Supporting materials

Text S1. Experimental investigation of the TMWs

In the initial study, the mineral composition of the TMWs samples was identified using powder X-ray diffraction (XRD) [1]. The diffraction patterns were obtained with a Scingtag X1 automatic powder diffractometer and analyzed using JADE software (Material Data Inc.) along with standard reference patterns. To estimate the relative amounts of phases in the Rietveld refinement, Siroquant software was used for full-profile XRD analysis[1], with an uncertainty of approximately ±5 wt% in the results.

As for the sequential extraction, 1 g of the sample was taken at a time, which was air-dried, sieved to less than 2 mm and extracted sequentially with different reagents (Figure S1) [1]. The extracts and residues were analyzed by inductively coupled plasma mass spectrometry (ICPMS) [2] and hydride generation atomic absorption spectrometry (HG-AAS) [3] to obtain the percentage of morphological fractions in different samples.

Text S2. Single machine learning algorithms

S2.1 Decision tree (DT)

DT is a tree structure based supervised learning technique that can solve classification problems as well as regression problems [4]. It progressively divides the dataset into subsets or nodes by selecting features and segmentation criteria with the aim of maximizing the purity of the target variable or minimizing the prediction error. When a termination node is reached or a termination condition is satisfied, the decision tree assigns a prediction value to each leaf node, which is usually the average of the values of the target variable of all training samples within that node [5]. Decision trees can handle regression prediction tasks in an intuitive and flexible way.

S2.2 Support vector regression (SVR)

SVR is a supervised regression model using the principles of support vector machines [6]. The core idea is to find the optimal hyperplane by maximizing the marginal distance between the data points and the hyperplane, allowing some of the data points to fall outside the margin. In addition, SVR utilizes nuclear techniques to implicitly perform nonlinear transformations in high-dimensional spaces to better fit complex data patterns [5].

S2.3 Ridge regression (Ridge)

Ridge regression is a technique used to analyze regression data for multiple covariances, which are usually found in models with a large number of parameters. When multicollinearity is present, although the least squares estimates are unbiased, their variance is large, resulting in a large gap between the predicted and true values. To address this problem, ridge regression controls the magnitude of the coefficients by applying a penalty to the coefficients and minimizing the sum of squares of the residuals after the penalty. By introducing a modest bias in the regression estimates, ridge regression is able to reduce the standard errors and provide more stable and reliable estimates [7].

S2.4 K-nearest neighbors (KNN)

KNN is one of the simplest machine learning algorithms [8]. Using supervised learning, KNN can be used to handle classification and regression problems on large sample data with short training time and low complexity [9]. The way it works is that when calculating the average, the nearest neighbor contributes more than the distant neighbor. If d is the distance between the node and its neighbor, then the weight of the neighbor is $1/d$ [10].

Text S3. Ensemble machine learning algorithms

S3.1 Random forest (RF)

RF uses random sampling of individual machine learning model predictions to construct multiple decision trees, which are then merged to generate ensemble predictions [11]. It constructs multiple trees by sampling a subset of the original dataset and randomly selecting a subset of features for node splitting [5]. A random subset of the training data and a random subset of the features are used to construct each tree. This randomization reduces overfitting and improves the generalizability of the model [12]. Instead of using only one predictor, the end result is to use the average of multiple predictors to improve accuracy, involving both classification and regression problems that can be solved with random forests [4].

S3.2 Extreme gradient boosting (XGBoost)

The extreme gradient boosting (XGB) algorithm combines bagging-boosting with feature stochasticity to effectively mitigate the overfitting problem [13]. The difference between XGB and traditional GBDT is the introduction of a new objective function which combines the model's loss function and regularization term to better control model complexity and improve generalization capabilities [5]. In this model, the decision tree divides the input data space into different regions by specific decision rules and achieves the integration effect by aggregating the outputs of multiple decision trees to build a model with complex features and interactions [14]. In addition, XGB employs an innovative sparse-aware learning algorithm to construct parallel trees, which improves the performance and accuracy of the model, especially when dealing with sparse datasets [15].

S3.3 Gradient boosting decision tree regression (GBDT)

GBDT uses a decision tree as the base learner and optimizes in the negative gradient direction of the loss function by a gradient boosting method to train the model incrementally [5]. The goal is to improve the efficiency of regression trees by integrating them with gradient boosting algorithms [16]. In the prediction phase, GBDT obtains the final output by aggregating the weighted votes of all trees. The contribution of each tree usually depends on its effect on the model, with better performing trees being assigned greater weights. GBDT is particularly well suited for solving high-dimensional, complex and non-linear regression problems [5].

