基于生成式对抗网络的机器学习预测 Re（Ⅶ）的表观扩散系数

冯佳星[1,2]，高学文[1,2]，徐 克[1,2]，伍 涛[1,2,*]

1. 湖州师范学院 工学院，浙江 湖州 313000；
2. 全省工业固废热解处置技术及智能化装备重点实验室，浙江 湖州 313000

摘要：表观扩散系数（D_a）是高放废物处置库安全评价的关键参数。然而，由于样本数有限、扩散机制不明确等问题，难以满足复杂地质条件下高精度预测的需求。本工作采用机器学习算法预测膨润土中 Re(Ⅶ)的 D_a 值。数据集包含 1 073 组实验样本和 26 个输入特征量。通过引入高斯噪声与生成式对抗网络（GAN）技术进行数据增强，最终将样本数扩充到 4 292 组。探讨了样本数对 D_a 预测精度的影响，并比较了集成算法 LGBM-XGBoost 与深度神经网络（DNN）算法对预测性能的影响。回归预测结果表明，LGBM-XGBoost 集成模型的预测性能优于 DNN 模型，最优模型的决定系数 R^2 为 0.94。通过沙普利可加性特征解释方法（SHAP）分析和特征重要性评估，发现总孔隙率与有效压实密度是影响 D_a 预测精度的主要因素。为了验证模型的泛化能力，采用贯穿扩散法测量了压实膨润土中 Re(Ⅶ)的 D_a 值，随着压实干密度从 1 800 kg/m^3 降低到 1 200 kg/m^3，D_a 值从 1.09×10^{-10} m^2/s 增加到 2.49×10^{-10} m^2/s。LGBM-XGBoost 模型预测的 D_a 相对标准偏差低于 17%，表明该模型在未见样本上保持稳定预测性能。该方法为高放废物地质处置安全性评价提供了一种潜在的预测方法和机理分析工具。

关键词：机器学习；放射性核素；表观扩散系数；膨润土；扩散实验

中图分类号：TL942.1 **文献标志码：**A **文章编号：**0253-9950（2025）06-0695-10

doi: 10.7538/hhx.2025.47.06.0695

Predicting Apparent Diffusion Coefficient of Re(Ⅶ) Using Machine Learning With Generative Adversarial Networks

FENG Jia-xing[1,2], GAO Xue-wen[1,2], XU Ke[1,2], WU Tao[1,2,*]

1. Department of Engineering, Huzhou University, Huzhou 313000, China;
2. Provincial Key Laboratory of Industrial Solid Waste Pyrolysis Disposal Technology and Intelligent Equipment, Huzhou 313000, China

Abstract: Diffusion is the predominant transportation mechanism of radionuclides in compacted bentonite, which is attributed to the low permeability, high swelling capacity, and strong adsorption characteristics. The apparent diffusion coefficient(D_a) is a crucial parameter in the safety evaluation of high-level radioactive waste repositories. However, it remains challenging to accurately predict the D_a value under complex geological conditions due to scarce experimental data and unclear diffusion

收稿日期：2025-06-28；**修订日期：**2025-11-08

基金项目：国家自然科学基金资助项目（12475340）

* **通信联系人：**伍 涛

mechanisms. In this study, machine learning models were employed to predict the D_a values of Re(Ⅶ) in compacted bentonites. The dataset included 1 073 experimental instances with 26 input features. Feature engineering techniques were applied to standardize the data, including outlier removal, logarithmic transformation, and max-min normalization. Data augmentation was performed using both Gaussian noise injection and the generative adversarial network(GAN) techniques, expanding the datasets to 4 292 instances with 26 input features. The influence of instance quantity on predictive accuracy was systematically analyzed, with comparative performance evaluation conducted between an integrated light gradient boosting machine-extreme gradient boosting(LGBM-XGBoost) algorithm and a deep neural network(DNN) architecture. It shows that the predictive accuracy increases with increasing quantity of instances. The predictive accuracy increases significantly after using Gaussian noise injection and GAN techniques. However, Gaussian noise injection resultes in a decrease of model robust. In addition, the LGBM-XGBoost model outperforms the DNN model in predictive accuracy, achieving a coefficient of determination(R^2) of 0.99 for training set, 0.94 for validation set, and 0.94 for test sets. 95% of the instance predictions fell within a factor of 2 of the experimental values. Shapley additive explanation(SHAP) and feature importance(FI) techniques were applied in the LGBM-XGBoost model to analyze the predictive contribution of input features. It shows that the total porosity and compaction dry density are the top-two contributors. To evaluate the model's generalization capability, through-diffusion experiments were conducted to measure the D_a values of Re(Ⅶ) in saturated compacted bentonite. The D_a values increase from 1.09×10^{-10} m^2/s to 2.49×10^{-10} m^2/s with decreasing compacted dry density from 1 800 kg/m^3 to 1 200 kg/m^3. The negative relationship between D_a and compacted dry density is consistent with the results of SHAP and FI analysis. It can be explained that the increase in total porosity facilitates Re(Ⅶ) diffusion. The LGBM-XGBoost model exhibites excellent generalization capability, with relative errors of D_a below 17%. This study establishes a potential predictive approach and mechanistic analysis tool for the safety assessment of high-level radioactive waste repositories.

