Key Drivers of Metal Removal in Constructed Wetlands Treating Acid Mine Drainage Revealed by Machine Learning

Jingkang Zhang[1,2], Xingjie Wang[2,3], Liyuan Ma[1,2,3*], Jianwei Zhou[1,2]

[1]*School of Environmental Studies, China University of Geosciences, Wuhan 430074, China*
[2]*Institute of Geological Survey, China University of Geosciences, Wuhan 430074, Hubei, China*
[3]*School of Engineering, Cardiff University, Cardiff CF243AA, United Kingdom*

Abstract

Constructed wetlands (CWs) have been used for treating acid mine drainage (AMD), yet their metal removal mechanisms remain unclear. Herein, machine learning (ML) was employed to predict multi-metal removal efficiencies, with XGBoost achieving highest accuracy ($R^2 > 0.8$) for total Fe, Mn, Al, and Zn removal. Feature importance analysis identified operation days (1–185) and inflow chemical oxygen demand (COD, 6.5–1027.6 mg/L) as dominant predictors. Partial dependence plots revealed interactions between predictors. Inflow parameters contributed 57.6% to metal removal, surpassing time series and wetland properties. This study provides data-driven insights for optimizing CWs in AMD treatment.

Keywords: Acid mine drainage; Machine learning; Constructed wetland; Metal removal efficiency

Introduction

Acid mine drainage (AMD), generated from sulfide mineral oxidation, poses severe environmental risks due to high metal content and acidity (Younger *et al.*, 2002; Stumm and Morgan, 2013; Blowes *et al.*, 2005). Constructed wetlands (CWs) offer sustainable AMD treatment through metal precipitation and biological processes (Jouini *et al.*, 2020), yet their performance is affected by a variety of factors. Traditional statistical methods fail to capture these complex interactions inherent in such systems, while machine learning (ML) has shown promise in decoding multivariate systems (Palansooriya *et al.*, 2022). Various ML methods, such as Random Forest (RF), Extreme Gradient Boosting (XGBoost), k-Nearest Neighbors (kNN) and Neural Networks (NN), have been utilized to monitor and map contaminants in soil (Wu *et al.*, 2013) and groundwater (Lopez *et al.*, 2021). However, ML applications in CWs for AMD treatment remain limited. This study aims to bridge this gap by developing five ML models to predict multi-metal removal efficiencies (total Fe, Mn, Al, Zn). As the quality of the dataset brought into a model profoundly affects the validity of the model (Briscoe and Marin, 2020; Kim *et al.*, 2022), it was crucial to ensure the robustness of the initial datasets. Therefore, we devoted much effort to the construction of the dataset and feature engineering to obtain a practice-oriented dataset. This data-driven ML approach elucidated the complex interactions in constructed wetlands, providing a deeper understanding of how varying parameters affect the removal efficiency of metals in AMD treatment.

Materials and methods

Data from 31 published studies (from 2006 to 2023) were collected, focusing on CWs treating AMD. Key parameters included wetland properties (length, width, height, plant type), inflow/outflow parameters (pH, COD, metals concentration), and time series (operation days). Missing data were imputed using RF, Histogram Gradient Boosting Regression (HGBR) and Hot Deck imputation, with outliers removed via Kolmogorov-Smirnov tests.

The final dataset comprised 354 data points with 29 features and 7 target variables (total Fe, Mn, Al, Zn, Ni, Co, Cr removal efficiencies). To simplify the ML model and improve its performance, feature filtering was performed based on feature correlation and ML-based feature importance analysis (Palansooriya *et al.*, 2022). Hierarchical clustering grouped correlated features based on Pearson correlation coefficients (PCC). Further, feature importance analysis was conducted with ML-based model to determine the significance of each feature in predicting the target variable (Zhu *et al.*, 2019). By integrating results from feature importance and correlation analysis, the most important feature within a cluster was selected as input features.

