Minerals Engineering 143 (2019) 105899

Contents lists available at ScienceDirect

Minerals Engineering

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

Machine learning application to automatically classify heavy minerals in river sand by using SEM/EDS data☆

Huizhen Hao[a,b], Ronghua Guo[c], Qing Gu[a,d,*], Xiumian Hu[c]

[a] Software Institute, Nanjing University, Nanjing 210023, China
[b] Department of Communication Engineering, Nanjing Institute of Technology, Nanjing 211167, China
[c] School of Earth Sciences and Engineering, Nanjing University, Nanjing 210023, China
[d] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China

ARTICLE INFO

Keywords:
Heavy mineral
Machine learning
Energy dispersive *X*-ray spectrometers
Sand
Classification
Sedimentology
Geology

ABSTRACT

Heavy minerals are generally trace components of sand or sandstone. Fast and accurate heavy mineral classification has become a necessity. Energy Dispersive *X*-ray Spectrometers (EDS) integrated with Scanning Electron Microscopy (SEM) were used to obtain rapid heavy mineral elemental compositions. However, mineral identification is challenging since there are wide ranges of spectral datasets for natural minerals. This study aimed to find a reliable, machine learning classifier for identifying various heavy minerals based on EDS data. After selecting 22 distinct heavy minerals from modern river sands, we obtained their elemental data by SEM/EDS. The elemental data from a total of 3067 mineral grains were collected under various instrumental conditions. We compared the classification performance of four classifiers (Decision Tree, Random Forest, Support Vector Machine, Bayesian Network). Our results indicated that machine learning methods, especially Random Forest, can be used as the most effective classifier for heavy mineral classification.

1. Introduction

Heavy minerals with densities higher than 2.89 g/cm^3 are generally considered minor components of sand or sandstones, typically forming 1 percent of the weight in the samples (Mange and Wright, 2007). Heavy-mineral analysis is an effective tool for studying the sedimentary provenance of siliciclastic rocks, reconstructing sedimentary source to sink routes, subdividing and correlating non-fossiliferous siliciclastic strata, and finds various uses in mining, exploration and forensic science (Mange and Wright, 2007). Many heavy minerals are diagnostic of particular sources and the factors effecting the distribution of heavy minerals in sediment are well understood (Garzanti and Andò, 2007). However, due to the lack of rapid and accurate analytical techniques, the identification of heavy minerals still relies on qualitative visual microscopical inspection. This traditional, qualitative, composition analysis of heavy-mineral assemblage is laborious, time-consuming, and requires a highly skilled operator, which greatly restricts the use of heavy mineral analysis (Vermeesch et al., 2017). Development of a fast and reliable automated system for heavy mineral identification would not only have great significance for the basic sedimentologic research, but also important applications in the hydrocarbon exploration and mining industries.

In recent years, some researchers have attempted to develop classification and identification systems for heavy minerals, but little progress has been made. Ando and Garzanti (2014) applied Raman spectroscopy as an innovative tool for the reliable identification of heavy minerals. However, this method was used as an auxiliary source for single mineral identifications, not automatic classification.

Currently, the Quantitative Evaluation of Minerals by SCANning electron microscopy (QEMSCAN), produced by the FEI Company, is the only commercially available method for heavy mineral analysis (Gu, 2003). Generally speaking, QEMSCAN is a good choice for this purpose since it is simple and adaptable. However, Nie and Peng (2014) compared the heavy mineral assays obtained by QEMSCAN and traditional optical resolution in Chinese loess and red clay, and found significant differences between results. For instance, QEMSCAN yielded estimates of hornblende content three times higher than those obtained from traditional optical methods. The accuracy of rutile abundance was two times higher, and the accuracy of tourmaline, garnet, zircon, and epidote by QEMSCAN delivered estimates that were 97%, 54%, 51%, and

☆ **Link to the code:** The Weka 3 software (https://www.cs.waikato.ac.nz/ml/weka/) was used to do machine learning analyses.

* Corresponding authors at: State Key Laboratory of Novel Software Technology, Nanjing University, Xianlin Street 163, Nanjing 210023, China (Q. Gu). School of Earth Sciences and Engineering, Nanjing University, Nanjing 210023, China (X. Hu).
E-mail addresses: guq@nju.edu.cn (Q. Gu), huxm@nju.edu.cn (X. Hu).

https://doi.org/10.1016/j.mineng.2019.105899
Received 24 February 2019; Received in revised form 3 July 2019; Accepted 7 August 2019
0892-6875/ © 2019 Elsevier Ltd. All rights reserved.

35% higher, respectively, than those obtained from traditional optical method. These discrepancies arise primarily because (1) QEMSCAN cannot distinguish between minerals with the same chemical composition (e.g., titanium oxides) or rock fragments; and (2) mineral composition is variable, which means the measured mineral energy spectrum may differ greatly from that of the standard mineral database used to calibrate the QEMSCAN software.

Energy-Dispersive *X*-ray Spectrometers (EDS) has also been used to deliver heavy mineral identifications. In this approach, an *X*-ray spectrum generated by the interaction of the electron beam with the atoms of the samples is used to bombard a solid sample, and the reflected intensity of the *X*-ray beam is detected by the scanning grid to produce an elemental distribution image or "map". EDS with Scanning Electron Microscopy (SEM/EDS) is performed to obtain the X-ray spectrum for each pixel of the corresponding image in the thin sections and then analyze the mineral composition by identifying the mineral of each pixel. Accordingly, SEM/EDS analysis represented an effective, reliable, intuitive and quantitative element-identification procedure which has been used widely in both rock and mineral analysis (Akkaş et al., 2015).