Key words: machine learning; radionuclide; apparent diffusion coefficient; bentonite; diffusion experiments

膨润土常被选作高放废物处置库工程屏障中的缓冲/回填材料，用于阻碍放射性核素向生物圈的迁移。研究表明，在压实膨润土中，核素的迁移机制以扩散为主导[1]。表观扩散系数(D_a)描述了放射性核素在膨润土中的扩散行为，是处置库安全评价中的关键参数[2]。在过去的几十年里，国内外学者为了获取该参数开展了大量研究[3-5]。目前，D_a主要采用扩散实验进行测定，但是由于实验周期长达数月甚至数年，导致测试成本高昂且效率低下[6]。在实际处置更为复杂环境中，D_a受到多种因素的共同影响，包括核素性质、膨润土理化性质和孔隙水化学组成等，这使得基于扩散模型的参数预测同样面临准确性不足、计算量巨大的技术挑战[7]。

机器学习模型具有强大的计算能力，可以处理复杂的高维数据[8]，有助于解决现有扩散模型存在的计算效率低的难题。日本原子能机构开发的放射性核素扩散数据库(JAEA-DDB)包含5 000多组实验样本数据和30多个特征量，为构建机器学习数据库奠定了坚实基础[9]。目前，机器学习方法已在核素扩散研究中得到广泛应用[10-15]。例如，Shi等[16]采用了轻量级梯度提升机(LGBM)和随机森林(RF)对D_a进行预测，为解决原始数据集中存在的数据缺失问题，采用回归插补法填补缺失数据，将实验样本数从850组增加至956组。研究表明随着输入特征量和样本数的增加，预测模型的性能显著提升，模型的决定系数R^2最高达0.98(R^2越接近1，表明模型的预测精度越高)。Gong等[17]利用决策树和神经网络等六种机器学习方法预测放射性核素在膨润土中的有效扩散系数(D_e)，数据集包含10个输入特征量，样本数为860组。其中RF模型表现最优，R^2最高达0.99。值得注意的是，生成式对抗网络(GAN)算法是当前研究的热点，已经成功应用于

样本数据的扩充[18]。本课题组采用GAN-人工神经网络(ANN)模型预测核素在压实膨润土中的D_e和吸附分配系数,数据集的输入特征量为26个,样本数为2 068组(包括1 034组实验样本和1 034组生成样本),得到模型的R^2最高达0.98[12]。通常,高精度机器学习模型依赖大规模高质量的数据集。然而,JAEA-DDB数据库中存在大量数据缺失问题,实际可用的实验数据较为有限,在一定程度上限制了机器学习模型预测精度的进一步提高。

本工作分别采用集成算法(LGBM-XGBoost)与深度神经网络(DNN)预测压实膨润土中核素的D_a。为了提高模型的预测精度,采用生成式对抗网络(GAN)并引入高斯噪声进行数据增强,将有限实验数据扩充翻倍,最终构建的数据集包括4 292组样本数(其中1 073组实验样本、3 219组生成样本)和26个输入特征量。通过五倍交叉验证对两个模型进行了训练、优化和测试,获得最优预测模型。此外,还采用贯穿扩散实验测量Re(Ⅶ)(代替放射性核素^{99}Tc(Ⅶ))在压实安吉膨润土中的D_a。其中安吉膨润土的蒙脱石含量为46%,测量得到的Re(Ⅶ)在安吉膨润土中的D_a作为模型的外推实验数据,用于评价模型的泛化能力。本工作希望开发具有高精度、强鲁棒性和可解释的核素D_a预测模型,为处置库的安全评价提供潜在预测工具。

1　材料与方法

1.1　扩散实验

1.1.1　材料和设备　膨润土粉末产自中国浙江省安吉县,未经进一步处理直接使用。乙酸铵(NH_4OAc)溶液、高铼酸铵(NH_4ReO_4)、氯化钠等试剂均购买自上海阿拉丁生化科技股份有限公司,上述试剂均为分析纯。Optima7000DV型电感耦合等离子体光谱仪(ICP-OES),美国珀金埃尔默公司。

1.1.2　贯穿扩散实验　采用贯穿扩散法测量Re(Ⅶ)(代替放射性核素^{99}Tc(Ⅶ))的D_a。扩散池由2个底盖、1个外套、2个密封圈、4个固定螺丝和2个不锈钢滤片组成。首先,用游标卡尺测量扩散池高度、内径、底盖凸面高度及2片不锈钢滤片厚度,计算目标体积。随后,根据目标体积与预设压实干密度,计算不同干密度下需称量的膨润土粉末质量;称取对应质量膨润土加入扩散池,将2片滤片分别置于膨润土上下两侧固定形态。最后,压实膨润土至底盖与扩散池外套完全贴合无空隙,拧紧4个固定螺丝,完成装样。由于核素在不锈钢滤片中的有效扩散系数为膨润土中的10%,同时其厚度为膨润土厚度的10%,因此本工作忽略了滤片效应[19]。采用微量HCl和NaOH溶液将NaCl溶液调节至pH=7.5±0.1。膨润土的压实干密度为1 200～1 800 kg/m^3,在25 ℃条件下,连接在保定兰格蠕动泵中,扩散池采用pH=7.5±0.1的0.5 mol/L NaCl溶液充分接触5周,使膨润土块达到水饱和状态。扩散池一侧(扩散距离x'=0)为200 mL Re(Ⅶ)源溶液,另一侧(x'=L,L为膨润土土块厚度)为10 mL 0.5 mol/L NaCl样品溶液。为了保持恒定浓度梯度,定期更换NaCl样品溶液。使用ICP-OES测量Re(Ⅶ)的浓度,从而获得Re(Ⅶ)的累积扩散总量和扩散通量随时间变化的实验数据。

1.1.3　扩散实验数据处理　有效扩散系数(D_e)和岩石容量因子(α)通过自编程序(FDP)拟合累积扩散总量(A_{cum})与时间(t)的关系计算得出,累积扩散总量(A_{cum})计算公式如式(1):

$$A_{cum}=SLc_0\left(\frac{D_e t}{L^2}-\frac{\alpha}{6}-\frac{2\alpha}{\pi^2}\sum_{n=1}^{\infty}\frac{(-1)^n}{n^2}\cdot\exp\left(-\frac{D_e n^2\pi^2 t}{L^2\alpha}\right)\right) \quad (1)$$

式中:S(m^2)、L(m)和c_0(mol/L)分别为压实膨润土块的横截面积、厚度和Re(Ⅶ)的初始浓度。扩散通量($J(L,t)$)的计算公式如式(2):

$$J(L,t)=\frac{1}{S}\cdot\frac{\partial A_{cum}}{\partial t} \quad (2)$$

1.2　机器学习

1.2.1　集成算法　Light Gradient Boosting Machine(LGBM)属于决策树算法,采用分布式高效梯度提升框架,训练速率更快、效率更高[20]。eXtreme Gradient Boosting(XGBoost)同样属于决策树算法,作为梯度提升算法的扩展,通过梯度提升技术提升模型准确性与泛化能力。它具有高效训练、正则化控制过拟合和处理大规模数据等优势[21]。集成机器学习模型可以结合两者的优势,提高模型的预测性能和鲁棒性,为单个模型中的偏差和方差提供了一个更好的解决方案[14,22]。本工作使用scikit包中的投票回归方法预测输出特征量,通过

