Nuclear Engineering and Technology 55 (2023) 1708–1717

Contents lists available at ScienceDirect

Nuclear Engineering and Technology

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

Original Article

Identification of Pb–Zn ore under the condition of low count rate detection of slim hole based on PGNAA technology

Haolong Huang [a], Pingkun Cai [a], Wenbao Jia [a, b, *], Yan Zhang [c]

[a] Department of Nuclear Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, 211106, China
[b] Collaborative Innovation Center of Radiation Medicine of Jiangsu Higher Education Institutions, Suzhou, 215000, China
[c] Engineering Research Center of Nuclear Technology Application, Ministry of Education, East China University of Technology, Nanchang, 330013, China

ARTICLE INFO

Article history:
Received 2 September 2022
Received in revised form
5 December 2022
Accepted 4 January 2023
Available online 6 January 2023

Keywords:
Prompt gamma neutron activation analysis (PGNAA)
Monte Carlo simulation
Lead-zinc ore
Borehole logging
Machine learning

ABSTRACT

The grade analysis of lead-zinc ore is the basis for the optimal development and utilization of deposits. In this study, a method combining Prompt Gamma Neutron Activation Analysis (PGNAA) technology and machine learning is proposed for lead-zinc mine borehole logging, which can identify lead-zinc ores of different grades and gangue in the formation, providing real-time grade information qualitatively and semi-quantitatively. Firstly, Monte Carlo simulation is used to obtain a gamma-ray spectrum data set for training and testing machine learning classification algorithms. These spectra are broadened, normalized and separated into inelastic scattering and capture spectra, and then used to fit different classifier models. When the comprehensive grade boundary of high- and low-grade ores is set to 5%, the evaluation metrics calculated by the 5-fold cross-validation show that the SVM (Support Vector Machine), KNN (K-Nearest Neighbor), GNB (Gaussian Naive Bayes) and RF (Random Forest) models can effectively distinguish lead-zinc ore from gangue. At the same time, the GNB model has achieved the optimal accuracy of 91.45% when identifying high- and low-grade ores, and the F_1 score for both types of ores is greater than 0.9.

1. Introduction

Lead and zinc metals are widely used in important fields such as electrical, mechanical, military and nuclear industries [1–5]. With the continuous development of the global economy, the demand for lead and zinc metal consumption in various industries is expanding, which puts forward higher requirements for improving the survey efficiency, development and utilization of lead-zinc ore deposits [6]. Huangshaping polymetallic deposit, located in Hunan Province, China, is a significant lead-zinc mineral resource base. According to the chemical composition analysis of the ore samples, the target elements for beneficiation and recovery are Pb and Zn, which are mainly in the form of galena and sphalerite, respectively; in addition, there are large amount of sulfide such as pyrite and pyrrhotite. Based on the features of chemical composition, it can be seen that the Huangshaping deposit is a lead-zinc primary polymetallic sulfide ore.

Timely information about mineral grade is crucial to deposit mining (typically on a time scale of minutes, depending on drilling speed). Due to the complexity of the chemical composition and distribution of ore in Huangshaping deposit, general borehole analysis methods will take long drilling sampling and large errors in subsequent assay results owing to the heterogeneity of the formation of the mineral deposit [7,8]. Prompt Gamma Neutron Activation Analysis (PGNAA) technology has the characteristics of online and non-destructive measurement, which has been applied to real-time elemental analysis of bulk samples like coal, cement, explosives, etc. [9–12] Mineral grade estimation with PGNAA is a Logging While Drilling (LWD) method. It collects information of characteristic gamma-ray emitted during the interaction between neutrons and nuclei of ore/gangue atoms in formation by way of inelastic scattering and neutron capture to achieve qualitative and quantitative analysis of target elements in the mineral deposit.

Since PGNAA can provide real-time composition information of an order of magnitude large volume of formation rocks in comparison with geophysical methods such as sonic logging, electromagnetic wave method, etc., it has been applied to grade estimation of copper and iron ore [13,14]. However, in the grade

* Corresponding author. Department of Nuclear Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, 211106, China.
E-mail address: jiawb@nuaa.edu.cn (W. Jia).

https://doi.org/10.1016/j.net.2023.01.005

control of lead-zinc ore, on account of the comparatively small radius of borehole, the small-sized detector used will cause a lower count rate and higher statistical deviation to have a serious impact on the measurement of element content. At the same time, the cross-sections of Pb and Zn are low, so it is difficult to analyze their characteristic peaks according to traditional methods. Direct quantitative analysis by spectrum processing will cause great uncertainties with less practical significance. In recent years, the cross-application of machine learning and spectroscopic analysis technology has become more and more extensive, which also proves that machine learning is an effective method for processing multi-dimensional and low SNR (Signal to Noise Ratio) data [15–17].

Therefore, this study proposes a mineral grade analysis method that combines PGNAA technology and machine learning. Considering the mineral-associated traits of lead-zinc ore in the mineral deposit, the spectrum of a certain number of ores of different grades and gangue is obtained through Monte Carlo simulation, and trained by machine learning classification algorithm to realize the identification and classification of lead-zinc ores of different grades and gangue. Qualitative and semi-quantitative real-time analysis of PGNAA will provide guidance for subsequent drilling sampling quantitative measurement and reduce time and economic costs caused by repeated sampling and multiple assays.

2. Materials and methods

2.1. Monte Carlo simulation of borehole logging

The borehole logging model was constructed using the MCNP-4C with the general structure is presented in Fig. 1 and can be briefly described as follow.

- A borehole filled with fresh water with a radius of 3.5 cm; the radius and height of formation model are 100 cm and 160 cm respectively, which is much larger than the size of the borehole that can be regarded as infinite comparatively; different minerals filled the whole formation.
- The D-T pulsed neutron source emits 14.1 MeV neutron with a pulse width and period of 50 μs and 2 ms respectively, which has the source intensity of 1×10^8 n/s; the radius and height of neutron tube are 1.9 cm and 22 cm respectively.
- Gamma-ray counts are recorded by a BGO scintillator detector with a radius and height of 1.9 cm and 15.24 cm respectively, and detector-source distance is 35 cm.
- The shield between the detector and the D-T pulsed neutron source is tungsten metal, with a radius and height of 2.2 cm and 10 cm respectively; it is set up to reduce the neutrons and gamma rays from the neutron source to the scintillator directly [14,18].
- The BGO scintillator is wrapped in a 3 mm thick aluminum shell. There is a 3 mm thick TC11 titanium alloy case outside the entire detection tube. In addition, at the BGO scintillator and tungsten metal position on the case, a layer of 2 mm thick fluororubber is covered, which is designed to reduce the capture gamma rays generated from the instrument case.
- The termination condition of the MCNP code is set by the history cutoff card (NPS = 1E+08), i.e., simulation will terminate after histories of 10^8 neutrons; according to our settings for the pulsed neutron source, this corresponds to a measuring time of 40 s for a certain depth logging point.