S3.4 Extremely randomized trees (ET)

ExtraTrees is a machine learning integration method closely related to decision trees, similar to other integration techniques such as Bootstrap aggregation and random forests [15]. This tree-based approach, which is widely used in several domains, improves the accuracy of the results and the generalization of the model by introducing a high degree of randomness in the attributes and cut-points when splitting the tree nodes in order to ensure that each decision tree maintains structural differences [17].

S3.5 Categorical boosting (CATBoost)

The category boosting tree algorithm is a GBDT optimization algorithm proposed in 2018, which can effectively handle category-based features and improve model performance without extensive preprocessing [18]. As the traditional one-shot coding may lead to dimensionality explosion, the CatBoost algorithm improves the features with a high number of categories by using a statistical method based on greedy objective-based approach, which is able to reduce the impact of noise as well as low-frequency category data on the data distribution [9].

S3.6 Natural gradient boosting (NGBoost)

Natural gradient boosting is a supervised learning technique that aims to achieve probabilistic prediction through gradient boosting and natural gradient algorithms [19]. The natural gradient method takes into account the distributional properties of the parameter space and adapts to its geometric structure, thus improving the stability of the algorithm during training. NGBoost not only outputs the predicted values in a regression problem,

but also directly outputs the probability distributions of the different predicted values, which makes it ideally suited for parameter optimization problems in engineering applications [20].

Table S1 Tuned hyper-parameters and their tuning ranges for ten ML algorithms

ML model	Hyper-parameter	Range
RF	n_estimators	[200, 300, 400, 500, 1000, 2000]
	max_depth	[None, 5, 10, 15, 20]
	min_samples_split	[2, 3, 5, 7]
	min_samples_leaf	[2, 8, 15, 20]
XGBoost	n_estimators	[200, 300, 400, 500, 1000, 2000]
	max_depth	[None, 5, 10, 15, 20]
	learning_rate	[0.01,0.1]
GBDT	n_estimators	[200, 300, 400, 500, 1000, 2000]
	max_depth	[None, 5, 10, 15, 20]
	learning_rate min_samples_leaf	[0.01,0.1] [2, 8, 15, 20]
ET	n_estimators	[200, 300, 400, 500, 1000, 2000]
	max_depth	[None, 5, 10, 15, 20]
	min_samples_split	[2, 3, 5, 7]
	min_samples_leaf	[2, 8, 15, 20]
CATBoost	Iterations	[200, 500, 1000, 2000]
	depth	[3, 6, 8, 10]
	learning_rate	[0.01,0.05,0.1,0.2]
NGBoost	n_estimators	[200, 300, 400, 500, 1000, 2000]
	learning_rate	[0.01,0.1]
	natural_gradient	/
DT	max_depth	[None, 5, 10, 15, 20]
	min_samples_split	[2, 3, 5, 7]
	min_samples_leaf	[2, 8, 15, 20]
SVR	C	[1,10,100]
	gamma	[0.02,0.1]
	kernel	/
Ridge	alpha	[1,2,4,6,8]
	tol	[0.00001,0.001,0.01,0.1,1]
	solver	/
KNN	leaf_size	[50,100,150,200,250]
	n_neighbors	[2,4,6,8,10]
	algorithm	/

Table S2 Optimal hyper-parameters and their R^2 values on the test set

Model	Optimal hyper-parameter	R^2
RF	n_estimators=200; max_depth=none; min_samples_split=2; min_samples_leaf=2	0.9421
XGBoost	n_estimators=200; max_depth=10; learning_rate=0.1	0.9402
GBDT	n_estimators=500; max_depth=20; learning_rate=0.01; min_samples_leaf=8	0.9433
ET	n_estimators=300; max_depth=20 min_samples_split=2; min_samples_leaf=2	0.9423
CATBoost	iterations=2000; depth=8; learning_rate=0.05	0.9389
NGBoost	n_estimators=300; learning_rate=0.01; natural_gradient=TRUE	0.9070
DT	max_depth=15; min_samples_split=7; min_samples_leaf=2;	0.9195
SVR	*C*=100; gamma=0.1; kernel=poly; algorithm=auto	0.4576
Ridge	Alpha=8; tol=0.00001; solver=auto	0.4092
KNN	leaf_size=50; n_neighbors=10;	0.3089