However, like QEMSCAN, the SEM/EDS method of mineral analysis does not always provide accurate results; possible causes include (1) variable mineral chemical compositions; (2) the same chemical composition of some minerals; and (3) the low resolution of *X*-ray spectrum, especially for mineral samples with line overlap or interference. All these make mineral identification based on SEM/EDS data very challenging. Thus, the need for new, computer-aided heavy-mineral identification techniques remains.

The fields of Machine Learning (ML) and Artificial Intelligence (AI) have recently seen a number of highly-publicised successes in the field of earth sciences (Bergen et al., 2019), especially in the areas of mineral processing (McCoy and Auret, 2019), mineral prospecting (Rodriguez-Galiano et al., 2015), geochemical anomaly identification (Zuo and Xiong (2018)), classification of volcanic ash particles (Shoji et al., 2018) and mineral recognition/classification (Carey et al., 2015).

At first, computer-aided techniques for mineral classification based on EDS data only checked look-up tables for identified minerals in scanning electron micrograph frames and employed a maximum likelihood classification (Tovey and Krinsley, 1991; Clelland and Fens, 1991; Flesche et al., 2000). With the development of artificial intelligence, researchers have attempted to use machine-learning methods to carry out automatic mineral identification. In a pioneering investigation, Ruisanchez et al. (1996) used a Kohonen self-organizing map to analyze the EDS data from 12 different minerals. Somewhat later, Gallagher and Deacon (2002) augmented Ruisanchez's self-organizing map approach to automated mineral identification based on their SEM/EDS data by adding three different multilayer perceptrons to the data-analysis procedure. These authors concluded that backpropagation and quasi-Newton algorithms performed well at mineral-identification tasks. Tsuji et al. (2010) also used the Kohonen self-organizing map approach to train a system that could automatically classify eight minerals based on electron probe data. Akkaş et al., 2015 evaluated the use of a decision-tree approach to the automatic classification of 10 different minerals based on SEM/EDS data and obtained a good result. Ishikawa and Gulick (2013) used backpropagation to train an artificial neural network on the Raman spectra data of igneous minerals which was successful in the automated identification of olivine, quartz, plagioclase, potassium feldspar, mica and pyroxene, with accuracy reaching to 83–100 percent. Some researchers have even combined image analysis (Backscattered Electron Imaging, BSE) with mineral Energy-Dispersive *X*-ray emission Spectrometers for mineral identification. For instance, Frei et al. (2005) and Keulen et al. (2012) described a Computer-Controlled Scanning Electron Microscopy (CCSEM) automatic particle analysis system based on Backscattered Electron and Energy Dispersive *X*-ray Spectrum data. The Geological Survey of Denmark and Greenland has employed this CCSEM system to determine the elemental chemistry of both individual minerals and rock samples.

From our review it is clear that most previous investigations in this area have focused on the use SEM/EDS data to identify different rock types (mainly magmatic and metamorphic rocks) in thin sections (Akkaş et al., 2015) and rarely considered the automated classification of heavy minerals from ancient sedimentary rocks or from modern river sands. In this study, we are trying to develop an effective, rapid, and reliable method to classify heavy minerals in river sands using SEM/EDS data. River sands were collected from three rivers in China, and heavy minerals of river sands were manually selected and analyzed using SEM/EDS. Firstly we established a standard EDS database for different heavy minerals based on manual verification. Secondly, different methods of machine learning were used to classify the heavy minerals with determination of the most effective method for automated mineral classification made via comparison of the test results. Lastly, the test data were classified according to the selected method, which is used to verify the validity and accuracy of the method in heavy mineral classification.

2. Samples and experimental methods

2.1. Sampling and sample preparation

In this study, heavy minerals were selected from three modern river sand samples, which were collected on active sand bars in the main trunk of Yangtze River (16A001, GPS: 32°10′0.02″N, 118°58′41.61″E, near Nanjing Qixia Mountain), Yarlung Zangbo (16A063, GPS: 29°19′13.5″N, 88°51′28.4″E, near Xigaze Cong Song village), and Pumchu (16B027, GPS: 28° 09′35.96″N, 87°20′45.87″E, near Chentang village). More information about the Tibetan river sands can be found in Guo et al. (2019). All heavy mineral grains were handpicked randomly, mounted in epoxy resin and polished to produce a smooth, flat surface. Microphotographs were obtained to reveal internal structures and to select spots for analysis (Fig. 1).

2.2. EDS analysis and data acquisition

In total, 2255 mineral grains from 22 types of heavy minerals were selected for SEM/EDS analysis, which was carried out using Oxford Aztec X-Max 150 X-ray spectrometry and the Zeiss Supra 55 Field Emission Scanning Electron Microscope at the State Key Laboratory of Mineral Deposit Research, Nanjing University. The working distance was 8.7 mm, with an acceleration voltage of 15 kv and a beam current of 60 nA.

The test times for each grain of heavy mineral by EDS were 90, 40 and 6 seconds, respectively. The data obtained in the 90-second test were regarded as the most reliable and were used as the standard for heavy mineral characterization in the training and test (or validation) sets. Data produced under the 40- and 6-second tests was used to evaluate the reliability of classification of heavy minerals. Some of minerals included in our analysed sample are very similar in chemical composition (e.g., zoisite, actinolite and epidote - all of which belong to the "epidote group"; orthoferrosilite, clinoenstatite, augite, and pigeonite - all of which belong to the "pyroxene group"). Accordingly, the system was only trained to identify minerals which have distinctly different chemical compositions (see Fig. 2).

The data obtained was divided into training and testing sets, and the training set contained the 22 most commonly used heavy minerals, as shown in Table 1.