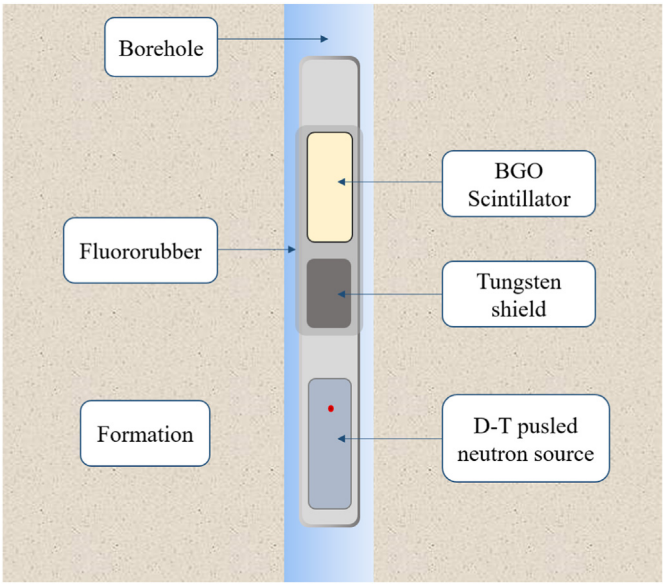

Fig. 1. Schematic diagram of borehole logging MCNP simulation model.

The energy deposition spectra in this model are calculated by F8 tally and then normalized according to the total count of gamma photons. At the same time, in order to simulate the fluctuation of the energy deposition of the physical detector, the method of Broadening During Simulation (BDS) is adopted [19]. This processing is carried out during the operation of the MCNP code, the deposited energy of the photon recorded by F8 tally will be the center of a Gaussian distribution, and a randomly sampled energy from this distribution to replace the original deposited energy will be recorded. The above process is implemented using the Gaussian Energy Broadening (GEB) option that comes with the MCNP code. The desired Full Width at Half Maximum (FWHM) is calculated by Eq. (1) and constants a, b and c (determined by the performance of the detector) are inputted by user. The FWHM parameters of the BGO scintillator used in this study are set a = 0.0218; b = 0.0593; c = 0.277.

$$FWHM = a + b\sqrt{E + cE^2} \tag{1}$$

In addition to simulating the energy broadening of the physical detector, this model also simulates the neutron emission period of the D-T pulsed neutron source. A distributed density function is deployed for the TME option of the SDEF card in the MCNP code and expressed by Eq. (2), then the neutron source will take 2000 μs as a period and the duration of neutron emission is 50 μs Therefore, the pure capture spectrum can be obtained by detector within 50–2000 μs, and the pure inelastic scattering spectrum is attained through deducting a certain proportion of capture spectrum from the spectrum acquired within 0–50 μs.

$$D(x) = \begin{cases} 1, 0 < x \le 50 \\ 0, 50 < x \le 2000 \end{cases} \tag{2}$$

2.2. Samples

The galena (PbS, accounting for about 83.07% of lead output) and sphalerite (ZnS, accounting for about 91.96% of zinc output) in the Huangshaping deposit are symbiotic ores, the average grades of lead and zinc in raw ores are 2.54% and 6.26% respectively. In terms of mineral-associated relationship, they are mainly embedded with sulfides such as pyrite (FeS_2) and pyrrhotite (FeS); Very few are

directly distributed in gangue dominated by calcite and quartz [20,21]. The results of microanalysis show that most of the sphalerite contains a high content of iron, which is a typical high-iron sphalerite type.

The gangue is mainly calcite ($CaCO_3$), followed by quartz (SiO_2), siderite ($FeCO_3$), almandine ($Fe_3Al_2(SiO_4)_3$), etc. Most of the above together form gangue aggregates. For example, some calcite and pyrite are symbiotic as carbonate combinations, and very few small particles of calcite and pyrite are irregularly embedded in the gap of sulfide.

Based on the above mineral distribution relationship and the assay results of part of mineral samples, a total of 220 groups of samples were set up in the simulation, including 110 groups each of lead-zinc ore and gangue. According to the content of Pb and Zn, the grade of Pb + Zn>5% is set to high-grade ore, Pb + Zn ≤ 5% is set to low-grade ore, and both the content of Pb and Zn in the gangue are less than 1%. The elemental composition of gangues and lead-zinc ores of different grades are listed in Table 1. Usually, ores with a comprehensive grade of lead and zinc less than 5% account for about 20–40% of all lead-zinc ore in a mining area, but considering that the imbalance between the two categories of the data set will affect the performance of the classification algorithm [22], the low-grade ore was "oversampled" in the simulations, so the low-grade ore and high-grade ore each account for 50%, that is, both ores have 55 sets of samples.

2.3. Machine learning classification algorithm

The spectra collected from Monte Carlo simulation will be used as a sample data set for training and testing classification algorithms. The identification process of gamma-ray spectrum of mineral samples is illustrated in Fig. 2. To start with, two different classifier models are obtained by fitting the classification algorithms with the training set. After that, the classification of lead-zinc ore and gangue is performed by classifier model 1 with the test set, and then classifier model 2 divides the ore samples into low- and high-grade categories. This study used four different machine learning classification algorithms [23], including Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Gaussian Naive Bayes (GNB) and Random Forest (RF). Their basic principles are described as follows.

2.3.1. Support Vector Machine

For a given sample data set, each sample has the same number of features with the associated labels ($y \in \{-1, 1\}$), where each feature is a dimension of a hyper-plane. SVM algorithm [24] aims at finding a hyper-plane (also known as "decision boundary" and can be defined by Eq. (3)) that divides the hyper-space into two classes as shown in Fig. 3 (for a binary classification, but it can be extended to multi-class problem as well) and tries to achieve maximum separation distance between two classes. This process will result in two hyper-planes parallel to the decision boundary and located on either side of it, called margin boundaries and can be given by Eq. (4). These red dots on the margin boundary in Fig. 3 are the support vectors.

$$w^T x + b = 0 \tag{3}$$

$$w^T x + b = \pm 1 \tag{4}$$

With the input data of x_i (gamma spectrum with several number of channels), the main goal of training a SVM classification model is to solve the following optimization problem:

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i \tag{5}$$

$$y_i \left(w^T \phi(x_i) + b \right) \geq 1 - \xi_i \tag{6}$$

$$\xi_i \geq 0, i = 1, \ldots, n \tag{7}$$

where ξ_i is the acceptable distance from the correct margin boundary and C (called Penalty term and greater than zero) controls the strength of the penalty (inversely proportional to the strength of the penalty). It will allow for some mis-classification due to fluctuations in gamma spectrum counts. Here training vectors x_i are mapped into a higher dimensional space by the kernel function ϕ as they cannot be classified linearly simply. In this study the RBF kernel function was used and as presented in Eq. (8).

$$K(x_i x_j) = \phi(x_i)^T \phi(x_i) = \exp\left(-\gamma \| x_i - x_j \|^2 \right), \gamma > 0 \tag{8}$$

After training, for any new set of spectrum data prediction of its class (+ or - , representing different grades of ore) is possible.

