Cite this: DOI: 10.1039/c0xx00000x

www.rsc.org/xxxxxx

ARTICLE TYPE

Classification of different types of slag samples by Laser-induced breakdown spectroscopy(LIBS) coupled with random forest based on variable importance(VIRF)

Hongsheng Tang,[a] Tianlong Zhang,[a*] Xiaofeng Yang[b] and Hua Li[a,c*]

Received (in XXX, XXX) Xth XXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX
DOI: 10.1039/b000000x

Laser induced breakdown spectroscopy(LIBS) technique coupled with random forest based on variable importance(VIRF) was proposed to perform classification analysis of slag samples. Three types of slag samples(open-hearth furnace slag, converter slag and high titanium slag) were identified and classified by random forest(RF) method with different pre-processing methods(normalized with maximum integrated intensity, first-order derivative and second-order derivative) and different input variables(200-300, 200-400, 200-500, 200-600, 200-700 and 200-800nm), the importance of input variable was employed to improve the classification performance of RF model for slag samples. Averaged OOB(out-of-bag) error, sensitivity, specificity and accuracy were calculated to evaluate the classification performance of RF model for slags. Normalized by maximum integrated intensity LIBS spectra(200-500nm) of slag samples as input variable was constructed the PLS-DA, SVM, RF and VIRF model for the classification analysis of slags. VIRF model shows a better classification performance than other three model. LIBS technique coupled with RF perhaps is a promising approach to achieve the online analysis and process control of slag and even industrial waste recycling.

1. Introduction

In steel-making industry, a large amounts of solid co-products is generated in the form of slag and sludge. The world's annual output of slag from iron and steel industries reaches almost 50 million tons.[1] Slag as an significant byproduct in steel industry plays a decisive role in ensuring smelting operation smoothly, steel quality, metal recovery and so on. There are different types of steel industry slags such as blast furnace slag (BF) also called iron slag, basic oxygen furnace slag (BOF), electric arc furnace acid slag (EAF), ladle furnace basic slag (LF) also called refining slag and so on. Slag can be divided into smelting slag, refining slag and synthetic slag according to the difference of metallurgical process; it can be also divided into acid slag, neutral slag and alkaline slag. The major component of slag include CaO, SiO_2, Al_2O_3, MgO, Fe_2O_3, TiO_2 and so on. Each type of slag has its typical chemical, mineralogical, and physical properties. Classification and identification of slag contributes to recycling and reuse of metallurgical waste. There are many significant applications on slag such as blast furnace slag can be used as a cement raw material, high phosphorus slag can be used as a fertilizer, vanadium slag can be used as raw materials for the refining of vanadium, and so on. V. Gupta[2] focused on the reuse of slag material as low cost adsorbents for water treatment.

Slag analysis was performed by different techniques[3,4] such as chemical analysis, X-ray fluorescence(XRF), inductively coupled plasma optical emission spectroscopy(ICP-OES), mass spectroscopy(MS) and so on. However, these approaches require complicated sample preparation and much analysis time, which fails to timely obtain the information of steel product, and even hinders their application for real time and fast analysis. Laser induced breakdown spectroscopy(LIBS) is a promising and prospect analytical technique based on laser plasma spectroscopy with the advantages of multiple elements simultaneous analysis for all types of the samples(solid, liquid, gas or aerosol).[5-7] At present, the LIBS technique has become the subject of metallurgical analysis.[8-10] The application of LIBS technique to metallurgical industry includes iron ore selection,[11,12] process control,[13,14] iron slag analysis[15-18] and so on.

