Study On the Composition of Glass Products Based on Soft Voting Model and Grey Correlation Analysis

Chenyang Li [1, #], Jiani Zhang [2, #, *]

[1] School of Electronic and Information Engineering, Changchun University of Science and Technology, Changchun, China, 130013

[2] School of Electro-Optical Engineering, Changchun University of Science and Technology, Changchun, China, 130013

* Corresponding Author Email: 1963171048@qq.com

[#] These authors contributed equally

Abstract. The inspection and identification of ancient glass composition has always been an important research content, this paper on the ancient glass as a medium, for its chemical composition, weathering characteristics and classification rules and other aspects of the study, the establishment of a soft voting integrated learning model to random forest algorithm, SVM algorithms and GBDT algorithms as a base classifier, the use of soft voting on the base classifier for the classification of the results of the integration of the learning, and then to the The chemical composition of unknown artefacts is analysed and classified, and the classification accuracy of soft voting is as high as 96.6%, and the coverage rate of ROC curve reaches 99.1%, which indicates that it is reasonable and effective to solve the problem with the integrated learning model. Finally, by changing the number of decision trees, the classification accuracy rate is found to be unchanged, which verifies that the sensitivity of this model is good. In addition, this paper establishes a grey correlation analysis model to generate grey correlation coefficient matrix heat map for the chemical composition of different categories of glass artifacts, the horizontal coordinate of the heat map is the parent sequence, the vertical coordinate is the subsequence, according to the heat map of the correlation between the chemical composition of different types of glass artifacts to summarize and analyze, and in order to validate the reasonableness of the model, but also its Spearman correlation analysis, the results of the analysis of the two models are similar, so it can be judged that the application of the model is reasonable and effective. The results of the two models are similar, so it can be judged that the results of the grey correlation analysis model applied are reasonable. Finally, a detailed comparison of the chemical composition correlations of the major categories and subcategories is carried out, which is a differential analysis of the correlations of different categories of chemical compositions.

Keywords: Glass composition, soft voting integrated learning, grey correlation analysis.

1. Introduction

Since ancient times, glass is an important and valuable evidence of trade between China and the West, glass was first introduced into China from West Asia and Egypt in the form of jewellery, and then China absorbed its technology, using local materials to improve it, so although the appearance of the differences is not great, but the chemical composition is quite different [1-3]. For example, lead-barium glass in the lead oxide, barium oxide content is higher, and China's own invention of the glass is to have this feature; potassium glass is mainly high in potassium as a combustion agent fired, mainly popular in Southeast Asia, India, and China's Lingnan region [4-6].

Ancient glass is more susceptible to environmental influences and weathering, in the process of chemical reactions, resulting in changes in the proportion of the components in the glass, thus affecting the judgement of the glass category [7-8]. There is a batch of archaeological samples with distinguished categories, and the following problems are proposed to be solved:

(1) According to a number of unknown categories of glass artifacts, their chemical composition is analysed separately to identify the type to which they belong, and the sensitivity of the classification results is analysed.

(2) To analyse the correlation between the chemical compositions of glass artefact samples of different categories, and to compare them, and to analyse the differences through the results of comparison.

2. Materials and Methods

2.1. Data acquisition

The data of this research paper comes from the 2022 National College Students Mathematical Modelling C questions.

2.2. Analysis method

Based on the existing chemical composition content of several unknown categories of glass artefacts, they are analysed and identified as belonging to the category. Due to the limitation of the results given by a single algorithm, multiple classification algorithms are considered to classify them, and then these algorithms are regarded as base classifiers, and integrated learning is carried out, and the classification results obtained after integrated learning are more accurate. Finally, by changing the number of decision trees to observe the accuracy of the classification results, and then analyse the sensitivity of the model [9].

For different categories of glass artefacts, the correlations between their chemical compositions are analysed, and the heat map is drawn after outputting the correlation matrix to analyse their characteristics. In order to verify the accuracy of the results, another method should be considered to analyse the correlation using another method to observe the difference in the results, and if the results are basically the same, then it is considered that the grey correlation analysis model can be better for the correlation analysis of different categories of glass artifacts in terms of the relationship between their chemical compositions. Finally, the differences between the correlations of chemical compositions between major categories and subcategories were analysed [10].

