Chemometrics and Intelligent Laboratory Systems 264 (2025) 105444

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Multifactorial analysis of fluorescence detection for soil total petroleum hydrocarbons using random forest and multiple linear regression

Gaoyong Shi [a,b,c], Ruifang Yang [b,c,*], Nanjing Zhao [a,b,c,**], Gaofang Yin [b,c], Wenqing Liu [a,b,c]

[a] College of Environmental Science and Optoelectronic Technology, University of Science and Technology of China, Hefei, 230026, China
[b] Key Laboratory of Environmental Optics and Technology, Anhui Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Hefei, 230031, China
[c] Key Laboratory of Optical Monitoring Technology for Environment of Anhui Province, Hefei, 230031, China

ARTICLE INFO

Keywords:
Soil
Total petroleum hydrocarbons
Random forest
Multiple linear regression
Influencing factors

ABSTRACT

This study combined random forest (RF) and multiple linear regression (MLR) approaches to analyze the influence of various factors on the fluorescence detection of total petroleum hydrocarbons (TPH) in soil. We considered the effects of soil moisture, organic matter, and minerals, and tested samples of three common soil types and varying concentrations of soil petroleum hydrocarbons using a self-developed fluorescence imaging technology. The fluorescence signals are greatly influenced by moisture, organic matter, and minerals, exhibiting distinct effects depending on the soil types and hydrocarbon concentrations. The RF model improves accuracy and consistency by constructing decision trees, making it appropriate for non-linear and high-dimensional data scenarios, although its underperformance in our study. The MLR model provides a comprehensive understanding of the linear relationships between variables, displaying better statistical performance and consistency in most cases of our experiment, with a coefficient of determination (R^2) above 0.8, and Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) all lower than those of the RF. Our research provides an important scientific basis for monitoring, evaluating, and managing soil petroleum hydrocarbon pollution, aiding in the formulation of effective soil pollution prevention strategies, and offers a foundation for further research into environmental risk assessment and soil remediation.

1. Introduction

The contamination of soil by petroleum hydrocarbons is a major concern in the realm of environmental conservation. and the associated hazards to ecosystems and human well-being cannot be disregarded [1]. Petroleum hydrocarbons are an intricate blend consisting of diverse hydrocarbons, such as alkanes, alkenes, and aromatics [2]. It is imperative to have precise and swift identification of the entire petroleum hydrocarbons present in soil in order to evaluate environmental hazards, supervise soil contamination, and execute efficient remediation strategies. Hence, accurate and rapid detection of the total amount of petroleum hydrocarbons in soil is crucial for assessing environmental risks, monitoring soil pollution, and implementing effective remediation strategies.

Various intricate environmental factors influence the detection of TPH in soil. Among these, moisture is a crucial component that affects the detection of soil petroleum hydrocarbons [3]. In soils with elevated moisture levels, the presence of water can cause petroleum hydrocarbons to evaporate, dissolve, or undergo degradation by microorganisms [4]. Changes in soil moisture can also modify the soil's pore structure and the diffusion rate of petroleum hydrocarbons within the soil, therefore leading to fluctuations in the total amount of hydrocarbons. Moreover, organic matter has a crucial role in the adsorption, dispersion, and migration [5,6] of petroleum hydrocarbons in soil. Humic and fulvic acids, major components of organic matter, can adsorb petroleum hydrocarbon molecules, thereby reducing their bioavailability. Minerals such as kaolinite, siliceous sandstone, phosphates, and sodium silicate can react with petroleum hydrocarbons to form insoluble precipitates,

* Corresponding author. Key Laboratory of Environmental Optics and Technology, Anhui Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Hefei, 230031, China.
** Corresponding author. Key Laboratory of Environmental Optics and Technology, Anhui Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Hefei, 230031, China.
E-mail addresses: rfyang@aiofm.ac.cn (R. Yang), njzhao@aiofm.ac.cn (N. Zhao).

https://doi.org/10.1016/j.chemolab.2025.105444
Received 7 August 2024; Received in revised form 12 May 2025; Accepted 22 May 2025
Available online 30 May 2025

reducing the activity of the hydrocarbons and thereby interfering with the detection results [7,8]. These environmental factors exhibit different effects under various soil types and petroleum hydrocarbon concentrations.

To analyze the impact of these factors on the fluorescence detection of TPH in soil and to establish accurate predictive models, our study employs methods based on RF [9,10] and MLR [11,12] for multi-factor analysis. Random forest, by constructing multiple decision trees and combining their results, can significantly enhance the model's accuracy and stability [13,14]. It excels in handling high-dimensional data, assessing feature importance, parallel processing, and managing missing values. MLR is based on the intuitive understanding of the linear relationship between a dependent variable and multiple independent variables, capable of handling non-linear and interrelated data, and offers good predictive performance and interpretability [15,16].

MLR models have been widely used in climate change research. Researchers used the MLR model to analyze the linear relationship between climate factors such as temperature, humidity, and precipitation, and successfully predicted the impact on agricultural product yields, thereby providing data support for agricultural management. Such research is of great significance to improving agricultural production efficiency and formulating agricultural policies [17]. The MLR model has also been used to predict the concentration distribution of polycyclic aromatic hydrocarbons in soil [18]. The MLR model still performs well in simplifying complex relationships and providing easy-to-interpret predictions, especially when dealing with simple environmental data. Similarly, RF models are widely used in the detection and prediction of environmental pollutants. For example, researchers used the RF model to successfully predict phenanthrene and fluoranthene in water bodies, especially when dealing with complex environmental data, RF can capture the nonlinear relationship between variables. A fast and accurate polycyclic aromatic hydrocarbon quantification method was established for real-time detection of water environmental pollution [19].

Francisco M. Canero et al. [20] compared the performance of an RF model created using feature selection approaches with that of Partial Least Squares Regression (PLSR) while measuring soil reflectance. The performance ratios for the best RF model with feature selection to the interquartile range were 1.93, 0.38, and 2.56, respectively. Meanwhile, the prediction accuracy of the PLSR model for organic matter, carbonates, and clay were 1.41, 0.29, and 1.81, respectively. R.K. Douglas et al. [21] conducted a study where they analyzed the visible-near infrared spectra of 85 soil samples that were wet and had not been treated, and were polluted with petroleum. The RF calibration model outperformed the PLSR model in predicting alkanes and polycyclic aromatic hydrocarbons, with an R^2 value of 0.71 and an RMSEP value of 0.99 mg/kg. In comparison, the PLSR model had an R^2 value of 0.36 and an RMSEP value of 66.66 mg/kg. Rehman et al. [22] utilized neural networks and MLR to forecast the erodibility of soil in the Malaysian Peninsula. Using correlation and principal component analysis, the coefficient of determination for MLR was found to be 0.446, with a mean squared error of 0.0000306. Agbaogun et al. [23] employed MLR to study the adsorption of lead, cadmium, and copper on five naturally occurring soils in Nigeria. By considering the maximum adsorption capacity of the metals, conventional soil characteristics, soil pH, initial metal concentration, and temperature, the researchers created a total of 255 distinct multiple linear regression (MLR) models. The optimal regression model had a mean absolute error of 0.158, a root mean squared error of 0.199, and a coefficient of determination of 0.66. Jingran Li and colleagues [24] employed a MLR model to analyze the variables that influence carbon emissions in China. The empirical research analyzed time series data of five indicators: social financing scale, fossil energy consumption ratio, urbanization level, energy processing and conversion efficiency, and per capita carbon emissions. The findings revealed a positive correlation between the scale of social financing and the proportion of fossil energy consumption with per capita CO_2 emissions. Conversely, urbanization level and energy processing and conversion efficiency showed a negative correlation with per capita carbon emissions. Hong Li and colleagues [25] effectively evaluated the levels of soil organic matter and total nitrogen using a combination of mid-infrared reflectance spectroscopy and multivariate regression analysis. The results showed that the MLR model combined with stability competitive adaptive reweighted sampling method provided the most accurate estimates for organic matter (predictive coefficient of determination of 0.72 and predictive residual of 1.89) and total nitrogen (predictive coefficient of determination of 0.84 and predictive residual of 2.50). Raed Jafar et al. [26] utilized water quality data collected between 2021 and 2022 from the drinking water lake in Latakia, Syria. Their objective was to assess the water quality index by examining the effectiveness of MLR and 19 other machine learning models. The results indicated that linear regression, least angle regression, and Bayesian ridge chain had strong performance in predicting the water quality index, achieving a coefficient of determination of 0.999 and a root mean squared error of 0.149. Huanzhi Wang et al. [27] explored the application of the RF model in predicting the presence of six heavy metals (lead, cadmium, chromium, arsenic, mercury, zinc) in soil by comparing it with a land use regression model. The RF model had a coefficient of determination of approximately 0.90, surpassing the land use regression model in terms of cross-validation R^2 and demonstrating a smaller root mean squared error.

The objective of our study is to examine how moisture, organic matter, and minerals in various soil types, as well as varied amounts of petroleum hydrocarbons, affect the ability to detect TPH in soil using fluorescence. It provides a scientific basis for the monitoring, assessment, and management of soil petroleum hydrocarbon pollution and offers more precise methodologies and recommendations for the traceability of petroleum hydrocarbons, environmental risk assessment, and remediation. Implementing this will aid in developing more efficient soil pollution prevention tactics, reducing the consequences of petroleum hydrocarbon pollution on the environment and human well-being. This study not only uncovers the complex mechanism behind the detection of soil petroleum hydrocarbon fluorescence, but also offers fresh insights and approaches for further research and practical applications in related domains.

This paper is broken into four pieces. The second portion will outline the procedure for preparing the experimental samples and the fundamentals of RF, MLR approaches, and fluorescence detection techniques. The third section will analyze and explain the data, and the final section will present a concise conclusion.

2. Material and methods

2.1. Material

2.1.1. Parameter overview

For the experiment, three specific soil types were chosen: red soil from Shaoguan, Guangdong (NSA6, Table S1), chestnut calcareous soil from Qinghai (GBW07497, Table S2), and red soil from Yingtan, Jiangxi (GBW07416b, Table S3). Guangdong Shaoguan red soil is a typical soil in the humid areas of southern China. Its acidic pH and high organic matter content enable it to simulate the petroleum hydrocarbon pollution of agricultural soils in the south. Jiangxi Yingtan red soil is a representative soil in central China. Its acidity and mineral content make it have high adsorption capacity. Qinghai chestnut soil is a soil in arid areas with low organic matter content and high mineral fixation capacity. This soil can simulate the petroleum hydrocarbon pollution in plateaus and arid areas, where the soil has poor water retention capacity and the migration and degradation of pollutants are limited.

Each soil type was tested under two different petroleum hydrocarbon concentrations (10 g/kg and 20 g/kg). The study focused on three key influencing factors: humidity, organic matter (humic acids, fulvic acids), and minerals (kaolin, siliceous sandstone, clay, sodium dihydrogen phosphate, and sodium silicate). By systematically altering these

parameters, the study aimed to understand their effects on the fluorescence signal in the context of soil contamination. The interactions between soil type, petroleum hydrocarbon concentration, and these influencing factors were studied elucidate the potential mechanisms affecting the fluorescence signal.

