Analytical Methods

PAPER

View Article Online
View Journal | View Issue

Cite this: *Anal. Methods*, 2019, **11**, 3419

Rapid quantitative analysis of the acidity of iron ore by the laser-induced breakdown spectroscopy (LIBS) technique coupled with variable importance measures-random forests (VIM-RF)†

Ping Wang,[a] Nan Li,[a] Chunhua Yan,[a] Yaozhou Feng,[a] Yu Ding,[c] Tianlong Zhang[a] and Hua Li*[ab]

Rapid and online analysis of the acidity of iron ore is extremely important for reasonable and efficient utilization of mineral resources. In this study, the laser-induced breakdown spectroscopy (LIBS) technique coupled with variable importance measures-random forests (VIM-RF) was proposed and applied for rapid and effective analysis of acidity in iron ore. LIBS spectra of 50 iron ore samples were collected, and the characteristic spectral lines of major elements (Ca, Mg, Si and Al) in iron ore samples were identified based on the National Institute of Standards and Technology (NIST) database. Different pre-processing methods, input variables and RF calibration model parameters were investigated and optimized by 5-fold cross validation (CV), and variable importance measurement (VIM) was used to optimize the input variables of the RF calibration model. In order to further verify the predictive ability and robustness of the VIM-RF calibration model, three calibration models of VIM-RF, partial least squares (PLS) and least squares support vector machine (LS-SVM) were applied for the quantitative analysis of acidity in iron ore, and the correlation coefficient (R^2) and root mean squared error (RMSE) were evaluation indices. The results show that the VIM-RF model exhibits an excellent predictive performance compared with the other two calibration models both for the calibration set and prediction set. Therefore, the LIBS technique combined with VIM-RF can achieve a rapid acidity analysis of iron ores, and it will provide a new method and technology for selection and quality control of iron ore in the metallurgical industry.

Received 3rd May 2019
Accepted 4th June 2019

DOI: 10.1039/c9ay00926d

rsc.li/methods

1. Introduction

Steel and iron play a significantly important role in the continuous development of the world economy, and their quality requirement is becoming more and more strict. Iron ore is the main raw material in the iron and steel industry, and its quality is very important for the sustained and stable development of the steel industry. The acidity of iron ore is mainly determined by the amount of CaO, MgO, Al_2O_3 and SiO_2, and can be calculated using the (CaO + MgO)/(Al_2O_3 + SiO_2) concentration ratio.[1] If the acidity is above 1, it can be identified as alkaline ore; otherwise, it is considered as acidic ore. Alkaline iron ore is widely applied in the metallurgy field due to its desulfurization and a lower iron-coke ratio. However, for an acidic iron ore with a high melting point, fluxing agents need to be added into the blast furnace to reduce the melting point, which ensures smooth operation in the metallurgical process. Thus, an acidic iron ore with a high melting point can not only cause energy waste, but also reduce the utilization rate of raw materials in blast furnace smelting. Overall, the accurate determination of the acidity of iron ore is not only helpful to ensure smooth operation of the smelting process in blast furnace smelting, but also improves the quality of metallurgical products.

Conventional analytical techniques for iron ore mainly include atomic absorption spectroscopy (AAS),[2] inductively coupled plasma optical emission spectroscopy (ICP-OES),[3] and inductively coupled plasma mass spectroscopy (ICP-MS).[4] Although these analytical techniques show good sensitivity and accuracy, they generally require complex sample preparation and a longer analysis time, which hinder their application in rapid and online analysis. Hence, a precise, simple and rapid analytical technique is necessary to online monitor the

[a]*Key Laboratory of Synthetic and Natural Functional Molecular Chemistry of Ministry of Education, College of Chemistry and Material Science, Northwest University, Xi'an, 710127, China*

[b]*College of Chemistry and Chemical Engineering, Xi'an Shiyou University, Xi'an 710065, China. E-mail: huali@nwu.edu.cn*

[c]*Jiangsu Key Laboratory of Big Data Analysis Technology, Nanjing University of Information Science &Technology, Nanjing, 210044, China*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c9ay00926d

composition and acidity of iron ore in the metallurgical industry. Laser-induced breakdown spectroscopy (LIBS) is a promising and prospective analytical tool with the advantages of rapid, simultaneous multi-element analysis and requiring no complex sample pretreatment,[5–7] and has received increasing attention from many research groups. The LIBS technique has become the subject of metallurgical analysis, and its application has increased considerably in the metallurgical industry, including iron ore,[8–10] steel materials[11–13] and steel slag,[14,15] especially in the rapid analysis and selection of iron ore. At present, there are a series of research studies on iron ore analysis by the LIBS technique, which mainly include classification analysis and quantitative determination. For rapid classification analysis of iron ore, Yan *et al.*[16] employed LIBS coupled with N-nearest neighbours to accurately identify four types of iron ores with different acidity. Sheng *et al.*[10] utilized LIBS combined with random forests (RF) to discriminate successfully ten iron ore grades.

Several methods have been proposed for the quantitative analysis of iron ore using the LIBS technique. Grant *et al.*[17] employed a univariate approach coupled with the LIBS technique to determine the major element (Ca, Si, Mg, Al and Ti) concentrations in iron ore samples. However, the accuracy of the univariate calibration method was severely influenced by matrix effects and spectral interference, which may destroy the relationship between the characteristic line intensity of elements and the element content. The multivariate calibration method has exhibited increasing applications in quantitative measurements of unknown samples with its advantages of overcoming complex matrix effects, reducing self-absorption of spectral lines and solving spectral interference. Recently, a series of multivariate methods have been proposed for quantitative analysis of iron ore by the LIBS technique, such as partial least squares (PLS),[18,19] principal component regression (PCR),[9,20] support vector machines (SVM)[21,22] and extreme learning machine (ELM).[23] Hao *et al.*[1] employed PLS as a common multivariate analysis method to determine the acidity in iron ore by the LIBS technique, and a good result (the average relative error (ARE) of the acidity was 3.65%, and the RMSE of the acidity was 0.0048) was obtained. Yaroshchyk[8] used PCR, PLS, multi-block PLS, and serial PLS to analyze the iron content in iron ore by the LIBS technique, and the results showed that PCR provided superior performance for Fe determination with an R^2 of 0.97 and RMSE of 2.2%. However, the stability and prediction accuracy of these calibration models are heavily influenced by randomly selected parameters, and they are prone to fall into the local optimum. Besides, the above-mentioned methods can only distinguish relationships between the input and output, but never find the interdependence among variables. In their modeling process, each variable can be given the same importance, and these models fail to discriminate the true and noise variables. Variable importance measurement (VIM) helps to select the best subset of variables, and is beneficial to obtain a simpler and more accurate model by the simplest way of removing redundant variables. Therefore, it is necessary to perform variable importance measurement to establish a simpler and more accurate model.