五折交叉验证结合网格搜索优化 VotingRegressor, LGBM 的权重设置为 0.6, XGBoost 为 0.4。LGBM-XGBoost 模型的预测结果计算公式如式(3):

$$\hat{y}=\sum_{i=1}^{n} y_i\omega_i \tag{3}$$

式中: $\hat{y}$ 表示集成算法预测结果; y_i 和 ω_i 分别表示第 i 个模型的预测结果和对应的权重。

1.2.2 深度神经网络 深度神经网络(DNN)是一种模拟人脑神经网络结构和功能的计算模型, 它利用反向传播算法通过迭代优化网络参数(包括权重和偏置), 从而减小预测值和实际值之间的误差[23]。该模型由多层相互连接的人工神经元构成, 包括输入层、隐藏层和输出层。数据流由输入层接收, 经过隐藏层处理, 最后到达输出层。每个神经元对其输入进行线性加权求和, 通过非线性激活函数如 ReLU 处理产生输出。通过这种层级结构, DNN 可以学习输入和输出之间的复杂映射关系。具体而言, 第 j 个神经元的输出(Q_j)、激活函数(f)和权重(W_{ji})之间的关系可由式(4)表示[24]:

$$Q_j=f\pi\left(\sum_i (W_{ji}\cdot X_i)+b_j\right) \tag{4}$$

式中: X_i 为前一层中第 i 个人工神经元的输出, b_j 为第 j 个人工神经元的偏置。

1.2.3 生成式对抗网络 生成式对抗网络(GAN)由两个神经网络: 生成器网络(G)和鉴别器网络(D)子模型组成[18]。生成器从具有概率分布 P_z 的随机噪声(z)中创建生成数据, 旨在模拟真实数据(x)的分布 $P_{\text{data}}(x)$, 而 D 通过最小化 $\lg(1-D(G(z)))$ 和最大化 $\lg(D(x))$ 来区分生成数据和真实数据。损失函数如式(5):

$$\min_G\max_D V(D,G)=E_{x,P_{\text{data}}(x)}[\lg D(x)]+E_{z,P_z(z)}[\lg(1-D(G(z)))] \tag{5}$$

式中: E 为期望, 用于计算对应数据分布下损失函数的平均值。因此, 训练过程会经历多次迭代, 直到鉴别器无法再区分真实数据样本和生成的数据样本。

1.2.4 数据集描述 本工作采用的放射性核素扩散数据集包括 1 086 个样本数和 26 个输入特征量, D_a 为模型的输出特征量[12](详见表 1)。26 个输入特征量分为三组: (1)膨润土性质, 包含 9 个输入特征量, 分别为蒙脱石质量分数(m)、混合物含砂量(R_s)、外比表面积(A_{ext})、阳离子交换容量(CEC)、土的颗粒密度(ρ_s)、混合物压实干密度(ρ_d)、有效压实密度(ρ_b)、总孔隙率(ε_{tot})和蒙脱石堆叠层数(n_c); (2)孔隙水性质, 包含 10 个输入特征量, 分别为离子强度(I)、$c(K^+)$、$c(Na^+)$、$c(Ca^{2+})$、$c(Cl^-)$、$c(HCO_3^-/CO_3^{2-})$、$c(SO_4^{2-})$、pH、T 和德拜长度(κ^{-1}); (3)放射性核素性质, 包含 7 个输入特征量, 分别为质子数(Z)、中子数(N')、分子质量(M)、离子电荷(z)、离子在水中的扩散系数(D_w)、水合离子半径(r)和摩尔电导率(λ)。其中 α、吸附分配系数(K_d)和 D_a 密切相关, 所以不包含在输入特征量中。采用马氏距离方法(阈值设为 10)剔除 13 个异常值后, 得到数据集Ⅰ(包含1 073 个实验样本)。为了扩充样本数量、提升数据集多样性, 从而提高模型预测精度, 本工作采用高斯噪声注入、GAN 生成等数据增强方法, 在数据集Ⅰ的基础上, 添加标准差为 0.1 的高斯白噪声, 构建了包含 2 146 个样本的数据集Ⅱ。进一步利用 GAN 对数据进行增强, 构建了包含4 292 个样本的数据集Ⅲ。为了提高模型的预测精度, 对数据偏度大于 2 的输入特征量以及量纲差异显著的输入特征量(D_w 及 κ^{-1})进行对数转换, 随后采用 min-max 归一化法对所有输入特征量进行标准化预处理。

1.2.5 数据集划分与评估 采用常用的数据集划分方法, 将数据集按照 80% 和 20% 的比例随机分为训练集和测试集。为了有效防止模型的过拟合, 训练过程中采用五折交叉验证。具体而言, 在每一轮交叉验证中, 训练集被随机分为五个子集, 其中 80% 的数据用于模型训练, 剩余 20% 的子集数据则作为验证集用于超参数的调优。最终模型性能的评估在独立的测试集上进行。模型的预测精度分析使用统计性能指标进行评估:

$$R^2=1-\frac{\sum_{i=1}^{N}\left(x_i^{\text{exp}}-x_i^{\text{pred}}\right)^2}{\sum_{i=1}^{N}\left(x_i^{\text{exp}}-x_{\text{ave}}^{\text{exp}}\right)^2} \tag{6}$$

$$\text{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i^{\text{exp}}-x_i^{\text{pred}})^2} \tag{7}$$

$$\text{MSE}=\frac{1}{n}\sum_{i=1}^{n}(x_i^{\text{exp}}-x_i^{\text{pred}})^2 \tag{8}$$

$$\text{MAE}=\frac{1}{n}\sum_{i=1}^{n}|x_i^{\text{exp}}-x_i^{\text{pred}}| \tag{9}$$