2.3.2. K-nearest neighbor

There is a training sample data set and each sample has a corresponding label, the K-Nearest Neighbor algorithm [25] stores instances of the them and predicts the new unlabeled point as the class with the most representatives within the k nearest neighbors of that point. Given a training set $\{(x_i = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(n)}), y_i)\}$, the distance of the new input x_j from each sample in the training set can be determined by Eq. (9) and prediction of its class can be expressed by Eq. (10).

$$L_2(x_i, x_j) = \left(\sum_{l=1}^{n} \left| x_i^{(l)} - x_j^{(l)} \right|^2 \right)^{\frac{1}{2}} \tag{9}$$

$$y = argmax \sum_{x_i \in N_k(x)} I(y_i = c_i) \tag{10}$$

where $N_k(x)$ is the k neighbors of the input x, c_i refers to the j th class, and $I()$ is the indicator function which takes value 0 if $y_i \neq c_i$ and value 1 $y_i = c_i$.

Usually, k is an integer that is not greater than 20 and is highly data-dependent. Its effect on the classification results as shown in Fig. 4, when $k = 5$, the unlabeled input is classified as class (+), whereas $k = 9$ it will be classified as class (−). The k value in this

Table 1
The elemental composition and content of lead-zinc ores of different grades and gangues.

Sample	Element Content (%)								
	Pb	Zn	Fe	Si	Al	Ca	S	C	O
Low grade ore	1–3.5	1–3.5	15–25	8–16	0–4	6–15	10–16	2–6	20–36
High grade ore	3–8	3–11	12–25	5–15	0–4	6–15	10–20	2–6	20–34
Gangue	0.1–1	0.1–1	7–21	4–25	0–4	7–20	0.5–4	3–10	43–48

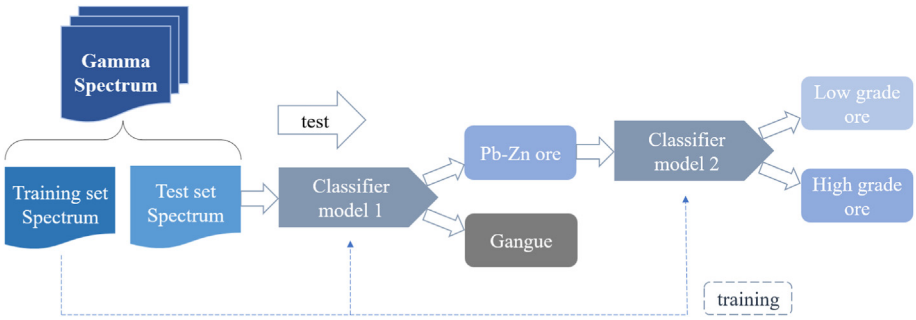

Fig. 2. Recognition procedure of machine learning classification algorithm.

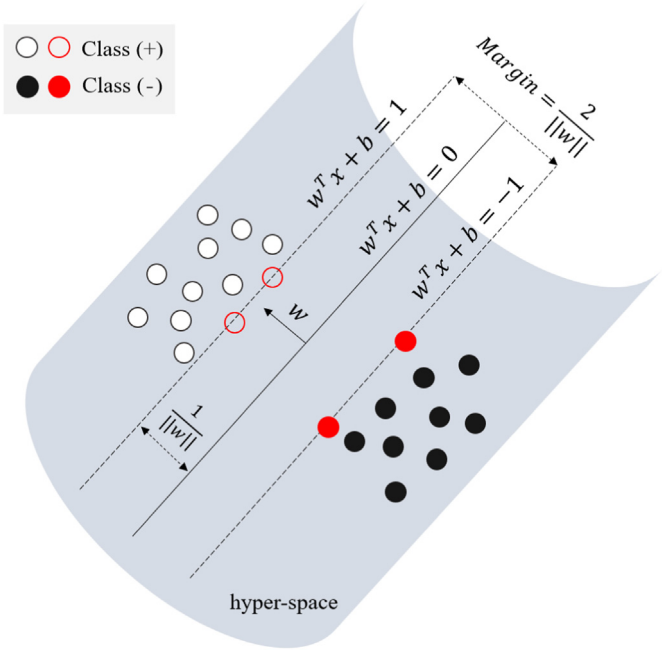

Fig. 3. Schematic diagram of SVM algorithm.

study determined by grid search technique as 3.

2.3.3. Gaussian Naive Bayes

Naive Bayes classifier [26] work based on the Bayesian rule and probability theorems. For a given input vector $(x_1, x_2, \ldots, x_n)$ with the class label y, the classifier assume that every pair of features are conditionally independent with each other. Thus, the probability of y can be calculated by a contingent probability as specified in Eq. (11).

$$P(y|x_1, x_2, \ldots, x_n) = \frac{P(x_1, x_2, \ldots, x_n|y)P(y)}{P(x_1, x_2, \ldots, x_n)} \tag{11}$$

where $(x_1, x_2, \ldots, x_n)$ denotes the features of the input vector. According to the independence of each feature, we can obtain the relationship as represented in Eq. (12), and then Eq. (11) will get the form as expressed in Eq. (13).

$$P(x_i|y, x_1, x_2, \ldots, x_n) = P(x_i|y) \tag{12}$$

$$P(y|x_1, x_2, \ldots, x_n) = \frac{\prod_{i=1}^{n} P(x_i|y)P(y)}{P(x_1, x_2, \ldots, x_n)} \tag{13}$$

As for a given instance, $P(x_1, x_2, \ldots, x_n)$ is constant. Therefore, we can use the classification rule as specified by Eq. (14).