2.3. Input-data labeling

Heavy minerals handpicked by Langfang Geological Service Co., LTD (www.lfcxdz.com) were further labeled according to the chemical components of major and trace elements measured by the EDS method. The chemical properties of each heavy mineral can be found in the

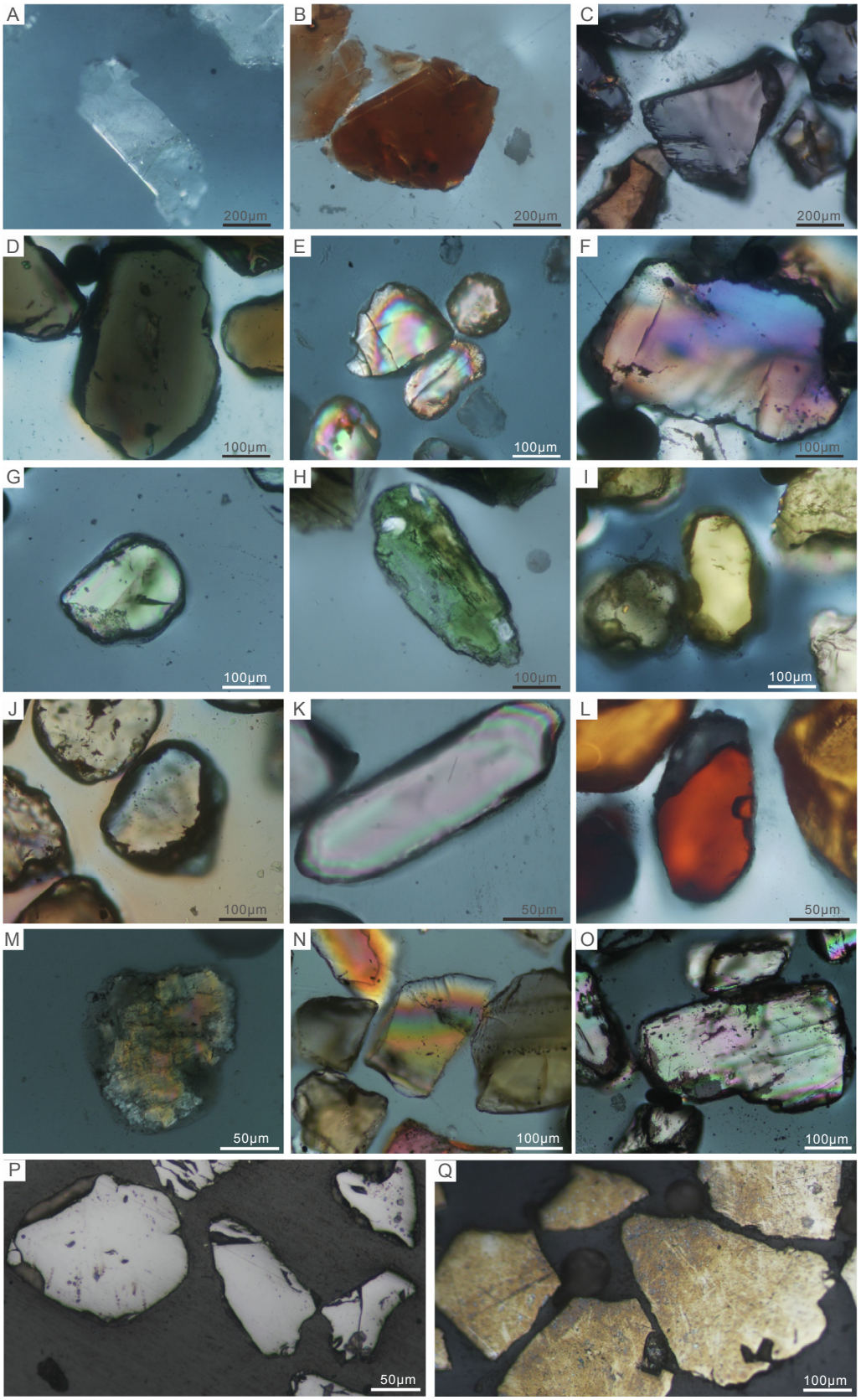

Fig. 1. Microphotograph of detrital heavy minerals. (A) muscovite, (B) biotite, (C) garnet, (D) tourmaline, (E) monazite, (F) apatite, (G) augite, (H) hornblende, (I) epidote, (J) sphene, (K) zircon, (L) rutile, (M) grossular, (N) staurolite, (O) actinolite, (P) magnetite, (Q) pyrite.

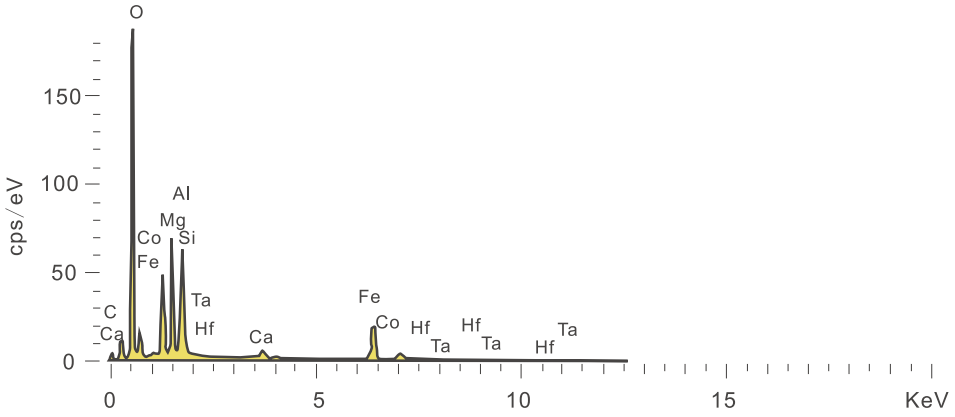

Fig. 2. An example of the X-ray spectrum of hornblende grain from the Yangtze river sand (sample 16A001). KeV (kiloelectron volt) in horizontal coordinate represents the unit of X-ray energy or the peak energy, while cps in vertical coordinate represents the unit of X-ray count or the peak intensity. cps = count per second; eV = electron volt.