式中: x_i^{exp} 和 x_i^{pred} 分别是离子的 D_a 的实验值和预测

表 1　数据集的特征量和样本数的详细信息

Table 1　Detailed information on feature quantities and sample sizes of dataset

物理量名称	物理量符号	最小值	最大值	平均值	标准差	偏度
蒙脱石质量分数	m /%	0.0	1.0	0.79	0.2	−1.0
含砂量	R_s /%	0	90	3.3	11.3	3.8
外比表面积	A_{ext}/(m²•g⁻¹)	26	112	34.8	13	2.9
阳离子交换容量	CEC/(meq, 以100 g吸附剂计)	11.3	121	89	21	−0.7
颗粒密度	ρ_s/(kg•m⁻³)	2 600	2 900	2 798	78	−0.3
压实干密度	ρ_d/(kg•m⁻³)	200	2 330	1 314	376	−0.1
有效压实密度	ρ_b/(kg•m⁻³)	200	2 100	1 291	366	0
总孔隙率	ε_{tot}	0.15	0.93	0.52	0.14	0.1
堆叠层数	n_c	8	54	38	10.8	−1.1
离子强度	I/(mol•L⁻¹)	0	6.6	0.3	1	5.6
钾离子浓度	$c(K^+)$/(mol•L⁻¹)	0	0.16	0	0	32.6
钠离子浓度	$c(Na^+)$/(mol•L⁻¹)	0	6.6	0.3	1	5.7
钙离子浓度	$c(Ca^{2+})$/(mol•L⁻¹)	0	0.37	0.01	0	7.5
氯离子浓度	$c(Cl^-)$/(mol•L⁻¹)	0	6.6	0.3	1	5.7
碳酸根/碳酸氢根浓度	$c(HCO_3^-/CO_3^{2-})$/(mol•L⁻¹)	0	0.01	0.0	0	4.3
硫酸根浓度	$10^{-3}c(SO_4^{2-})$/(mol•L⁻¹)	0	40	2.1	0	4.2
pH值	pH	3	13.4	7.7	2.1	1.1
温度	T/℃	5	90	27.7	14.1	1.7
德拜长度	κ^{-1}	1.2	327 000	126 080	150 331	0.4
质子数	Z	6	95	40	28	0.6
中子数	N'	6	146	55	45	0.8
分子质量	M/(g•mol⁻¹)	22	484	117	97	1
电荷数	z	−4	5	0.4	1.6	0.5
水中离子扩散系数	$10^{-10}D_w$/(m²•s⁻¹)	1	57.4	16.4	8.6	1.2
水合离子半径	r/Å(1 Å=0.1 nm)	0.5	23.4	2.1	2.5	6.9
离子摩尔电导率	λ/(m²•S•mol⁻¹)	0	0.04	0.01	0.01	3.4
表观扩散系数	$10^{-13}D_a$/(m²•s⁻¹)	0.002 5	1 072	811	127	3.3

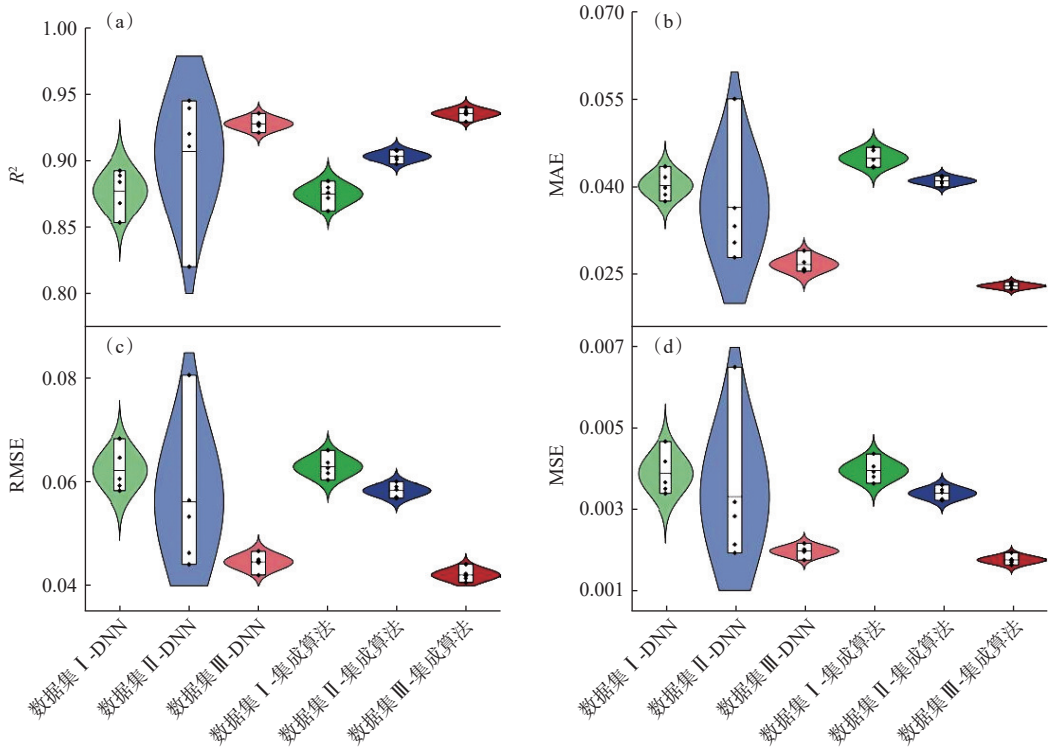
值; x_{ave}^{exp} 表示 D_a 的平均实验值。R^2 值接近 1 和均方根误差(RMSE)、均方误差(MSE)、平均绝对误差(MAE)值接近 0 表示模型具有较高的预测精度。

2　结果与讨论

2.1　模型开发与测试

分别采用集成算法(LGBM-XGBoost) 和 DNN 模型对放射性核素在压实膨润土中的表观扩散系数(D_a)进行预测分析。图 1 为三个数据集中测试集的五折交叉验证预测结果。模型的超参数设置详见表 2。