Classification and identification of slag has been the subject of most government environmental agencies in the world.[19,20] The classification of slags by LIBS can be fulfilled depends on difference of its major component and corresponding concentration. In other words, it was completed by the difference comes from spectra integrated intensity and wavelength of LIBS spectra. The LIBS technique combined with chemometrics methods is an effective approach to improve the classification performance of slags. Zhang et al[20] employed LIBS coupled with partial least squares discriminant analysis(PLS-DA) to classify open-hearth furnace slag and high titanium slag. However, PLS-DA has some disadvantages such as low classification accuracy and overfitting. Random Forest(RF), a new classification algorithm based on multiple classifier, was proposed by Leo Breiman[21]. It is an ensemble of unpruned classification tree created by using bootstrap samples of the training data and random feature selection in tree introduction. Prediction was completed by the majority vote of multiple classifiers to determine the final category for the test samples. The train dataset was used to construct the multiple classifier, and the final category of the predictive sample are determined by the major vote of the classification results for each classifier. It has proved that RF classifier has a good tolerance for noise, as well as avoid over-fitting phenomenon.[22] In addition, LIBS combined with RF was applied for identification and classification of rocks, pen ink and iron ore samples.[23,24] LIBS combined with RF also could be used for the quantitative analysis of multiple elements in fourteen steel samples.[25]

The present work explores the combination of LIBS technique and RF based on variable importance(VIRF) for classification analysis of slags. A series of 60 slag samples were compressed into pellets and prepared for LIBS measurement. Three types of slag samples(open-hearth furnace slag, converter slag and high titanium slag) were identified and classified by RF method with different pre-processing methods (normalized with maximum integrated intensity, first-order derivative and second-order derivative) and different input variables(200-300, 200-400, 200-500, 200-600, 200-700 and 200-800nm), the variable importance was employed to improve classification performance of RF model for slag samples. Averaged OOB(out-of-bag) error, sensitivity, specificity and accuracy were calculated to evaluate the classification performance of RF model for slags.

2. Methodology

2.1. LIBS setup and acquisition conditions

The detailed description of LIBS setup was shown in the previous works.[29] A Q-switched Nd: YAG laser(λ=1064nm, 10 ns pulse FWHM, 80 mJ/pulse, repetition rate of 5 Hz) was used to generate the plasma in air at atmosphere pressure on the pellets. The pulse laser beam was focused onto the slag sample surface vertically by a 50mm focal-distance lens, which was generated a spot of about 0.2 mm diameter. The emission from the plasma created was collected with a 4-mm aperture, with a 7mm focus fused silica collimator placed at 45° angel with respect to the laser pules and a distance of 3 cm from the sample, and then focused into an optical fiber, which was coupled to the entrance of the Echelle spectrometer Aryelle 400(LTB, German). The spectrometer provides a constant spectral resolution (CSR) of 6000 over a wavelength range 200-800nm displayable in a single spectrum. An electron-multiplying CCD(EMCCD) camera (QImaging, UV enhanced, 1004 ×1002 Pixels, USA) coupled to the spectrometer was used for detection of the dispersed light. The overall linear dispersion of the spectrometer camera system ranges from 37 pm/pixel(at 220nm) to 133 pm/pixel (at 800nm). To prevent the EMCCD from detecting the early plasma continuum, a mechanical chopper is used in front of the entrance slit. The experiments were carried out under atmosphere condition, and the gate width of spectrometer was set to 2 ms. The detector was set to 1.5 µs delay time between the laser pulse in order to prevent the detection of bremsstrahlung radiation.

2.2. Slag samples and LIBS measurements

A total of 60 slag samples for three types of slag(open-hearth furnace slag(OHFS), converter slag(CS) and high titanium slag(HTS)) were provided by Pangang group Chengdu ore & steel Co., Ltd(China) in the form of slag powder. Each slag sample was homogenized to produce a very fine powder until all of the powder passed through a 200-mesh stainless steel sieve using a ball grinding mill. Table 1 lists the concentration(wt. %) of major component of three types of slag samples. There are 20 slag samples for each type of slags. The slag pellet was made with a tablet press at 400 Mpa for 5 min. LIBS spectra of 20 different position of each sample surface are gathered. In order to decrease the effects of shot to shot fluctuations, each measure spectrum was obtained by accumulation of 50 laser pulses. The total of the spectra for 60 slag sample was 1200(20 LIBS spectra for each slag sample). The training set and test set were selected by Kennard-Stone algorithm[26] with the ratio of the number of train samples to the whole data is 0.7. The number of training samples is 42, and the number of test samples is 18. The data processing and classification analysis of slag samples were performed on Matlab2007a(Mathworks).