3. Results and Analyses

3.1. Glass chemical composition classification model

3.1.1 Model selection

According to the chemical composition of the unknown category of glass artifacts for analysis, and to identify the type of its belonging to the general choice of Random Forest algorithm for analysis, after reviewing the literature, there is also literature using SVM, GBDT and other algorithms to carry out similar analyses, so this paper will be the integration of multiple machine learning algorithms on the unknown category of glass artifacts to analyse the chemical composition of the chemical composition of the category and to identify the category.

Random Forest is an algorithm that integrates multiple trees through the idea of integrated learning, and its basic unit is a decision tree, combining multiple decision trees together, each time the data set is randomly selected with put back, and at the same time, randomly selected part of the features as inputs, so the algorithm is known as the Random Forest algorithm, which is a Bagging algorithm that uses the decision tree as an estimator, and the result is superior to the classification made by a single decision tree. The result is better than the classification result made by a single decision tree.

SVM algorithm itself is a binary classification algorithm, which is an extension of the perceptron algorithm model, the idea is: in the perceptron model, you can find more than one hyperplane that can be classified to separate the data and optimise it by keeping all the points as far away from the hyperplane as possible. Therefore, as long as the points that are closer to the hyperplane are kept as far away from this hyperplane as possible, the model classification effect will be better.

Relevant formulas:

$$W^T + b = \pm 1 \tag{1}$$

$$y \in \{1, -1\} \tag{2}$$

$$\frac{|(W^T+b)|}{\|W\|_2} = \frac{1}{\|W\|_2} \tag{3}$$

In this paper, we mainly apply the nonlinear SVM algorithm with the following objective function:

$$\begin{cases} min_a(\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i\, a_j y_i y_j (\oint x_i \cdot \oint x_j) - \sum_{i=1}^{n} a_i) \\ s.t.\sum^{n}{}_{\substack{j=1 \\ 0\le a_i \le C}} \;\; a_i y_i = 0 \end{cases} \tag{4}$$

Replacing the inner product $\oint x_i \cdot \oint x_j$ with the kernel function to compute the optimal ai,the values of w and b for the hyperplane can be derived.

The GBDT algorithm is a gradient boosting decision tree, the output of which is an accumulation of several decision trees, each of which is a fit to the residuals of the predictions of the previous combination of decision trees, which is called a classification tree when used to solve classification problems.

The key to this process is to find the decision trees. GBDT considers the negative gradient of the loss function in the current model as an approximation of the residuals of the boosted tree algorithm to fit the trees. First initialise a weak classifier, calculate the negative gradient value of the loss function, then use the data set to fit the next round of models, repeat the calculation of the negative gradient value and the fitting process, and construct the gradient boosting tree using m base models.

3.1.2 Integration Learning Based on Soft Voting

Voting (Voting) is one of the integration learning algorithms widely used in classification algorithms, which is mainly divided into Hard Voting (Hard Voting) and Soft Voting (Soft Voting). Hard Voting is the voting of each classifier with the same weight, and its principle is the principle of majority voting: if a classification result of the base classifier is more than half of the results, the integration algorithm selects the result, and if there is no half of the results, then there is no output; the principle of Soft Voting is also the principle of majority voting, but the weights of the votes of the base classifiers can be defined by themselves, and when the classification effect of the base classifiers has a large difference, the use of Soft Voting should be chosen to give a greater weight to the base classifier with a better classification performance. When the classification effect of each base classifier differs greatly, soft voting should be chosen to give the base classifier with better classification performance a larger weight, so as to optimise the classification results (Figure 1).