The experiment utilized a crude oil fractional specimen (J42002) as the petroleum hydrocarbon. Initially, a certain amount of standard soil powder was weighed and sifted through a 100-mesh sieve, then a predetermined amount of crude oil was added. Afterwards, the crude oil was thoroughly mixed with the standard soil, placed in a mixer to ensure even mixing, and ultimately, soil samples with petroleum hydrocarbon concentrations of 10 g/kg and 20 g/kg were prepared. During the sample preparation stage, the soil, minerals and organic matter we selected are all national standard substances. The selection of these substances helps to simulate the matrix effects that may occur in actual samples and provide an accurate reference benchmark for the fluorescence signal.

2.1.2. Humidity

The moisture content in the soil has a direct influence on the ability of plant roots to absorb water and on the circumstances necessary for their growth. Optimal moisture levels enhance the absorption of nutrients and water by plants, as well as support the growth and metabolic functions of microorganisms, which in turn promote the breakdown of organic matter and the cycling of nutrients [28]. Soil moisture has a substantial impact on the physical characteristics of soil, including its density, porosity, and aeration. Elevated humidity can impede the evaporation of petroleum hydrocarbons in the soil and modify the adsorption properties of the soil surface, which can impact the strength of the bond between petroleum hydrocarbons and soil particles. In addition, optimal moisture levels promote microbial activity, which speeds up the breakdown of petroleum hydrocarbons [29].

The experiment involved collecting soil samples with variable percentages of petroleum hydrocarbons. Different volumes of deionized water were then added to these samples to create moisture levels of 0 %, 10 %, 20 %, 25 %, 30 %, 35 %, and 40 %, as depicted in Fig. 1. Once the samples were prepared, they were wrapped with a sealing film and stored in a refrigerator to ensure that enough water evaporated before testing. The equation for determining soil moisture is as follows:

$$W = \frac{M_1 - M_0}{M_0} \tag{1}$$

where M_1 indicates the weight of the soil after it has been dried, M_0 represents the weight of the soil when it contains water, and W represents the ratio of the water content in the soil to the dry weight of the soil, which is a measure of the soil moisture.

2.1.3. Organic matter

Organic matter in soil, which consists of plant and animal leftovers, microbial metabolic products, and their transformed forms, is a crucial element of soil that significantly influences its physical, chemical, and biological characteristics. Organic matter contributes to the development of a solid granular structure in the soil, which enhances its ability to allow air to circulate and retain water, while also boosting its resistance to erosion. During the process of decomposition, organic matter releases vital nutrients, including nitrogen, phosphorous, and potassium, which are necessary for the growth of plants. This continuous release of nutrients ensures a constant supply for plant growth [30]. Additionally, organic matter can regulate soil pH, improve the soil environment, and interact with soil minerals to form humus, increasing the soil's cation exchange capacity and thereby enhancing its fertility retention. Organic matter in soil has the ability to adsorb petroleum hydrocarbon molecules present in samples, which can change how easily they move and how accessible they are to living organisms in the soil. By stimulating soil microbial activity, the addition of organic matter can accelerate the degradation of petroleum hydrocarbons [31]. Furthermore, organic matter has the potential to undergo chemical reactions with petroleum hydrocarbons, resulting in changes to their chemical characteristics and environmental dynamics.

To explore the impact of soil organic matter on the fluorescence signals of soil petroleum hydrocarbons, two common types of organic matter were chosen for analysis: humic acid and fulvic acid. Within the laboratory setting, different quantities of humic acid and fulvic acid solids were measured and pulverized into a fine powder utilizing a 100-mesh sieve. The soil petroleum hydrocarbon samples were combined with varying quantities of organic matter to form samples with mass fractions of 0 %, 0.5 %, 1 %, 2 %, 3 %, 5 %, and 10 %.

We first used the XK96-4E micro-oscillator to oscillate for 15 min, and then used a high-throughput soil grinding instrument with a running frequency of 40 Hz, each time running for 5 min, interrupted for 1 min, and cycled 10 times. And then we obtained 5 parallel samples for a certain concentration of samples, and the relative standard deviation of the measured signal values was less than 5 %, or even smaller.

2.1.4. Minerals

Soil minerals, which are the non-organic constituents of soil, mainly comprise a variety of mineral particles and mineral compounds. They are not only fundamental to the structure and texture of soil but also a

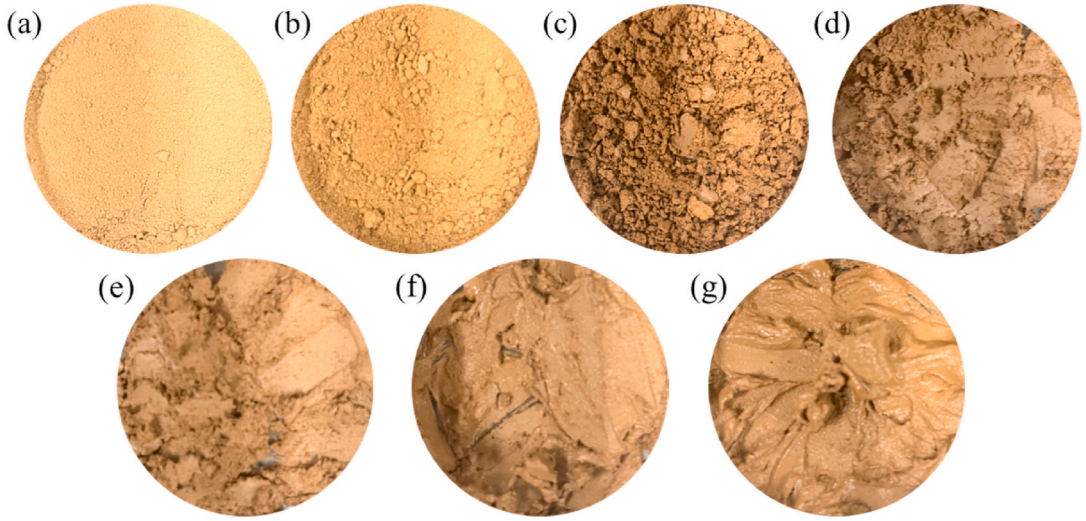

Fig. 1. Soil petroleum hydrocarbon samples at different humidity. (a)0 %; (b)10 %; (c)20 %; (d)25 %; (e)30 %; (f)35 %; (g)40 %.

vital source of essential elements for plant growth. Common minerals include quartz, feldspar, mica, montmorillonite, kaolinite, as well as iron and aluminum oxides. The weathering and disintegration of these minerals release components important for plant growth, such as potassium, calcium, magnesium, iron, and phosphorus.

Their existence can also impact soil characteristics like as pH, cation exchange capacity, and buffering capacity, therefore altering the accessibility of nutrients to plants. The presence of clay minerals and iron/aluminum oxides in soil petroleum hydrocarbon samples can impact the movement and breakdown rate of hydrocarbons in the soil [32]. Clay minerals have a large specific surface area and strong adsorption capacity, especially for the physical adsorption of petroleum hydrocarbons. Their specific surface area and microporous structure provide abundant adsorption sites for petroleum hydrocarbons, and petroleum hydrocarbon molecules bind to the surface of clay minerals through weak interactions such as van der Waals forces and hydrogen bonds. Many petroleum hydrocarbons are highly hydrophobic and easily interact with the hydrophobic parts of clay minerals, which enhances adsorption and slows down the migration of petroleum hydrocarbons in soil. In addition, clay minerals can retain petroleum hydrocarbons for a long time, especially in the soil surface, making the degradation of petroleum hydrocarbons more difficult. In the retention state, it is difficult for degrading microorganisms to access pollutants, which slows down the degradation process. Iron/aluminum oxides interact with petroleum hydrocarbons through surface chemical reactions. The oxide surface is positively charged and can electrostatically interact with the negatively charged groups (such as carboxyl and hydroxyl) in petroleum hydrocarbons, thereby enhancing the adsorption of petroleum hydrocarbons. In soils where iron oxides are present, petroleum hydrocarbons may participate in redox reactions, especially the oxidation of aromatic hydrocarbons, changing their chemical structure and making them easier or more difficult to degrade. Specific minerals can function as catalysts, facilitating the oxidation and biodegradation of petroleum hydrocarbons. The presence of minerals can also impact the composition and behavior of soil microorganisms, thereby indirectly impacting the breakdown of petroleum hydrocarbons.

To explore the influence of soil minerals on the fluorescence signals of soil petroleum hydrocarbons, a total of five minerals were chosen for analysis: kaolinite, siliceous sandstone, clay, sodium dihydrogen phosphate, and sodium silicate. Table 1 presents the typical values and uncertainty for siliceous sandstone, while Table 2 shows the typical values and standard deviations for clay. Within the laboratory, different quantities of these five minerals were measured and pulverized into a fine powder utilizing a 100-mesh sieve. They were mixed with varying quantities of soil petroleum hydrocarbon samples to generate mineral-soil petroleum hydrocarbon samples with mass fractions of 0 %, 10 %, 20 %, 30 %, 40 %, and 50 %.

2.2. Methods

2.2.1. Random forest

Random forest is a typical classifier comprised of numerous decision trees, which enhances the model's accuracy and stability by mixing the predicted outcomes of multiple decision trees. Random selection of samples and features during the development of a RF guarantees the diversity of the base classifiers, hence enhancing the RF's classification performance [33,34]. The RF technique is built upon the Bagging method [35], which is a type of parallel ensemble learning. When using RFs, the addition of randomness in column variables on top of row variable randomness helps reduce the occurrence of overfitting in RF regression predictions. The main principle of the Bagging approach is to use the Bootstrap resampling method [36] to extract several samples from the original sample, construct a decision tree for each sample, and then integrate the predictions of multiple decision trees to reach the final prediction result by voting. Fig. 2 illustrates the process of generating a RF.

During the construction of a RF, numerous training sets are generated by sampling with replacement from the original dataset. Each training set is then used to train a tree. As each sampling is done randomly, the data that is not chosen in each training set is considered as the Out of Bag (OOB) data for that specific tree. These data have not been included in the training process of the current tree, so they can be utilized to assess the model's capacity to generalize. During the training of individual decision trees, the common approach for determining the optimal attribute split is as follows. When working with discrete qualities, the formula for determining information gain [37] is as follows:

$$Inf\ o_{Gain} = Entropy - \sum_{i \in I} p_i \cdot Entropy_i \tag{2}$$

where *Entropy* represents the entropy of the parent node, and $Entropy_i$ represents the entropy of node i. The greater the entropy, the more information the node contains, and the less pure it is; p_i represents the ratio of the data volume of child node i to the data volume of the parent node. The greater the *Info_Gain*, the smaller the entropy after the split, making the child nodes purer and the classification more effective. Therefore, the attribute with the highest *Info_Gain* is selected as the split attribute.