RF (Zhao *et al.*, 2023), XGBoost, Support Vector Regression (SVR) (Palansooriya *et al.*, 2022), kNN (Yin *et al.*, 2024) and Artificial Neural Network (ANN), were selected for this study and built based on Python 3.9.7. StandardScaler in Scikit-Learn (version 1.4.1.post1) was used to standardize the input features. Following data standardization, 80% of data were randomly extracted from each input dataset and used for model training, while the remaining 20% were used for testing (Yin *et al.*, 2024; Zhang *et al.*, 2023). The method of grid search with cross-validation (5-fold) was employed during the initial training process to conduct hyperparameter tunning, aiming to enhance model performance and mitigate the risk of overfitting (Yan *et al.*, 2021; Bergstra and Bengio, 2012; Zhu *et al.*, 2023). The coefficient of determination (R^2) and root-mean-square error (RMSE) were utilized to compare the prediction accuracy and quantify the prediction performance (Hu *et al.*, 2022). Feature importance was assessed using SHapley Additive exPlanations (SHAP). Partial dependence plots (PDPs) were utilized to visualize the interaction effects between key predictors (e.g., inflow COD and operation days) on metal removal efficiency. Fig. 1 shows the framework and detailed steps for this study.

Results and discussion

Across the initial dataset, missing data were identified for 19 variables. Subsequent the complete dataset (Dataset A) is obtained by missing data filling based on RF, HGBR and Hot Deck imputation methods. Following the completion of dataset filling, an extensive feature analysis ensued, which included PCC, hierarchical clustering (Fig. 2a) and model-based feature importance analysis (Fig. 2b). **Inflow COD** was discovered to be the most important feature for predicting metal removal efficiencies. To refine the dataset and enhance model generalization while reducing computational complexity, a feature filtering process which involved integrating the outcomes of hierarchical clustering with feature importance assessments was conducted. Following this procedure, six representative features were selected from Dataset A to form Dataset B. Despite the apparent optimality of dataset B from an ML perspective, it is crucial to acknowledge that the selection of input features based solely on their correlation and importance may not consistently adhere to domain expertise and real-world necessities. Therefore, through the

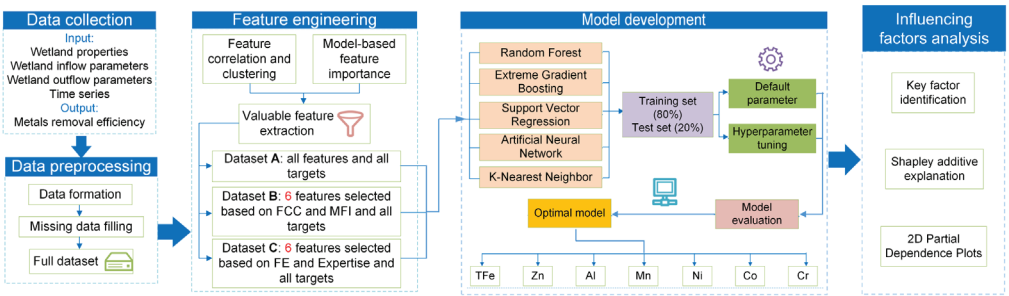

Figure 1 *The flowchart provides a detailed overview of the strategy employed for predicting the efficiencies of metals removal in AMD treated by constructed wetlands using a machine learning framework. Note: FCC: feature correlation and clustering; MFI: model-based feature importance; FE: feature engineering; TFe: total Fe.*

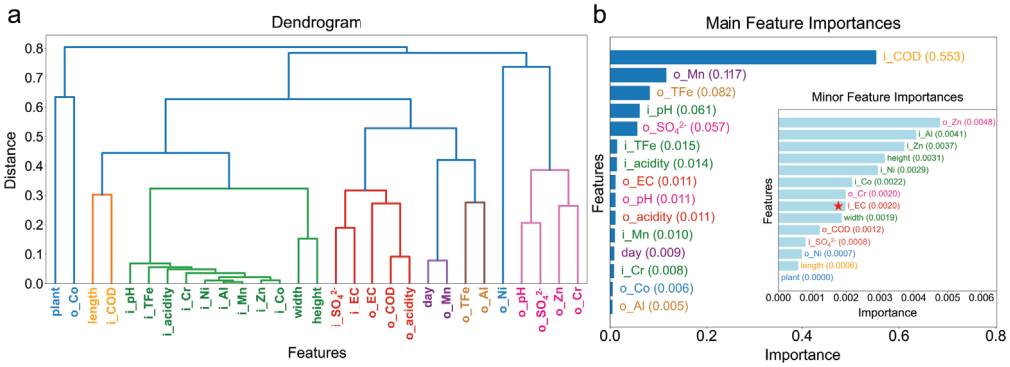

Figure 2 Input feature analysis: (a) hierarchical clustering and (b) feature importance from the XGBoost model. Note: the prefix "i_" represents inflow parameters, while the prefix "o_" represents outflow parameters.

integration of particle experiment conditions and feature analysis results, Dataset C was built. Three datasets were defined:

Dataset A: All features and targets (full dataset).