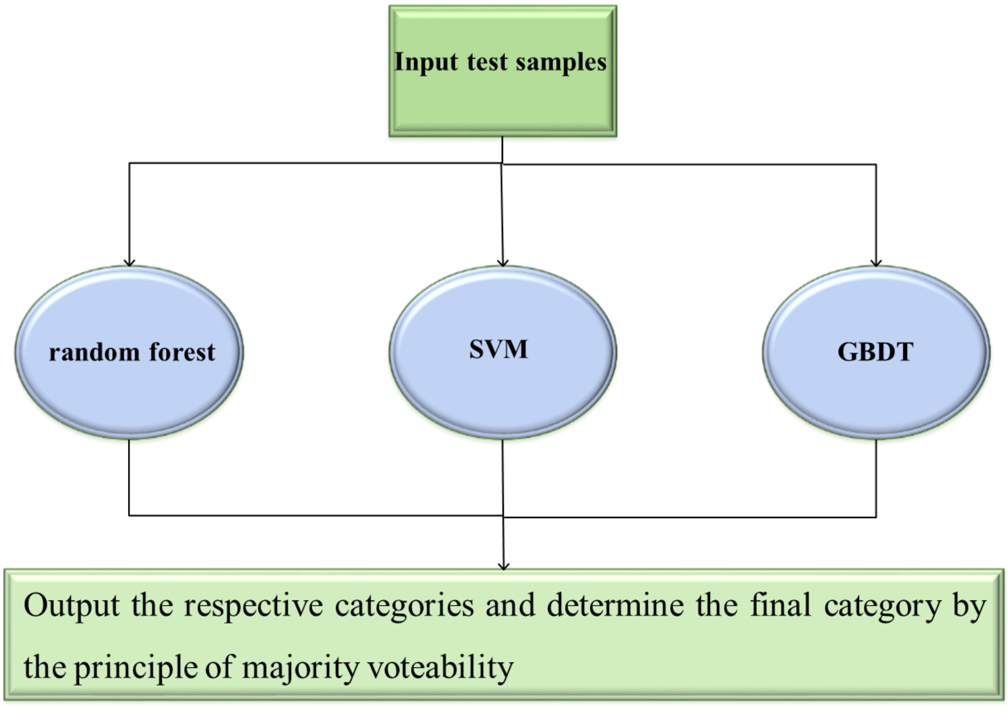

Figure. 1 Schematic diagram of integrated learning soft voting

Firstly, the three base classifiers are fitted to the data and reconciled, and the better parameters are selected to derive the performance, i.e., the accuracy, of the three base classifiers on the dataset, as shown in Table 1 below:

Table 1 Accuracy of base classifiers

sorter	Training set accuracy	Test Set Accuracy	Combined accuracy	**AUC value**
Random forest	0.903	0.913	0.911	0.822
SVM	0.844	1	0.952	0.889
GDBT	0.865	1	0.954	0.942

As can be seen from the data in the table, the closest effect between the training set and the test set is the random forest, and the SVM algorithm is slightly inferior, but the overall performance is good.

In order to judge the results of classification models more accurately, this paper introduces the AUC value for reference, the AUC value is the area covered by the ROC curve, which can comprehensively take into account the recall, precision and accuracy, and the classification models with an AUC value of 0.8 or more are generally considered acceptable. The results show that all the above three classification models are acceptable (Figure. 2).

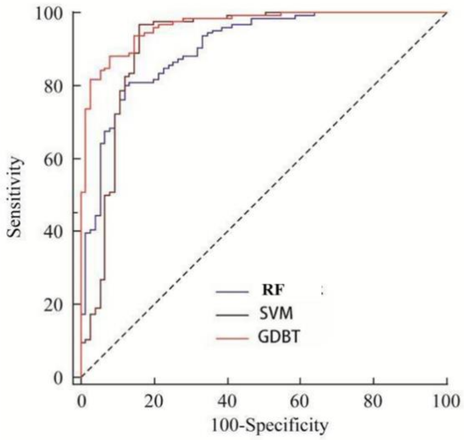

Figure 2 ROC curves of the three base classifiers

Since the effect of each base classifier is different, this paper applies to soft voting, which is based on the combined accuracy of the classifiers on the overall data to determine the weight of each base classifier, the formula is as follows：

$$W_i = \frac{Accuracy_i}{\sum_{i=1}^{3} Accuracy_i} \tag{5}$$

The results of the soft vote classification are shown in Table 2 below:

Table 2 Classification results of unknown artefacts

Artifact number	Classification of projections	Subclassification of projections
A1	High Potassium Glass	Low calcium glass
A2	Lead Barium Glass	Low barium glass
A3	Lead Barium Glass	Low barium glass
A4	Lead Barium Glass	Low barium glass
A5	Lead Barium Glass	Low barium glass
A6	High Potassium Glass	Low calcium glass
A7	High Potassium Glass	Low calcium glass
A8	Lead Barium Glass	High barium glass

The accuracy of Soft Voting is shown in Table 3 below.

Table 3 Accuracy of Soft Voting

sorter	Training set accuracy	Test Set Accuracy	Combined accuracy	**AUC value**
Soft Voting	0.902	1	0.969	0.991

It can be seen that Soft Voting (Soft Voting) performs excellently in classification, indicating that the integrated learning algorithm solves the problem very well and classifies the unknown glass artefacts effectively and reasonably.