2.2.2. Multiple linear regression

Multiple linear regression is a statistical technique used to model the relationship between a dependent variable and numerous independent variables. It quantifies the extent of impact that several independent variables have on the dependent variable. The MLR model is obtained through a sequence of calculations that yield a regression equation [38]. Unlike the basic linear regression, which involves just one set of data for the independent variable, MLR deals with the relationship between the dependent variable and several explanatory variables within a comprehensive regression plane [39,40]. Fig. 3 depicts the disparities between simple regression and multiple regression.

Once an MLR model has been established, it is essential to analyze the statistical significance of the linear effects of the independent variables on the dependent variable. This analysis aids in identifying independent variables that have a substantial impact on the dependent variable and assess the importance of each independent variable's impact on the dependent variable. Thus, the general form of a MLR model can be expressed as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 \beta_2 + \ldots + \beta_p x_p + \varepsilon \tag{3}$$

Where y represents the dependent variable, $x_1, x_2, \ldots, x_p$ are the independent variables, $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the regression coefficients to be estimated, and ε denotes the random error term, which has a mean of zero.

If there are n sets of observational data $\left(x_{i1}, x_{i2}, \ldots, x_{ip}; y_i\right)$ where $i = 1, 2, \ldots, n,$ then the general form of the linear model can be expressed as:

Table 1
Standard values and uncertainties for siliceous sandstone.

	Mass Fraction ($\times 10^{-2}$)										
	SiO_2	Al_2O_3	Fe_2O3	TiO_2	CaO	MgO	K_2O	Na_2O	MnO	P_2O_5	Cr_2O_3
Standard values	94.41	3.20	0.088	0.019	0.094	0.025	1.26	0.47	0.0011	0.007	0.00034
Uncertainties	0.2	0.07	0.009	0.002	0.009	0.004	0.05	0.05	0.0002	–	0.00005

Table 2
Standard values and standard deviations for clay.

	Mass Fraction ($\times 10^{-2}$)							
	SiO_2	Al_2O_3	Fe_2O3	CaO	MgO	K_2O	Na_2O	TiO_2
Standard Values	49.98	26.27	10.55	0.13	0.46	0.79	0.06	0.70
Standard deviations	0.1	0.06	0.14	0.01	0.09	0.02	0.005	0.12
	P_2O_5	MnO	SO_3	Cl^-	L.O.I.	FeO	H_2O^+	CO_2
Standard Values	0.14	0.052	0.049	0.0041	10.62	(0.080)	(9.64)	(0.041)
Standard deviations	0.02	0.008	0.017	0.0008	0.16			

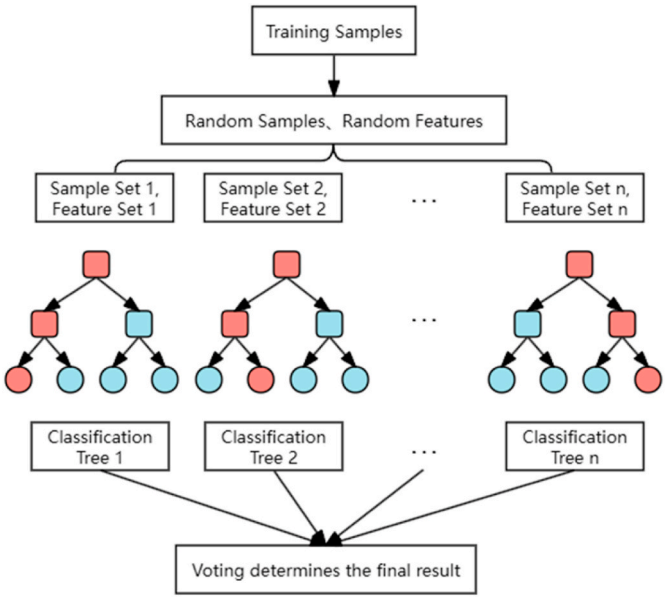

Fig. 2. The generation process of RF. It builds multiple decision trees during training and outputs the class mode or mean prediction of each tree.

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_p x_{1p} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_p x_{2p} + \varepsilon_2 \\ \quad\quad \ldots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_p x_{np} + \varepsilon_n \end{cases} \tag{4}$$

When expressed in matrix form, the linear model becomes more streamlined and can be represented as:

$$y = X\beta + \varepsilon \tag{5}$$

$$where\ y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \varepsilon = \begin{bmatrix} \epsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{6}$$

After establishing a MLR model, it is crucial to estimate the unknown parameters. The typical method used for this purpose is Ordinary Least Squares (OLS). Suppose $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_{n-1}$ are the OLS estimates for the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_{n-1}$, then the observed values of y can be expressed as:

$$y_k = \widehat{\beta}_0 + \widehat{\beta}_1 x_{k1} + \ldots + \widehat{\beta}_{n1} x_{kn1} + \varepsilon_k \tag{7}$$

$$e_k = y_k - \widehat{y}_k \tag{8}$$

where $k = 1, 2, \ldots, N_0$ represents the estimated value of the error ε_k.

In accordance with the least squares method, the objective is to minimize the sum of squared deviations, denoted as W, between the observed values and the predicted values. That is, there is a minimal value of $W = \sum_{k=1}^{n} [y_k - \widehat{y}_k]^2$. By solving the matrix equation, the least squares estimate of the regression coefficients β can be obtained as:

$$\widehat{\beta} = (X^T X)^{-1} X^T Y \tag{9}$$

Error metrics are used to evaluate the performance of a model, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). These metrics provide different insights into the accuracy and reliability of the regression model:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \widehat{y}_i| \tag{10}$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (\widehat{y}_i - y_i)^2 \tag{11}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\widehat{y}_i - y_i)^2} \tag{12}$$

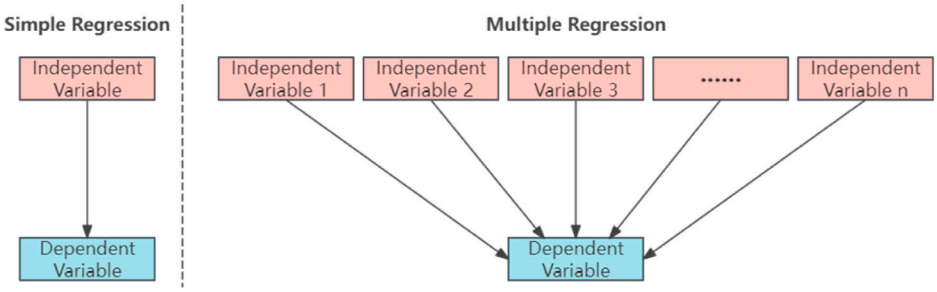

Fig. 3. The difference between univariate regression and multiple linear regression. Simple linear regression involves only one set of independent variable data, while mlr describes the regression problem between a dependent variable and multiple independent variables.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{13}$$

Where n is the number of data points, y_i is the actual value, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the mean of the actual values. Lower values of MSE and RMSE indicate smaller errors, suggesting that the model is more accurate and effective. The R^2, closer to 1 indicates better model performance, as it shows how well the variance in the response variable is explained by the independent variables.

2.2.3. Detection technology

The experiment employed a self-developed soil petroleum hydrocarbon fluorescence imaging monitoring system, capable of rapidly and non-destructively detecting the TPH in soil. The system excites petroleum hydrocarbon pollutants in the soil with an ultraviolet light source, producing a fluorescent signal that is swiftly captured and analyzed by a CCD camera. By measuring the intensity and characteristic parameters of these fluorescent signals, the presence and concentration of petroleum hydrocarbon pollutants can be accurately determined.

The system is composed of four primary components: an excitation light source module, fluorescence imaging module, main control unit, and data processing unit. Fig. 4 demonstrates that the light generated by the excitation light source module is collimated through a lens and focused onto the sample. The fluorescence emitted by the sample is captured by the camera lens after passing through a refractive structure and filter, entering the fluorescence collection system. The fluorescence imaging module is composed of a CCD detector. The main control unit is responsible for controlling the system's circuitry, including the driving circuits for the excitation light source and imaging module, and the data processing unit carries out qualitative and quantitative analysis of soil petroleum hydrocarbon pollution. Petroleum pollutants encompass crude oil and its processed products, including gasoline, diesel, kerosene, and lubricating oil. The excitation light source in our system is a narrow-band ultraviolet LED with a central wavelength of 285 nm, which operates at high power and intensity. We used the Otsu method to automatically obtain the threshold of the image and binarize the image, and then perform a dot multiplication operation on the binarized image and the original image. The core goal of this method is to maximize the inter-class variance, that is, the grayscale value difference between the foreground class and the background class, so as to maximize the distinction between the foreground and the background. The biggest advantage of the Otsu method is that it is automated and unsupervised, and can automatically select the optimal threshold based on the grayscale distribution of the image. This not only improves the accuracy of image segmentation, but also greatly improves the efficiency of data processing [41]. Compared to traditional detection methods, fluorescence imaging offers significant advantages in terms of sensitivity, speed, and non-destructive testing, as shown in Table 3.

3. Results and discussion

3.1. Signal values of each substance under different soil types and sample concentrations

We tested samples of three standard soil types and two petroleum hydrocarbon concentrations to study the effects of humidity, organic matter (humic acid, fulvic acid), and minerals (kaolinite, siliceous sandstone, clay, sodium dihydrogen phosphate, sodium silicate) on the fluorescence signal values. Fig. 5 displays the results. The humidity curve in Fig. 5(a), the kaolinite curve in Fig. 5(d), the siliceous sandstone curve in Fig. 5(e), the sodium dihydrogen phosphate curve in Fig. 5(g), and the sodium silicate curve in Fig. 5(h) indicate that within a certain range, higher humidity or mass fractions lead to higher signal values, and the signal values of samples with higher concentrations are generally higher than those of samples with lower concentrations. When humidity increases, water can effectively enhance the solubility and dispersion of fluorescent substances, thereby improving the excitation and emission process of fluorescent substances. By acting on fluorescent molecules, water reduces the energy loss of fluorescent molecules, enhances their luminescence efficiency, and thus increases the intensity of fluorescent signals [42]. In addition, increased humidity may change the interaction between fluorescent molecules and their surroundings, reduce the probability of non-radiative transitions, and convert more excitation energy into light signals, thereby enhancing the signal. Minerals (such as kaolin, siliceous sandstone, sodium dihydrogen phosphate, and sodium silicate) have specific light scattering properties. When their concentration increases, they may enhance the light scattering effect in the sample and make the excitation light more evenly distributed on the fluorescent molecules. This can improve the utilization efficiency of the excitation light, thereby enhancing the

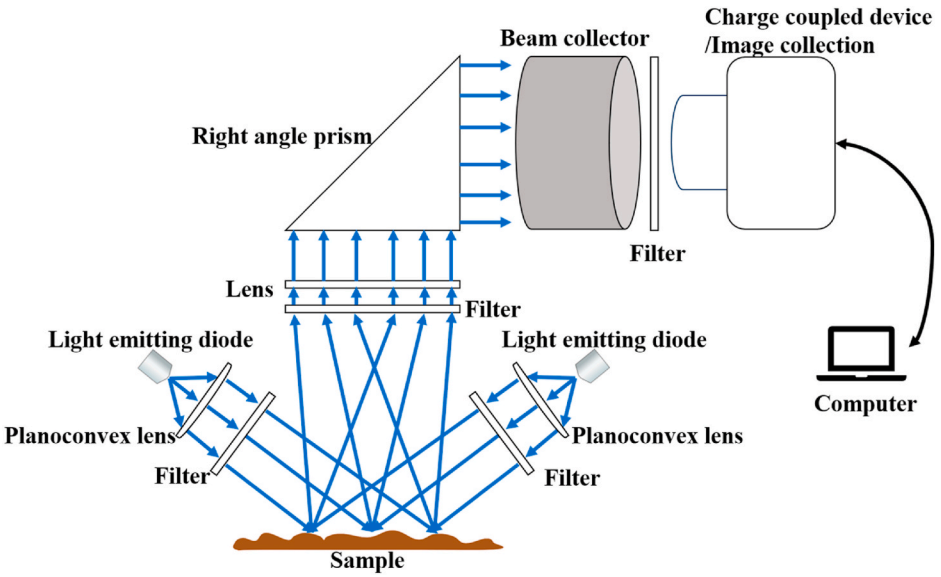

Fig. 4. Schematic diagram of the experimental system. It contains light-emitting diode, planoconvex lens, optical filter, right angle prism, beam collector, charge coupled device, sample, and computer.