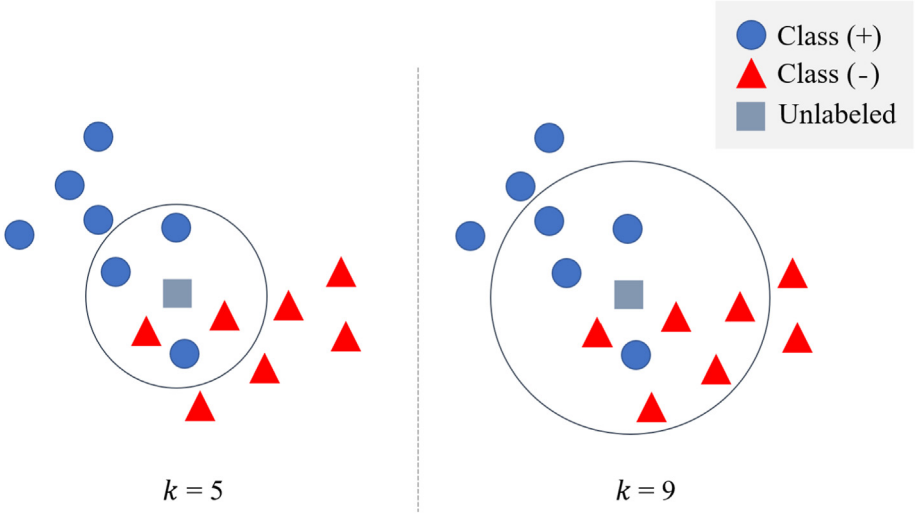

Fig. 4. Schematic diagram of KNN algorithm.

$$P(y|x_1, x_2, \ldots, x_n) \propto \prod_{i=1}^{n} P(x_i|y)P(y)$$

$$\Downarrow$$

$$\hat{y} = \underset{y}{argmax} \prod_{i=1}^{n} P(x_i|y)P(y) \tag{14}$$

Gaussian Naive Bayes classification [27] is a case of Naive Bayes method with an assumption of having a Gaussian distribution on all features given the class label as presented in Eq. (15). This assumption is suitable for the energy deposition process of gamma photons.

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \tag{15}$$

2.3.4. Random Forest

A series of subsets are obtained from the original sample data set by bootstrap method as input data for different decision trees. The decision tree model devotes to predicting the value of a target variable by learning simple decision rules inferred from the data features. Random Forest [28] is a classical ensemble learning model that contains multiple decision classification trees as shown in Fig. 5, and the results are collected by randomly selecting the features of each decision tree, and finally a majority voting form is used for each specific problem to ensure the stability and accuracy of the prediction results.

3. Results and discussion

3.1. Data processing

The gamma-ray spectrum of partial gangue and lead-zinc ores of different grades obtained by Monte Carlo simulations is presented in Fig. 6, which records a total of 512 channels of gamma photon counts in the energy range of 0–10.00 MeV. It can be observed although there is no obvious elemental characteristic peak, there are significant differences between ore and gangue in the energy range of 0.70–0.90 MeV and 2.20–2.40 MeV, and the count contributions of these two intervals are derived from: the inelastic scattering of Ca (0.77 MeV) and Pb (0.80 MeV), the capture of S (0.84 MeV) as well as the inelastic scattering of S (2.23 MeV), the capture of Ca (1.94 MeV) and S (2.38 MeV), respectively. when the elemental composition of formation materials change, the macroscopic thermal neutron absorption cross-section will vary with it, which will affect the recorded gamma photon yield that is manifested as a certain difference in the total count of the energy range of 5.60–6.60 MeV, and the count contribution of this part comes from the capture of Fe (5.92 MeV, 6.02 MeV) and Ca (6.42 MeV), the inelastic scattering of O (6.13 MeV).

The inelastic scattering and capture spectrum of partial gangue and lead-zinc ores of different grades as shown in Figs. 7 and 8 respectively. The mostly reactions are inelastic scattering because it is difficult to set up a neutron moderator around the detection tube in small-sized borehole, and the neutron moderation is mainly through the water in the borehole and the formation material itself, which reduces the probability of neutron capture. At the same time, due to the small-sized detector uesed, the higher the energy, the worse the statistics of the spectrum, and the fluctuation of the gamma photon counts is dramatic when the energy is greater than 8.00 MeV, so this part will be eliminated in the subsequent input data processing for the classification algorithm.

Typically, in common PGNAA online detection of industrial material, such as cross-band PGNNA analyzers, the expected count rate is usually between 40 and 80 kcps (using larger size BGO scintillators and DC mode neutron sources). The count rate of simulated spectrum in this study is around 7 kcps, which is far from the normally required for quantitative analysis of spectrum in PGNAA industrial assays. Therefore, it is important to choose a suitable classification algorithm to analyze this type of gamma-ray spectrum.

For the identification of lead-zinc ores and gangues, a complete 512 channels gamma-ray spectrum was selected as the input data of the classification algorithm, that is, the data set contains 220 vectors of 512 dimensions. For the classification of high- and low-grade lead-zinc ores, the following four spectrum data were selected to fit the algorithms and compared the performance of classifier models.

1. 512 channels of gamma photon energy deposition spectrum in the energy range of 0–10.00 MeV
2. 395 channels of inelastic scattering spectrum in the energy range of 0.50–8.00 MeV
3. 395 channels of capture spectrum in the energy range of 0.50–8.00 MeV
4. 790 channels of inelastic scattering spectrum + capture spectrum in the energy range of 0.50–8.00 MeV

3.2. Model evaluation

In order to compare and select the best performance classification algorithm, the input data set will be divided into training set and test set. After each time the classifier model is fitted with the training set, the performance of the model is tested using the test

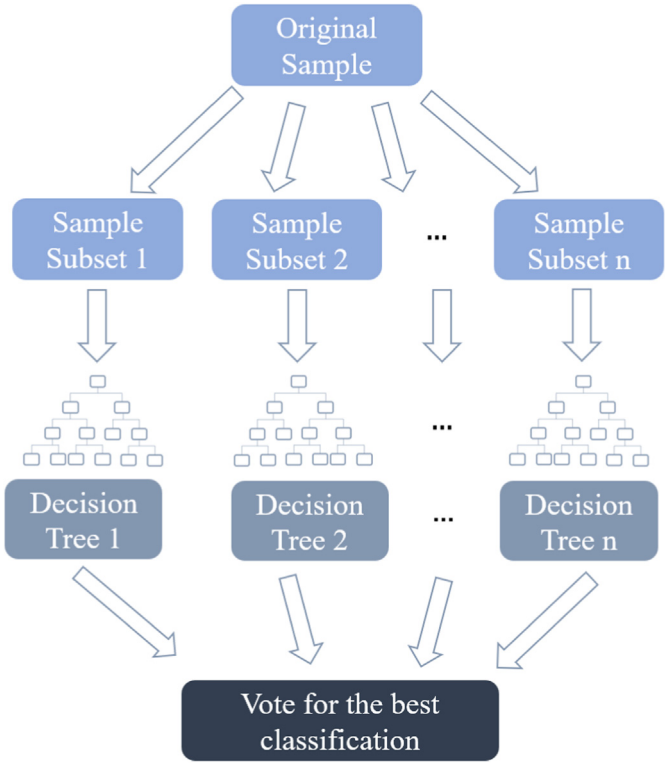

Fig. 5. Schematic diagram of RF algorithm.