Table 1 The concentration(wt%) of the major components in slag samples

slags	Fe_2O_3	SiO_2	TiO_2	CaO	MgO	MnO_2	Al_2O_3
open-hearth furnace slag	16.15-19.66	16.41-20.22	0.40-1.06	20.56-27.81	15.68-21.27	12.35-15.85	6.18-7.80
converter slag	13.96-14.88	12.05-12.66	--	47.95-48.38	8.01-8.38	--	0.69-0.95
high titanium slag	2.90-5.98	2.12-4.20	81.57-85.57	0.43-0.77	2.52-3.05	1.07-1.45	3.29-3.97

2.3 Random forest(RF)

RF is an advanced classification and regression method based on statistical learning theory. A resampling technique based on bootstrap method was used to continuously generates training and test samples; the training sample generates multiple classification tree form with random forest, the final predictions results based on the combination are received by a simple majority voting of the single classification tree. The process of RF training model described in previous work.[27] There are two significant parameters in RF: (1) the number of ensemble trees in the forest (**ntree**) and (2) the number of peaks randomly selected as the candidates for splitting at each node ($\mathbf{m_{try}}$). The ideal random forest model not only has higher classification accuracy and stability, but also has higher efficiency. Theoretically, the generalization error of the classifier tends to a finite upper bound when **ntree** reaches a certain value. In other words, if **ntree** is increased above the optimum, there is a general increase in the computational expense, but the results do not improve significantly. In this work, **ntree** was set as 500. $\mathbf{m_{try}}$ is one of the most major characteristic through each division that introduces random nodes for randomly selected attributes. It was assumed that there were M attributes(wavelengths) in the training sample, and $\mathbf{m_{try}}$ attributes were extracted randomly as candidate attribute between each of the internal nodes in the decision tree ($\mathbf{m_{try}} \ll M$). **ntree** and $\mathbf{m_{try}}$ for the RF model can be optimized by the OOB error estimation. Moreover, $\mathbf{m_{try}} = \sqrt{M}$ was found to be the best choice based on the OOB error rate.[21]

A significant characteristic of RF is able to calculate and rank the importance of each variable(the LIBS spectral peaks). For each tree grown in RF, OOB was put down and the number of votes cast for the correct class was counted. The value of variable m in OOB was randomly permuted and these samples

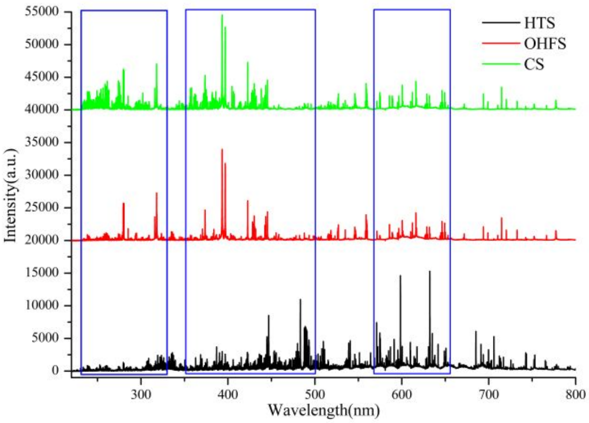


Fig. 1 Representative LIBS spectra of slag samples, black line for HTS; red line for OHFS; green line for CS.

were put down the tree. The numbers of votes were counted for the correct class in the variable-m-ranked OOB data; and then, we again count the number of votes for the correct class in the untouched OOB data. Subtracting the two counts, and averaging this number over all the trees in the forest is the raw importance score for the variable. Finally, the important scores were computed depending on the correlations between the trees, in other words, it was obtained depend on the contribution of input variable to classification result. The process of random forest based on variable importance(VIRF) model as follows: (1) Suppose there is original spectra ***A***, and original spectra is used to construct random forest classification model. (2) Calculate variable importance of each variable for classification analysis based on OOB error. (3) Remove the variable of variable importance is zero, and generate a new input variable ***B***. (4) Input variable ***B*** is used to construct VIRF classification model.