Random forests (RF) as a novel multivariate method based on multiple decision trees were proposed by Leo Breiman,[24] and can well deal with the above-mentioned shortcomings in the modeling process. They are ensembles of unpruned decision trees created using bootstrap samples of the training data and random feature selection. In RF modeling, the bootstrap sample set is used to construct multiple decision trees, and the final predictive results are determined by taking the average of the predictions of all the individual decision trees from the forest. It can distinguish nonlinear approximation of relationships among variables, and rank the importance of variables based on its inbuilt VIM. The RF method has several significant features, including: having a good tolerance for noise, VIM even in the presence of high levels of noise or spurious information, and general resistance to over-fitting. In recent years, VIM-RF combined with LIBS have been extensively applied in many fields. Tang *et al.*[25] presented VIM-RF combined with LIBS to perform the classification analysis of slag samples. Tian *et al.*[26] proposed VIM-RF combined with LIBS to classify wines with different production regions, and satisfactory classification results were obtained with a classification accuracy of 100% for the tested samples. Although the studies above made a great advance in classification analysis, there have been few reports on the quantitative analysis by VIM-RF combined with LIBS.

The present work explores the combination of LIBS technology with the VIM-RF model for the rapid analysis of the acidity of iron ore. At first, LIBS spectra of 50 iron ore samples were collected, and the NIST database was utilized to identify the major elements in iron ore. Different pre-processing technologies, input variables and RF model parameters were investigated and optimized by 5-fold CV to construct an optimized calibration model for acidity analysis, and the input variables of the RF model were optimized by variable importance measurement (VIM). Then the VIM-RF model was applied for the quantitative analysis of CaO, MgO, SiO_2, Al_2O_3 and acidity in iron ore, and the corresponding results were compared with those of PLS and LS-SVM.

2. Materials and methods

2.1 Sample preparation

A total of 50 iron ore samples were investigated in the present work. Five standard iron ore powder samples of 1# (GBW07822), 8# (GBW07824), 15# (GBW07826), 22# (GBW07828) and 29# (GBW07830) were purchased from the Institute of Geo physical and Geo-chemical Exploration (China) in this study. The rest of the iron ore samples were prepared by mixing different contents of analytically pure reagents (Al_2O_3 and SiO_2) with five standard iron ore powder samples. Table 1 lists the major element concentration and certified acidity of 50 iron ore samples by X-ray fluorescence (XRF). All analytical samples were prepared into pellets with 0.6 g iron ore powder and 0.4 g polyvinyl alcohol (PVA) as a binding material that was used to avoid the scattering of analytical samples by the pulse laser. All pellets were prepared with a tablet press at 20 MPa for 3 min.

Table 1 The major element concentration (wt%) and the certified acidity of iron ore samples

No.	CaO	MgO	SiO_2	Al_2O_3	Acidity
1#	2.8400	1.6800	60.8600	3.5700	0.0702
2#	2.8105	1.6625	60.2276	4.5700	0.0690
3#	2.7810	1.6451	59.5966	5.5700	0.0679
4#	2.6044	1.5406	55.8118	11.5700	0.0615
5#	2.6949	1.5942	62.8600	3.3876	0.0647
6#	2.4772	1.4654	65.8600	3.1140	0.0572
7#	2.1144	1.2508	70.8600	2.6579	0.0458
8#	2.0000	2.2200	33.9300	2.2700	0.1166
9#	1.9795	2.1973	33.5824	3.2700	0.1133
10#	1.9591	2.1746	33.2354	4.2700	0.1102
11#	1.8363	2.0383	31.1527	10.2700	0.0935
12#	1.9395	2.1528	35.9300	2.2013	0.1073
13#	1.8487	2.0520	38.9300	2.0983	0.0951
14#	1.6973	1.8840	43.9300	1.9264	0.0781
15#	1.3600	3.6200	11.4800	0.9900	0.3994
16#	1.3463	3.5834	11.3641	1.9900	0.3692
17#	1.3325	3.5469	11.2483	2.9900	0.3427
18#	1.2501	3.3275	10.5524	8.9900	0.2342
19#	1.3293	3.5383	13.4800	0.9676	0.3369
20#	1.2832	3.4156	16.4800	0.9341	0.2698
21#	1.2064	3.2111	21.4800	0.8782	0.1976
22#	0.1800	0.2800	10.9300	1.0200	0.0385
23#	0.1782	0.2772	10.8196	2.0200	0.0355
24#	0.1764	0.2743	10.7094	3.0200	0.0328
25#	0.1654	0.2574	10.0464	9.0200	0.0222
26#	0.1760	0.2737	12.9300	0.9971	0.0323
27#	0.1699	0.2643	15.9300	0.9628	0.0257
28#	0.1560	0.2486	20.9300	0.9055	0.0185
29#	0.1400	0.2200	5.0500	0.9900	0.0596
30#	0.1386	0.2178	4.9990	1.9900	0.0510
31#	0.1372	0.2156	4.9481	2.9900	0.0444
32#	0.1287	0.2022	4.6420	8.9900	0.0243
33#	0.1371	0.2154	7.0500	0.9692	0.0440
34#	0.1326	0.2084	10.0500	0.9379	0.0310
35#	0.1253	0.1968	15.0500	0.8857	0.0202
36#	2.8253	1.6713	60.5452	4.0700	0.0696
37#	2.7958	1.6539	59.9134	5.0700	0.0685
38#	2.7221	1.6103	58.3341	7.5700	0.0657
39#	1.9898	2.2086	33.7562	2.7700	0.1149
40#	1.9693	2.1859	33.4088	3.7700	0.1118
41#	1.9181	2.1291	32.5405	6.2700	0.1043
42#	1.3531	3.6016	11.4217	1.4900	0.3837
43#	1.3394	3.5651	11.3059	2.4900	0.3555
44#	1.3051	3.4738	11.0162	4.9900	0.2986
45#	0.1791	0.2786	10.8745	1.5200	0.0369
46#	0.1773	0.2758	10.7642	2.5200	0.0341
47#	0.1727	0.2687	10.4884	5.0200	0.0285
48#	0.1393	0.2189	5.0244	1.4900	0.0550
49#	0.1379	0.2167	4.9734	2.4900	0.0475
50#	0.1343	0.2111	4.8460	4.9900	0.0351