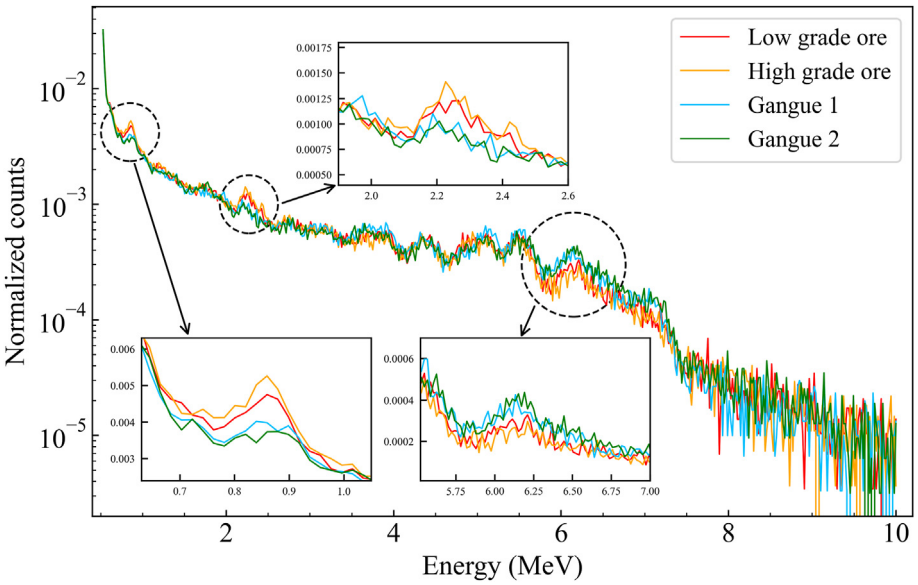

Fig. 6. Gamma-ray spectrum of partial lead-zinc ores of different grades and gangue.

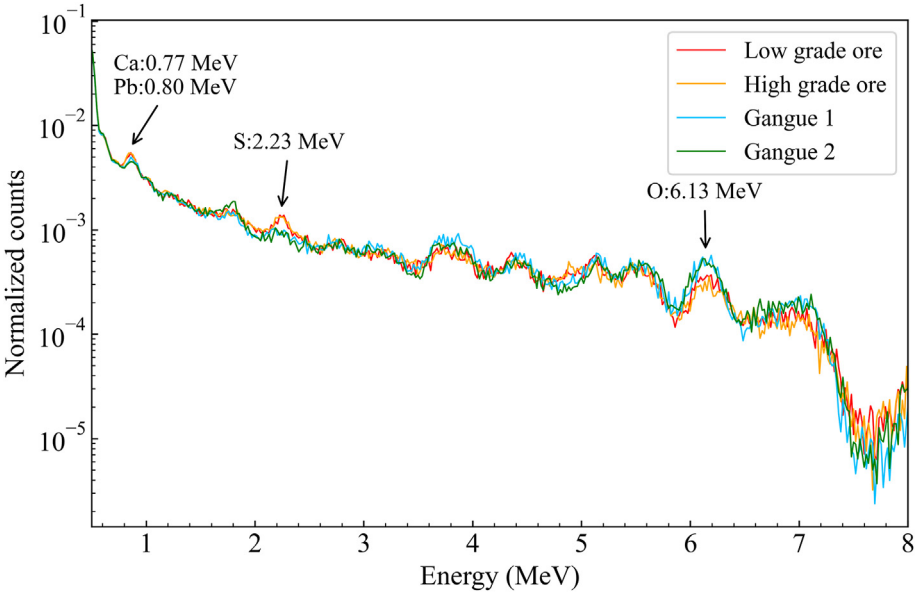

Fig. 7. Inelastic scattering spectrum of partial lead-zinc ores of different grades and gangue.

data set, and the parameters TP, FP, TN and FN are calculated. They represent the number of true positive, false positive, true negative and false negative in the test results, respectively, and the following evaluation metrics are calculated to measure the performance of the classifier model [29]:

The Accuracy is defined as a ratio between the correct predictions to the total number of tests and can be calculated by Eq. (15).

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \times 100\% \tag{15}$$

The Recall (also known as TPR: true positive rate or sensitivity) and Specificity (also known as TNR: true negative rate) represent the proportion of all positive samples and negative samples successfully recognized by the classifier model, respectively. When the number of samples varies greatly between the two categories, the high Accuracy may be confusing. The Recall and Specificity can reflect the probability that two types of samples will be successfully identified as well as are presented using Eqs. (16) and (17), respectively.

$$Recall = TPR = Sensitivity = \frac{TP}{TP + FN} \times 100\% \tag{16}$$

$$Specificity = TNR = \frac{TN}{TN + FP} \times 100\% \tag{17}$$

The positive predictive value (PPV) and negative predictive value (NPV) represents the proportions of predicted true positives and predicted true negatives in the predicted positives and negatives, respectively. For the predicted results, these two metrics

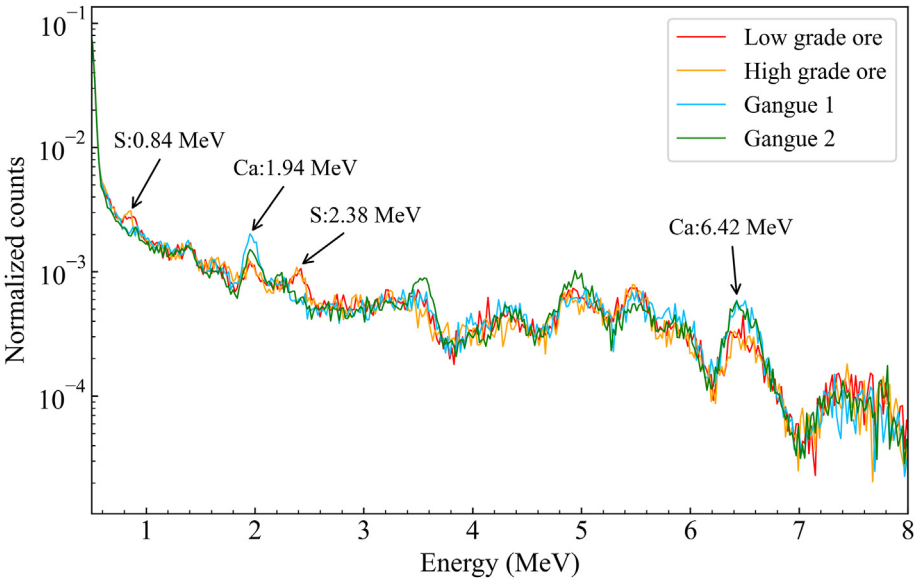

Fig. 8. Capture spectrum of partial lead-zinc ores of different grades and gangue.