The averaged OOB error, sensitivity, specificity and accuracy are the statistical parameters to evaluate the performance of RF model for slag samples. Averaged OOB error are calculated by an estimate of the error rate (ER) for classification analysis of RF.[26] The sensitivity is the percentage of the samples of a category accepted by the class model. The specificity is the percentage of the samples of the categories which are different from the modeled one, rejected by the class model. The accuracy of classification procedure is expressed as fraction of correctly classified samples to the total samples.[28]

3. Results and discussion

3.1. LIBS spectra of three types of slag samples

Fig 1 shows the averaged spectrum of three types of slags in the range of 200-800nm, which includes the emission lines of the major component in slag. Slag is complex sample containing many chemical elements and thus related to LIBS spectra characterized by hundreds of atomic lines. There is obvious difference between the averaged LIBS spectrum with the range of 200-500nm and 580-650nm on three types of slags, which contributes to classification and identification of three types of slags. The differences among three types of slags come from the concentration of TiO_2, CaO, MnO_2 and Al_2O_3. Spectral lines of major element(Ca, Si, Al, Mg, Fe, Mn and Ti) in slag sample were detected and identified based on NIST atomic database.[29] For high titanium slag, the concentration of its major component-

Table 2 The RF training model for the classification of slags with different pre-processing method

Pre-processing method	Averaged OOB error	Sensitivity	Specificity	Accuracy
without pre-processing	0.0050	0.9952	0.9976	0.9968
Normalization	0.0001	1.0000	1.0000	1.0000
first-order derivative	0.0011	1.0000	1.0000	1.0000
second-order derivative	0.0016	1.0000	1.0000	1.0000

Table 3 The RF training model for the classification of slags with different input variables

Input variable	Averaged OOB error	Sensitivity	Specificity	Accuracy
200-300nm	0.0097	0.9905	0.9952	0.9937
200-400nm	0.0012	1.0000	1.0000	1.0000
200-500nm	0.0002	1.0000	1.0000	1.0000
200-600nm	0.0046	0.9952	0.9972	0.9968
200-700nm	0.0052	0.9952	0.9976	0.9968
200-800nm	0.0050	0.9952	0.9976	0.9968

TiO_2 is over than 80%; the convert slag is given priority to CaO, and there are no TiO_2 and MnO_2 in convert slag; the proportion of MgO and MnO_2 of open-hearth slag is larger than high titanium slag and convert slag. Therefore, classification analysis of slags could be brought out by the differences of specific components for each type slag.

3.2. RF model with different pre-processing methods and input variables

An excellent training model with appropriate pre-processing methods and input variables is essential for the classification analysis of slag using LIBS and RF. In order to improve the classification performance of RF model, the input variables addressed by pre-processing method(i.e. normalization) was used to decrease the differences comes from laser pulse energy fluctuations and increase the comparability among the different types of slags. Normalization is an effective method for eliminating the differences comes from laser pulse energy fluctuations, and derivation can be used to increase the comparability among the different types of slags. In this work, the RF training model for the classification of slags with the whole spectra(200-800nm) as input variable by different pre-processing methods was shown in Table 2. As seen from Table 2, sensitivity, specificity and accuracy of RF training model with three pre-processing methods(Normalized by maximum integrated intensity, First-order derivative and Second-order derivative) were higher than the RF model without pre-processing, meanwhile, the averaged OOB error of RF training model with three pre-processing methods is lower than the RF model without pre-processing. For three pre-processing method, all of three RF models show well classification performance, sensitivity, specificity and accuracy were 1.0000, 1.0000 and 1.0000, respectively; however, the averaged OOB error of RF model with normalized by maximum integrated intensity is lower than with first-order derivative and second-order derivative. Therefore, the LIBS spectra with normalized by maximum integrated intensity as input variable were employed to construct RF training model for the classification analysis of three types of slags.

Input variables are also significant for the training model of slags. In this work, the RF training model for the classification of

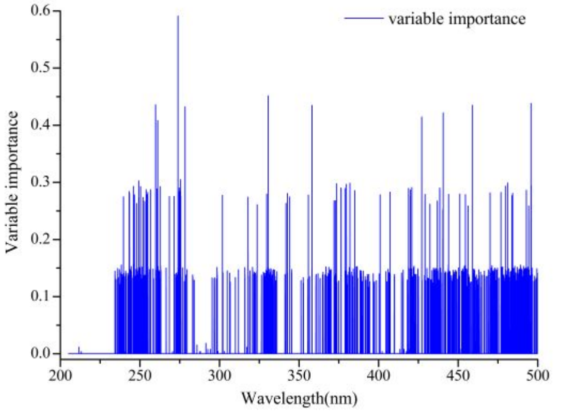


Fig. 2 The relationship between the variable importance of RF model and LIBS spectra(200-500nm) of slags