Table 1
Types and analyzed numbers of heavy minerals from the river sands used in this study.

Mineral numbering in this study	Abbreviation	Mineral Name	Mineral numbers of 90-second data	Mineral numbers of 40-second data	Mineral numbers of 6-second data
1	Act	actinolite	50	–	–
2	Alm	almandine	203	56	29
3	Ap	apatite	178	29	30
4	Aug	augite	60	5	12
5	Bt	biotite	32	43	18
6	Cen	clinoenstatite	24	11	1
7	Di	diopside	16	13	6
8	Ep	epidote	289	32	29
9	Fs	orthoferrosilite	31	–	–
10	Grs	grossular	14	3	3
11	Hbl	hornblende	334	42	37
12	Mag	magnetite	136	–	–
13	Mnz	monazite	103	50	20
14	Ms	muscovite	60	37	29
15	Py	pyrite	102	–	–
16	Rt	rutile	109	38	20
17	Spn	sphene	33	56	24
18	St	staurolite	42	–	2
19	Tur	tourmaline	112	43	21
20	Zo	zoisite	43	4	8
21	Zrn	zircon	248	30	31
22	Pgt	pigeonite	36	–	–
Total number			2255	492	320

Mineral Database (http://www.webmineral.com/). Table 2 summarizes 12 types of major and trace elements of 22 types of heavy minerals showing the values of characterized elements of each mineral. Each EDS-analyzed grain was further characterized by chemical properties with the X-ray data, see the dataset related to this study (Mendeley Data, https://doi.org/10.17632/t6t82b2h7h.2).

3. Heavy mineral classification

3.1. Selecting decision attributes

The chemical composition of heavy minerals was obtained by SEM/EDS. Generally, the major element is greater than 10 wt% while minor and trace elements have mass concentrations of between 1 and 10 wt% and less than 1 wt%, respectively (see Goldstein et al., 2003; Newbury and Ritchie, 2015). However, due to factors such as the uncertainty of standard mineral composition and the analysis error of the original data, the overall accuracy is usually closer to ± 2 percent. Thus, an element content of more than 2 wt% was set as the minor element threshold. Otherwise, elements with weight percent lower than 2 were regarded as trace elements. In addition, because the trace element content in some heavy minerals is too low to be detected in EDS analysis, our analysis focused on the EDS data of the major and minor elements with concentration greater than 2 wt% of the heavy minerals.

Our EDS data indicated that major elements varied among minerals with positive detections in certain minerals but none in other minerals. Therefore, we counted all nonzero elements in each mineral. If the nonzero elements accounted for 80 percent of the total analyzed heavy mineral grains we considered the element to be a decision attribute of the heavy mineral; otherwise, it was regarded as a non-decision attribute.

In our study, each heavy mineral was analyzed separately to determine its decision attributes according to its major and trace-element composition (see the dataset related to this study, Mendeley Data, https://doi.org/10.17632/t6t82b2h7h.2). All decision attributes were combined to obtain a total attribute set. Overall, 22 decision attributes (Table 2) made up the decision-attribute set for the 22 heavy minerals. However, we obtained a preliminary classification based on these 22 decision attributes, and found the error rates of actinolite, augite, epidote and grossular to be extremely high. Subsequent analysis of these four heavy minerals revealed that these high error rates resulted from the stoichiometric formulae of actinolite which was similar to both augite and epidote with grossular. Additional elements (including Cr_2O_3, Pm_2O_3, Ta_2O_5, WO_3) were originally taken as non-decision attributes, due to their low in content ratio. Therefore, it was necessary to include these additional elements in the decision-attribute set; which increased the range of this set to 26. Detailed decision attributes of all heavy minerals employed in this investigation were listed in Table 3. In this paper, we used the abbreviations of 42 A, 26 A, and 12 A to present 42 decision attributes, 26 decision attributes and 12 decision attributes, respectively.

3.2. Selecting the classifier

In recent decades, a large number of machine learning classification methods have been developed, such as Decision Trees, Random Forest, Support Vector Machine, Bayesian Classifiers, Artificial Neural Networks and Ensemble Learning. Among them, these former four methods have been used widely in mineral analyses (Carey et al., 2015).

(a) Decision Tree (DT) (Safavian and Landgrebe, 1991; Friedl and Brodley, 1997; Akkaş et al., 2015) represents a mapping relationship between attributes and values and has good robustness to noise. The Decision Tree is composed of nodes, branches and leaves. Each divergent path represents a possible attribute value, each leaf node corresponds to a path from the root node to the leaf node, and every leaf node represents a possible result of classification. The classification algorithm of the Decision Tree includes the learning process to establish a tree model and the classification process based on a tree model. In our learning process a tree was constructed using the top-down approach.

Table 2
Elemental values variation of 12 types of major and trace elements for guideline of manual labeling of 22 types of heavy minerals used in this study.