2.2 LIBS spectra collection

LIBS measurement was carried out using a Q-switched Nd:YAG laser (LOTIS, TII2131, Belarus) with a wavelength of 1064 nm, pulse energy of 83 mJ, repetition rate of 5 Hz and pulse duration of 10 ns. Iron ore samples were placed on a micro-auto *xyz* translation stage, and the laser beam was focused onto the iron ore sample surface with a 50 mm focal length plano-convex lens. Plasma emission was collected by employing a lens coupled to an optical fiber (with a 1000 nm core diameter and 0.22 numerical aperture) and detected using an Echelle spectrometer (ARYELLE-Butterfly, LTB400, Germany) equipped with an electron multiplying charge-coupled device (EMCCD) camera (QImaging, UV enhanced, 1004 × 1002 Pixels, USA), providing a constant spectral resolution (CSR) of 6000 over a wavelength range of 220–800 nm. The iron ore samples were measured directly at atmospheric pressure in air, and the optimal delay time was set to 3 μs. To reduce the influence of laser energy fluctuation on spectral intensities, each measured spectrum was collected by the accumulation of 50 laser pulses. In this study, the analytical spectra were the average of 50 spectra for each iron ore sample, and 50 (one LIBS spectra for each iron ore sample) analytical spectra were acquired from the 50 iron ore samples. Among these 50 iron ore samples, 35 samples (1–35#) were randomly selected as the calibration set, and the other 15 samples (36–50#) were selected as the prediction set. Matlab codes of the RF, PLS and LS-SVM algorithm can be obtained from references,[27–29] and all the calculations were implemented using MATLAB (version 2014a, Mathworks).

2.3 Random forests (RF)

Random forests are an advanced method of machine learning, and a decision tree ensemble algorithm constructs a set of statistical approximations to make a predictive model.[30,31] It combines the bootstrapping[32] and the random feature selection technology. In the training stage, b bootstrap sample sets as calibration sets were randomly drawn from n samples to construct b regression trees. In this step, approximately one-third of the dataset as the predictive set is left out in the calibration dataset to calibrate the performance of each tree. m_{try} ($m_{try} < p$) is randomly chosen from the p variable number of original data at each node of the tree (m_{try} represents the number of peaks at each node). The final predictive results for test samples are obtained by taking the average of the predictions of all the individual trees.

There are two important RF model parameters: (i) the number of trees in the forest (n_{tree}), and (ii) the number of peaks randomly selected as the candidates for splitting at each node of the tree (m_{try}). In the RF model, the default value of m_{try} is $\sqrt{M}$[33] (M represents the total number of spectral points in a spectrum). The default value is usually the case to obtain better results, but may not be able to achieve the best results. RF model parameters of n_{tree} and m_{try} can be optimized based on the out-of-bag (OOB) estimation and OOB error was used as the evaluation index. As a result, n_{tree} and m_{try} were 500 and 226 based on the OOB error ($m_{try} = \sqrt{M}$, $M = 51\,234$).

2.4 Variable importance measurements (VIM)

Input variables play a crucial role in modeling, and can be classified as informative variables and uninformative variables. Informative variables can improve the prediction performance of models, while uninformative variables may even have a harmful influence on analysis results and increase the calculation time. The RF model can identify the best subset

based on its inbuilt VIM between many variables (practically the most difficult part of modeling).

The computation of the VIM based on the RF consists of the following steps:

(1) Assuming that there is an original spectrum A, which can be used to establish the RF model;

(2) Generating B subsets of variables, and building B RF sub-models based on the obtained variable importance;

(3) Calculating the RMSECV of each sub-model;

(4) Comparing the RMSECV of the new sub-model with the RMSECV of the previous sub-model, and if the RMSECV of the new sub-model is larger than the RMSECV of the previous sub-model, the procedure will be repeated until no further improvement in the minimum RMSECV of the new sub-models is obtained;

(5) Obtaining the optimal variable importance;

(6) Constructing the optimal VIM-RF model based on the optimal variable importance.

A flow chart of the optimal VIM-RF model construction is described in Fig. 1.

3. Results and discussion

3.1 LIBS spectra analysis of iron ores

Fig. 2 shows the average LIBS spectra of the 1# iron ore sample in the spectral region of 245–510 nm, and the characteristic spectral lines of Ca, Mg, Si and Al elements in iron ore samples were identified based on the National Institute of Standards and Technology (NIST) database,[34] and are summarized in Table 2. Iron ore as a complex mineral mainly includes CaO, MgO, SiO_2 and Al_2O_3, and its acidity can be calculated with the CaO, MgO, SiO_2 and Al_2O_3 content. In other words, the acidity of iron ore can be directly determined using the spectral line intensity of relevant elements. But due to some uncertain factors, such as instrumental, environmental and complex matrix effects, the relationship between the acidity of iron ore and single spectral line intensity may not be satisfactory in quantitative analysis of LIBS. Beyond that, as can be seen from Fig. 2, there is much background information and noise around the analysis lines in the raw LIBS spectra. Consequently, some LIBS spectral pretreatment is necessary to construct a good RF calibration model.

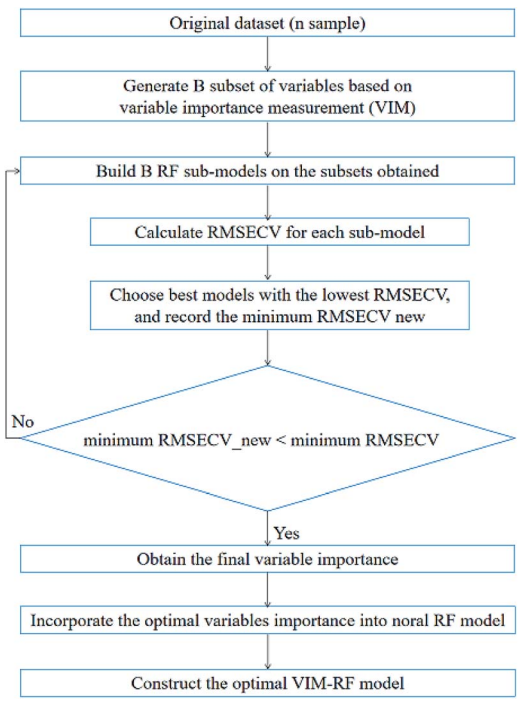

Fig. 1 The flow chart of the optimal VIM-RF model construction.