当引入高斯白噪声后, 样本数从 1 073(数据集Ⅰ)增加到 2 146(数据集Ⅱ)时, 除 DNN 算法的鲁棒性有所下降外, 所有数据集的性能指标均得到了显著改善。LGBM-XGBoost 算法凭借自身的优势表现出更强的鲁棒性。进一步采用 GAN 技术进行数据增强, 当样本数增加到 4 292(数据集Ⅲ)时, LGBM-XGBoost 和 DNN 算法的性能指标均获得进一步提高, 在测试集中 R^2 分别为 0.94 和 0.93。实验结果表明, 数据集增强能够有效提升机器学习模型处理数据内部的非线性和复杂特征的能力, 从而显著改善预测性能和模型的鲁棒性[12]。LGBM-XGBoost 的算法整体表现优于 DNN 算法。作为集成模型, LGBM-XCBoost 通过决策树的组合机制能够更有效地捕获数据中的复杂的模式和潜在关系[14]。



图中黑色曲线表示性能指标分布的核密度估计曲线，黑色散点代表各次交叉验证结果，箱图中的黑色横线代表平均值，箱体高度反映了机器学习模型的鲁棒性（箱体越扁代表模型的鲁棒性能越好）

（a）——R^2，（b）——MAE，（c）——RMSE，（d）——MSE

图 1 使用五倍交叉验证对三个数据集中测试集的机器学习模型预测结果

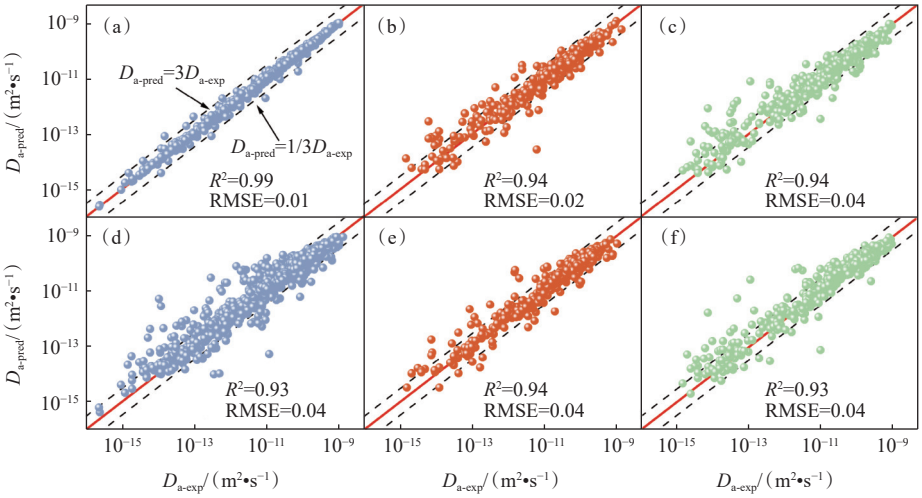
Fig. 1 Machine learning model prediction results of test sets in three datasets tested using five fold cross-validation

表 2 模型超参数设置表

Table 2 Model hyperparameter setting table

LGBM-XGBoost				DNN	
LGBM		XGBoost			
Max_depth	−1	Max_depth	3	N-epoch	5 000
Learning_rate	0.05	Eta	0.05	Learning_rate	0.001
Num_leaves	30	N_estimators	1 000	Hidden layers	3
Min_data_in_leaf	20	Gamma	0.01	Kernel_Regularizer_L2	0.01
Feature_fraction	0.2	Lambda	0.09	Number of neurons	32
Bagging_freq	30	Subsample	0.51	Activation_function	ReLU
Bagging_seed	28	Reg_alpha	0.04		
Bagging_fraction	0.43	Min_child_weight	6		
Lambda_l1	0.03	Colsample_bytree	0.5		
Lambda_l2	0.05				

基于数据集Ⅲ展现出最优的预测性能，本工作采用LGBM-XGBoost和DNN的最优模型进行实验和预测D_a值的回归分析（图2）。如图2所示，图2（a, d）、（b, e）、（c, f）中数据点分别代表训练集、验证集和测试集的回归预测结果。数据点分布越接近对角线，表明模型D_a预测值（$D_{a\text{-}pred}$）与实验值（$D_{a\text{-}exp}$）吻合度越高，反映出模型在模拟放射性核素扩散过程方面具有更好的性能表现。定量分析显示，LGBM-XGBoost（R^2=0.94）的预测性能优于DNN（R^2=0.93）。进一步对数据集Ⅲ的预测误差进行统计分析发现：DNN模型中，预测值偏离真实值达3倍以上的样本占比6.55%（281个），偏离2倍以上的样本占比12.44%（534个）；相比之下，LGBM-XGBoost模型的预测误差更低，