Mineral numbering	Abbre-viation	Al_2O_3*	CaO	FeO	K_2O	MgO	MnO	Na_2O	P_2O_5	SiO_2	SO_3	TiO_2	ZrO_2
1	Act	< 2	22–24	6–15		8–16				50–55			
2	Alm	15–25	< 9	18–41		< 18				30–50			
3	Ap		51–55						40–44				
4	Aug	< 4	17–25	6–16		8–17				49–55			
5	Bt	14–18		16–23	3–10	7–15				36–44		2–14	
6	Cen	< 6		5–17		27–36				53–58			
7	Di	< 6	21–24	2–6		14–19				52–55			
8	Ep	10–28	12–28	6–19						35–50			
9	Fs		< 7	28–48			< 11			38–50			
10	Grs	20–31	22–35	4–14		< 4				38–41			
11	Hbl	4–18	10–13	10–28		3–17				40–53			
12	Mag			69–100								< 25	
13	Mnz	< 2	< 2						25–38				
14	Ms	19–39		< 4	8–14	< 3				49–66			
15	Py			28–48							50–71		
16	Rt											98–100	
17	Spn		26–29							30–32		34–39	
18	St	40–58		10–20						25–41			
19	Tur	30–42	< 3	5–17		2–12		2–4		39–43			
20	Zo	17–33	11–35	< 4						37–55			
21	Zrn									27–32			61–69
22	Pgt		5–13	7–17		13–22				50–58			

* In weight (%). Same as other elements.

In our classification process the attribute values of the unknown samples were compared with the Decision Tree nodes to determine the path for classification. Common algorithms include CART and C4.5.

(b) Support Vector Machine (SVM) (Cortes and Vapnik, 1995) is based on the Vapnik–Chervonenkis dimension of statistical learning theory and the structural risk minimization principle. It has been widely used in many fields, especially for solving small size of samples dataset, nonlinearity and high-dimensional pattern recognition. The aim of Support Vector Machine is to find an hyperplane that optimizes the edge between the two nearest sample points. The sample points on the maximized edge boundary are called support vectors, and the middle section of the edge is the optimal classification hyperplane.

(c) Random Forest (RF) (Breiman, 2001; Pal, 2005) is a classifier with multiple Decision Trees. It corrects the overfitting of the training set by constructing a multitude of Decision Trees and outputting the mean prediction (regression) of the individual trees. Decision Trees tend to have high variance when they utilize different training and test sets of the same data, since they tend to overfit on training data which leads to poor performance on unseen data. Random Forest is a popular ensemble method that chooses only a subsample of the feature space to make the trees de-correlated and prunes the trees by setting stopping criteria for node splits.

(d) Bayesian Network (BN) (Pearl, 1988) is a probabilistic graphical model that optimizes a set of variables and their conditional dependencies via a directed acyclic graph. The basic principle assumes that an attribute on a given class is independent of the others, and the unknown sample is predicted as the one with the largest posterior probability.

3.3. Performance measures

To evaluate the performance of the classifier, we employed two methods: a 10-fold cross-validation and a special test set. In 10-fold cross-validation on trainning set (Kohavi, 1995), the original samples were divided randomly into ten equal-sized subsets, of which a single subset was retained as the validation set for testing the model; the remaining nine subsets were used as training data. This process was repeated ten times, with each of the ten subsets used exactly once as the validation data. The average of the ten results (see below) was used as an estimate of the algorithm accuracy. A special test set was set up to test the generalization ability of the classifier for different data and whether the classifier constructed by the data at the 90-second test time can be used to identify the data at the 40- and 6-second test times.

To determine suitable performance measures for the classifier, we mainly use accuracy, the kappa coefficient and mean absolute error.

Accuracy is the ratio of the correct number of all classifications to

Table 3
Heavy mineral decision attributes used in this study.

Attribute Name	42 A*	26 A	12A	Attribute Name	42 A	26 A	12A	Attribute Name	42A	26 A	12A
Ag_2O	√			La_2O_3	√	√		SiO_2	√	√	√
Al_2O_3	√	√	√	MgO	√	√	√	Sm_2O_3	√	√	
Au_2O_3	√			MnO	√	√	√	SO_3	√	√	√
Br_2O_5	√			Na_2O	√	√	√	Ta_2O_5	√	√	
CaO	√	√	√	NaO	√			ThO_2	√	√	
Ce_2O_3	√	√		Nb_2O_5	√			TiO_2	√	√	√
CoO	√	√		Nd_2O_3	√	√		Tl_2O	√		
Cr_2O_3	√	√		OsO_2	√			UO_3	√		
Dy_2O_3	√			P_2O_5	√	√	√	V_2O_5	√	√	
FeO	√	√	√	Pm_2O_3	√	√		WO_3	√	√	
Gd_2O_3	√			Pr_2O_3	√	√		Y_2O_3	√		
HfO_2	√	√		PtO_2	√			Yb_2O_3	√	√	
IrO_2	√			Rb_2O	√			ZrO_2	√	√	√
K_2O	√	√	√	Sc_2O_3	√			ZnO	√		

* The abbreviations of 42 A, 26 A, and 12 A present 42 decision attributes, 26 decision attributes and 12 decision attributes, respectively.

Table 4
Classification performance of the dataset set with 26 decision attributes by applying 10-fold cross-validation.

Performance measure	Bayesian Networks	Random Forest	Decision Tree	Support Vector Machine
Accuracy (%)	98.05	98.63	97.29	97.43
Kappa coefficient	0.98	0.99	0.97	0.97
Mean absolute error	0.002	0.004	0.003	0.002

Table 5
Classification performance of heavy mineral dataset with 12 and 26 decision attributes.

Performance measure	Bayesian Network		Random Forest		Decision Tree		Support Vector Machine	
	Value	Difference *	Value	Difference	Value	Difference	Value	Difference
Accuracy (%)	97.29	−0.75	97.97	−0.66	96.54	−0.75	96.23	−1.20
Kappa coefficient	0.97	−0.008	0.98	−0.007	0.9623	−0.008	0.96	−0.013
MAE	0.0028	0.0006	0.0055	0.0015	0.0036	0.0006	0.0034	0.0011

* The difference between 12 and 26 decision attributes.

the total categories, which describes the ability of the model to correctly predict the class labels.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

where TP = True positive; FP = False positive; TN = True negative; FN = False negative.