3.1.2 Sensitivity Analysis

In order to verify the sensitivity of the classification model, this paper changes the number of decision trees to observe the change of the correct rate of grouping, and the results are shown in Figure 3 below:

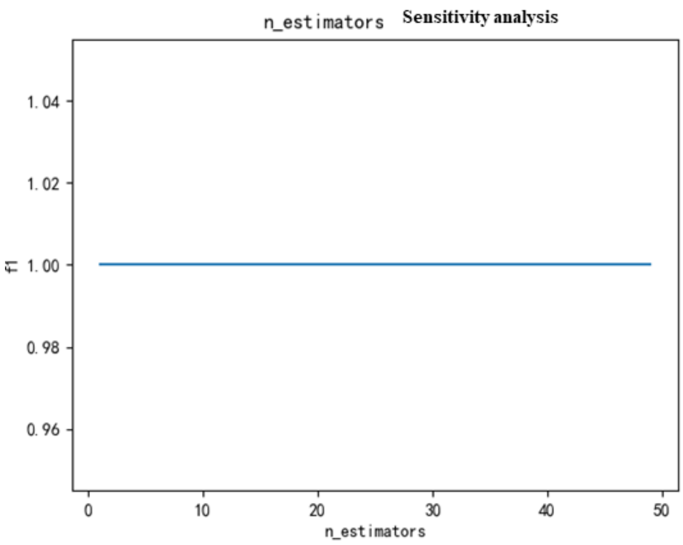

Figure. 3 Sensitivity of the integrated learning classification model

With the change in the number of decision trees, it was found that the accuracy did not change significantly, proving that the sensitivity of the integrated learning model used was better.

3.2. Application of grey correlation analysis to the chemical composition of different categories of samples

Grey correlation analysis is a quantitative method to measure the development and change of a system. The main basis for judging whether the connection is close or not is the similarity of the geometry of the sequence curves. The closer the geometry of the curves, the greater the correlation, and vice versa.

Step1: Determine the analysis sequence.

There are two types of analytical series, one is called the parent series (reference series), that is, the data series reflecting the development of the overall behavioural characteristics of the system, which is equivalent to the dependent variable in the regression analysis; the other is called the subsequence series (comparative series), that is, the data series consisting of factors affecting the development of the system, which is equivalent to the independent variable in the regression analysis. The 14 chemical components in the middle order take turns to be the mother series, and the remaining 13 are the sub-sequences.

Step2: Calculate the correlation coefficient.

Since the existing data have the same dimension, the process of dimensionless quantification is omitted, and the degree of correlation between the subsequence and the parent series, i.e., the correlation coefficient, is calculated directly. Calculate the minimum difference and maximum difference b of the parent-child sequence to get the correlation degree between each element of the subsequence and the corresponding element of the parent sequence:

$$\gamma\big(x_0(k), x_i(k)\big) = \frac{a+\rho b}{|x_0(k)-x_i|(k)+\rho b}, \forall i, k, \dots \tag{6}$$

where ρ is the resolution factor.

Step4: Calculate the mean value of the correlation coefficient, i.e., the degree of correlation between the indicator and the president of the system.

$$\gamma(x_0, x_i) = \frac{\sum_{k=1}^{n} \gamma\big(x_0(k), x_i(k)\big)}{n} \tag{7}$$

For different categories of glass artefact samples, the correlation between their chemical compositions was analysed, and in order to make the results more intuitive, heat maps were plotted for different categories of glass artefacts after outputting the grey correlation matrix, as shown in Figures 4, 5, 6, 7, 8 and 9 below.

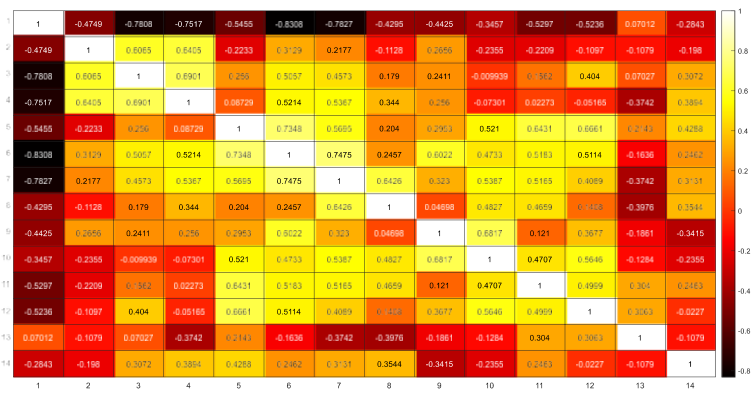

Figure. 4 Thermogram of grey correlation coefficient matrix for high potassium glasses

In high potassium glasses, silica has a low overall correlation with all other chemical components, and iron oxide shows a good correlation with most of the chemical components.