Table 3
Comparison of fluorescence imaging detection technology and other methods in terms of sensitivity, specificity, cost and practical usability.

Technology	Sensitivity	Specificity	Cost	Practical Usability	Disadvantages
Fluorescence Imaging	High	Medium	Low	Identify target molecules with fluorescent properties and display images in real time without complicated sample pretreatment, which is suitable for real-time monitoring of large-scale sites.	Only specific target analytes are excited to emit fluorescence.
Chromatography	High	High	Medium	GC: analysis of various organic or inorganic samples with boiling points below 400 °C; HPLC: separation and analysis of high boiling point, thermally unstable, and biological samples.	Difficult to characterize the separated components; Slow analysis speed.
Mass spectrometry	High	High	High	Identify molecules through their mass-to-charge ratio, different molecules and even isotopes or isomers can be distinguished.	Complex samples may affect ionization and reduce quantitative accuracy; It is bulky and sensitive to environmental conditions such as power supply, temperature and humidity, and cannot realize real-time data monitoring of the actual site.
Spectroscopy	High	High	High	The analysis speed is fast and the operation is simple; multiple elements or compounds can be determined simultaneously with good selectivity.	Spectral quantitative analysis requires that the composition and structural state of the standard sample should be basically consistent with the sample being analyzed, which is often difficult.

fluorescence signal [43].

The decrease in signal values under the NSA-6 soil observed in Fig. 5 (b) for humic acid may be attributed to the higher concentration of humic acid, which results in enhanced light absorption. This, in turn, leads to a reduction in the amount of light reflected back and consequently lowers the signal values. In addition, when humic acid molecules are present in large concentrations, they can interact with each other and cause self-absorption phenomena. This would lead to a greater absorption of the excitation light, hence further reducing the signal levels.

The first increase in signal observed in the GBW07416b and GBW07497 soils could potentially be attributed to the addition of humic acid, which enhances the solubility of petroleum hydrocarbons in the soil samples, hence rendering the hydrocarbons more easily visible. The ensuing decline in signal may be attributed to a chemical interaction occurring between the humic acid and petroleum hydrocarbons, resulting in the conversion of some hydrocarbons into less detectable molecules, hence diminishing the signal strength. The subsequent rise in signal could be attributed to the re-additional of humic acid, which once again modifies the chemical conditions of the soil samples, hence enhancing the detectability of some components of the petroleum hydrocarbons.

The drop in signal values observed in the NSA-6 soil for Fig. 5(c), which pertains to fulvic acid, can be attributed to the LED light source stimulating the fulvic acid molecules in the soil petroleum hydrocarbon samples. As the quantity of fulvic acid rises, the number of fulvic acid molecules in the soil also increases, resulting in heightened competition for excited-state transitions. Consequently, the fulvic acid molecules absorb and convert a greater amount of light energy, which leads to a decrease in the number of photons that return to their original condition on the ground. Consequently, this causes a reduction in the measured signal values. The causes behind the alterations in signal in the GBW07416b and GBW07497 soils are comparable to those seen in the presence of humic acid.

Fig. 5(f) shows that in the case of the NSA-6 and GBW07497 soils, an increase in clay concentration results in a decrease in the fluorescence signal due to the strong light absorption by clay particles. Within the GBW07416b soil, with a petroleum hydrocarbon concentration of 10 g/kg, the presence of a tiny quantity of clay particles aids in the dispersion of the petroleum hydrocarbon molecules. This dispersion leads to an increase in the observable surface area of the hydrocarbons and enhances the fluorescence signal. As the mass fraction of clay further increases, clay particles begin to adsorb and envelop a greater number of petroleum hydrocarbon molecules, resulting in a reduction in fluorescence signal. This covering effect masks the fluorescence emission of the petroleum hydrocarbon molecules, resulting in a decrease in the signal. For the samples with a petroleum hydrocarbon concentration of 20 g/kg, clay particles immediately begin to adsorb and cover a large amount of petroleum hydrocarbon molecules from the start, resulting in a weakened fluorescence signal. This shielding effect is more pronounced in higher concentration samples because clay particles can rapidly cover a large number of petroleum hydrocarbon molecules.

3.2. *Influencing factors analysis——Random forest*

In the RF method, the influencing factors of humidity, humic acid, fulvic acid, kaolin, siliceous sandstone, clay, sodium dihydrogen phosphate and sodium silicate and their corresponding signal values were used as input variables. The number of decision trees, the maximum depth of the tree and the learning rate were determined to be 100, 14 and 0.1 respectively by using the grid search method and cross validation method.

Training and prediction with these parameters improved the overall accuracy of the model, reduced the error rate, and effectively reduced the risk of overfitting. 80 % of the dataset is used for training and 20 % for testing. In addition, we used 5-fold cross validation to reduce overfitting and enhance the generalization ability of the model when evaluating the model. The proportion of each influencing factor quantified by the RF method is shown in Table 4. The mean standard errors of the evaluation indicators R^2, MAE, MSE and RMSE are calculated to be 0.14, 16.05, 5580.24 and 25.07 respectively.

The proportion of each influencing factor under the RF method as shown in Fig. 6, was determined by the 'MeanDecreaseGini' metric. This metric quantifies the total decrease in node impurity, weighted by the proportion of samples passing through the node, averaged over all trees. A higher value indicates a greater contribution of the variable to the model's predictive power, effectively ranking the importance of each feature. T-tests revealed that in NSA-6 soil, humidity had significantly higher importance than other factors at both 10 g/kg and 20 g/kg concentrations ($p < 0.05$). In GBW07416b soil, kaolin showed the highest importance at 10 g/kg, while humidity was significantly higher at 20 g/kg ($p < 0.05$). For GBW07497 soil, humidity was the most important factor at 10 g/kg, whereas clay showed the highest importance at 20 g/kg ($p < 0.05$). These findings confirm the varying significance of the factors across different soil types and concentrations. Based on the literature, soil moisture and minerals are key factors affecting petroleum hydrocarbon concentrations and demonstrated a more significant direct impact in our analysis. While organic matter's impact is generally indirect [44], the findings indicate it can have a smaller direct effect under specific soil conditions, highlighting the complex interactions in soil systems.

Based on Table 4, the contributions of each influencing factor under

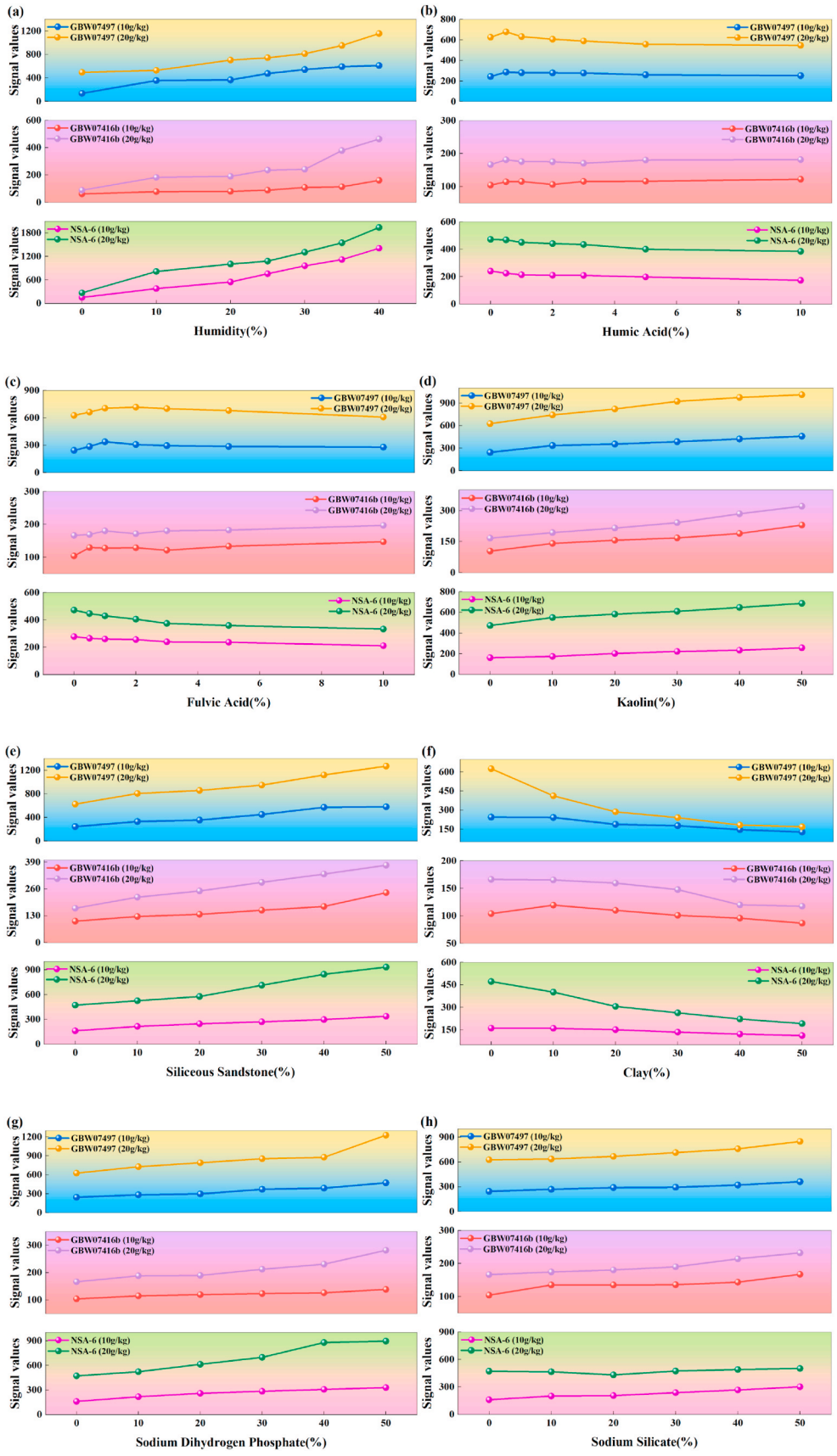

Fig. 5. Signal values of each substance under different soil types and sample concentrations. (a)Humidity; (b)Humic Acid; (c)Fulvic Acid; (d)Kaolin; (e)Siliceous Sandstone; (f)Clay; (g)Sodium Dihydrogen Phosphate; and (h)Sodium Silicate. NSA-6 is the number of red soil in Shaoguan, Guangdong; GBW07497 is the number of chestnut soil in Qinghai; GBW07416b is the number of red soil in Yingtan, Jiangxi. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 4
The proportion of each influencing factor under the RF method. It contains three types of soil and two petroleum hydrocarbon concentrations, showing the proportion and error of each concentration.