Dataset B: Six most important features (inflow COD, inflow pH, outflow EC, outflow Mn, outflow TFe, outflow SO_4^{2-}).

Dataset C: Six practical features (inflow COD, inflow pH, inflow acidity, inflow EC, wetland height, operation days), selected based on monitoring feasibility and domain expertise, alongside the assessment of feature importance and correlation.

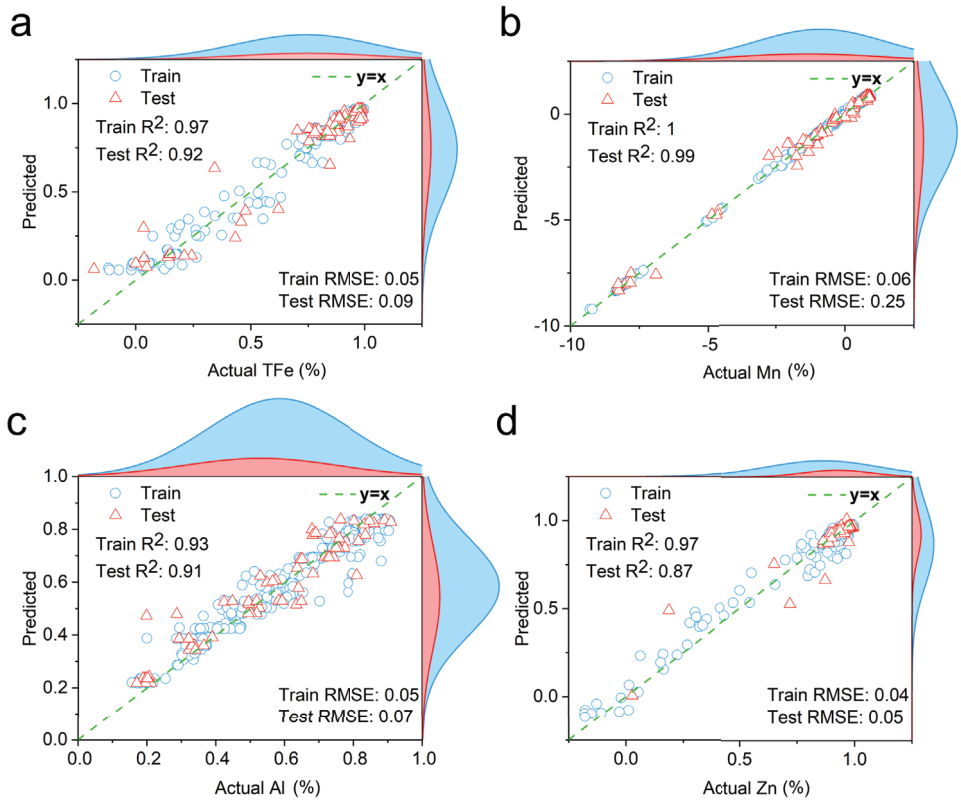

Figure 3 The predictive performance demonstration of the optimal model XGBoost for (a) TFe removal efficiency, (b) Mn removal efficiency, (c) Al removal efficiency, and (d) Zn removal efficiency based on dataset C, evaluated using R2 and RMSE as assessment metrics. RMSE = root-mean-square error.

In assessing the fundamental predictive performance of different models, Dataset A, encompassing all inputs from the full dataset was utilized. This strategy aimed to mitigate the potential reduction in predictive accuracy resulting from the exclusion of valuable features by Datasets B and C.

Five ML models (RF, XGBoost, SVR, kNN, ANN) were evaluated for predicting metal removal efficiencies (Fe, Mn, Al, Zn). XGBoost achieved the highest accuracy (R^2 > 0.8) across all datasets. Dataset A, with comprehensive features, yielded the best predictions, while datasets B and C showed slight declines due to feature selection. XGBoost demonstrated robust performance on the test set, particularly for TFe and Mn removal (Fig 3).