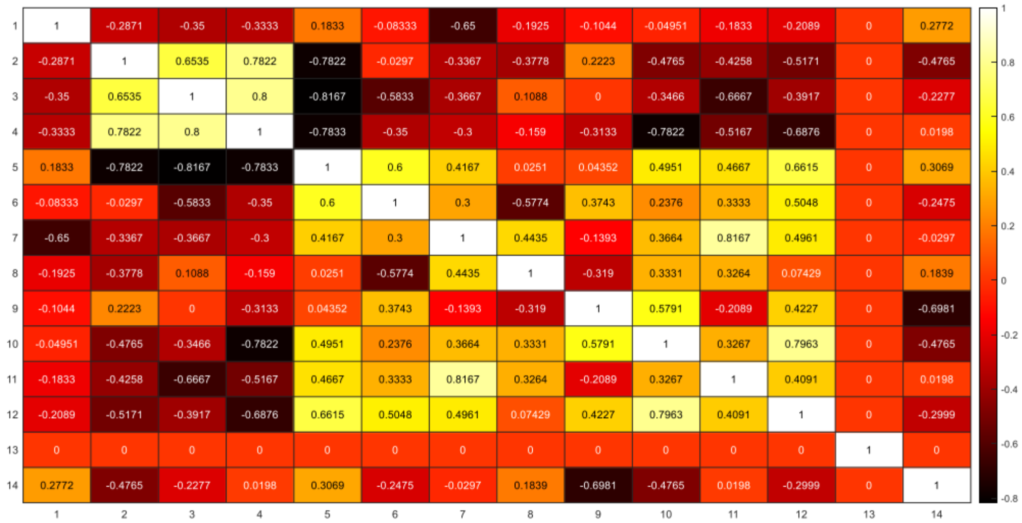

Figure 5 Thermogram of grey correlation coefficient matrix for high-calcium glass

There is little overall correlation between the chemical compositions of high-calcium glass, and somewhat more correlation between the five chemical compositions of magnesium oxide and aluminium oxide.

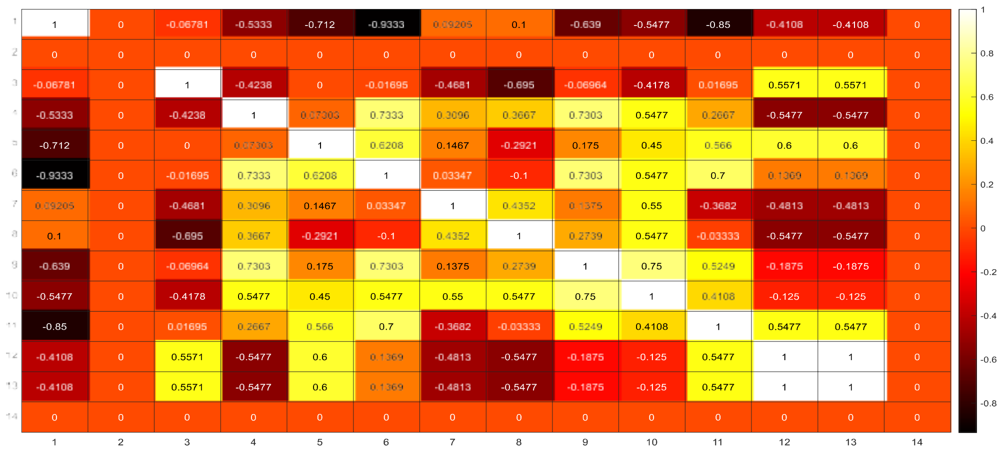

Figure 6 Thermogram of grey correlation coefficient matrix for low-calcium glass

The chemical composition in low-calcium glass is more relevant overall, except for silica, sodium oxide and sulphur dioxide.