The kappa coefficient is used to determine the classification difference between the specified classifier and the random classification.

$$K = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the relative observed agreement among raters, p_e is the hypothetical probability of chance agreement.

For categories k, number of items N and n_{ki} the number of times rater i predicted category k: $p_e = \frac{1}{N^2}\sum_k n_{k1}n_{k2}$.

The mean absolute error (MAE) is the average of the absolute values of the deviations between the predicted value and the true value. It represents the deviation of any predicted values in the set.

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|f(x_i) - y_i|$$

where y_i is the true value and $f(x_i)$ is the predicted value, m is the number of the predictions.

4. Experimental results

We proposed three major methodological parameters that can be addressed by the automated classification of heavy minerals in river sands using SEM/EDS: (1) Which is the best classification method for the heavy minerals in river sands using EDS data? (2) Which elements of heavy minerals can be used as decision attributes for effective classification? (3) Is there any classification difference between the shorter test times (40 and 6 s) and the standard test time?

4.1. Classifier comparison

The EDS data from 2255 single heavy minerals from 22 types of heavy minerals were obtained under the 90-second test. Twenty-six decision attributes were made according to the proposed methods. The EDS data were classified by using Decision Tree, Random Forest, Support Vector Machine and Bayesian Network algorithms. For each algorithm a 10-fold cross-validation was used to evaluate the classification performance of each classifier, and accuracy, kappa coefficient, and mean absolute error were calculated to conduct the quantitative performance measures.

The test results show that the classification of heavy minerals by four classifiers was nearly identical, with an accuracy greater than 97.29% and a kappa coefficient greater than 0.97, which is highly consistent with the expert label; the average absolute error is less than 0.004. Confusion matrixes were supplied in the dataset related to this study, Mendeley Data, https://doi.org/10.17632/t6t82b2h7h.2.

The best classifier is the Random Forest, of which the classification accuracy is 98.63%, the kappa coefficient is 0.99, and the MAE is 0.004. The worst classifier is the Decision Tree, of which the classification accuracy is 97.29%, the kappa coefficient is 0.97, and the MAE is 0.003.

4.2. Detection of decision attributes

In order to compare different attributes automatically selected by their values and by manual characterized elements as showed in Table 2, we reduced the 26 decision attributes of 22 heavy minerals to 12 decision attributes (Table 2). Then, the four classifiers of the Decision Tree, Random Forest, Support Vector Machine and Bayesian Network were used to classify the heavy minerals according to the 12 decision attributes. The 10-fold cross-validation was used to evaluate the classification results. The accuracy, kappa coefficient, and mean absolute error were calculated to measure the quantitative performance. The results are shown in Table 5. Confusion matrixes were supplied in the dataset related to this study, Mendeley Data, https://doi.org/10.17632/t6t82b2h7h.2.

The results based on 12 decision attributes show that the Random Forest is the most effective classifier, with a classification accuracy of 97.97%, kappa coefficient of 0.98, and MAE of 0.0055. Compared with the results of 26 decision attributes, the classification performance is slightly lower, and the classification accuracy decreases by 0.66%. The classification accuracy of the other three classifiers decreased to a different extent but ranged between 0.66% and 0.75%.

The classification results based on 12 decision attributes are very close to those based on 26 decision attributes (the difference is less than 2%).

4.3. Evaluation of test time

As data collected at longer intervals exhibit less pronounced noise, it produces more precise characteristics of the heavy minerals. However, qualitative heavy mineral analysis requires large numbers of grains, usually at least 400. Hence, it is necessary to simplify the test process and quickly complete classification and identification of heavy minerals. We produced heavy mineral classification based on EDS data in 40-second and 6-second tests in order to compare these data with the 90-second test data, thus further to evaluate the effects of the test time

Table 6
Classification performance of heavy mineral dataset on the 40-second test data with 12 decision attributes.

Performance measure	Bayesian Network		Random Forest		Decision Tree		Support Vector Machine	
	Value	Difference *	Value	Difference	Value	Difference	Value	Difference
Accuracy (%)	96.54	−1.50	97.97	−0.66	96.95	−0.34	95.94	−1.49
Kappa coefficient	0.960	−0.017	0.980	−0.007	0.970	−0.004	0.960	−0.016
MAE	0.004	0.001	0.008	0.004	0.005	0.002	0.004	0.001

* The difference between 40 and 90 s.

Table 7
Classification performance of heavy mineral dataset on the 6-second test data with 12 decision attributes.

Performance measure	Bayesian Network		Random Forest		Decision Tree		Support Vector Machine	
	Value	Differentials*	Value	Differentials	Value	Differentials	Value	Differentials
Accuracy (%)	90.31	−7.74	97.81	−0.81	87.81	−9.48	86.88	−10.55
Kappa coefficient	0.890	−0.084	0.980	−0.009	0.868	−0.103	0.860	−0.115
MAE	0.009	0.007	0.018	0.014	0.015	0.012	0.012	0.010

* The difference between 6 and 90 s.

on heavy mineral classification.

The 2255 EDS data produced under the 90-second test interval were used to train the four classifiers. A total of 492 EDS data of 16 heavy mineral types were included in the 40-second test and 320 EDS data of 17 heavy mineral types in the 6-second test. Both of these data sets were subjected to analysis by the trained (90-second) classifiers. Because six heavy minerals types were not analyzed in 40-second and 6-second test, three elemental components (PmO_3, Ta_2O_5, and Yb_2O_3) were missing from those EDS data. Thus, only 23 decision attributes were used in conjunction the 40- and 6-second test data. Results are shown in Tables 6 and 7 for the 40-second test and 6-second test, respectively. Although the attributes and number of grains of 40-second and 6-second test sets are not as exactly same as those of 90-second test sets, our results did not appear to be effected by this discrepancy. We leave further consideration of this question to be explored in the near future.