Influential Factors	NSA-6		GBW07416b		GBW07497	
Concentrations	10 g/kg	20 g/kg	10 g/kg	20 g/kg	10 g/kg	20 g/kg
Humidity	93.32 %	79.95 %	14.04 %	55.37 %	50.53 %	15.14 %
Humic Acid	0.09 %	0.04 %	1.72 %	0.48 %	1.15 %	0.39 %
Fulvic Acid	0.48 %	1.13 %	8.23 %	0.95 %	1.96 %	0.84 %
Kaolin	0.34 %	2.93 %	49.86 %	18.49 %	11.66 %	10.03 %
Siliceous Sandstone	0.11 %	0.53 %	3.53 %	4.17 %	2.15 %	2.66 %
Clay	2.73 %	6.15 %	4.37 %	7.24 %	18.41 %	52.17 %
Sodium Dihydrogen Phosphate	2.10 %	9.12 %	4.85 %	9.93 %	11.25 %	16.15 %
Sodium Silicate	0.81 %	0.14 %	13.40 %	3.38 %	2.88 %	2.62 %
R^2	0.9102	0.4914	−0.0373	0.1845	0.0311	0.3201
MAE	38.07	114.19	24.67	39.10	74.78	112.10
MSE	2542.76	27947.13	1613.32	3568.08	13129.28	32614.00
RMSE	50.43	167.17	40.17	59.73	114.58	180.59

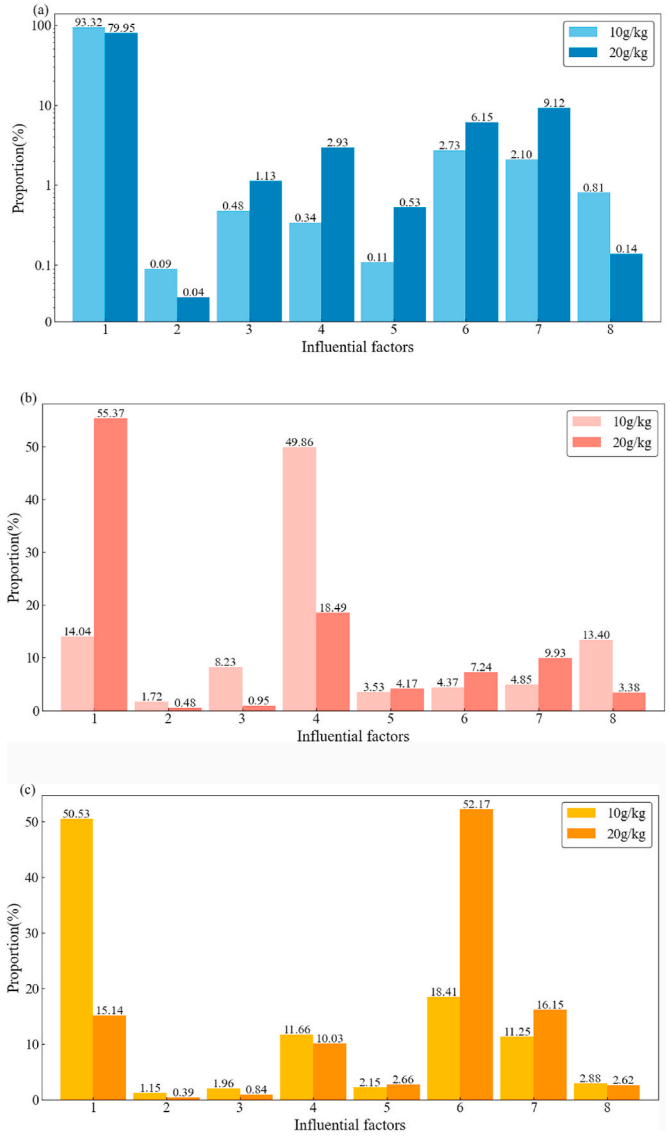

Fig. 6. The proportion of each influencing factor under the RF method. (a)NSA-6; (b)GBW07416b; and (c)GBW07497. The horizontal axis 1-8 represents humidity, humic acid, fulvic acid, kaolin, siliceous sandstone, clay, sodium dihydrogen phosphate, and sodium silicate.

different soil types and sample concentrations were calculated as 51.4 %, 0.6 %, 2.3 %, 15.6 %, 2.2 %, 15.2 %, 8.9 %, and 3.9 %, respectively. Consequently, the top three factors with the highest contributions were initially identified as humidity, kaolinite, and clay. Although the RF model achieved the optimal performance (R^2 = 0.9102) in the NSA-6 soil at a concentration of 10 g/kg, its performance significantly declined at 20 g/kg (R^2 = 0.4914). Similarly, low R^2 values were observed in the GBW07416b and GBW07497, indicating that the RF model does not demonstrate consistent applicability across all soil types and petroleum hydrocarbon concentrations. Furthermore, the large variations in model performance across different soils and concentrations were reflected by the mean MAE values (ranging from 38.07 to 114.19) and RMSE values (ranging from 50.43 to 180.59).

3.3. *Influencing factors analysis——Multiple linear regression*

Taking into account the problem of multicollinearity, the principal component regression method was used to reduce the dimension of problem variables with high collinearity, which effectively reduced the impact of collinearity. The retained principal components explained the following proportion of variance in the dataset: 13.88 %, 13.59 %, 13.59 %, 13.59 %, 13.59 %, 13.42 %, 13.37 %, and 4.97 %. Collectively, these components ensure a significant amount of information was preserved. We also used the variance inflation factor (VIF) to perform a multicollinearity test, and the results were all less than 2, indicating no multicollinearity. To verify the homoskedasticity of the residuals, we performed the Breusch-Pagan test, and the p-value of the test was greater than 0.05, indicating that there was no heteroskedasticity problem. The proportion of each influencing factor quantified by the MLR method is shown in Table 5. The mean standard errors of the evaluation indicators R^2, MAE, MSE and RMSE are calculated to be 0.02, 5.82, 641.95 and 7.95 respectively.

The proportion of each influencing factor under the MLR method as shown in Fig. 7, was derived from the standardized coefficients. These coefficients indicate the relative importance of each predictor variable in explaining the variation of the target variable. They are calculated by transforming the original coefficients using the standard deviations of the predictor and target variables, allowing for comparison across variables with different units or scales. Similarly, t-tests revealed that in NSA-6 soil, humidity had significantly higher importance than other factors at both 10 g/kg and 20 g/kg concentrations ($p < 0.05$). In GBW07416b soil, fulvic acid showed the highest importance at 10 g/kg, while humidity was significantly higher at 20 g/kg ($p < 0.05$). For GBW07497 soil, humidity was the most important factor at 10 g/kg, whereas siliceous sandstone showed the highest importance at 20 g/kg ($p < 0.05$). The contributions of each influencing factor under different soil types and sample concentrations were calculated as 27.3 %, 7.7 %, 18.1 %, 10.7 %, 15.1 %, 6.3 %, 9.5 % and 5.2 %, respectively. Thus, the top three factors with the highest contributions were initially identified

Table 5
The proportion of each influencing factor under the MLR method. It contains three types of soil and two petroleum hydrocarbon concentrations, showing the proportion and error of each concentration.

Influential Factors	NSA-6		GBW07416b		GBW07497	
Concentrations	10 g/kg	20 g/kg	10 g/kg	20 g/kg	10 g/kg	20 g/kg
Humidity	61.36 %	39.77 %	2.26 %	21.20 %	25.80 %	13.29 %
Humic Acid	0.12 %	6.51 %	14.02 %	12.78 %	3.44 %	9.52 %
Fulvic Acid	14.04 %	15.40 %	36.23 %	18.35 %	14.92 %	9.84 %
Kaolin	2.61 %	6.67 %	16.97 %	12.73 %	12.40 %	13.10 %
Siliceous Sandstone	6.54 %	11.86 %	15.90 %	17.98 %	19.71 %	18.66 %
Clay	3.99 %	6.61 %	1.49 %	2.81 %	7.67 %	15.51 %
Sodium Dihydrogen Phosphate	6.87 %	11.78 %	4.59 %	8.61 %	10.91 %	14.16 %
Sodium Silicate	4.47 %	1.40 %	8.53 %	5.54 %	5.15 %	5.92 %
R^2	0.9648	0.9723	0.8215	0.8596	0.9365	0.9186
MAE	28.16	33.60	9.45	15.51	19.53	48.97
MSE	2092.93	2894.52	198.93	660.59	825.20	4304.49
RMSE	45.75	53.90	14.10	25.70	28.73	65.91