characterize the accuracy of the prediction results of positive and negative samples as well as can be defined in Eqs. (18) and (19), respectively.

$$PPV = Precision = \frac{TP}{TP + FP} \times 100\% \tag{18}$$

$$NPV = \frac{TN}{TN + FN} \times 100\% \tag{19}$$

F_1 score is the harmonic mean of Precision and Recall, usually only for positive samples; the recognition effect of low-grade ores is as vital as high-grade ores in this study, so the F_1 score is also calculated for negative samples (low-grade ores), denoted as $F_1(-)$. They are specified using Eqs. (20) and (21) respectively.

$$F_1 = 2 \times \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2TP}{2TP + FP + FN} \tag{20}$$

$$F_1(-) = \frac{2TN}{2TN + FP + FN} \tag{21}$$

Matthews Correlation Coefficient (MCC) [30] is mainly used to measure the binary classification problem as represented in Eq. (22). It takes into account TP, TN, FP and FN as a relatively equilibrium indicator, which can also be used in case of unbalanced samples. The value range of MCC is [−1,1]. The value of 1 indicates an optimal prediction, 0 represents that the predicted values are not as good as the results of random prediction, and −1 means that completely inconsistent with the true values.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{22}$$

Considering that the sample set of simulated data usually still be difficult to summarize the real and comprehensive situation of the mineral deposit, in order to avoid the result deviation caused by overfitting during training procedure and make full use of the limited data, the P-repeated K-fold cross-validation method is adopted to obtain the evaluation metrics of the classifier model [31]. The process of the 10-repeated 5-fold cross-validation method using in this study is illustrated in Fig. 9. Taking the identification test of high- and low-grade ores as an example, 110 samples are divided into five subsets after scrambling the data order through random seeds in each round of validation. Each subset takes turns as a test set, and the rest is a training set to fit the model, that is, the training set accounts for four-fifths of a total of 88 samples, and the test set accounts for one-fifth of a total of 22 samples. a total of 5 × 10 rounds of validation were carried out, and the means of evaluation metrics in all the validation were calculated.

3.2.1. Identification of Pb–Zn ore and gangue

In the identification of lead-zinc ore and gangue (The classifier model 1 as shown in Fig. 2), the positive sample is lead-zinc ore, and the negative sample is gangue. The Accuracy of different classifier models were calculated by 5-fold cross-validation and all were 100%, which indicates that they can successfully distinguish lead-zinc ore from gangue. Furthermore, the lead and zinc grades of the gangues were both below 1% in our simulations, which were set based on actual industrial grades of lead-zinc sulfide ores, so this model can satisfy the industrial requirements.

3.2.2. Identification of high- and low-grade Pb–Zn ore

In the identification of high- and low-grade lead-zinc ore (The classifier model 2 as presented in Fig. 2), the positive sample is high-grade ore, and the negative sample is low-grade ore. The evaluation metrics of different classifier models were calculated by 10-repeated 5-fold cross-validation as listed in Tables 2–5 respectively, where the comparison of the two types of F_1 and MCC values is shown in Fig. 10, and the sample data set used were the gamma photon energy deposition spectrum, inelastic scattering spectrum, capture spectrum and synthetical spectrum of inelastic scattering + capture respectively mentioned in Section 3.1. Except for the RF model using input data set of inelastic spectra obtained an Accuracy of only 78.82%, the Accuracy of the other models to identify high- and low-grade ores were higher than 80%. The probability of successful identification of low-grade ores was always higher than that of high-grade ores (Recall < Specificity), while the accuracy of the prediction results was the opposite (PPV > NPV), indicating that other interfering elements in the formation (mainly associated metals such as Fe, Al, Ca, etc.) have a more serious influence on the spectrum of low-grade ores, causing them

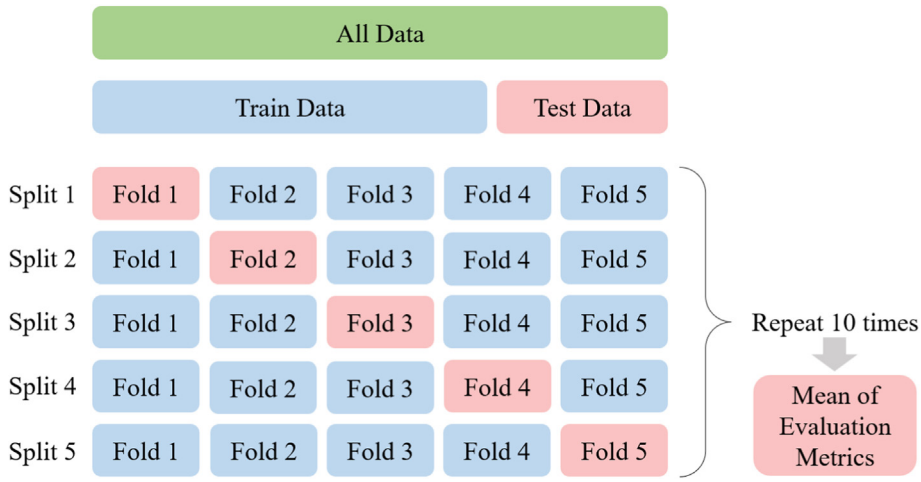

Fig. 9. Schematic diagram of 10-repeated 5-fold cross-validation.

Table 2
Comparison of evaluation metrics of different classification algorithms when identifying high- and low-grade Pb–Zn ore (gamma photon energy deposition spectrum).

Classifier Model	SVM	KNN	GNB	RF
Accuracy (%)	86.45	82.09	86.55	82.27
Recall (%)	86.88	77.81	82.82	76.73
Specificity (%)	86.51	87.49	89.89	88.05
PPV (%)	86.07	86.10	89.72	86.93
NPV (%)	86.25	79.49	84.38	79.79

Table 3
Comparison of evaluation metrics of different classification algorithms when identifying high- and low-grade Pb–Zn ore (inelastic scattering spectrum).

Classifier Model	SVM	KNN	GNB	RF
Accuracy (%)	85.27	80.09	88.00	78.82
Recall (%)	81.23	67.32	82.55	71.71
Specificity (%)	89.15	93.80	92.93	85.04
PPV (%)	87.84	91.46	92.03	84.51
NPV (%)	82.84	74.26	85.23	74.75

Table 4
Comparison of evaluation metrics of different classification algorithms when identifying high- and low-grade lead ore (capture spectrum).

Classifier Model	SVM	KNN	GNB	RF
Accuracy (%)	81.27	81.55	91.36	80.00
Recall (%)	77.29	80.27	85.47	70.85
Specificity (%)	85.05	82.84	96.91	88.90
PPV (%)	83.57	83.38	96.47	85.83
NPV (%)	79.46	80.40	87.31	76.14

Table 5
Comparison of evaluation metrics of different classification algorithms when identifying high- and low-grade Pb–Zn ore (synthetical spectrum of inelastic scattering + capture).