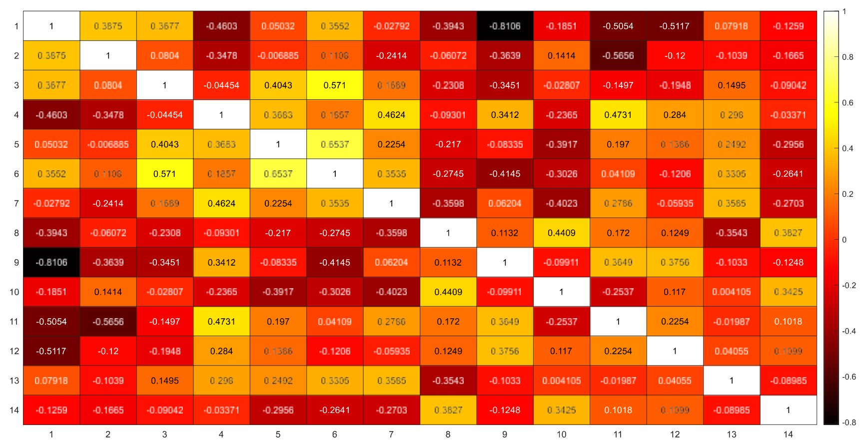

Figure 7 Heat map of grey correlation coefficient matrix of lead barium glass

The correlation between the chemical components of lead-barium glass is moderate overall, and the difference in correlation is small regardless of which chemical component is in the parent series.

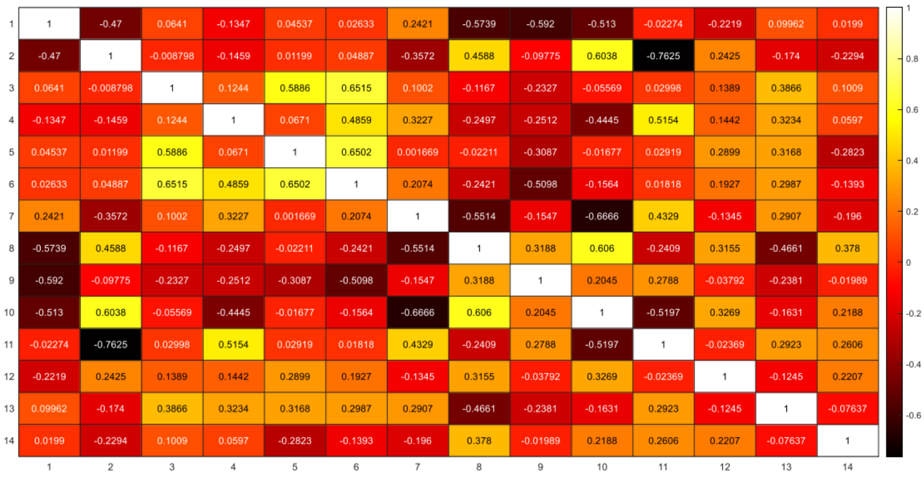

Figure. 8 Thermogram of grey correlation coefficient matrix of high barium glass

The correlation becomes significantly smaller when the six chemical components such as alumina and iron oxide in the high barium glass are used as the parent sequence.