(a, b, c)——LGBM-XGBoost; (d, e, f)——DNN

(a, d)——训练集; (b, e)——验证集; (c, f)——测试集

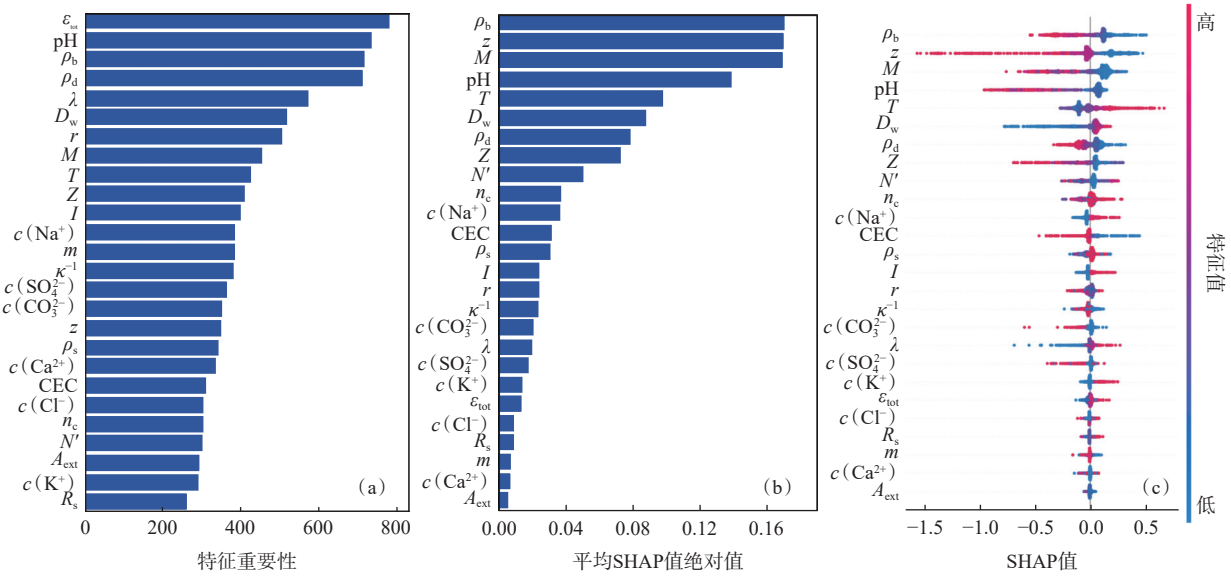
图 2　基于数据集Ⅲ的实验与预测表观扩散系数回归图

Fig. 2　Experimental and predicted apparent diffusion coefficient regression graphs based on dataset Ⅲ

95% 的样本数的预测值与真实值的比值小于2倍。其中，2倍以上偏离样本占比4.96%(213个)，3倍以上偏离样本仅占2.63%(113个)。这一结果充分证明，LGBM-XGBoost算法较DNN算法具有更高的预测精度。

2.2　模型可解释性分析

特征重要性和沙普利可加性特征解释方法(Shapley additive explanations, SHAP)技术常用在机器学习模型中解决“黑箱”问题，通过对输入特征量的重要性排序来揭示核素扩散机理[11]。图3通过特征重要性和SHAP分析技术揭示输入特征量对预测结果的贡献程度。其中，纵坐标按照特征重要性从高到低排列。图3(c)为SHAP分析蜂窝图。横坐标表示输入特征量对预测结果的影响大小与方向：SHAP值为正表示输入正向推动，为负表示为反向推动；点的颜色表示输入特征量的大小，颜色越深，特征量越大。综合考虑特征重要性和平均SHAP值绝对值分析表明：



(a)——特征重要性，(b)——平均SHAP值绝对值，(c)——SHAP值蜂窝图

图 3　基于数据集Ⅲ的LGBM-XGBoost模型的全局性分析

Fig. 3　Global analysis of LGBM-XGBoost model based on dataset Ⅲ

ε_{tot} 和 ρ_b 分别是影响 D_a 的最关键因素。SHAP 分析与特征重要性分析结果不一致，这反映了不同机器学习算法在特征重要性评估及预测机制上的固有差异。机理分析说明：较高的孔隙率意味着膨润土中存在更多相互连接的孔隙空间，有利于放射性核素扩散[11]。较大的有效压实密度导致膨润土颗粒之间的孔隙减小、结合更紧密，从而抑制放射性核素的扩散[11]。值得注意的是，当含砂量为 0 时，有效压实密度 ρ_b 与压实干密度 ρ_d 完全相等；在含砂量不为 0 的情况下，二者仍呈显著正相关关系。

2.3 贯穿扩散实验及模型应用

2.3.1 安吉膨润土特性及 Re 浓度 N_2-BET 法测定安吉膨润土比表面积（A_{ext}）为 60.3 m^2/g[25]；XRD 结合 Rietveld 精修定量分析显示，其矿物组成为蒙脱石（46%，质量分数，下同）、石英（33%）、正长石（10%）、微斜长石（8%）和方解石（3%）[25]；CEC 测定结果为 76 meq/100 g（吸附剂，下同），主要可交换阳离子及交换容量为：K^+（1 meq/100 g）、Na^+（24 meq/100 g）、$1/2Ca^{2+}$（44 meq/100 g）、$1/2Mg^{2+}$（8 meq/100 g）；激光散射法测定其特征粒径 d_{50} 为 11.6 μm[25]。Re(Ⅶ)溶液制备：将高铼酸铵溶解于 0.5 mol/L NaCl 溶液中，经 ICP-OES 测定得到浓度为 1.3×10^{-3} mol/L 的 Re(Ⅶ)溶液。

2.3.2 贯穿扩散实验 采用贯穿扩散法测定了 Re(Ⅶ)在安吉膨润土中的 D_a 值。图 4 为实验测得的累积扩散总量 A_{cum} 和扩散通量 $J(L, t)$ 随时间变化的曲线。由图 4 可以看出，A_{cum} 和 $J(L, t)$ 的变化规律均呈现出核素扩散过程典型的过渡态和稳态两个阶段特征。在过渡态阶段（0～3 d），A_{cum} 随着时间增加呈缓慢增加的趋势；进入稳态阶段（3 d 后），A_{cum} 与时间呈线性关系。而 $J(L, t)$ 在过渡态阶段快速上升，到达稳态阶段后则保持不变。实验数据表明：Re(Ⅶ)的扩散过程约需 3～4 d 完成从过渡态到达稳态的转变。此外，当膨润土的 ρ_d 由 1 200 kg/m^3 增加至 1 800 kg/m^3，A_{cum} 和 $J(L, t)$ 呈现明显下降趋势。这是因为膨润土的孔

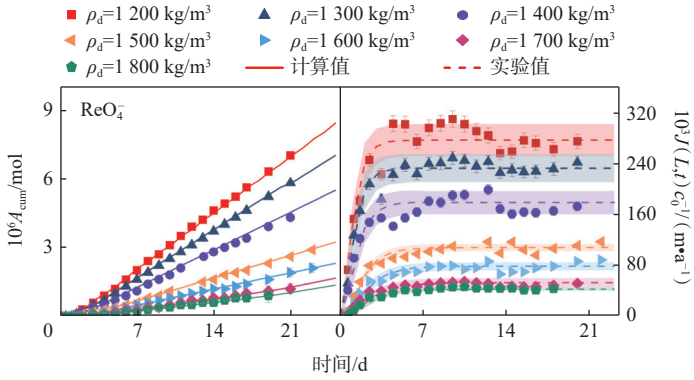


图 4 Re(Ⅶ)在安吉膨润土中的累积扩散总量(a)和扩散通量(b)