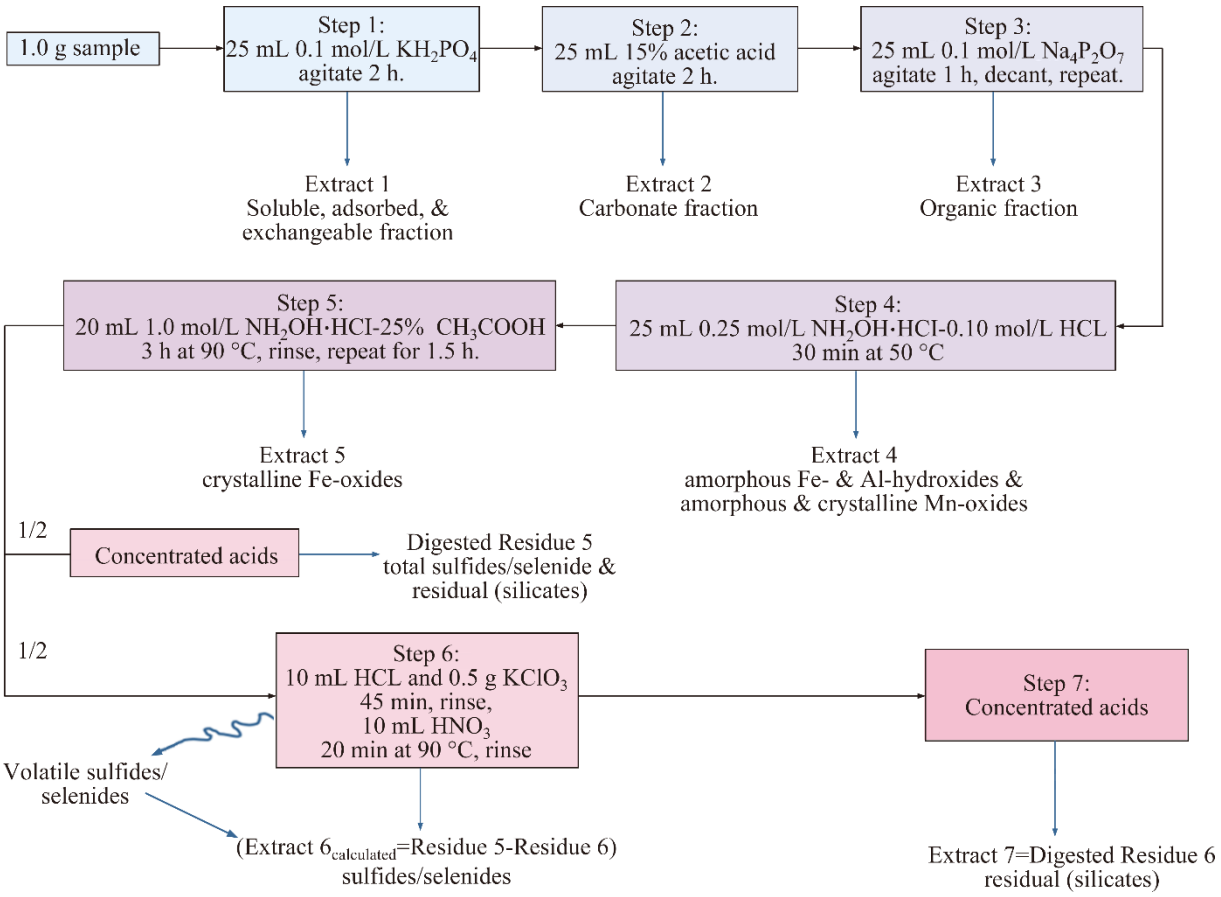

Figure S1 Schematic of sequential extraction procedure

References

[1] PIATAK N M, SEAL R R II, HAMMARSTROM J M. Sequential extraction results and mineralogy of mine waste and stream sediments associated with metal mines in Vermont, Maine, and New Zealand[R]. Reston, Virginia, USA: U.S. Geological Survey, 2007.

[2] LAMOTHE P J, MEIER A L, WILSON S A. The determination of forty-four elements in aqueous samples by inductively coupled plasma-mass spectrometry[R]. Denver, Colorado, USA: U.S. Geological Survey, 1999.

[3] ZHANG Yan-lin, ADELOJU S B. Flow injection–hydride generation atomic absorption spectrometric determination of selenium, arsenic and bismuth [J]. Talanta, 2008, 76(4): 724−730. DOI: 10.1016/j.talanta.2008.03.056.