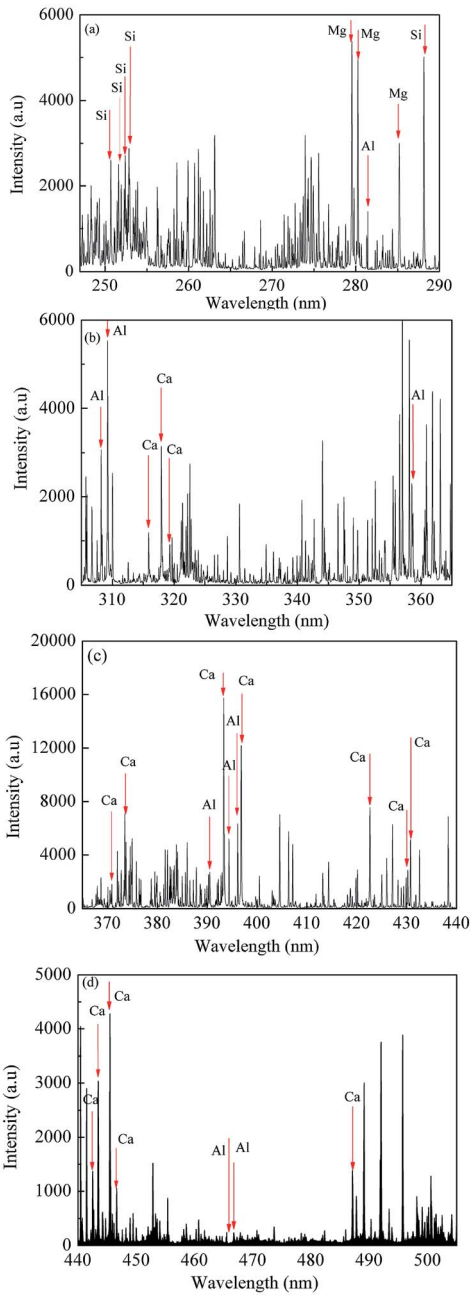

Fig. 2 The average LIBS spectrum of the 1# iron ore sample in the spectral regions of (a) 245–290 nm, (b) 305–365 nm, (c) 365–440 nm, and (d) 440–510 nm.

Table 2 Spectral lines for the quantitative analysis of the acidity of iron ore

Element	Wavelength (nm)
Si	250.69, 251.61, 252.41, 252.85, 288.15
Al	281.62, 308.22, 309.27, 358.66, 390.06, 394.40, 396.15
Mg	279.55, 280.27, 285.21
Ca	315.89, 317.93, 318.13, 370.60, 373.69, 393.37, 396.85, 422.67, 430.25, 430.77, 442.54, 443.57, 445.47, 445.66, 487.81

3.2 The selection of pretreatment methods of LIBS spectra on iron ore

In this study, several pretreatment methods based on first derivatives,[35] second derivatives,[35] wavelet transform[36] (WT) and normalization were utilized to address original LIBS spectra and enhance the RF model accuracy. The predictive performance of different pretreatment methods is evaluated by 5-fold cross-validation (CV) and two evaluation indices of correlation coefficient (R^2) and root mean squared error (RMSE). Fig. 3 shows the RMSECV and R_{CV}^2 values of the RF model with LIBS spectra addressed by different pretreatment methods as input variables. It can be seen from Fig. 3 that the minimum RMSECV and maximum R_{CV}^2 are achieved using the second derivative for CaO, MgO, SiO_2 and acidity analysis of iron ore. However, for Al_2O_3 analysis of iron ore, the first derivative combined with normalization shows a better predictive performance with a smaller RMSECV value and a higher R_{CV}^2 value. Three pretreatment methods of normalization, WT and first derivative show a relatively poor performance, and the obtained RMSECV and R_{CV}^2 values were worse than those of the raw spectra. Thus, the second derivative was selected for CaO, MgO, SiO_2 and acidity analysis in iron ore, and the first derivative combined with normalization was used for Al_2O_3 analysis in iron ore.

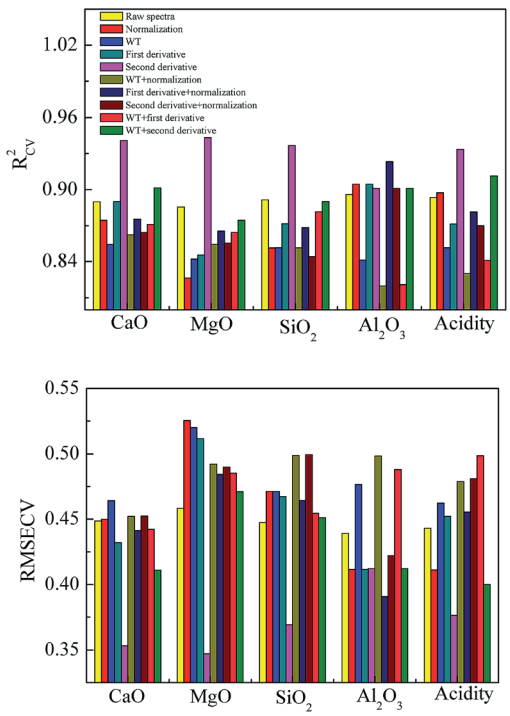

Fig. 3 R_{CV}^2 and RMSECV obtained with different pretreatment methods: bars correspond to raw spectra followed by normalization, WT, first derivative, second derivative, WT + normalization, first derivative + normalization, second derivative + normalization, WT + first derivative, and WT + second derivative.

After LIBS spectral pretreatment, the predictive performance of the RF calibration model was investigated by using 5-fold cross-validation (CV) and the prediction set (shown in Table 3). For 5-fold CV, the R_{CV}^2 values of the RF calibration model were above 0.9200 for CaO, MgO, SiO_2, Al_2O_3 and acidity analysis; for the prediction set, the R_P^2 values were over 0.9300 for CaO, MgO and SiO_2 analysis. However, the R_P^2 values of Al_2O_3 and acidity analysis were only 0.9023 and 0.8649 in the prediction set. In addition, the whole LIBS spectrum of iron ore samples contains 51 260 variables in this study, and it is a complex LIBS spectrum with much characteristic information, interference spectra and noise. Besides, a few input variables might result in informative variables loss. Hence, it is necessary to select an effective input variable of the RF calibration model for Al_2O_3 and acidity analysis in iron ore.

3.3 Optimization of RF model input variables

Seven different wavelength regions containing typical emission lines of the Ca, Mg, Si and Al elements were necessary to optimize and obtain appropriate input variables to reduce the modelling time and improve the predictive ability of the RF model (shown in Fig. 4). As can be seen from Fig. 4, compared to other spectral regions, the RF models with the regions of 220–450 nm and 220–500 nm as input variables achieve the highest R_{CV}^2 and the smallest RMSECV for Al_2O_3 and acidity, respectively. The RF models with the raw spectra (200–800 nm) as input variables have a significantly better performance than the other spectral regions for CaO, MgO, and SiO_2 analysis. In addition, the optimal wavelength region produces a slight reduction in RMSE and increase in R^2 values for the RF model in Al_2O_3 and acidity analysis.