The classification results based on the data of the 40-second test showed that the classification accuracy decreased by 0.34–1.5%, whereas the MAE increased by 0.001 to 0.004. Confusion matrixes were supplied in the dataset related to this study, Mendeley Data, https://doi.org/10.17632/t6t82b2h7h.2. These results indicated that the classification accuracy based on EDS data produced under the 40-second test did not change appreciably. In addition, the kappa coefficient is between 0.96 and 0.98, which decreased by 0.004 to 0.017, indicating that there is no substantive difference between the classification performance of heavy mineral EDS data produced between the 40-second and 90-second tests.

As with the 90-second test, the Random Forest is the most effective classifier among the four classifiers under the 40-second test, with a classification accuracy of 96.54% and a kappa coefficient of 0.96. Compared to the 90-second test, the classification accuracy decreased by 0.66%, the MAE increased by 0.004, and the kappa coefficient decreased by 0.007%. Overall, the classification performance of the 40-second test is consistent with that of the 90-second test.

The classification based on the EDS data obtained under the 6-second test showed that there is a significant difference in the classification accuracy between the heavy mineral EDS data obtained under the 6-second and 90-second tests. The accuracy rate decreased by 0.8–10.55%, while the MAE increased by 0.007–0.014, and the kappa coefficient variation range was also large, decreasing by 0.009 to 0.115.

As with the classification based on EDS data produced under the 40-second and 90-second tests, for the 6-second data testing, the Random Forest is the most effective classifier among the four classifiers, with an overall classification accuracy of 97.81%, a kappa coefficient of 0.98, and an MAE of 0.018, which is consistent with classification based on the EDS data produced under the 90-second test. However, the classification accuracy of the other three classifiers fluctuated greatly, especially the Support Vector Machine, for which the classification accuracy rate decreased by more than 10%.

5. Conclusion

This paper documented the performances of a selection of machine learning methods for the heavy mineral classification of a single, composite sample of river sands by using SEM/EDS data. Through the comparison of test results generated by different classifiers, different decision attributes and different test times, we drew the following conclusions:

(1) The 26 elemental compositions of heavy minerals as the decision attribute can distinguish the 22 most common heavy minerals in our composite river sands and could be used as a basis for the classification of this sample.
(2) Our results indicate that machine learning methods, including Decision Tree, Random Forest, Support Vector Machine and Bayesian Network, can be used as effective classifiers for heavy mineral classification. Among these, the Random Forest approach delivered marginally better results than all others tested.
(3) The results based on 12 decision attributes by using the Random Forest are comparable with a classification accuracy of 97.97% relative to those based on 26 decision attributes.
(4) With the application of the machine learning methods, the classification performance on the EDS data of the heavy minerals produced under the 6-second test is comparable to that obtained in the longer (40 and 90 s) test times. This may provide a method for reducing the test time and improving the test efficiency of the heavy mineral analysis.

Through the best classification method (Random Forest), decision attribute (26 elements for 90-second testing and 23 elements for 40-second testing and 6-second testing) and the shortest testing time (6 s), the potential for precise, fast and efficient heavy mineral auto-identification by SEM/EDS data seems promising. In the classification of sand, other types of heavy minerals besides the 22 studied types of heavy minerals need to be included, which may require the use of deep-learning algorithms. In heavy mineral classification, minerals with similar chemical components, such as garnet group, pyroxene group and amphibole group, cannot be precisely distinguished. This true mineral identification (as opposed to mineral category identification) may

require the incorporation of other information (e.g., optical images) in the overall machine learning system design.

Computer code availability

The used software is Weka 3 (120.3 MB), available in https://www.cs.waikato.ac.nz/ml/weka/. The code developer was Machine Learning Group from the Department of Statistics, University of Waikato, New Zealand (address: Private Bag 3105, Hamilton 3240 New Zealand). The software was developed in English. The latest official releases of Weka require Java 8 or later. Weka is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization. In this work, we used the tools in the Weka 3 to do data classification and result evaluation.

Dataset related to this paper

The dataset related to this study is published in Mendeley Data, https://doi.org/10.17632/t6t82b2h7h.2. Among them, three tables include the EDS data of analyzed elements including 90-second elemental data of 2255 grains, 40-second elemental data of 492 grains and 6- second elemental data of 320 grains, respectively. The Tables 4–7 include the confusion matrixes based on datasets of different analyzing times and different decision attributes. This dataset is linkable at the online version of this paper in the journal website https://www.sciencedirect.com/journal/minerals-engineering. Cite this dataset: Hao, Huizhen; Guo, Ruohua; Gu, Qing; Hu, Xiumian (2019), "Dataset for: Machine learning application to automatically classify heavy minerals in river sand by using SEM/EDS data", Mendeley Data, V2, https://doi.org/10.17632/t6t82b2h7h.2.

Declaration of Competing Interest

The authors have no completing interests to declare.

Acknowledgements

Mrs Juan Li was thanked for helping in the SEM/EDS lab. Prof. Dr. Norman MacLeod in Nanjing University was acknowledged for English revising. We are grateful to helpful advice and constructive suggestions from two anonymous reviewers and to the editor Dr. Kristian E. Waters for the careful handling. This study was supported by the National Natural Science Foundation of China Project (41972111) and Second Tibetan Plateau Scientific Expedition and Research Program (STEP), Ministry of Science and Technology, China (2019QZKK020604).