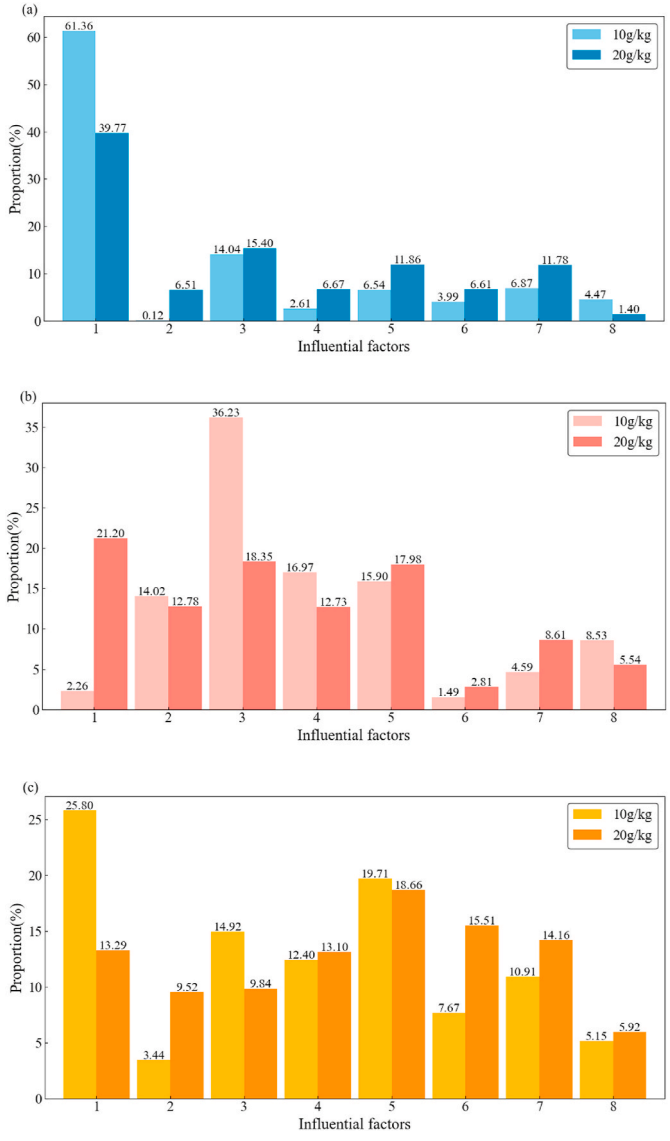

Fig. 7. The proportion of each influencing factor under the MLR method. (a) NSA-6; (b)GBW07416b; and (c)GBW07497. The horizontal axis 1-8 represents humidity, humic acid, fulvic acid, kaolin, siliceous sandstone, clay, sodium dihydrogen phosphate, and sodium silicate.

as humidity, fulvic acid, and siliceous sandstone. Overall, the R^2 values in the MLR model were higher, particularly at the 20 g/kg concentration in the NSA-6 and GBW07416b soils, which were 0.9723 and 0.8596, respectively, indicating a good fit of the model. The MAE and RMSE values were generally lower than those of the RF model, indicating smaller errors.

3.4. Statistical analysis

To compare the two models, we added paired sample *t*-test in terms of R^2, MAE, MSE, and RMSE in Table 6. The standard error of mean (SEM) and 95 % confidence interval of each model are also compared in Table 7. Fig. 8 shows the error bars of the two models in R^2, MAE and RMSE. In terms of these indicators, MLR significantly outperforms RF. Although the MSE did not reach statistical significance, the MSE value of MLR was still lower than that of RF, indicating that its overall performance was better than that of RF.

Comparing the two methods, the RF approach may have some utility in specific contexts, such as for certain soil types (e.g., NSA-6) and under low concentration conditions. This is likely due to its ability to capture nonlinear patterns and complex interactions through its ensemble of decision trees. However, the MLR method demonstrates superior model fit and error metrics performance across most soil types and concentrations. The MLR model's straightforward and interpretable nature also makes it more effective for understanding the direct relationships between variables and their impact on soil petroleum hydrocarbon concentrations. In conclusion, while RF might offer some benefits in isolated scenarios with nonlinear relationships, MLR provides more reliable and consistent results across different scenarios, particularly when

Table 6
T-test for RF and MLR.

indicator	t-Statistic	p-Value	Significance level (α = 0.05)	Conclusion
R^2	4.721	0.005238	$p < 0.05$	The difference is significant, and the goodness of fit of MLR is better than that of RF.
MAE	−3.491	0.017444	$p < 0.05$	The difference is significant, and the MAE of MLR is significantly lower than the RF.
MSE	−2.326	0.067545	$p > 0.05$	The difference was not significant, and the MSE was not statistically different.
RMSE	−3.248	0.022757	$p < 0.05$	The difference is significant, and the RMSE of MLR is significantly lower than the RF.

Table 7
Standard error of mean and 95 % confidence intervals of RF and MLR.

indicator	RF		MLR	
	SEM	confidence intervals	SEM	confidence intervals
R^2	0.142	(0.038, 0.595)	0.024	(0.864, 0.960)
MAE	16.050	(35.693, 98.610)	5.819	(14.465, 37.275)
MSE	5580.235	(-10936.944, 10937.577)	641.949	(571.223, 3087.664)
RMSE	25.067	(-48.815, 49.448)	7.946	(23.441, 54.589)

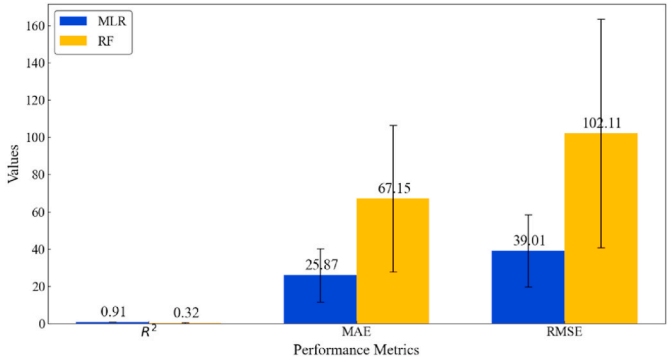

Fig. 8. Error bars for R^2, MAE, and RMSE in the two models.

considering the comprehensive model fit and error metrics.

3.5. Sustainability evaluation by greenness, blueness, and whiteness

Assessing the sustainability of analytical methods is crucial for understanding their environmental and economic impacts [45,46]. In modern analytical chemistry, influenced by green chemistry and sustainable development [47,48], evaluating greenness, blueness, and whiteness has become essential for determining the superiority and innovation of analytical methods [49]. Five tools—NEMI, ESA, ComplexGAPI, AGREE, and RGB12—were employed to assess and compare the whiteness and greenness of the proposed methods with those of pharmacopeial methods [50].

The "greenness" assessment focuses on environmental impact, including reagent usage, waste generation, and energy consumption [51,52]. While this study combines RF and MLR models to effectively reduce the use of chemical reagents and solvents compared to traditional methods, we acknowledge that the lack of a quantitative green assessment is a limitation. By employing non-contact data collection from soil samples and an efficient computational framework, the proposed approach significantly reduces chemical waste in pollutant concentration analysis. Unlike traditional chromatographic and spectrophotometric techniques, this method diminishes reagent consumption and environmental pollution [53,54], highlighting its green analytical advantages for environmental protection. Future research could investigate more comprehensive quantitative assessments to further validate the environmental benefits of the proposed approach.

The "blueness" assessment emphasizes methodological efficiency in data processing and energy consumption [55,56]. The proposed approach excels in handling large-scale datasets. The RF model, with parallel computing and an optimized training process, efficiently processes vast soil sample data, meeting diverse analytical needs. Meanwhile, the MLR model suits rapid analysis of smaller datasets. Overall, the method has significant advantages in energy efficiency and large-scale data processing [57,58], making it ideal for real-time monitoring and environmental emergency response systems.

The "whiteness" assessment focuses on the transparency and interpretability of the analytical method, especially regarding traceability and user comprehension of model output [59,60]. Although RF is generally complex as an integrated learning method, feature importance analysis clearly shows each variable's impact on petroleum hydrocarbons' fluorescence signal, enhancing the method's interpretability. MLR provides a concise linear regression model, making some analysis results more intuitive. By combining RF and MLR, we balance high accuracy and easy interpretability, ensuring model transparency.

In conclusion, this study combines greenness, blueness, and whiteness evaluations to optimize the analytical method in terms of environmental impact, data processing efficiency, and method transparency. The proposed method enhances analytical accuracy while fulfilling sustainable development requirements, demonstrating broad application potential.

4. Future works and limitations

In our future studies, we will explore the interaction between soil type, petroleum hydrocarbon concentration, and other potential factors. Besides, we will focus on several key steps to address the challenges associated with fluorescence-based detection. First, we plan to conduct experiments under varying soil moisture and contaminant concentrations to evaluate their impact on fluorescence signals. Additionally, we will collect data from different geographic regions and soil types to explore how natural soil heterogeneity affects fluorescence responses. To further enhance model accuracy, we will integrate field data with multivariate modeling techniques to account for the effects of moisture variations and contaminant heterogeneity. Finally, extended field experiments will be performed to monitor temporal changes in environmental conditions and fluorescence signals, ensuring the long-term stability and reliability of the models. These efforts will not only improve our understanding of the real-world complexities impacting fluorescence detection but also strengthen the theoretical foundation for its in-situ applications.

Concerning limitations, fluorescence imaging has good detection effects for higher concentrations of pollution, but its sensitivity is generally lower than that of chromatography, especially in the case of low concentrations of petroleum hydrocarbon pollution, and may not provide quantitative data as accurate as chromatography. The findings suggest that certain factors, such as humidity and minerals, significantly impact petroleum hydrocarbon behavior. These insights could inform future research aimed at enhancing fluorescence detection sensitivity. For instance, optimizing environmental conditions, such as maintaining optimal humidity levels or accounting for mineral interactions, might improve detection outcomes. Additionally, the effects of organic matter, though less direct, indicate that understanding and managing soil chemistry could also play a role in sensitivity improvements. These directions offer valuable avenues for future research focused on refining fluorescence detection techniques for lower concentrations.

5. Conclusion

The aim of our research was to investigate how the environmental conditions, including humidity, organic matter, and minerals, impact the fluorescence detection of TPH in soil. The impacts of humidity, organic matter, and minerals on the fluorescence signals of three common soil types were studied in experiments. The experiments were conducted using two different concentrations of petroleum hydrocarbons. The results indicate significant differences in the influence of humidity, organic matter, and minerals on fluorescence detection across different soil types and concentrations of petroleum hydrocarbons. Humidity in the NSA-6 soil, organic matter (fulvic acid) and minerals (kaolinite) in the GBW07416b soil, and minerals (clay and siliceous sandstone) in the GBW07497 soil had notable effects on the fluorescence signals of petroleum hydrocarbons. When comparing the results of the RF and MLR approaches, it was found that the RF method outperformed in dealing with nonlinear relationships and complex interactions,

especially when the concentration was low. In contrast, MLR demonstrated superior model fit and error metrics performance across most soil types and concentrations, with better overall statistical performance and consistency in prediction errors than RF. However, the choice of MLR might limit the ability to capture potential complex nonlinear interactions compared to more flexible models like RF. While MLR provides a good balance between simplicity and performance in many cases, it is important to consider its assumptions and limitations, especially when dealing with highly nonlinear data. Future research could explore hybrid models or alternative approaches that combine the computational efficiency of MLR with the nonlinear modeling capabilities of other techniques.

This research, by revealing the mechanisms via which humidity, organic matter, and minerals affect the fluorescence detection of total soil petroleum hydrocarbons, provides a scientific basis for environmental risk assessments and pollution management. The findings not only help in developing more effective soil pollution prevention strategies to mitigate the impact of petroleum hydrocarbons on the environment and human health but also offer new ideas and methods for research and practice in related fields. This will promote the development of soil petroleum hydrocarbon monitoring technologies and enhance the efficiency and accuracy of environmental protection efforts.