Classifier Model	SVM	KNN	GNB	RF
Accuracy (%)	82.82	86.82	91.45	81.27
Recall (%)	82.13	85.06	85.00	72.34
Specificity (%)	83.91	88.11	97.52	89.88
PPV (%)	83.49	87.75	97.13	87.65
NPV (%)	82.13	86.00	87.19	76.95

to be more easily mis-classified than high-grade one.

Regardless of the data set used, the recognition effect of the GNB model is better than that of other models. When using synthetical spectrum of inelastic scattering + capture as input data, GNB obtained the optimal accuracy: 91.45% and the highest MCC value: 0.8326, meanwhile, Recall and Specificity was 85.00% and 97.52% respectively, and the F_1 score for both types of samples was greater than 0.9. The above results show that the algorithm can identify high- and low-grade ores very well, and the accuracy of the identification results is credible.

4. Conclusion

In this study, a lead-zinc ore identification method combining PGNAA and machine learning is proposed. In the small-sized borehole logging of lead-zinc mine model simulated by the MCNP code, the small size of the detector and the small cross-section of lead and zinc jointly result in a low count rate of the gamma-ray spectrum and susceptibility to interference from other elements in the formation. The machine learning classification algorithm solves this series of problems well, indicating where the classification features of gamma-ray spectra in field logging might be, which is the core concern of classifier model training using gamma-ray spectra of a real setup.

We used four types of gamma-ray spectrum as input data in turn and tested four different classification algorithms, the evaluation metrics obtained from cross-validation show that each classifier model can completely distinguish lead-zinc ore from gangue, and the Accuracy of the best effective GBN model (using the synthetical spectrum of inelastic scattering + capture as input data) in identifying high- and low-grade ores has reached 91.45%; meanwhile, the measuring time of each depth logging point is 40 s, realizing the qualitative and semi-quantitative real-time analysis of lead-zinc ore in small-sized borehole, which will provide reference for drilling sampling and assay as well as improve the exploration efficiency of lead-zinc mineral deposits.

However, this study addresses lead-zinc deposits of the high-iron sulfide type, where the gangue in the formation is mainly $CaCO_3$, SiO_2 and $FeCO_3$, etc. For other types of mineral deposits such as lead-zinc oxide ores, or in association with other metallic ores, the performance of the classifier model will vary and depend on the characteristics of the specific spectrum data. In addition, the boundaries for high- and low-grade lead-zinc ores in this study were set based on the geological report of resource exploration in

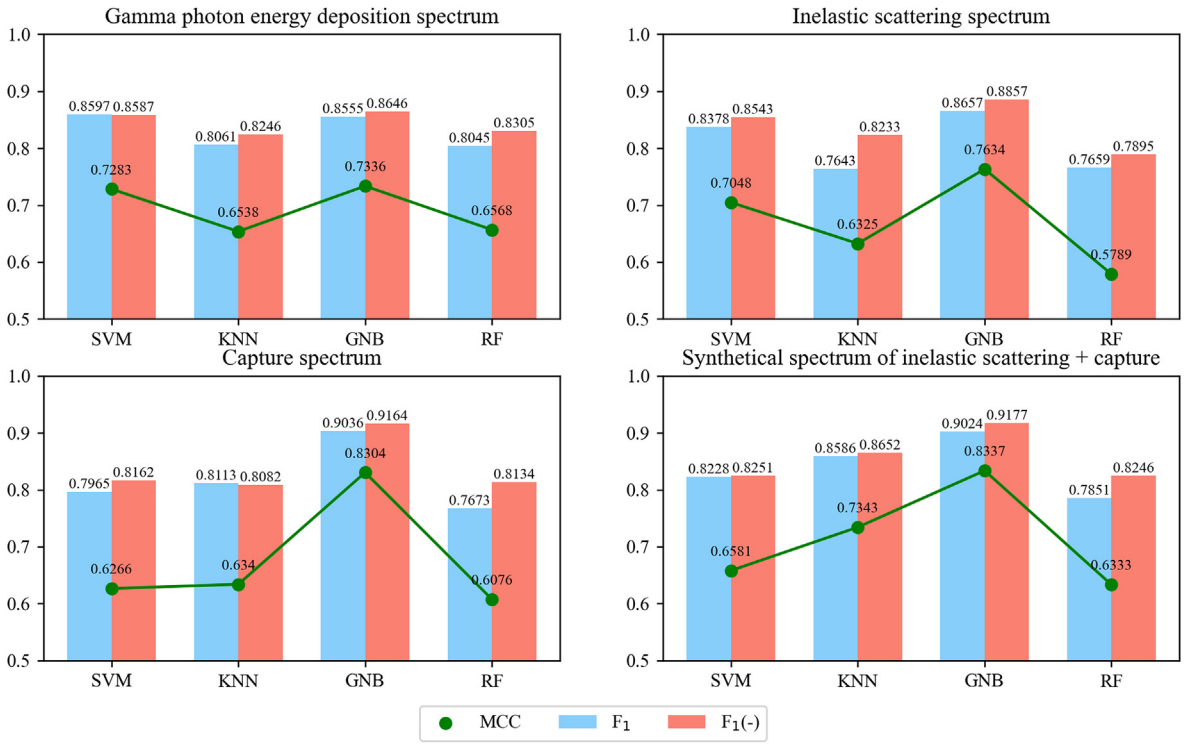

Fig. 10. Comparison of two types of F_1 and MCC values of different classification algorithms when identifying high- and low-grade Pb–Zn ore.

the Huangshaping deposit. For mineral deposits of different types and geologies in other regions, or the actual prospecting needs, more and different grade boundaries can be set to further test the recognition effect of the classification algorithm.

Declaration of competing interest

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

Acknowledgements

This work was supported by the NSAF (Grant No. U1930125), the National Natural Science Foundation of China (11975121 and 41904160).