In accordance with previously published research, both Fe and Mn were identified as two key metals requiring particular attention in the treatment of AMD using constructed wetlands (Chen *et al.*, 2023; Singh and Chakraborty, 2020). During the process of Fe conversion to hydroxides, the transformation of aluminum often accompanies (Singh and Chakraborty, 2020). Additionally, research indicated that aluminum played a significant role in plant growth and can mitigate the toxicity of metals such as Fe, Mn and H^+ in acidic soils (Nguegang *et al.*, 2022). Furthermore, it was found that the concentration of Zn significantly exceeds the standard limits. Therefore, this study focuses on predicting the removal efficiency and analyzing the influential factors of TFe, Mn, Al and Zn. This aligns with the emphasis on key pollutant metals in relevant published studies (Singh and Chakraborty, 2020).

To quantitatively decipher the factors influencing the prediction of metal removal efficiencies, we employed the SHAP analysis on optimal model to reflect the importance of these factors (Fig. 4a, b). The comparison of feature importance rankings between the two analysis methods reveals discrepancies, but inflow COD and operation days were both considered to be the key factors. PDPs revealed nonlinear relationships between predictors and metal removal efficiency (Fig 5). For example, Fe removal efficiency peaked at low COD (<300 mg/L) and declined after 80 days of operation. Mn removal efficiency was highly sensitive to COD, with negative removal observed at high COD (>800 mg/L). These findings highlight the importance of optimizing COD levels and operational duration for effective metal removal.

Conclusions

In this study, we utilized ML to predict and analyze multi-metal removal efficiencies in constructed wetlands treating AMD. The main findings are summarized as follows:

- Five ML models were developed, with the **XGBoost** model emerging as the most effective, achieving high predictive accuracy (R^2 > 0.8) for the removal efficiency of total iron, manganese, aluminum and zinc.

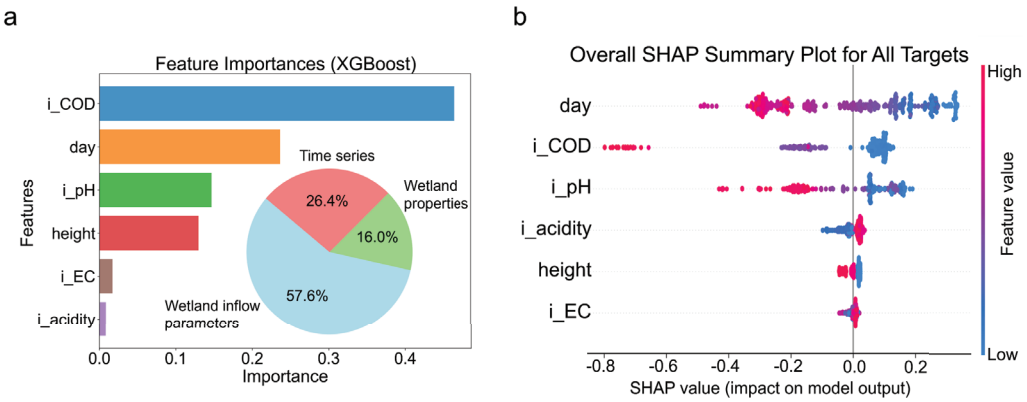

Figure 4 *Influential factors analysis based on dataset C and the optimal XGBoost model: (a) feature importance assessment based on XGBoost model and (b) Shapley additive explanation method.*

- Detailed feature analysis using the XGBoost model identified **operation days** (1–185) and **inflow COD** (6.523–1027.631 mg/L) as significant predictors of metal removal efficiency. These factors were found to have a substantial impact on the effectiveness of the wetland treatment process.
- The empirical categories for metal removal efficiency, ranked by importance, were wetland inflow parameters in first place, followed by time series, and wetland properties in last place. Inflow parameters were quantified to exert the highest influence on metal removal efficiency at 57.6%.
- Partial dependence plots elucidated the non-linear relationships between key predictors and metal removal efficiencies. This analysis revealed that specific ranges of operation days and COD levels are critical for optimizing the removal processes, providing actionable insights for the monitoring and management of constructed wetlands.