Next, we give some guiding suggestions. For areas with high humidity, it is recommended to use a drainage system to reduce the impact of moisture on the migration and dissolution of petroleum hydrocarbons; for areas with high organic matter content, the fixation of petroleum hydrocarbons can be reduced by regulating soil organic matter; for soils with a greater impact of minerals, mineral regulators can be used to reduce the adsorption of petroleum hydrocarbons and improve the remediation effect. For low-concentration pollution, it is recommended to use biological remediation methods that optimize environmental conditions, while for medium and high-concentration pollution, a combination of physical and chemical remediation (such as elution, pyrolysis) and biological remediation can achieve more efficient degradation. In the context of this study, the analysis of influencing factors provides some insights that indirectly support this recommendation. For instance, at lower concentrations of petroleum hydrocarbon pollution, the relatively simpler interaction of factors such as humidity and minerals may create conditions where biological processes can effectively degrade contaminants. However, as concentrations increase, the complex interplay of multiple factors indicates a more challenging environment that likely requires a multi-faceted remediation strategy. While this study does not directly investigate remediation techniques, the understanding of how different factors influence petroleum hydrocarbon concentrations can inform the selection of appropriate remediation methods. In addition, it is recommended that the government invest in an intelligent soil pollution monitoring platform, combining fluorescence detection technology with environmental data to achieve real-time monitoring and early warning, and establish an early warning mechanism in high-risk areas to initiate remediation measures in a timely manner.

CRediT authorship contribution statement

Gaoyong Shi: Writing – review & editing, Writing – original draft, Resources, Methodology, Data curation. **Ruifang Yang:** Writing – review & editing, Investigation, Funding acquisition, Data curation. **Nanjing Zhao:** Writing – review & editing, Investigation. **Gaofang Yin:** Methodology, Investigation. **Wenqing Liu:** Writing – review & editing, Investigation.

Author statement

We declare that it is an original work and all authors read and approved final version of the manuscript. The manuscript is being submitted for the first time for publication.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was performed with financial support from the Key Technologies Research and Development Program (CN) No. 2022YFC3700902、No. 2020YFC1807204-1, the Science and Technology Service Network Initiative (CN) No. KFJ-STS-QYZD-2021-04-001-4.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2025.105444.

Data availability

The data that has been used is confidential.

References

[1] C. Zhang, D. Wu, H. Ren, Bioremediation of oil contaminated soil using agricultural wastes via microbial consortium, Sci. Rep. 10 (2020) 9188, https://doi.org/10.1038/s41598-020-66169-5.
[2] Saiful, S. Hasima, N. Kamila, Rahmi, Cellulose acetate from palm oil bunch waste for forward osmosis membrane in desalination of brackish water, Results Eng. 15 (2022), https://doi.org/10.1016/j.rineng.2022.100611.
[3] U. Roland, F. Holzer, F.D. Kopinke, Ambivalent role of water in thermodesorption of hydrocarbons from contaminated soil, Environ. Sci. Technol. 45 (2) (2011) 732–737, https://doi.org/10.1021/es102778h.
[4] D. Li, X. You, On optimal condition of plant–microbial remediation of petroleum hydrocarbon polluted soil, Soil Sediment Contam. 30 (1) (2020) 35–57, https://doi.org/10.1080/15320383.2020.1787328.
[5] K. Luko-Sulato, S. Mounier, L.M. Furlan, J.S. Govone, G.T. Bueno, V. Rosolen, Spatial variation of soil organic matter and metal mobility in wetland soils: implications for biogeochemical processes in lateritic landscape, CATENA 237 (2024), https://doi.org/10.1016/j.catena.2024.107810.
[6] J. Xu, Y. Rong, L. Liu, W. Bai, J. Dai, Efficient fenton oriented oxidation of petroleum hydrocarbons in soil by regulating hydrophilic functional groups in soil organic matter, J. Environ. Chem. Eng. 12 (1) (2024), https://doi.org/10.1016/j.jece.2023.111772.
[7] Y. Yang, D. Liu, X. Liang, X. Li, Influence of mineral species on oil–soil interfacial interaction in petroleum-contaminated soils, Chin. J. Chem. Eng. 61 (9) (2023) 147–156, https://doi.org/10.1016/j.cjche.2023.02.015.
[8] Y. Wang, Y. Huang, P. Xi, X. Qiao, J. Chen, X. Cai, Interrelated effects of soils and compounds on persulfate oxidation of petroleum hydrocarbons in soils, J. Hazard. Mater. 408 (2020), https://doi.org/10.1016/j.jhazmat.2020.124845.
[9] O. Campesato, Python 3 for machine learning, Mercury Learning Inform. (2020).
[10] S. Ratnasingam, J. Muñoz-Lopez, Distance correlation-based feature selection in random forest, Entropy 25 (9) (2023), https://doi.org/10.3390/e25091250.
[11] L. Lou, H. Xu, Humidity research based on multiple regression and time-varying exponential smoothing model, Academic J. Math. Sci. 5 (2) (2024), https://doi.org/10.25236/AJMS.2024.050201.
[12] G. Hesamian, F. Torkian, A. Johannssen, N. Chukhrova, A learning system-based soft multiple linear regression model, Intell. Syst. Appl. 22 (2024) 200378, https://doi.org/10.1016/j.iswa.2024.200378.
[13] H. Xia, J. Tang, L. Aljerf, T. Wang, J. Qiao, Q. Xu, Q. Wang, P. Ukaogo, Investigation on dioxins emission characteristic during complete maintenance operating period of municipal solid waste incineration, Environ. Pollut. 318 (2022) 120949, https://doi.org/10.1016/j.envpol.2022.120949.
[14] Y. Liang, J. Tang, H. Xia, L. Aljerf, B. Gao, M.L. Akele, Three-dimensional numerical modeling and analysis for the municipal solid-waste incineration of the grate furnace for particulate-matter generation, Sustainability 15 (16) (2023), https://doi.org/10.3390/su151612337.
[15] H. Xia, J. Tang, L. Aljerf, C. Cui, B. Gao, P.O. Ukaogo, Dioxin emission modeling using feature selection and simplified DFR with residual error fitting for the grate-based MSWI process, Waste Manag. 168 (2023) 256–271, https://doi.org/10.1016/j.wasman.2023.05.056.