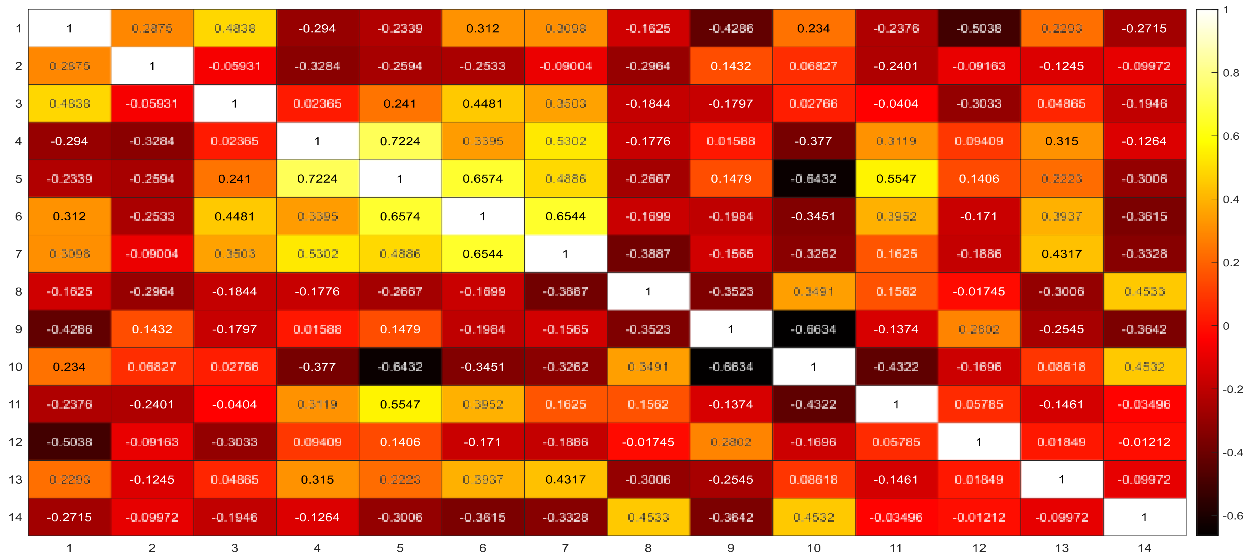

Figure 9 Heat map of grey correlation coefficient matrix for low barium glass

The correlation between the chemical compositions of low-barium glass is small overall, which is especially obvious when barium oxide is the parent sequence.

In order to test whether the method is reasonable, this paper does Spearman correlation analysis on the chemical composition content in different categories of glass artifacts and observes whether the results are consistent with the results of grey correlation analysis.

Spearman's correlation coefficient (Spearman) is calculated as.

$$r_s = 1 - \frac{6\sum_{i=1}^{n} {d_i}^2}{n(n^2-1)} \quad (8)$$

Where n is the number of samples and d represents the grade difference between the data The results of the calculations are shown in Tables 4 and 5 below (in the case of high potassium glass):

Table 4 Spearman's correlation coefficient table

K	silicon dioxide	sodium oxide	potassium oxide	calcium oxide	magnesium oxide	aluminium oxide	iron oxide	eggplant
silicon dioxide	1	-0.475	-0.781	-0.752	-0.546	-0.831	-0.783	-0.430
sodium oxide	-0.475	1.000	0.606	0.640	-0.223	0.313	0.218	-0.113
potassium oxide	-0.781	0.606	1.000	0.690	0.256	0.506	0.457	0.179
calcium oxide	-0.752	0.640	0.690	1.000	0.087	0.521	0.537	0.344
magnesium oxide	-0.546	-0.223	0.256	0.087	1.000	0.735	0.569	0.204
aluminium oxide	-0.831	0.313	0.506	0.521	0.735	1.000	0.748	0.246
iron oxide	-0.783	0.218	0.457	0.537	0.569	0.748	1.000	0.643
eggplant	-0.430	-0.113	0.179	0.344	0.204	0.246	0.643	1.000
lead oxide	-0.443	0.266	0.241	0.256	0.295	0.602	0.323	0.047
barium oxide	-0.346	-0.236	-0.010	-0.073	0.521	0.473	0.539	0.483
phosphorus pentoxide	-0.530	-0.221	0.156	0.023	0.643	0.518	0.517	0.466
strontium oxide	-0.524	-0.110	0.404	-0.052	0.666	0.511	0.409	0.141
tin oxide	0.070	-0.108	0.070	-0.374	0.214	-0.164	-0.374	-0.398
sulfur dioxide	-0.284	-0.198	0.307	0.389	0.429	0.246	0.313	0.354

Table 5 Spearman's correlation coefficient

K	lead oxide	barium oxide	phosphorus pentoxide	strontium oxide	tin oxide	sulfur dioxide
silicon dioxide	-0.443	-0.346	-0.530	-0.524	0.070	-0.284
sodium oxide	0.266	-0.236	-0.221	-0.110	-0.108	-0.198
potassium oxide	0.241	-0.010	0.156	0.404	0.070	0.307
calcium oxide	0.256	-0.073	0.023	-0.052	-0.374	0.389
magnesium oxide	0.295	0.521	0.643	0.666	0.214	0.429
aluminium oxide	0.602	0.473	0.518	0.511	-0.164	0.246
iron oxide	0.323	0.539	0.517	0.409	-0.374	0.313
eggplant	0.047	0.483	0.466	0.141	-0.398	0.354
lead oxide	1.000	0.682	0.121	0.368	-0.186	-0.341
barium oxide	0.682	1.000	0.471	0.565	-0.128	-0.236
phosphorus pentoxide	0.121	0.471	1.000	0.500	0.304	0.246
strontium oxide	0.368	0.565	0.500	1.000	0.306	-0.023
tin oxide	-0.186	-0.128	0.304	0.306	1.000	-0.108
sulfur dioxide	-0.341	-0.236	0.246	-0.023	-0.108	1.000

Upon comparison, the two methods were found to have similar conclusions, and the results of the applied grey correlation analysis can be judged to be reasonable.