The findings offer a foundation for further research and practical applications aimed at enhancing the performance of constructed wetlands in treating acid mine drainage.

References

Bergstra J, Bengio Y (2012) Random Search for Hyper-Parameter Optimization. J Mach Learn Res 13:281–305

Blowes D, Ptacek C, Jambor J, Weisener CJEG (2005) The geochemistry of acid mine. 9:149

Briscoe J, Marin O (2020) Looking at neurodevelopment through a big data lens. Science 369 (6510). doi:10.1126/science.aaz8627

Chen D, Wang G, Chen C, Feng Z, Jiang Y, Yu H, Li M, Chao Y, Tang Y, Wang S, Qiu R (2023) The interplay between microalgae and toxic metal(loid)s: mechanisms and implications in AMD phycoremediation coupled with

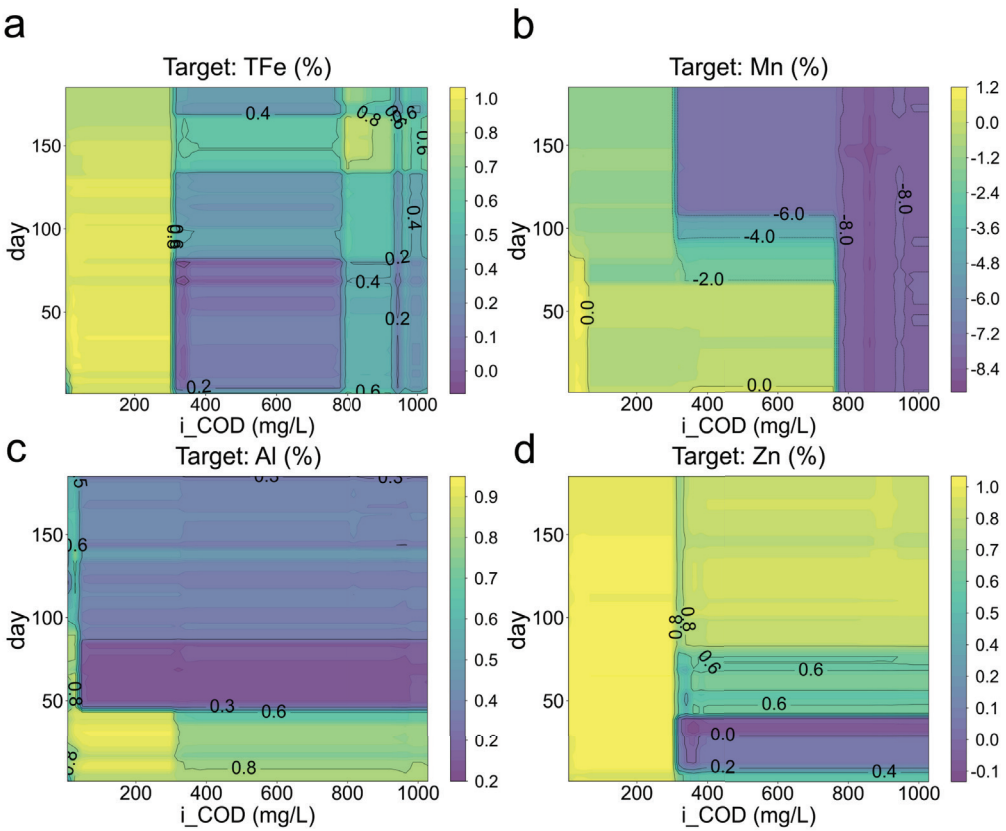

Figure 5 The interaction between inflow COD and operation days was analyzed to assess its impact on the (a) TFe removal efficiency, (b) Mn removal efficiency, (c) Al removal efficiency, and (d) Zn removal efficiency.