[16] T. Wang, J. Tang, L. Aljerf, J. Qiao, M. Alajlani, Emission reduction optimization of multiple flue gas pollutants in Municipal solid waste incineration power plant, Fuel 381 (2025) 133382, https://doi.org/10.1016/j.fuel.2024.133382.
[17] G. Babatunde, A.A. Emmanuel, O.R. Oluwaseun, O.B. Bunmi, A.E. Precious, Impact of climatic change on agricultural product yield using k-means and multiple linear regressions, Int. J. Education Manag. Eng. (2019), https://doi.org/10.5815/IJEME.2019.03.02.
[18] Z. Zhao, H. Zeng, J. Wu, L. Zhang, Concentrations, sources and potential ecological risks of polycyclic aromatic hydrocarbons in soils from Tajikistan, Int. J. Environ. Pollut. 61 (1) (2017), https://doi.org/10.1504/IJEP.2017.082696.
[19] M. Guo, M. Li, H. Fu, Y. Zhang, T. Chen, H. Tang, T. Zhang, H. Li, Quantitative analysis of polycyclic aromatic hydrocarbons (PAHs) in water by surface-enhanced Raman spectroscopy (SERS) combined with Random Forest, Spectrochim. Acta, Part A 287 (1) (2023) 122057, https://doi.org/10.1016/j.saa.2022.122057.
[20] F.M. Canero, V. Rodriguez-Galiano, D. Aragones, Machine learning and feature selection for soil spectroscopy. An evaluation of Random Forest wrappers to predict soil organic matter, clay, and carbonates, Heliyon 10 (9) (2024) e30228, https://doi.org/10.1016/j.heliyon.2024.e30228.
[21] R.K. Douglas, S. Nawar, M.C. Alamar, F. Coulon, A.M. Mouazen, Rapid detection of alkanes and polycyclic aromatic hydrocarbons in oil-contaminated soil with visible near-infrared spectroscopy, Eur. J. Soil Sci. 70 (1) (2019) 140–150, https://doi.org/10.1111/ejss.12567.
[22] M.A. Rehman, N.A. Rahman, A.N.H. Ibrahim, N.A. Kamal, A. Ahmad, Estimation of soil erodibility in Peninsular Malaysia: a case study using multiple linear regression and artificial neural networks, Heliyon 10 (7) (2024) e28854, https://doi.org/10.1016/j.heliyon.2024.e28854.
[23] B.K. Agbaogun, B.I. Olu-Owolabi, H. Buddenbaum, K. Fischer, Adaptive neuro-fuzzy inference system (ANFIS) and multiple linear regression (MLR) modelling of Cu, Cd, and Pb adsorption onto tropical soils, Environ. Sci. Pollut. Res. 30 (2023) 31085–31101, https://doi.org/10.1007/s11356-022-24296-8.
[24] J. Li, J. Li, Analysis of the current situation and influencing factors of China's carbon emissions—based on the multiple linear regression model, Financ. Eng. Risk Manag. 6 (10) (2023), https://doi.org/10.23977/ferm.2023.061006.
[25] H. Li, J. Wang, J. Zhang, T. Liu, G.E. Acquah, H. Yuan, Combining variable selection and multiple linear regression for soil organic matter and total nitrogen estimation by DRIFT-MIR spectroscopy, Agronomy 12 (3) (2022) 638, https://doi.org/10.3390/agronomy12030638.
[26] R. Jafar, A. Awad, I. Hatem, K. Jafar, E. Awad, I. Shahrour, Multiple linear regression and machine learning for predicting the drinking water quality index in Al-Seine Lake, Smart Cities 6 (5) (2023) 2807–2827, https://doi.org/10.3390/smartcities6050126.
[27] H. Wang, Q. Yilihamu, M. Yuan, H. Bai, H. Xu, J. Wu, Prediction models of soil heavy metal(loid)s concentration for agricultural land in Dongli: a comparison of regression and random forest, Ecol. Indic. 119 (2020), https://doi.org/10.1016/j.ecolind.2020.106801.
[28] S. An, X. Chen, F. Li, S. Wang, M. Shen, X. Luo, S. Ren, H. Zhao, Y. Li, L. Xu, Long-term species-level observations indicate the critical role of soil moisture in regulating China's grassland productivity relative to phenological and climatic factors, Sci. Total Environ. 929 (2024), https://doi.org/10.1016/j.scitotenv.2024.172553.
[29] Y. Sun, Y. Liu, G. Yue, J. Cao, C. Li, J. Ma, Vapor-phase biodegradation and natural attenuation of petroleum VOCs in the unsaturated zone: a microcosm study, Chemosphere 336 (2023), https://doi.org/10.1016/j.chemosphere.2023.139275.
[30] C. Chen, Y. Dong, A. Thompson, Electron transfer, atom exchange, and transformation of iron minerals in soils: the influence of soil organic matter, Environ. Sci. Technol. 57 (29) (2023) 10696–10707, https://doi.org/10.1021/acs.est.3c01876.
[31] Y. Chen, P.W. Liu, L. Whang, Y. Wu, S. Cheng, Effect of soil organic matter on petroleum hydrocarbon degradation in diesel/fuel oil-contaminated soil, J. Biosci. Bioeng. 129 (5) (2020) 603–612, https://doi.org/10.1016/j.jbiosc.2019.12.001.
[32] I. Afzal, A. Kuznetsova, J. Foght, A. Ulrich, T. Siddique, Crystalline iron oxide mineral (magnetite) accelerates methane production from petroleum hydrocarbon biodegradation, Environ. Pollut. 363 (1) (2023), https://doi.org/10.1016/j.envpol.2024.125065.
[33] H. Xia, J. Tang, L. Aljerf, J. Chen, Unveiling dioxin dynamics: a whole-process simulation study of municipal solid waste incineration, Sci. Total Environ. 954 (2024) 176241, https://doi.org/10.1016/j.scitotenv.2024.176241.
[34] T. Wang, J. Tang, H. Xia, L. Aljerf, R. Zhang, H. Tian, M.L. Akele, Intelligent optimal control of furnace temperature for the municipal solid waste incineration process using multi-loop controller and particle swarm optimization, Expert Syst. Appl. 257 (2024) 125015, https://doi.org/10.1016/j.eswa.2024.125015.
[35] C. Wang, J. Du, X. Fan, High-dimensional correlation matrix estimation for general continuous data with Bagging technique, Mach. Learn. 111 (2022) 2905–2927, https://doi.org/10.1007/s10994-022-06138-3.
[36] I.S. Dar, S. Chand, Bootstrap-quantile ridge estimator for linear regression with applications, PLoS One 19 (4) (2024), https://doi.org/10.1371/journal.pone.0302221.
[37] B. Zhang, Z. Wang, H. Li, Z. Lei, J. Cheng, S. Gao, Information gain-based multi-objective evolutionary algorithm for feature selection, Inf. Sci. 677 (2024), https://doi.org/10.1016/j.ins.2024.120901.
[38] J. Zhang, R.L. Rardin, J.R. Chimka, Budget constrained model selection for multiple linear regression, Commun. Stat. Simulat. Comput. 52 (11) (2021) 5537–5549, https://doi.org/10.1080/03610918.2021.1991956.
[39] H. Xia, T. Jian, L. Aljerf, T. Wang, B. Gao, Q. Xu, Q. Wang, P. Ukaogo, Assessment of PCDD/Fs formation and emission characteristics at a municipal solid waste incinerator for one year, Sci. Total Environ. 883 (2023) 163705, https://doi.org/10.1016/j.scitotenv.2023.163705.
[40] H. Xia, T. Jian, L. Aljerf, T. Wang, B. Gao, M. Alajlani, AI-based tree modeling for multi-point dioxin concentrations in municipal solid waste incineration, J. Hazard Mater. 480 (2024) 135834, https://doi.org/10.1016/j.jhazmat.2024.135834.
[41] G. Shi, R. Yang, N. Zhao, G. Yin, J. Yang, Y. Jiang, W. Liu, Rapid detection of total petroleum hydrocarbons in soil using advanced fluorescence imaging techniques, ACS Omega 9 (27) (2024) 29350–29359, https://doi.org/10.1021/acsomega.4c01298.
[42] Y. Gu, Z. Zuo, Z. Zhang, C. Shi, X. Gao, J. Lu, Algorithmic study of total petroleum hydrocarbons in contaminated soil by three-dimensional excitation-emission matrix fluorescence spectroscopy, Chinese Optics 13 (4) (2020) 852–864, https://doi.org/10.37188/CO.2019-0216.
[43] Yu I. Pikovskii, L.A. Korotkov, M.A. Smirnova, R.G. Kovach, Laboratory analytical methods for the determination of the hydrocarbon status of soils (a review), Eurasian Soil Sci. 50 (2017) 1125–1137, https://doi.org/10.1134/S1064229317100076.
[44] L. Wang, Y. Cheng, R. Naidu, M. Bowman, The key factors for the fate and transport of petroleum hydrocarbons in soil with related in/ex situ measurement methods: an overview, Front. Environ. Sci. 9 (2021), https://doi.org/10.3389/fenvs.2021.756404.
[45] M.K. Halim, O.M. Badran, A.E.F. Abbas, Greenness, blueness and whiteness evaluated-chemometric approach enhanced by Latin hypercube technique for the analysis of lidocaine, diclofenac and carcinogenic impurity 2,6-dimethylaniline, Sustain. Chem. Pharm. 38 (2024) 101463, https://doi.org/10.1016/j.scp.2024.101463.
[46] S.A. Mahmoud, A.E.F. Abbas, N.S. Katamesh, Greenness, whiteness, and blueness assessment with spider chart solvents evaluation of HPTLC-densitometric method for quantifying a triple combination anti-Helicobacter pylori therapy, Sustain. Chem. Pharm. 37 (2024) 101412, https://doi.org/10.1016/j.scp.2023.101412.
[47] M.K. Halim, O.M. Badran, A.E.F. Abbas, Sustainable chemometric methods boosted by Latin hypercube technique for quantifying the recently FDA-approved combination of bupivacaine and meloxicam in the presence of bupivacaine carcinogenic impurity: comprehensive greenness, blueness, and whiteness assessments, Microchem. J. 200 (2024) 110276, https://doi.org/10.1016/j.microc.2024.110276.
[48] A.A. El-Masry, A.E.F. Abbas, Y.A. Salem, A dual methodology employing ion-pair chromatography and built-in UV spectrophotometry for quantifying recently approved combination of mometasone and indacaterol in a novel combined metered dose inhaler: assessing the greenness, carbon footprint, blueness, and whiteness, BMC Chem. 18 (143) (2024), https://doi.org/10.1186/s13065-024-01242-y.
[49] N.S. katamesh, A.E.F. Abbas, M.K. Halim, M.A. Abdel-Lateef, S.A. Mahmoud, Green micellar UPLC and complementary eco-friendly spectroscopic techniques for simultaneous analysis of anti-COVID drugs: a comprehensive evaluation of greenness, blueness, and whiteness, BMC Chem. 18 (149) (2024), https://doi.org/10.1186/s13065-024-01254-8.
[50] A.E.F. Abbas, M. Gamal, I.A. Naguib, M.K. Halim, B.A.M. Said, M.M. Ghoneim, M. M.A. Mansour, Y.A. Salem, Sustainable quantification of glycopyrronium, indacaterol, and mometasone along with two genotoxic impurities in a recently approved fixed-dose breezhaler formulations and biological fluids: a machine learning-augmented UV-spectroscopic approach, Microchem. J. 206 (2024) 111586, https://doi.org/10.1016/j.microc.2024.111586.
[51] A.E.F. Abbas, N.A. Abdelshafi, M. Gamal, M.K. Halim, B.A.M. Said, I.A. Naguib, M. M.A. Mansour, S. Morshedy, Y.A. Salem, Simultaneously quantifying a novel five-component anti- migraine formulation containing ergotamine, propyphenazone, caffeine, camylofin, and mecloxamine using UV spectrophotometry and chemometric models, BMC Chem. 18 (233) (2024), https://doi.org/10.1186/s13065-024-01339-4.
[52] A.A. El-Masry, S.A. Elsabour, A.E.F. Abbas, Y.A. Salem, Implantation of impressive chromatographic and built-in UV spectrophotometry approaches for sustainable Estimation of olopatadine and mometasone in pharmaceuticals; eco-scale and BAGI applications, Spectrochim. Acta, Part A 327 (2025) 125409, https://doi.org/10.1016/j.saa.2024.125409.
[53] Y.A. Salem, A.E.F. Abbas, A.E. Salem, A.A. Abdella, A.A. El-Masry, Multi-assessed green sustainable chromatographic resolution of nicotine and caffeine; application to in-vitro release from a new quick mist mouth spray co-formula, BMC Chem. 18 (200) (2024), https://doi.org/10.1186/s13065-024-01306-z.
[54] N.S. Katamesh, A.E.F. Abbas, S.A. Mahmoud, Four chemometric models enhanced by Latin hypercube sampling design for quantification of anti-COVID drugs: sustainability profiling through multiple greenness, carbon footprint, blueness, and whiteness metrics, BMC Chem. 18 (54) (2024), https://doi.org/10.1186/s13065-024-01158-7.
[55] K.A.M. Attia, A.E.F. Abbas, A. Ei-Olemy, N.A. Abdelshafi, S.M. Eid, A recycled-material-based electrochemical eco-sensor for sensitive detection of antischistosomal drug residues in bovine-derived food samples, BioChip J. 18 (2024) 257–274, https://doi.org/10.1007/s13206-024-00144-4.
[56] K.A.M. Attia, A. Ei-Olemy, A.E.F. Abbas, S.M. Eid, A sustainable data processing approach using ultraviolet-spectroscopy as a powerful spectral resolution tool for simultaneously estimating newly approved eye solution in the presence of extremely carcinogenic impurity aided with various greenness and whiteness assessment perspectives: application to aqueous humor, J. Chem. Res. 47 (5) (2023), https://doi.org/10.1177/17475198231195811.
[57] A.E.F. Abbas, M. Gamal, I.A. Naguib, M.K. Halim, B.A.M. Said, M.M. Ghoneim, M. M.A. Mansour, Y.A. Salem, Application of machine learning assisted multi-variate UV spectrophotometric models augmented by kennard stone clustering algorithm

for quantifying recently approved nasal spray combination of mometasone and olopatadine along with two genotoxic impurities: comprehensive sustainability assessment, BMC Chem. 19 (98) (2025), https://doi.org/10.1186/s13065-025-01391-8.

[58] A.E.F. Abbas, M. Gamal, I.A. Naguib, M.K. Halim, B.A.M. Said, M.M. Ghoneim, M. M.A. Mansour, Y.A. Salem, Electrochemical platform with Ag/ZnO nanorods for green, blue, and white determination of the newly approved drug roxadustat in pharmaceuticals and plasma: NQS assessment and UN-SDGs alignment, Microchem. J. 212 (2025) 113326, https://doi.org/10.1016/j.microc.2025.113326.

[59] K.A.M. Attia, A. El-Olemy, S.M. Eid, A.E.F. Abbas, A green-and-white integrative analytical strategy combining univariate and chemometric techniques for quantifying recently approved multi-drug eye solution and potentially cancer-causing impurities: application to the aqueous humor, J. AOAC Int. 107 (1) (2024) 146–157, https://doi.org/10.1093/jaoacint/qsad087.

[60] K.A.M. Attia, A. Serag, S.M. Eid, A.E.F. Abbas, A new chemometrically assisted UV spectrophotometric method for simultaneous determination of tamsulosin and dutasteride in their pharmaceutical mixture, J. AOAC Int. 105 (6) (2022) 1755–1761, https://doi.org/10.1093/jaoacint/qsac080.