We then implemented the variable importance measurement of the RF model together with cross-validation. The variable importance of each LIBS spectral intensity is shown in Fig. 5. As can be seen, most of the variable importance of the

Table 3 The results of the RF model with the full LIBS spectrum addressed for quantitative analysis of CaO, MgO, SiO_2, Al_2O_3 and acidity

	5-fold CV		Prediction	
Components	R_{CV}^2	RMSECV (wt%)	R_P^2	RMSEP (wt%)
CaO	0.9408	0.3533	0.9837	0.1611
MgO	0.9432	0.3471	0.9352	0.3634
SiO_2	0.9366	0.3691	0.9455	5.7809
Al_2O_3	0.9233	0.3908	0.9023	0.5959
Acidity	0.9300	0.3765	0.8649	0.0624

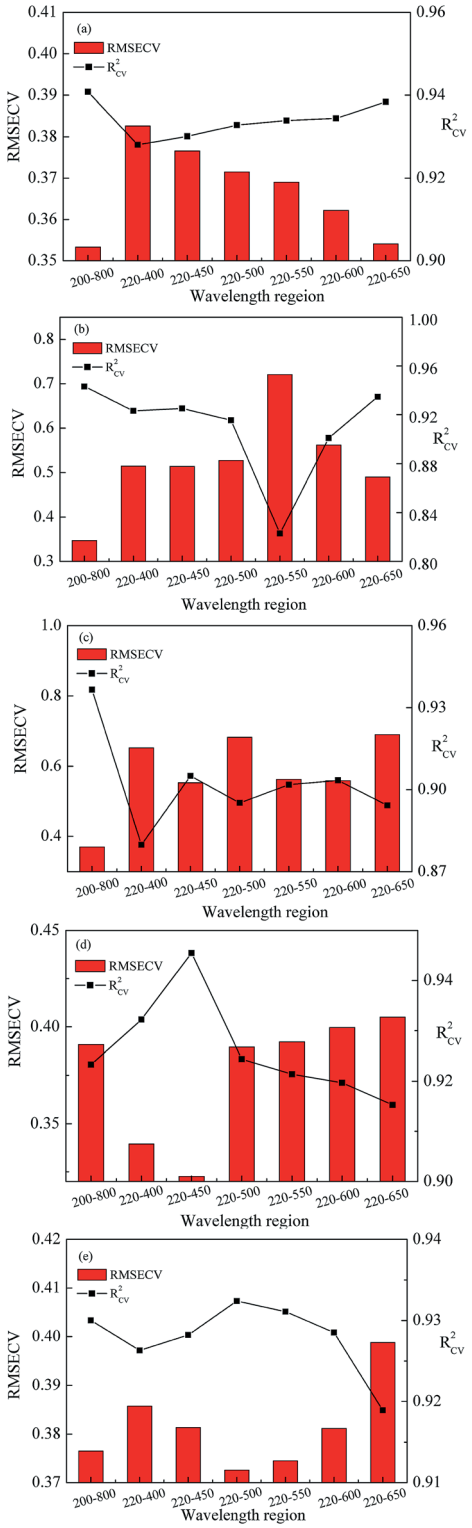

Fig. 4 Influence of different wavelength ranges on the RF model (a: CaO, b: MgO, c: SiO_2, d: Al_2O_3, and e: acidity).

original RF model is mainly distributed in the range of 0–0.1 corresponding to LIBS wavelengths of **220–500 nm** for acidity analysis. Fig. 6 presents the variations of the R_{CV}^2 and RMSECV with different variable importance threshold values for CaO, MgO, SiO_2, and Al_2O_3 analysis. Table 4 lists the effect of different variable importance on the VIM-RF model for MgO analysis. The minimum RMSECV and maximum R_{CV}^2 corresponded to the optimum variable importance. For MgO, the RMSECV decreased continuously, R_{CV}^2 increased gradually along with the threshold value of variable importance in the range of 0–0.08, the RMSECV reached a minimum value, and R_{CV}^2 reached a maximum value when the threshold value was set to 0.08. However, the RMSECV increased continually, and R_{CV}^2 decreased continually when the threshold value was over 0.08. The VIM-RF modeling time was reduced from 39.46 s (full spectrum) to 0.126 s (variable importance threshold value was 0.08) for MgO analysis. Therefore, the LIBS spectra with a variable importance threshold value of 0.08 as an input variable were selected to construct the VIM-RF model. However, for the CaO, SiO_2 **and Al_2O_3** analysis, the effects of different variable importance on the VIM-RF model for CaO, SiO_2 **and Al_2O_3** analysis are listed in Tables S1–S3.† No improvement was achieved after variable importance measurement. A relatively low R_{CV}^2 and a relatively high RMSECV value of the VIM-RF model are obtained compared to the full spectral model. For CaO, although the modeling time was reduced from 40.25 s to 0.665 s, the R_{CV}^2 value was decreased from 0.9408 to 0.9328, and RMSECV was enhanced from 0.3533 to 0.3557 wt%; for SiO_2, although the modeling time was reduced from 40.12 s to 0.365 s, the R_{CV}^2 value was reduced from 0.9366 to 0.9148 and RMSECV was increased from 0.3691 to 0.3999 wt%; for Al_2O_3, although the modeling time was reduced from 19.76 s to 2.260 s, the R_{CV}^2 value was decreased from 0.9455 to 0.8911 and RMSECV was increased from 0.3266 to 0.4551 wt%. This means that variable importance measurement in the RF modeling process may lose some important information for CaO, SiO_2 **and Al_2O_3** analysis, and is not suitable for the quantitative analysis of all oxides in iron ore.

For acidity analysis, Fig. 7 and Table 5 present the influence of different variable importance on the VIM-RF model. When the feature spectral bands (220–500 nm) were used as input variables for the RF model, the $R_{CV}^2$and RMSECV were 0.9324 and 0.3726 wt%, respectively. As soon as the variable importance threshold value was set to be 0, the obtained R_{CV}^2 and RMSECV were 0.9381 and 0.3615 wt%, respectively. At the same time, the VIM-RF modelling time was decreased from

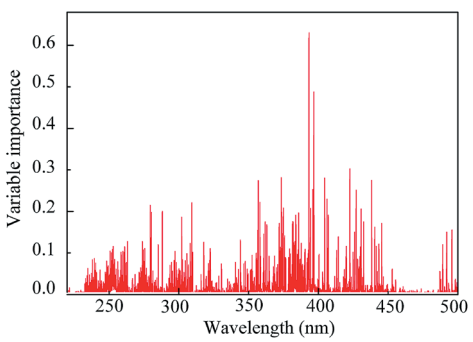

Fig. 5 The relationship between the variable importance of the RF model and LIBS wavelength.