Table 4 Classification result of PLS-DA, SVM, RF and VIRF model for slags

RF model	OOB error	Sensitivity	Specificity	Accuracy
PLS-DA	0.0733	0.9253	0.9189	0.9227
SVM	0.0511	0.9364	0.9344	0.9353
RF	0.0333	0.9667	0.9833	0.9778
VIRF	0.0111	0.9889	0.9944	0.9926

slags with different input variables(200-300, 200-400, 200-500, 200-600, 200-700 and 200-800nm) was investigated by averaged OOB error, sensitivity, specificity and accuracy. Table 3 shows the averaged OOB error, sensitivity, specificity and accuracy for RF training model for the classification of slags with different input variables. As seen from Table 3, it can be obtained a well RF training model with different input variables. All of sensitivity, specificity with six RF model were over than 0.9900, and all of averaged OOB error were under 0.01. However, the averaged OOB error of RF training model with the input variable(200-500nm) is lowest, and the same time, it takes the relative less time and improves the RF efficiency of classification analysis based on RF. The LIBS spectra with the range of 500-800nm doesn't contribute to the classification for slags, there is no change for sensitivity, specificity and accuracy of RF model with the input variables(200-600, 200-700 and 200-800nm). Although there are a rich spectral information for the whole spectra(200-800nm), it takes a longer time to construct the RF training model. For the LIBS spectra of 200-300 and 200-400nm, there are less LIBS spectral information, and fails to obtain a better classification result of RF model for slags. Therefore, the LIBS spectra with range of 200-500nm were selected as input variable to construct the RF training model for slag samples.

3.3. RF model with variable importance for the classification of slags

Variable importance of RF model can be obtained by using OOB data. The greater the variable importance of RF model is obtained, the better the classification performance of slags, and vice versa. Fig 2 shows the relationship between the variable importance of RF model and LIBS spectra(200-500nm) of slags. The majority variable importance of LIBS spectra(200-500nm) on RF model for classification analysis of slags is 0-0.15. Some variable importance is 0, in other words, the LIBS spectral peaks of variable importance equals to zero that doesn't contribute to the classification of slags. Hence, we can remove the LIBS spectral peaks of variable importance equals to zero, and extract the spectral peaks of variable importance over zero in order to improve the classification performance of slags based on RF. In order to validation the classification abilities of VIRF model for slags, we compared the VIRF model with PLS-DA, support vector machine(SVM) and RF method. Input variables of these four methods for training model are the LIBS spectrum(200-500nm). For the training model based on PLS-DA, the best latent variables optimized by 5-fold cross-validation is 10. When the PLS-DA model was trained upon the training set, the averaged classification accuracy is 97%. For SVM training model, the best parameters selected by genetic algorithm(GA) were used as input for an epsilon classification SVM with a radial basis function(RBF) kernel. The optimum parameters were set as: penalty parameter C = 98.25 and kernel parameter of RBF g = 0.08. The averaged classification accuracy is 95% for optimized SVM training model. Table 4 lists the classification result of PLS-DA, SVM, RF and VIRF model. The classification performance of VIRF model is better than conventional RF model. Sensitivity, specificity and accuracy of VIRF model is larger than other three model, meanwhile, its averaged OOB error is less than other three model. Hence, VIRF model shows a better classification performance for slag samples.

Conclusions

In summary, LIBS technique coupled with RF has been successfully used for the classification of 60 slag samples. Normalized by maximum integrated intensity LIBS spectra(200-500nm) of slag samples as input variable was constructed the PLS-DA, SVM, RF and VIRF model for the classification analysis of slags. VIRF model shows a better classification performance than other three model. It has been confirms that LIBS technique coupled with RF is promising approach to achieve the online analysis and process control of slag and even industrial waste recycling.