References

Akkaş, E., Akin, L., Çubukçu, H.E., Artuner, H., 2015. Application of decision tree algorithm for classification and identification of natural minerals using SEM-EDS. Comput. Geosci. 80, 38–48.
Ando, S., Garzanti, E., 2014. Raman spectroscopy in heavy-mineral studies. Geol. Soc. London Spec. Public. 386 (1), 395–412.
Bergen, K.J., Johnson, P.A., de Hoop, M.V., Beroza, G.C., 2019. Machine learning for data-driven discovery in solid Earth geoscience. Science 363 (6433), eaau0323.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Carey, C., Boucher, T., Mahadevan, S., Bartholomew, P., Dyar, M.D., 2015. Machine learning tools for mineral recognition and classification from Raman spectroscopy. J. Raman Spectrosc. 46, 894–903.
Clelland, W.D., Fens, T.W., 1991. Automated rock characterization with SEM/image-analysis technique. SPE Form. Eval. 437–443.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.
Flesche, H., Nielsen, A., Larsen, R., 2000. Supervised mineral classification with semi-automatic training and validation set generation in scanning electron microscope energy dispersive X-ray spectroscopy images of thin sections. Math. Geol. 32 (3), 337–366.
Frei, D., Knudsen, C., McLimans, R.K., Bernstein, S., 2005. Fully automated analysis of chemical and physical properties of individual mineral species in heavy mineral sands by computer controlled scanning electron microscopy (CCSEM). Heavy minerals 2005. Soc. Mining, Metall., Explor. 103–108.
Friedl, M.A., Brodley, C.E., 1997. Decision tree classification of land cover from remotely sensed data. Remote Sens. Environ. 61 (3), 399–409.
Gallagher, M., Deacon, P., 2002. Neural networks and the classification of mineralogical samples using x-ray spectra. Int. Conf. Neural Inform. Process. 2683–2687.
Garzanti, E., Andò S., 2007. Chapter 20 Heavy mineral concentration in modern sands: implications for provenance interpretation. Developments in Sedimentology, 58. Elsevier, p. 517–545.
Goldstein, J.I., Newbury, D.E., Echlin, P., Joy, D.C., Romig Jr., A.D., Lyman, C.E., Fiori, C., Lifshin, E., 2003. Scanning electron microscopy and X-ray microanalysis: A text for biologists, materials scientists, and geologists. Plenum Press.
Gu, Y., 2003. Automated scanning electron microscope based mineral liberation analysis an introduction to JKMRC/FEI mineral liberation analyser. J. Miner. Mater. Characteriz. Eng. 2 (01), 33–41.
Guo, R.H., Hu, X.M., Garzanti, E., Lai, W., Yan, B., 2019. Different response of detrital mineral to magmatic and metamorphic event in modern river sand. Earth Sci. Rev in revision.
Ishikawa, S.T., Gulick, V.C., 2013. An automated mineral classifier using Raman spectra. Comput. Geosci. 54 (4), 259–268.
Keulen, N., Frei, D., Riisager, P., Knudsen, C., 2012. Analysis of heavy minerals in sediments by computer-controlled scanning electron microscopy (CCSEM): principles and applications. Mineral. Assoc. Canada Short Course 42, 167–184.
Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. Int. Joint Conf. Artif. Intell. 14 (2), 1137–1145.
Mange, M.A., Wright, D.T., 2007. Heavy minerals in use. Elsevier, Amsterdam.
McCoy, J.T., Auret, L., 2019. Machine learning applications in minerals processing: a review. Miner. Eng. 132, 95–109.
Newbury, D., Ritchie, N., 2015. Performing elemental microanalysis with high accuracy and high precision by scanning electron microscopy/silicon drift detector energy-dispersive X-ray spectrometry (SEM/SDD-EDS). J. Mater. Sci. 50, 493–518.
Nie, J.S., Peng, W.B., 2014. Automated SEM-EDS heavy mineral analysis reveals no provenance shift between glacial loess and interglacial paleosol on the Chinese Loess Plateau. Aeolian Res. 13, 71–75.
Pal, M., 2005. Random forest classifier for remote sensing classification. Int. J. Remote Sens. 26 (1), 217–222.
Pearl, J., 1988. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, San Francisco, CA, USA.
Rodriguez-Galiano, V.F., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M., 2015. Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machine. Ore Geol. Rev. 71, 804–818.
Ruisanchez, I., Potokar, P., Zupan, J., Smolej, V., 1996. Classification of energy dispersion X-ray spectra of mineralogical samples by artificial neural networks. J. Chem. Inf. Comput. Sci. 36 (2), 214–220.
Safavian, S.R., Landgrebe, D., 1991. A survey of decision tree classifier methodology. IEEE Transac. Syst., Man, Cybern. 21 (3), 660–674.
Shoji, D., Noguchi, R., Otsuki, S., Hino, H., 2018. Classification of volcanic ash particles using a convolutional neural network and probability. Sci. Rep. 8, 8111.
Tovey, N.K., Krinsley, D.H., 1991. Mineralogical mapping of scanning electron micrographs. Sed. Geol. 75, 109–123.
Tsuji, T., Yamaguchi, H., Ishii, T., Matsuoka, T., 2010. Mineral classification from quantitative X-ray maps using neural network: application to volcanic rocks. Isl. Arc 19 (1), 105–119.
Vermeesch, P., Rittner, M., Petrou, E., Omma, J., Mattinson, C., Garzanti, E., 2017. High throughput petrochronology and sedimentary provenance analysis by automated phase mapping and LAICPMS. Geochem., Geophys., Geosyst. 18 (11), 4096–4109.
Zuo, R., Xiong, Y., 2018. Big data analytics of identifying geochemical anomalies supported by machine learning methods. Nat. Resour. Res. 27, 5–13.