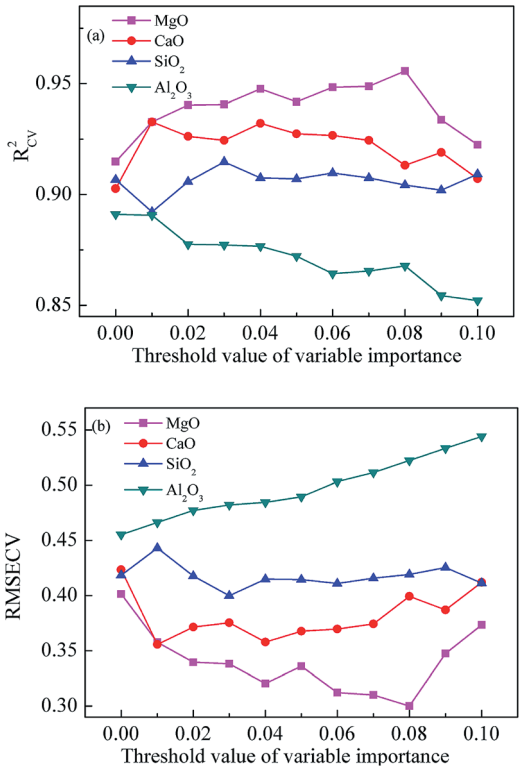

Fig. 6 The relationship between the variable importance threshold value and prediction performance of the VIM-RF model (a: R_{CV}^2; b: RMSECV).

27.13 s (the spectral range of 200–500 nm) to 2.737 s. When the variable importance threshold value continues to increase up to 0.0007 (variable number of 107), the best prediction performance of the VIM-RF model was achieved, and the largest R_{CV}^2 and lowest RMSECV were 0.9554 and 0.3123 wt%, respectively. The VIM-RF modeling time was only 0.106 s. With a variable importance threshold value of 0.0007, the R_{CV}^2 value of the VIM-RF model was increased from 0.9324 (the spectral range of 200–500 nm) to 0.9554, and the RMSE value was reduced from 0.3726 wt% to 0.3123 wt%. It is worth mentioning that there are only 107 input variables for acidity

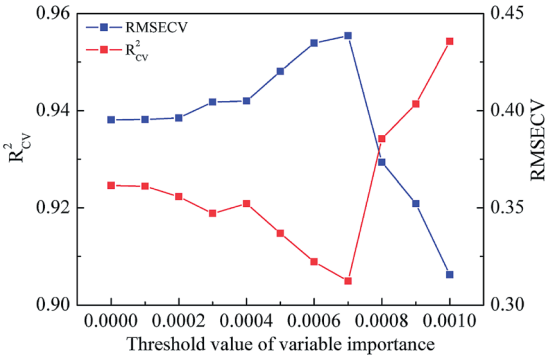

Fig. 7 The relationship between variable importance and R_{CV}^2 and RMSECV for acidity analysis.

analysis based on the VIM-RF model. Thus, the constructed VIM-RF model can explicitly make full use of informative variables, and discard the influence of uninformative variables. What' more, the number of remaining variables was greatly reduced after variable importance extraction.

3.4 Quantitative analysis of acidity of iron ore

The traditional artificial neural network (ANN) training process is quite slow and easy to fall into local optimum, and SVM requires long training time and high computational complexity. Thus, in order to validate the predictive ability of the VIM-RF model for acidity analysis in iron ore in this article, we compared the VIM-RF method with the PLS and LS-SVM methods using the same preprocessing conditions and input variables. For the PLS calibration model, the optimal latent variables obtained by 5-fold CV for CaO, MgO, SiO_2, Al_2O_3 and acidity were 5, 6, 5, 9 and 9, respectively. For the LS-SVM calibration model, two hyperparameters C and γ were optimized by the grid-search method and 5-fold CV. The optimum parameters of LS-SVM were set as: the (C, γ) are (80, 14 404) for CaO, (69, 43 669) for MgO, (163, 23 742) for SiO_2, (178, 47 096) for Al_2O_3 and (166, 34 604) for acidity. The 5-fold CV was used to internally validate the predictive performance of optimized VIM-RF, PLS and LS-SVM models (shown in Fig. 8). It can be

Table 4 The prediction performance of the VIM-RF model with different variable importance for MgO analysis

Variable importance	Variable numbers	R_{CV}^2	RMSECV (wt%)	Modeling time (s)
Full spectrum	51 234	0.9432	0.3471	39.46
0	3355	0.9149	0.4015	2.250
0.01	824	0.9328	0.3576	0.559
0.02	538	0.9404	0.3395	0.380
0.03	410	0.9406	0.3381	0.279
0.04	389	0.9476	0.3204	0.264
0.05	360	0.9418	0.3359	0.249
0.06	315	0.9483	0.3122	0.241
0.07	217	0.9488	0.3100	0.144
0.08	145	0.9557	0.3001	0.126
0.09	86	0.9337	0.3474	0.088
0.10	61	0.9225	0.3733	0.069

Table 5 Variable numbers, RMSECV, R_{CV}^2 values and modeling time with different variable importance for acidity analysis

Variable importance	Variable numbers	R_{CV}^2	RMSECV (wt%)	Modeling time (s)
220–500	30 955	0.9324	0.3726	27.13
0	3383	0.9381	0.3615	2.737
0.0001	316	0.9382	0.3612	0.296
0.0002	211	0.9385	0.3558	0.209
0.0003	191	0.9418	0.3472	0.190
0.0004	171	0.9420	0.3522	0.176
0.0005	153	0.9481	0.3369	0.162
0.0006	126	0.9539	0.3222	0.139
0.0007	107	0.9554	0.3123	0.106
0.0008	90	0.9294	0.3855	0.093
0.0009	77	0.9209	0.4034	0.077
0.001	68	0.9062	0.4356	0.062

seen from Fig. 8 that the R_{CV}^2 values of the VIM-RF model were all over 0.93, and the RMSECV value of the VIM-RF model was relatively lower than that of the other two models. The VIM-RF model shows a significantly better performance than PLS and LS-SVM for quantitative analysis of CaO, MgO, SiO_2, Al_2O_3 and acidity in iron ore.