Fe/Mn mineralization. J Hazard Mater 454:131498. doi:10.1016/j.jhazmat.2023.131498

Hu S, Liu G, Zhang J, Yan J, Zhou H, Yan X (2022) Linking electron ionization mass spectra of organic chemicals to toxicity endpoints through machine learning and experimentation. J Hazard Mater 431:128558. doi:10.1016/j.jhazmat.2022.128558

Jouini M, Benzaazoua M, Neculita CM, Genty T (2020) Performances of stabilization/solidification process of acid mine drainage passive treatment residues: Assessment of the environmental and mechanical behaviors. J Environ Manage 269:110764. doi:10.1016/j.jenvman.2020.110764

Kim T, Shin J, Lee D, Kim Y, Na E, Park JH, Lim C, Cha Y (2022) Simultaneous feature engineering and interpretation: Forecasting harmful algal blooms using a deep learning approach. Water Res 215:118289. doi:10.1016/j.watres.2022.118289

Lopez AM, Wells A, Fendorf S (2021) Soil and Aquifer Properties Combine as Predictors of Groundwater Uranium Concentrations within the Central Valley, California. Environ Sci Technol 55 (1):352–361. doi:10.1021/acs.est.0c05591

Nguegang B, Masindi V, Msagati Makudali TA, Tekere M (2022) Effective treatment of acid mine drainage using a combination of MgO-nanoparticles and a series of constructed wetlands planted with Vetiveria zizanioides: A hybrid and stepwise approach. J Environ Manage 310:114751. doi:10.1016/j.jenvman.2022.114751

Palansooriya KN, Li J, Dissanayake PD, Suvarna M, Li L, Yuan X, Sarkar B, Tsang DCW, Rinklebe J, Wang X, Ok YS (2022) Prediction of Soil Heavy Metal Immobilization by Biochar Using Machine Learning. Environ Sci Technol 56 (7):4187–4198. doi:10.1021/acs.est.1c08302

Singh S, Chakraborty S (2020) Performance of organic substrate amended constructed wetland treating acid mine drainage (AMD) of North-Eastern India. J Hazard Mater 397:122719. doi:10.1016/j.jhazmat.2020.122719

Stumm W, Morgan JJ (2013) Aquatic chemistry: chemical equilibria and rates in natural waters. John Wiley & Sons,

Wu G, Kechavarzi C, Li X, Wu S, Pollard SJT, Sui H, Coulon F (2013) Machine learning models for predicting PAHs bioavailability in compost amended soils. Chem Eng J 223:747–754. doi:10.1016/j.cej.2013.02.122

Yan J, Yan X, Hu S, Zhu H, Yan B (2021) Comprehensive Interrogation on Acetylcholinesterase Inhibition by Ionic Liquids Using Machine Learning and Molecular Modeling. Environ Sci Technol 55 (21):14720–14731. doi:10.1021/acs.est.1c02960

Yin M, Zhang X, Li F, Yan X, Zhou X, Ran Q, Jiang K, Borch T, Fang L (2024) Multitask Deep Learning Enabling a Synergy for Cadmium and Methane Mitigation with Biochar Amendments in Paddy Soils. Environ Sci Technol 58 (3):1771–1782. doi:10.1021/acs.est.3c07568

Younger PL, Banwart SA, Hedin RS, Younger PL, Banwart SA, Hedin RS (2002) Mine water hydrology. Springer,

Zhang Y, Feng Y, Ren Z, Zuo R, Zhang T, Li Y, Wang Y, Liu Z, Sun Z, Han Y, Feng L, Aghbashlo M, Tabatabaei M, Pan J (2023) Tree-based machine learning model for visualizing complex relationships between biochar properties and anaerobic digestion. Bioresour Technol 374:128746. doi:10.1016/j.biortech.2023.128746

Zhao W, Ma J, Liu Q, Dou L, Qu Y, Shi H, Sun Y, Chen H, Tian Y, Wu F (2023) Accurate Prediction of Soil Heavy Metal Pollution Using an Improved Machine Learning Method: A Case Study in the Pearl River Delta, China. Environ Sci Technol 57 (46):17751–17761. doi:10.1021/acs.est.2c07561

Zhu JJ, Yang M, Ren ZJ (2023) Machine Learning in Environmental Research: Common Pitfalls and Best Practices. Environ Sci Technol 57 (46):17671–17689. doi:10.1021/acs.est.3c00026

Zhu X, Li Y, Wang X (2019) Machine learning prediction of biochar yield and carbon contents in biochar based on biomass characteristics and pyrolysis conditions. Bioresour Technol 288:121527. doi:10.1016/j.biortech.2019.121527

Abstracts and Extended Abstracts – Reviewed