References

[1] H.Y. Chen, A.J. Li, D.E. Finlow, The lead and lead-acid battery industries during 2002 and 2007 in China, J. Power Sources 191 (2009) 22–27, https://doi.org/10.1016/j.jpowsour.2008.12.140.
[2] Z. Guo, Y. Ma, X. Dong, J. Huang, Y. Wang, Y. Xia, Environmentally friendly and flexible aqueous zinc battery using an organic cathode, Angew Chem. Int. Ed. Engl. 57 (2018) 11737–11741, https://doi.org/10.1002/anie.201807121.
[3] A. Verbič, M. Gorjanc, B. Simončič, Zinc oxide for functional textile coatings: recent advances, Coatings 9 (2019), https://doi.org/10.3390/coatings9090550.
[4] K.S. Nair, M. Mittal, K. Lal, R. Mahanti, C. Sivaramakrishnan, Development of rapidly solidified (RS) magnesium–aluminium–zinc alloy, Mater. Sci. Eng., A 304 (2001) 520–523.
[5] J.P. McCaffrey, H. Shen, B. Downton, E. Mainegra-Hing, Radiation attenuation by lead and nonlead materials used in radiation shielding garments, Med. Phys. 34 (2007) 530–537, https://doi.org/10.1118/1.2426404.
[6] G.M. Mudd, S.M. Jowitt, T.T. Werner, The world's lead-zinc mineral resources: scarcity, data, issues and opportunities, Ore Geol. Rev. 80 (2017) 1160–1190, https://doi.org/10.1016/j.oregeorev.2016.08.010.
[7] J. Charbucinski, J. Malos, A. Rojc, C. Smith, Prompt gamma neutron activation analysis method and instrumentation for copper grade estimation in large diameter blast holes, Appl. Radiat. Isot. 59 (2003) 197–203, https://doi.org/10.1016/s0969-8043(03)00163-5.
[8] J. Charbucinski, O. Duran, R. Freraut, N. Heresi, I. Pineyro, The application of PGNAA borehole logging for copper grade estimation at Chuquicamata mine, Appl. Radiat. Isot. 60 (2004) 771–777, https://doi.org/10.1016/j.apradiso.2003.12.007.
[9] M. Borsaru Z.J, Application of PGNAA for bulk coal samples in a 4p geometry, Appl. Radiat. Isot. 54 (3) (2001) 519–526, https://doi.org/10.1016/S0969-8043(99)00276-6.
[10] A.A. Naqvi, M.M. Nagadi, S. Kidwai, R. Khateeb ur, M. Maslehuddin, Search of a prompt gamma ray for chlorine analysis in a Portland cement sample, Nucl. Instrum. Methods Phys. Res. Sect. A Accel. Spectrom. Detect. Assoc. Equip. 533 (2004) 591–597, https://doi.org/10.1016/j.nima.2004.06.132.
[11] K. Hossny, S. Magdi, A.Y. Soliman, A.H. Hossny, Detecting explosives by PGNAA using KNN Regressors and decision tree classifier: a proof of concept, Prog. Nucl. Energy 124 (2020), https://doi.org/10.1016/j.pnucene.2020.103332.
[12] K. Oh, Neutronic design of pulsed neutron facility (PNF) for PGNAA studies of biological samples, Nucl. Eng. Technol. (2022), https://doi.org/10.1016/j.net.2021.07.024.
[13] K. Trofimczyk, S. Saraswatibhatla, C. Smith, Spectrometric nuclear logging as a tool for real-time, downhole assay – case studies using SIROLOG PGNAA, in: 11th SAGA Biennial Technical Meeting and Exhibition, 2009, https://doi.org/10.3997/2214-4609-pdb.241.trofimczyk_paper2ples.
[14] L. Tian, F. Zhang, J. Liu, X. Wang, Y. Ti, Monte Carlo simulation of Cu, Ni and Fe grade determination in borehole by PGNAA technique, J. Radioanal. Nucl. Chem. 315 (2018) 51–56, https://doi.org/10.1007/s10967-017-5636-9.
[15] W. Nunes, A. Da Silva, V. Crispim, R. Schirru, Explosives detection using prompt-gamma neutron activation and neural networks, Appl. Radiat. Isot. 56 (2002) 937–943.
[16] S.M. Galib, P.K. Bhowmik, A.V. Avachat, H.K. Lee, A comparative study of machine learning methods for automated identification of radioisotopes using NaI gamma-ray spectra, Nucl. Eng. Technol. 53 (2021) 4072–4079, https://doi.org/10.1016/j.net.2021.06.020.
[17] K. Mark, C.J. Sullivan, An automated isotope identification and quantification algorithm for isotope mixtures in low-resolution gamma-ray spectra, Radiat. Phys. Chem. 155 (2019) 281–286.
[18] F. Zhang, L. Tian, J. Liu, et al., Numerical simulation on scintillator detector response for determining element content in PGNAA system, J. Radioanal. Nucl. Chem. 311 (2017) 1309–1314, https://doi.org/10.1007/s10967-016-5034-8.
[19] W. Metwally, R. Gardner, A. Sood, Gaussian broadening of MCNP pulse height spectra, Trans. Am. Nucl. Soc. 91 (2004) 789–790.
[20] T. Ding, T. Tan, J. Wang, D. Ma, J. Lu, R. Zhang, B. Wu, Ore genesis of the Huangshaping skarn W–Mo–Pb–Zn deposit, southern Hunan Province, China: insights from in situ LA-MC-ICP-MS sulphur isotopic compositions, Geol. Mag. 159 (2022) 981–995, https://doi.org/10.1017/

s0016756822000188.
[21] T. Ding, D. Ma, J. Lu, R. Zhang, S.S. Zhang, Pb, and Sr isotope geochemistry and genesis of Pb–Zn mineralization in the Huangshaping polymetallic ore deposit of southern Hunan Province, China, Ore Geol. Rev. 77 (2016) 117–132, https://doi.org/10.1016/j.oregeorev.2016.02.010.
[22] D. Ramyachitra, P. Manikandan, Imbalanced dataset classification and solutions: a review, Int. J. Comput. Bus. Res. (IJCBR) 5 (2014) 1–29.
[23] S. Qi, W. Zhao, Y. Chen, et al., Comparison of machine learning approaches for radioisotope identification using NaI (Tl) gamma-ray spectrum, Appl. Radiat. Isot. 186 (2022), 110212, https://doi.org/10.1016/j.apradiso.2022.110212.
[24] C.W. Hsu, C.C. Chang, C.J. Lin, A Practical Guide to Support Vector Classification, 2003, pp. 1396–1440.
[25] N.S. Altman, An introduction to kernel and nearest-neighbor nonparametric regression, Am. Statistician 46 (1992) 175–185, https://doi.org/10.1080/00031305.1992.10475879.
[26] M. Scutari, Naive bayes classifiers, in: ICC 2022 - IEEE International Conference on Communications, 2022.
[27] L. Ali, S.U. Khan, N.A. Golilarz, et al., A feature-driven decision support system for heart failure prediction based on statistical model and Gaussian naive bayes, Comput. Math. Methods Med. (2019), https://doi.org/10.1155/2019/6314328.
[28] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[29] A. Tharwat, Classification assessment methods, Appl. Comput. Info. 17 (2021) 168–192, https://doi.org/10.1016/j.aci.2018.08.003.
[30] D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genom. 21 (2020) 6, https://doi.org/10.1186/s12864-019-6413-7.
[31] P. Refaeilzadeh, L. Tang, H. Liu, Cross-Validation, Encyclopedia of Database Systems, 2016, pp. 1–7.