To further validate the predictive ability of VIM-RF, PLS and LS-SVM models, the external validation of the three calibration models was implemented by using the prediction set (shown in Table 6). As we can see in Table 6, the average R^2 value of the VIM-RF model is 0.95, with a relatively low RMSE value; the average R^2 value of the PLS model is 0.82; the average R^2 value of the LS-SVM model is 0.88. All these results show that the VIM-RF model has a significantly better performance than PLS and LS-SVM. What's more, for acidity analysis, the R^2 and RMSE values of the VIM-RF model are obviously better than those of PLS and LS-SVM. This is because, on one hand, the content of SiO_2 is greater than that of CaO, MgO and Al_2O_3 in iron ore, and the analysis lines of Si with low excitation energies at upper levels and high transition probabilities were easily subjected to self-absorption effects. The VIM-RF model can automatically exclude irrelevant variables as well as the variables subjected to spectral interference and severe self-absorption by variable importance measurement. On the other hand, the acidity of the iron ore was directly calculated by LIBS coupled with the VIM-RF model, rather than using the intensity ratio of the characteristic lines of the four elements or the concentration ratio of the sum of CaO and MgO to the sum of Al_2O_3 and SiO_2 was used as an input variable to construct the quantitative analysis model of acidity (shown in Table 7). A comparison between the VIM-RF model and concentration ratio (CaO + MgO)/(Al_2O_3 + SiO_2) was performed. As can been seen from Table 7, the R^2 and RMSE values of acidity obtained with the VIM-RF model were higher than those of acidity obtained with the concentration ratio (CaO + MgO)/(Al_2O_3 + SiO_2). The acidity was calculated directly using the relationship between the acidity value and LIBS spectral intensity using the VIM-RF model, and larger R^2 and smaller RMSE values were produced by VIM-RF. Since the spectral line intensity of a single element is easily affected by the emission lines of other strong elements and the complex matrix effect, a large RMSE is produced for the acidity obtained with the concentrations ratio (CaO + MgO)/(Al_2O_3 + SiO_2). Hence, the accuracy of acidity quantitative analysis is significantly improved by using LIBS combined with the VIM-RF model.

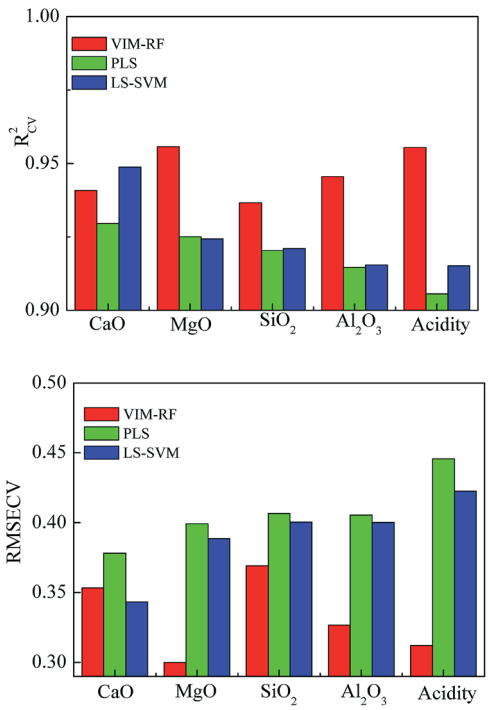

Fig. 8 Predictive performance of PLS, LS-SVM and VIM-RF models based on 5-fold cross-validation.

This proposed method can achieve rapid and real-time online analysis of the acidity of iron ore in the metallurgical industry, and it is very beneficial to control metallurgical materials, process analysis and selection and quality of iron ore. This approach in this study not only shortens analysis time and improves production efficiency, but also overcomes the shortcomings of the waste of raw materials and energy in the mineral industry. Therefore, we believe that this approach will be highly valued in the metallurgical field and among mining entrepreneurs.

Table 6 Performance comparison between VIM-RF, PLS and LS-SVM

	VIM-RF		PLS		LS-SVM	
Components	R^2	RMSE (wt%)	R^2	RMSE (wt%)	R^2	RMSE (wt%)
CaO	0.9837	0.1611	0.9594	0.2269	0.9918	0.1329
MgO	0.9848	0.1738	0.7914	0.5854	0.8697	0.4131
SiO_2	0.9455	5.7809	0.8719	7.8018	0.8739	6.5724
Al_2O_3	0.9354	0.5138	0.7843	1.3453	0.8047	0.9897
Acidity	0.9103	0.0554	0.7166	0.0628	0.8899	0.0536

Table 7 A predictive ability comparison of different predictive methods of acidity

The predictive methods of acidity	R^2	RMSE (wt%)
VIM-RF	0.9103	0.0554
Concentration ratio (CaO + MgO)/(Al_2O_3 + SiO_2)	0.8726	0.0622

4. Conclusion

LIBS combined with the VIM-RF model was successfully applied for rapid analysis of the acidity of iron ore. LIBS spectra of 50 iron ore samples were acquired by the LIBS technique, and the characteristic lines of major elements (Ca, Mg, Si and Al) in iron ore were identified using the NIST database. To obtain a better acidity quantitative result, the LIBS spectra were subjected to nine different pre-processing methods (normalization, WT, first derivative, second derivative, WT-normalization, first derivative-normalization, second derivative-normalization, WT-first derivative, and WT-second derivative), and the LIBS spectra processed using the second derivative as the input variable were used to construct the RF calibration model for acidity analysis. The effect of different variable importance values (from 0 to 0.001) on the VIM-RF model for acidity analysis was investigated, and R_{CV}^2 and RMSECV were used as the assessment criteria. The LIBS spectra processed using the second derivative with a variable importance value of 0.0007 as the input variable were used to construct the optimal VIM-RF model for acidity analysis. Then the optimized VIM-RF model was used for acidity analysis, and the obtained results were compared with those obtained with PLS and LS-SVM models. The obtained results showed that the VIM-RF model showed better predictive ability, with RMSE = 0.0554 wt%, and R^2 = 0.9103 for acidity analysis in the prediction set. The obtained results sufficiently demonstrated that LIBS coupled with VIM-RF is a practical technique for rapid and on-line analysis of acidity of iron ore and it will provide a new method and technology for selection and quality control of iron ore in the metallurgical industry.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 21873076, 21675123 and 21605123), Natural Science Basic Research Plan in Shaanxi Province of China (No. 2018JQ2013), Scientific Research Plan Projects of Shaanxi Education Department (No. 17JK0780), Northwest University Graduate Innovation and Creativity Funds (No. YZZ17126) and Natural Science Foundation of the Jiangsu Higher Education Institutions of China (17KJB535002).