Fig. 4 Accumulated mass(a) and flux(b) for Re(Ⅶ) in Anji bentonite

表 3 Re(Ⅶ)在安吉膨润土中的扩散参数

Table 3 Diffusion parameters of Re(Ⅶ) in Anji bentonite

阴离子	ρ_d/(kg•m^{-3})	α	$10^{10}D_a$/(m^2•s^{-1})	ε_{tot}	K_d/(m^3•kg^{-1})[25]	$10^{11}D_e$/(m^2•s^{-1})
ReO_4^-	1 200	0.45±0.040	2.49±0.314	0.532	0	11.2±1.00
	1 300	0.40±0.040	2.38±0.327	0.493	0	9.50±0.90
	1 400	0.36±0.030	2.03±0.258	0.454	0	7.30±0.70
	1 500	0.30±0.010	1.48±0.015	0.415	0	4.45±0.43
	1 600	0.26±0.004	1.23±0.005	0.376	0	3.20±0.25
	1 700	0.20±0.001	1.09±0.002	0.337	0	2.17±0.29
	1 800	0.16±0.001	1.09±0.001	0.298	0	1.75±0.10

注：n=3

隙率会随着 ρ_d 增大而减小, 从而导致 Re(Ⅶ)通过膨润土块的扩散总量减少。表 3 汇总了 Re(Ⅶ)的有效扩散系数(D_e)、D_a 和岩石容量因子(α)。其中, 表观扩散系数与总孔隙率的表达式如式(10)、(11):

$$D_a = \frac{D_e}{\alpha} \tag{10}$$

$$\varepsilon_{tot} = 1 - \frac{\rho_d}{\rho_s} \tag{11}$$

式中: ε_{tot} 为总孔隙率; ρ_d 为膨润土压实干密度; ρ_s 为膨润土颗粒密度。

Re(Ⅶ)的扩散参数随着压实干密度的增加而降低(表 3)。当 ρ_d 从 1 200 kg/m^3 增大至 1 800 kg/m^3, Re(Ⅶ)的 α 从 0.45 左右降至 0.16 左右。相应地, D_e 由约 11.2×10^{-11} m^2/s 降到约 1.75×10^{-11} m^2/s, 该结果与已发表研究[26]相吻合。同时, D_a 由约 2.49×10^{-10} m^2/s 降至约 1.09×10^{-10} m^2/s。值得注意的是, α 值始终小于 ε_{tot}, 这表明 Re(Ⅶ)阴离子没有吸附在膨润土表面[26]。这种现象是由于表面呈电负性的膨润土颗粒对 Re(Ⅶ)阴离子的排斥作用, 导致 Re(Ⅶ)无法进入膨润土的层间孔隙, 此时 α 表征的是实际可参与扩散过程的有效孔隙率。但 ρ_d 大于 1 700 kg/m^3, D_e 和 D_a 的减小可以忽略。

2.3.3　模型应用　为了评估模型的泛化能力, 将基于数据集Ⅲ构建的 LGBM-XGBoost 模型用于预测安吉膨润土中 Re(Ⅶ)的 D_a(图 5)。图 5 中两条黑色虚线分别代表 Re(Ⅶ)的 D_a 模型预测值($D_{a\text{-}pred}$)是实验值($D_{a\text{-}exp}$)的 1.2 倍和 0.8 倍的参考线, 用于量化模型预测结果与实验数据的偏离程度。数据点越靠近实线, 表明模型预测结果与实验值高度一致。可以看出, 这些数据点均落在两条黑色虚线构成的区间内, 该结果表明模型在未见样本上的预测相对标准偏差可控制在 17% 以内, 具有良好的泛化能力。

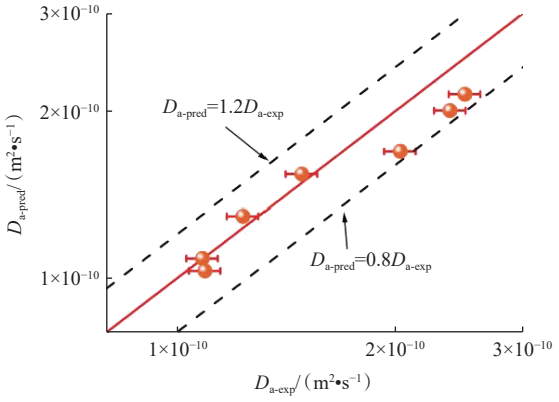


图 5　Re(Ⅶ) 在压实安吉膨润土中的表观扩散系数预测值与实验值的散点图

Fig. 5　Scatter plot of predicted and experimental apparent diffusion coefficient of Re(Ⅶ) in Anji bentonite

3　结　论

通过构建 LGBM-XGBoost 和 DNN 两种机器学习模型, 预测核素的表观扩散系数(D_a)。采用高斯噪声注入和 GAN 算法进行数据增强, 可以得出以下结论。

(1) 数据增强有利于提高模型预测性能和鲁棒性。LGBM-XGBoost 算法优于 DNN 模型, 数据增强使得 R^2 由 0.87 增加到 0.94, 基于数据集Ⅲ(包含 4 292 组样本数和 26 个输入特征量)的 D_a 预测精度最高。

(2) 采用特征重要性和 SHAP 分析探讨了核素扩散机理和规律, 发现总孔隙率与有效压实密度是影响 D_a 预测精度的主要因素, 这是由于较高的孔隙率意味着膨润土中存在更多相互连接的孔隙空间, 从而促进放射性核素的扩散迁移。

(3) 为了验证模型泛化能力, 还开展了贯穿扩散实验研究 Re(Ⅶ)(代替放射性核素^{99}Tc(Ⅶ))在安吉膨润土中的扩散行为。结果表明 Re(Ⅶ)的 D_a 值随着压实干密度的增加而降低。LGBM-XGBoost 模型预测的 D_a 相对标准偏差低于 17%, 表明该模型在未见样本上保持稳定预测性能。

本研究通过机器学习与数据增强技术的结合, 实现了对核素在膨润土中 D_a 值的高精度预测, 揭示了多物理场耦合下核素扩散过程中的关键参数。需要说明的是, 当前模型主要基于贯穿扩散法获得的阴离子核素扩散数据构建, 对类似体系具有较高的预测可靠性, 当应用于极端条件或不同矿物组成的缓冲材料时, 其适用性仍有待进一步验证。总体而言, 该模型为地质处置库安全评价提供了有力的数字化支撑, 也为其他屏障材料性能的预测提供了方法论参考。

参考文献:

[1] 王祥云,陈涛,王春丽,等.若干重要放射性核素在北山花岗岩及高庙子膨润土中的吸附和扩散研究[J].中国科学:化学,2020,50(11):1585-1599.