Acknowledgments

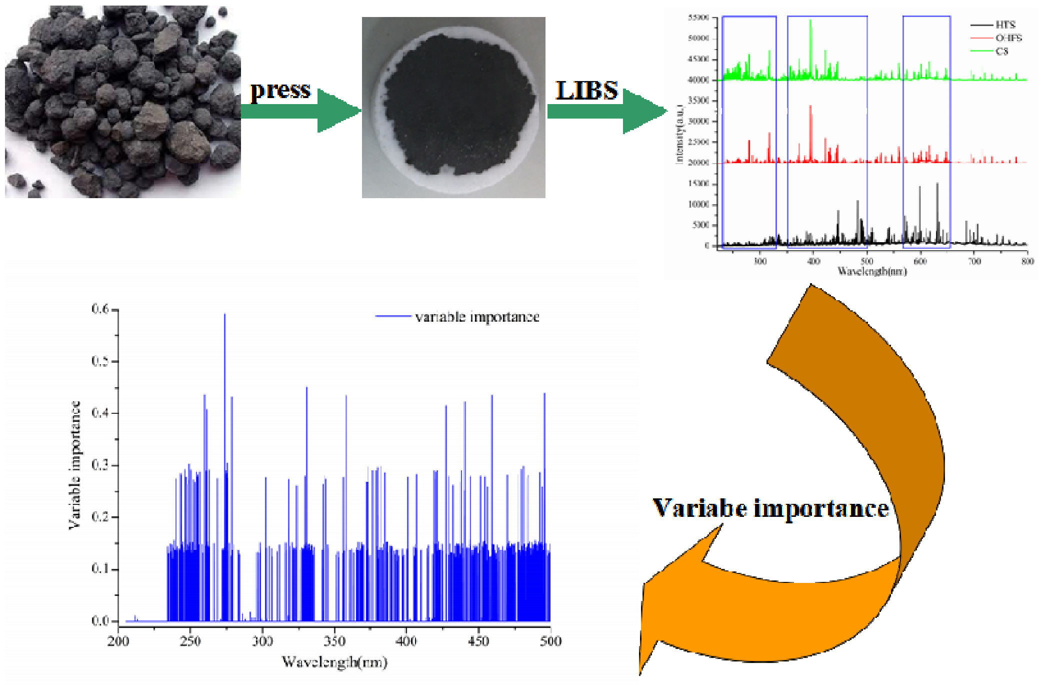
The work was supported by the National Major Scientific Instruments and Equipment Development Projects of China(No. 2011YQ030113), National Natural Science Foundation of China (No. 21375105), Research Fund for the Doctoral Program of Higher Education of China (No. 20126101110019), and Scientific Research Plan Projects of Shaanxi Education Department(No. 14JK1746).

Notes and references

[a] *Institute of Analytical Science, College of Chemistry & Material Science, Northwest University, Xi'an 710069, P. R. China. Tel:86-29-88382217; E-mail: huali@nwu.edu.cn*
[b] *College of Chemistry & Material Science, Northwest University, Xi'an 710069, P. R. China.*
[c] *College of Chemistry and Chemical Engineering, Xi'an Shiyou University, Xi'an 710065, P.R. China*

1. D. M. Proctor, K. A. Fehling, E. C. Shay, J. L. Wittenborn, J. J. Green, C. Avent, R. D. Bigham, M. Connolly, B. Lee, T. O. Shepker and M. A. Zak, *Environ. Sci. Technol.* 2000, **34**: 1576-1582.
2. V. K. Gupta, P. J. M. Carrott, M. M. L. Ribeiro Carrott and A. B. Shuhas, *Crit. Rev. Environ. Sci. Technol.* 2009, **39** : 783- 842