References

1 Z. Q. Hao, C. M. Li, M. Shen, X. Y. Yang, K. H. Li, L. B. Guo, X. Y. Li, Y. F. Lu and X. Y. Zeng, *Opt. Express*, 2015, **23**, 7795–7801.
2 J. Merten and B. Johnson, *Spectrochim. Acta, Part B*, 2018, **149**, 124–131.
3 J. Giersz, K. Jankowski, A. Ramsza and E. Reszkec, *Spectrochim. Acta, Part B*, 2018, **147**, 51–58.
4 S. Veyseh and A. Niazi, *Talanta*, 2016, **147**, 117–123.
5 F. J. Fortes, J. Moros, P. Lucena, L. M. Cabalín and J. J. Laserna, *Anal. Chem.*, 2012, **85**, 640–669.
6 J. Cortez and C. Pasquini, *Anal. Chem.*, 2013, **85**, 1547–1554.
7 N. L. Lanza, S. M. Clegg, R. C. Wiens, R. E. Mcinroy, H. E. Newsom and M. D. Deans, *Appl. Opt.*, 2012, **51**, 74–82.
8 P. Yaroshchyk, D. L. Death and S. J. Spencer, *J. Anal. At. Spectrom.*, 2012, **27**, 92–98.
9 D. L. Death, A. P. Cunningham and L. J. Pollard, *Spectrochim. Acta, Part B*, 2009, **64**, 1048–1058.
10 L. W. Sheng, T. L. Zhang, G. H. Niu, K. Wang, H. S. Tang, Y. X. Duand and H. Li, *J. Anal. At. Spectrom.*, 2015, **30**, 453–458.
11 L. Liang, T. L. Zhang, K. Wang, H. S. Tang, X. F. Yang, X. Q. Zhu, Y. X. Duan and H. Li, *Appl. Opt.*, 2014, **53**, 544–552.
12 T. L. Zhang, D. H. Xia, H. S. Tang, X. F. Yang and H. Li, *Chemometr. Intell. Lab. Syst.*, 2016, **157**, 196–201.
13 T. L. Zhang, L. Liang, K. Wang, H. S. Tang, X. F. Yang, Y. X. Duan and H. Li, *J. Anal. At. Spectrom.*, 2014, **29**, 2323–2329.
14 T. L. Zhang, S. Wu, J. Dong, J. Wei, K. Wang, H. S. Tang, X. F. Yang and H. Li, *J. Anal. At. Spectrom.*, 2015, **30**, 368–374.
15 V. Sturm, R. Fleige, M. de Kanter, R. Leitner, K. Pilz, D. Fischer, G. Hubmer and R. Noll, *Anal. Chem.*, 2014, **86**, 9687–9692.
16 C. H. Yan, Z. M. Wang, F. Q. Ruan, J. X. Ma, T. L. Zhang, H. S. Tang and H. Li, *Anal. Methods*, 2016, **8**, 6216–6221.
17 K. J. Grant, G. L. Paul and J. A. O'Neill, *Appl. Spectrosc.*, 1991, **45**, 701–705.

18 S. Makvandi, M. Ghasemzadeh-Barvarz, G. Beaudoin, E. C. Grunsky, M. B. Mc Clenaghan, C. Duchesne and E. Boutroy, *Ore Geol. Rev.*, 2016, **78**, 388–408.
19 U. König, T. Degen and N. Norberg, *Powder Diffr.*, 2014, **29**, S78–S83.
20 D. L. Death, A. P. Cunningham and L. J. Pollard, *Spectrochim. Acta, Part B*, 2008, **63**, 763–769.
21 C. L. Bérubéa, G. R. Olivo, M. Chouteau, S. Perrouty, P. Shamsipour, R. J. Enkin, W. A. Morris, L. Feltrin and R. Thiémonge, *Ore Geol. Rev.*, 2018, **96**, 130–145.
22 Y. M. Guo, L. B. Guo, Z. Q. Hao, Y. Tang, S. X. Ma, Q. D. Zeng, S. S. Tang, X. Y. Li, Y. F. Lu and X. Y. Zeng, *J. Anal. At. Spectrom.*, 2018, **33**, 1330–1335.
23 Y. Ding, F. Yan, G. Yang, H. X. Chen and Z. S. Song, *Anal. Methods*, 2018, **10**, 1074–1079.
24 L. Breiman, *Mach. Learn.*, 2001, **45**, 5–32.
25 H. S. Tang, T. L. Zhang, X. F. Yang and H. Li, *Anal. Methods*, 2015, **7**, 9171–9176.
26 Y. Tian, C. H. Yan, T. L. Zhang, H. S. Tang, H. Li, J. L. Yu, J. Bernard, L. Chen, S. Martin, N. Delepine-Gilon, J. Bocková, P. Veis, Y. P. Chen and J. Yu, *Spectrochim. Acta, Part B*, 2017, **135**, 91–101.
27 L. Breiman, A. Cutler, *Random Forest*, https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.
28 H. D. Li, Q. S. Xu and Y. Z. Liang, *Chemom. Intell. Lab. Syst.*, 2018, **176**, 34–43.
29 C. C. Chang and C. J. Lin, *LIBSVM – A Library for Support Vector Machines*, http://www.csie.ntu.edu.tw/cjlin/libsvm/.
30 L. Auret and C. Aldrich, *Miner. Eng.*, 2012, **35**, 27–42.
31 S. Janitza, G. Tutz and A. L. Boulesteix, *Comput. Stat. Data Anal.*, 2016, **96**, 57–73.
32 A. C. Cheng and M. Yeager, *J. Struct. Biol.*, 2007, **158**, 19–32.
33 A. Liaw and M. Wiener, *R. News*, 2002, **2/3**, 18–22.
34 https://physics.nist.gov/PhysRefData/Handbook/periodictable.htm.
35 Y. X. Yu, H. Y. Yu, L. B. Guo, J. Li, Y. W. Chu, Y. Tang, S. S. Tang and F. Wang, *Anal. Methods*, 2018, **10**, 3224–3231.
36 Y. M. Guo, L. M. Deng, X. Y. Yang, J. M. Li, K. H. Li, Z. H. Zhu, L. B. Guo, X. Y. Li, Y. F. Lu and X. Y. Zeng, *J. Anal. At. Spectrom.*, 2017, **32**, 2401–2406.