et al., Unveiling the migration of Cr and Cd to biochar from pyrolysis of manure and sludge using machine learning. Sci. Total Environ. **885**, 163895 (2023). https://doi.org/10.1016/j.scitotenv.2023.163895
42. M. Suvarna, P. Preikschas, J. Pérez Ramírez, Identifying descriptors for promoted rhodium-based catalysts for higher alcohol synthesis via machine learning. ACS Catal. **12**, 15373–15385 (2022). https://doi.org/10.1021/acscatal.2c04349

43. H. Liu, X.X. Li, Y. Yuan et al., Prediction of the first $2^+$ states properties for atomic nuclei using light gradient boosting machine. Nucl. Sci. Tech. **36**, 21 (2025). https://doi.org/10.1007/s41365-024-01613-z
44. J. Zhang, Z. Long, Z. Ren et al., Application of machine learning in ultrasonic pretreatment of sewage sludge: prediction and optimization. Environ. Res. **263**, 120108 (2024). https://doi.org/10.1016/j.envres.2024.120108
45. L.R. Van Loon, J. Mibus, A modified version of Archie's law to estimate effective diffusion coefficients of radionuclides in argillaceous rocks and its application in safety analysis studies. Appl. Geochem. **59**, 85–94 (2015). https://doi.org/10.1016/j.apgeochem.2015.04.002
46. H. Aromaa, M. Voutilainen, J. Ikonen et al., Through diffusion experiments to study the diffusion and sorption of HTO, $^{36}$CI, $^{133}$Ba and $^{134}$Cs in crystalline rock. J. Contam. Hydrol. **222**, 101–111 (2019). https://doi.org/10.1016/j.jconhyd.2019.03.002
47. Y. Tachi, K. Yotsuji, Diffusion and sorption of $Cs^+$, $Na^+$, $I^-$ and HTO in compacted sodium montmorillonite as a function of porewater salinity: integrated sorption and diffusion model. Geochim. Cosmochim. Acta **132**, 75–93 (2014). https://doi.org/10.1016/j.gca.2014.02.004
48. M. Glaus, S. Frick, L. Van Loon, A coherent approach for cation surface diffusion in clay minerals and cation sorption models: diffusion of $Cs^+$ and $Eu^{3+}$ in compacted illite as case examples. Geochim. Cosmochim. Acta **274**, 79–96 (2020). https://doi.org/10.1016/j.gca.2020.01.054
49. P. Chen, L.R. Van Loon, S. Koch et al., Reactive transport modeling of diffusive mobility and retention of $TcO_4^-$in Opalinus clay. Appl. Clay Sci. **251**, 107327 (2024). https://doi.org/10.1016/j.clay.2024.107327
50. T. Kozaki, N. Saito, A. Fujishima et al., Activation energy for diffusion of chloride ions in compacted sodium montmorillonite. J. Contam. Hydrol. **35**, 67–75 (1998). https://doi.org/10.1016/S0169-7722(98)00116-8
51. L. Van Loon, W. Müller, K. Iijima, Activation energies of the self-diffusion of HTO, $^{22}Na^+$ and $^{36}Cl^-$ in a highly compacted argillaceous rock (Opalinus clay). Appl. Geochem. **20**, 961–972 (2005). https://doi.org/10.1016/j.apgeochem.2004.10.007
52. M. Descostes, I. Pointeau, J. Radwan et al., Adsorption and retarded diffusion of $Eu^{III}$–$EDTA^-$ through hard clay rock. J. Hydrol. **544**, 125–132 (2017). https://doi.org/10.1016/j.jhydrol.2016.11.014
53. R. Dagnelie, P. Arnoux, J. Radwan et al., Perturbation induced by EDTA on HDO, $Br^-$ and $Eu^{III}$ diffusion in a large-scale clay rock sample. Appl. Clay Sci. **105**, 142–149 (2015). https://doi.org/10.1016/j.clay.2014.12.004
54. M. García Gutiérrez, M. Mingarro, T. Missana, Influence of temperature and dry density coupled effects on HTO, $^{36}$Cl, $^{85}$Sr and $^{133}$Ba diffusion through compacted bentonite. Prog. Nucl. Energy **176**, 105407 (2024). https://doi.org/10.1016/j.pnucene.2024.105407
55. C. Joseph, J. Mibus, P. Trepte et al., Long-term diffusion of U(VI) in bentonite: Dependence on density. Sci. Total Environ. **575**, 207–218 (2017). https://doi.org/10.1016/j.scitotenv.2016.10.005
56. K. Furukawa, Y. Takahashi, H. Sato, Effect of the formation of EDTA complexes on the diffusion of metal ions in water. Geochim. Cosmochim. Acta **71**, 4416–4424 (2007). https://doi.org/10.1016/j.gca.2007.07.009