3.2.2 Differences in chemical compositional correlations between different categories

The variability of chemical compositional correlations between different categories is shown in Table 6.

Table 6 Variability of chemical constituent correlations between different categories

Category of comparison	Differences in affiliation
High-potassium glass and lead-barium glass	By comparing the grey correlation heat map of high-potassium glass and lead-barium glass, it is found that the overall correlation of the chemical content in high-potassium glass is higher, while the correlation of the chemical content in lead-barium glass is a little bit weaker in comparison, but the effect can be comparable to that of high-potassium glass when potassium oxide and calcium oxide are made into the mother sequence. However, when potassium oxide and calcium oxide are used as the parent series, the effect is comparable to that of high-potassium glass.
High-calcium glass and low-calcium glass	In contrast, the overall correlation of the chemical composition content of high-calcium glass is not as high as that of low-calcium glass, which is particularly obvious when sodium oxide, potassium oxide, calcium oxide, phosphorus pentoxide and tin oxide are used as the parent series, and the correlation of low-calcium glass becomes significantly lower.
High barium glass and low calcium glass	High-barium glass and low-barium glass: The correlation between the chemical composition contents in high-barium glass and low-barium glass is very similar overall, with the exception of copper oxide as the parent series, where the correlation is significantly higher, and the rest of the fluctuations are relatively small.

4. Conclusions

In this paper, we use an integrated learning algorithm, using Random Forest, SVM algorithm and BDGT algorithm as the base classifiers, assigning weights to different algorithms according to their accuracies, and conducting soft voting, which produces more reasonable classification results and improves the fault tolerance. The soft-voting integrated learning model used in this paper is also applicable to all cases where classification is required, and to the scoring of metrics, e.g., credit risk scoring models.

References

[1] Dmytro B,Kateryna T,Tetiana H, et al. Study of Humidity During Sand Dewatering Using a Cone-Shaped Installation [J]. Key Engineering Materials,2023,6920.

[2] Dmytro T,Ilgar D,Vitalii N, et al. Relationship Between Properties of Floating Systems and Flammable Liquids in the Stopping Their Burning Technology[J]. Key Engineering Materials,2023,6921.

[3] Olga S,Andrey S,Olena C, et al. Fire-Resistant Coatings, Obtained by Layer-by-Layer Assembly, in the System of Silicic Acid Gel – Diammonium Hydrogen Phosphate – Urea[J]. Key Engineering Materials,2023,6921.

[4] Olga S,Olena T,M. A B, et al. Study of Phosphorus-Containing Silica Coatings Based on Liquid Glass for Fire Protection of Textile Materials [J]. Key Engineering Materials,2023,6921.

[5] Wu X,Dong F,Xintong J, et al. 27-Gauge vitrectomy vs. 25-gauge vitrectomy in the management of proliferative diabetic retinopathy with preoperative intravitreal injection of conbercept.[J]. Experimental and therapeutic medicine,2023,26(4).

[6] D F, M J H, P R T. Low-coherence and broadband confocal refractometry: reducing the measurement time [J]. Measurement Science and Technology,2023,34(12).

[7] Nandin M,YunSeong J,Ahmed E, et al. Double hits with bioactive nanozyme based on cobalt-doped nanoglass for acute and diabetic wound therapies through anti-inflammatory and pro-angiogenic functions [J]. Bioactive Materials,2024,31.

[8] Jinjin X,Zihao Y,Zhen W, et al. Peridynamic simulation for the damage patterns of glass considering the influence of rate-dependence and pre-defects[J]. Engineering Fracture Mechanics,2023,291.

[9] Zhen-Song C,Jing-Yi L,Jiang-Tao W, et al. BIM-enabled decision optimization analysis for architectural glass material selection considering sustainability[J]. Information Sciences,2023,647.

[10] Fadwa A,Émilie G,Géraldine A, et al. State diagram of whole milk for freeze-drying application [J]. Journal of Crystal Growth,2023,622.