[4] NALAYINI C, KATIRAVAN J, GEETHA S, et al. A novel dual optimized IDS to detect DDoS attack in SDN using hyper tuned RFE and deep grid network [J]. Cyber Security and Applications, 2024, 2: 100042. DOI: 10.1016/j.csa.2024.100042.
[5] GUO Li-sheng, XU Xin, NIU Cen-cen, et al. Machine learning-based prediction and experimental validation of heavy metal adsorption capacity of bentonite [J]. The Science of the Total Environment, 2024, 926: 171986. DOI: 10.1016/j.scitotenv.2024.171986.
[6] SAHU N, AZAD C, KUMAR U. Study and prediction of photocurrent density with external validation using machine learning models [J]. International Journal of Hydrogen Energy, 2024, 92: 1335−1355. DOI: 10.1016/j.ijhydene.2024.10.339.
[7] YANG Hong-rui, HUANG Kuan, ZHANG Kai, et al. Predicting heavy metal adsorption on soil with machine learning and mapping global distribution of soil adsorption capacities [J]. Environmental Science & Technology, 2021, 55(20): 14316−14328. DOI: 10.1021/acs.est.1c02479.
[8] CHENG S Y, PARK S, TRIVEDI M M. Multi-spectral and multi-perspective video arrays for driver body tracking and activity analysis [J]. Computer Vision and Image Understanding, 2007, 106(2, 3): 245−257. DOI: 10.1016/j.cviu.2006.08.010.
[9] HUANG Jing, PENG Yang, HU Lin. A multilayer stacking method base on RFE-SHAP feature selection strategy for recognition of driver's mental load and emotional state [J]. Expert Systems with Applications, 2024, 238: 121729. DOI: 10.1016/j.eswa.2023.121729.
[10] ALFADDA A, RAHMAN S, PIPATTANASOMPORN M. Solar irradiance forecast using aerosols measurements: A data driven approach [J]. Solar Energy, 2018, 170: 924−939. DOI: 10.1016/j.solener.2018.05.089.
[11] DORADO-GUERRA D Y, CORZO-PÉREZ G, PAREDES-ARQUIOLA J, et al. Machine learning models to predict nitrate concentration in a river basin [J]. Environmental Research Communications, 2022, 4(12): 125012. DOI: 10.1088/2515-7620/acabb7.
[12] VERMA D, BHATT U M, VIDYARTHI A. A machine learning framework for predictive electron density modelling to enhance 3D NAND flash memory performance [J]. e-Prime-Advances in Electrical Engineering, Electronics and Energy, 2024, 10: 100790. DOI: 10.1016/j.prime.2024.100790.
[13] HABIBI A, DELAVAR M R, SADEGHIAN M S, et al. A hybrid of ensemble machine learning models with RFE and Boruta wrapper-based algorithms for flash flood susceptibility assessment [J]. International Journal of Applied Earth Observation and Geoinformation, 2023, 122: 103401. DOI: 10.1016/j.jag.2023.103401.
[14] PIRAEI R, AFZALI S H, NIAZKAR M. Assessment of XGBoost to estimate total sediment loads in rivers [J]. Water Resources Management, 2023, 37(13): 5289−5306. DOI: 10.1007/s11269-023-03606-w.
[15] BARKHORDARI M S, ZHOU Na-na, LI Ke-chao, et al. Interpretable machine learning for predicting heavy metal removal efficiency in electrokinetic soil remediation [J]. Journal of Environmental Chemical Engineering, 2024, 12(6): 114330. DOI: 10.1016/j.jece.2024.114330.
[16] ARUNKUMAR P M, BALAJI N, MADHANKUMAR S, et al. Prediction of red chilli drying performance in solar dryer with natural energy storage element using machine learning models [J]. Journal of Energy Storage, 2024, 101: 113825. DOI: 10.1016/j.est.2024.113825.
[17] WANG Zi-jian, GUO Lin, GONG Hui-li, et al. Land subsidence simulation based on extremely randomized trees combined with Monte Carlo algorithm [J]. Computers & Geosciences, 2023, 178: 105415. DOI: 10.1016/j.cageo.2023.105415.
[18] LUO Mi, WANG Yi-fu, XIE Yun-hong, et al. Combination of feature selection and CatBoost for prediction: The first application to the estimation of aboveground biomass [J]. Forests, 2021, 12(2): 216. DOI: 10.3390/f12020216.
[19] DEGTYAREV V V, HICKS S J, FERREIRA F P V, et al. Probabilistic resistance predictions of laterally restrained cellular steel beams by natural gradient boosting [J]. Thin-Walled Structures, 2024, 205: 112367. DOI: 10.1016/j.tws.2024.112367.
[20] FU Jian-qin, LI Hao, SUN Xi-lei, et al. Many-objective optimization for overall performance of an electric sport utility vehicle under multiple temperature conditions based on natural gradient boosting model [J]. Energy, 2024, 304: 132078. DOI: 10.1016/j.energy.2024.132078.