3．L. Marjanovic, R. I. McCrindle, B. M. Botha and H. J. Potgieter, *Anal. Bioanal. Chem.*, 2004, **379**, 104-107.
4．S. Leroy, R. Simon, L. Bertrand, A. Williams, E. Foya and Ph. Dillmanna, *J. Anal. At. Spectrom.*, 2011, **26**, 1078-1087.
5．D. A. Cremers and R. C. Chinni. *Appl. Spectrosc. Rev.*, 2009, **44**, 457-506.
6．F. J. Fortes and J. J. Laserna. *Spectrochim. Acta, Part B,* 2010, **65**, 975-990.
7．D. W. Hahn and N. Omenetto. *Appl. Spectrosc.*, 2012, **66**, 347- 419.
8．S. J. J. Tsai, S. Y. Chen, Y. S. Chung and P. C. Tseng, *Anal. Chem.*, 2006, **78**, 7432-7439.
9．L. M. Cabalín, A. González, J. Ruiz and J. J. Laserna, *Spectrochim. Acta, Part B*, 2010, **65**, 680-687.
10．R. Noll, V. Sturm, U. Aydin, D. Eilers, C. Gehlen, M. Höhne, A. Lamott, J. Makowe and J. Vrenegor, *Spectrochim. Acta, Part B*, 2008, **63**, 1159-1166.
11．D. L. Death, A. P. Cunningham and L. J. Pollard, *Spectrochim. Acta, Part B*, 2008, **63**, 763-769.
12．P. Yaroshchyk, D. L. Death and S. J. Spencer, *J. Anal. At. Spectrom.*, 2012, **27**, 92-98.
13．R. Noll, H. Bette, A. Brysch, M. Kraushaar, I. Monch, L. Peter and V. Sturm, *Spectrochim. Acta, Part B*, 2001, **56**, 637-649.
14．L. Peter, V. Sturm and R. Noll, *Appl. Opt.*, 2003, **42**, 6199-6204.
15．M. A. Gondal, T. Hussain, Z. H. Yamani and A. H. Bakry, *J. Environ. Sci. Health, Part A: Toxic/Hazard. Subst. Environ. Eng.*, 2007, **42**, 767-775.
16．M Kraushaar, R Noll and H. U. Schmitz, *Appl. Spectrosc.*, 2003, **57**, 1282-1287.
17．V. Sturm, H. U. Schmitz, T. Reuter, R. Fleige and R. Noll, *Spectrochim. Acta, Part B*, 2008, **63**, 1167-1170.
18．B. Praher, R. Rössler, E. Arenholz, J. Heitz and J. D. Pedarnig, *Anal. Bioanal. Chem.*, 2011, **400**, 3367-3375.
19．B. Stumpe, T. Engel, B. Steinweg and B. Marschner, *Environ. Sci. Technol.,* 2012, **46**: 3964-3972.
20．T. Zhang, S. Wu, J. Dong, J. Wei, K. Wang, H. Tang, X. Yang and H Li, *J. Anal. At. Spectrom.*, 2015, **30**: 368-374
21．L. Breiman, *Mach. Learn.,* 2001, **45**: 5-32.
22．L. Zheng, D. G. Watson, B. F. Johnston, R. L. Clark, R. Edrada-Ebel, W. Elseheri, *Anal. Chim. Acta.,* 2009, **642**: 257-265.
23．J. Remus and K. S. Dunsin, *Appl. Opt.,* 2012, **51**, B49-B56.
24．T. Zhang, L. Liang, K. Wang, H. Tang, X. Yang, Y. Duan and H. Li, *J. Anal. At. Spectrom.,* 2014, **29**: 2323- 2329.
25．L. Sheng, T. Zhang, G. Niu, K. Wang, H. Tang, Y. Duan and H. Li, *J. Anal. At. Spectrom.*, 2015, **30**: 453-458.
26．E. Bouveresse, C. Hartmann, D. L. Massart, I. R. Last and K. A. Prebble, *Anal. Chem.,* 1996, **68**, 982-990.
27．A. Liaw and M. Wiener, *R News*, 2002, **2**, 18-22.
28．V. K. Unnikrishnan, K. S. Choudhari, Suresh D. Kulkarni, Rajesh Nayak, V. B. Kartha and C. Santhosh*, RSC Advances,* 2013, **3**, 25872-25880
29．A. Kramida, Yu. Ralchenko, J. Reader, NIST ASD Team, NIST Atomic Spectra Database (ver. 5.0), [online], National Institute of Standards and Technology, Gaithersburg, MD, 2012. (Available: http://physics.nist.gov/asd).



Laser induced breakdown spectroscopy(LIBS) technique coupled with random forest based on variable importance(VIRF) was presented for classification analysis of slag samples(open-hearth furnace slag, converter slag and high titanium slag).