[2] Chen Z, Wang S, Hou H, et al. China's progress in radionuclide migration study over the past decade (2010–2021): sorption, transport and radioactive colloid[J]. Chin Chem Lett, 2022, 33(7): 3405-3412.

[3] Cormenzana J L, García-Gutiérrez M, Missana T, et al.

Simultaneous estimation of effective and apparent diffusion coefficients in compacted bentonite[J]. J Contam Hydrol, 2003, 61(1-4): 63-72.

[4] Gooddy D C, Kinniburgh D G, Barker J A. A rapid method for determining apparent diffusion coefficients in Chalk and other consolidated porous media[J]. J Hydrol, 2007, 343(1-2): 97-103.

[5] Zoia A, Latrille C. Estimating apparent diffusion coefficient and tortuosity in packed sand columns by tracers experiments[J]. J Por Media, 2011, 14(6): 507-520.

[6] Joseph C, Mibus J, Trepte P, et al. Long-term diffusion of U(Ⅵ) in bentonite: dependence on density[J]. Sci Total Environ, 2017, 575: 207-218.

[7] Wu T, Yang Y, Wang Z, et al. Anion diffusion in compacted clays by pore-scale simulation and experiments[J]. Water Resour Res, 2020, 56(11): e2019WR027037.

[8] 尹晚秋,薄涛,赵玉宝,等.铀、锆以及铀锆合金精确原子间势的深度学习[J].核化学与放射化学,2024,46(5):450-461.

[9] Tochigi Y, Tachi Y. Development of diffusion database of buffer materials and rocks-expansion and application method of foreign buffer materials, JAEA-Data/Code 2009–029[R]. Tokai: Japan Atomic Energy Agency (JAEA), 2010: 29.

[10] Feng Z Y, Tian J L, Wu T, et al. Unveiling the Re, Cr, and I diffusion in saturated compacted bentonite using machine-learning methods[J]. Nucl Sci Tech, 2024, 35(6): 93.

[11] Wu T, Tian J, Shi X, et al. Predicting anion diffusion in bentonite using hybrid machine learning model and correlation of physical quantities[J]. Sci Total Environ, 2024, 946: 174363.

[12] Feng J, Gao X, Xu K, et al. Predicting distribution coefficient and effective diffusion coefficient of radionuclides in bentonite: multi-output neural network simulation and diffusion experimental study[J]. J Hazard Mater, 2025, 490: 137787.

[13] Shi X, Zhang P, Feng J, et al. Improving hydraulic conductivity prediction of bentonite using machine learning with generative adversarial network-based data augmentation[J]. Constr Build Mater, 2025, 462: 139962.

[14] Tian J L, Feng J X, Shen J C, et al. Prediction of radionuclide diffusion enabled by missing data imputation and ensemble machine learning[J]. Nucl Sci Tech, 2025, 36(10): 181.

[15] Pamungkas N S, Putra Z P, Pratama H A, et al. Supervised machine learning-based categorization and prediction of uranium adsorption capacity on various process parameters[J]. J Hazard Mater Adv, 2025, 17: 100523.

[16] Shi X, Tian J, Shen J, et al. Application of machine learning in predicting the apparent diffusion coefficient of Se(Ⅳ) in compacted bentonite[J]. J Radioanal Nucl Chem, 2024, 333(11): 5811-5821.

[17] Gong S Y, Yang X, Zhang K M, et al. Estimation of effective diffusion coefficient of radionuclides in bentonite by machine learning method[J]. Ann Nucl Energy, 2025, 214: 111223.

[18] Tran N T, Tran V H, Nguyen N B, et al. On data augmentation for GAN training[J]. IEEE Trans Image Process, 2021, 30: 1882-1897.

[19] Glaus M A, Rossé R, van Loon L R, et al. Tracer diffusion in sintered stainless steel filters: measurement of effective diffusion coefficients and implications for diffusion studies with compacted clays[J]. Clays Clay Miner, 2008, 56(6): 677-685.

[20] Zhang J, Mucs D, Norinder U, et al. LightGBM: an effective and scalable algorithm for prediction of chemical toxicity-application to the Tox21 and mutagenicity data sets[J]. J Chem Inf Model, 2019, 59(10): 4150-4158.

[21] Mbah O M, Madueke C I, Umunakwe R, et al. Extreme gradient boosting: a machine learning technique for daily global solar radiation forecasting on tilted surfaces[J]. J Eng Sci, 2022, 9(2): E1-E6.

[22] 段忠义,肖昆,杨亚新,等.基于集成学习的松辽盆地砂岩型铀矿地层岩性自动识别研究[J].原子能科学技术,2023, 57(12):2443-2454.

[23] Li Z, Montomoli F. Aleatory uncertainty quantification based on multi-fidelity deep neural networks[J]. Reliab Eng Syst Saf, 2024, 245: 109975.

[24] Mohammadi Golafshani E, Kim T, Behnood A, et al. Sustainable mix design of recycled aggregate concrete using artificial intelligence[J]. J Clean Prod, 2024, 442: 140994.

[25] Feng Z, Gao Z, Wang Y, et al. Application of machine learning to study the effective diffusion coefficient of Re(Ⅶ) in compacted bentonite[J]. Appl Clay Sci, 2023, 243: 107076.

[26] Wang H, Wu T, Chen J, et al. Through-and out-diffusion of Se(Ⅳ) and Re(Ⅶ) in compacted bentonite[J]. Adv Mater Res, 2014, 953-954: 614-620.