A model for identifying glass artifacts based on SOM clustering analysis and random forest algorithm

Ji Ma [1,#,*], Qian Chen [1,#], Haoxuan Li [1,#], Yongqi Chen [2,#], Yuteng Lu [3,#], Hong Yang [3,#]

[1] School of Materials Science and Chemical Engineering, Harbin University of Science and Technology, Harbin, China, 150080

[2.]School of Measurement and Control Technology and Communication Engineering, Harbin University of Science and Technology, Harbin, China, 150080

[3.]School of Electrical and Electronic Engineering, Harbin University of Science and Technology, Harbin, China, 150080

* Corresponding Author Email: m18595998612@163.com

[#]These authors contributed equally.

Abstract. With the deepening of human awareness of heritage conservation today, the category identification of excavated glass artifacts is particularly important. This paper draws on data related to the chemical composition of some glass artifacts. The data were analyzed by Mann-Whitney U analysis, SOM cluster analysis, and random forest algorithm, and a model was developed to accurately identify the categories of glass artifacts based on their chemical composition. The results show that the three chemical components of potassium oxide, barium oxide, and lead oxide have the greatest influence on the weathering and corrosion of glass artifacts through Mann-Whitney U analysis; the SOM cluster analysis model shows that the glass artifacts with high potassium can be divided into two subclasses, and the glass artifacts with lead and barium can be divided into three subclasses. Finally, we combined the results of the existing analysis and used the random forest algorithm to establish a model for accurate identification of glass artifacts based on their chemical composition. The sensitivity test shows that the model has high robustness and accuracy. This method will play an important role in the accurate identification of glass artifacts of unknown categories in the future.

Keywords: Glass Artifacts, Mann-Whitney U Analysis, SOM Cluster Analysis, Random Forest Algorithm.

1. Introduction

As an important proof of the friendship between the East and the West, glass was introduced into our country as early as the Roman era. In the subsequent development process, it was combined with our native technology to form native glass varieties, mainly lead-barium glass. However, ancient glass was buried underground for a long time, and due to the humidity, as well as the influence of environmental factors such as soil microorganisms and chemical substances [1], which makes the surface of glass products prone to weathering phenomena affecting the composition and character of glass, resulting in changes in the composition of glass, thus increasing the difficulty of judging the type to which it belongs.

Current research on glass focuses on the following aspects: geographic location, materials, and manufacturing processes [2] and also on the analytical identification of the main components of glass [3]. Among the methods used to study the main chemical components of glass are the following: proton excited x-fluorescence analysis (PIXE), neutron activation (INAA), and x-ray fluorescence analysis. Sven A. E. Johansson's team[4] proposed the particle-induced x-ray fluorescence analysis (PIXE) method, which is capable of analyzing elements with low content and low concentration and has the advantages of rapidity, small sampling, nondestructiveness, and high sensitivity, However, the disadvantage of this method is that it can only analyze the elemental content of the sample and not the structure; the INAA method studied by Robert J. Speakman's team[5] has a high accuracy and

the main microelements can be detected. However, it has a large error for some elements whose half-lives of radioisotopes are too short for radiation generation and cannot be measured accurately; A. M. DE FRANCESCO team [6] uses the XRF method to determine the source of archaeological obsidian in the Mediterranean region, which uses the physical principle to detect the elements of the material and can be analyzed qualitatively and quantitatively, with high analytical speed and non-destructive characteristics. However, other can only measure elements but not compounds, and the measured samples must meet many conditions to be allowed, such as smooth surface and homogeneous composition. These studies above mainly applied to the research methods such as correlation analysis, regression analysis, one-way ANOVA [7] in statistics, while the random forest model in machine learning, SOM cluster analysis model and other related models are less applied in this field. The random forest algorithm in machine learning can efficiently process the data and make accurate predictions when there are missing values in the original data, and the SOM algorithm can efficiently analyze the data when processing the chemical composition of glass artifacts to improve the analysis efficiency by saving time in the data analysis process. In this paper, Mann-Whitney U [8] analysis, SOM clustering analysis [9] and random forest algorithm [10] are used to analyze the data related to the chemical composition of glass artifacts, and then achieve accurate prediction of the type of glass artifacts.

2. Model construction

2.1. Mann-Whitney U test model construction

In order to more accurately identify the three main chemical components with the greatest variability in the glass artifacts, this paper first introduced the Mann-Whitney U[11] analysis using glass type as the variable value, which is based on nonparametric rank-sum hypothesis testing for two identical totals except for the overall mean, and then testing whether they are significantly different, and finally finding the three main chemical components with the greatest variability in the glass artifacts. The three main chemical components with the greatest variability were found. The specific process is as follows.

2.1.1 Establishing assumptions

E0: There is no difference between two independent groups of samples, and the variables in the two groups have the same distribution pattern.

E1: There is no difference between the two independent groups of samples, and the variables in the two groups have different distribution patterns.

2.1.2 Perform statistical variable definition

The Mann-Whitney U test uses the U statistic by taking the U_1、U_2 the smaller of the two values. U Statistical correlation is defined as follows:

$$U_1 = n_1 n_2 \pm \frac{n_1(n_1+1)}{2} - R_1 \tag{1}$$

$$U_2 = n_1 n_2 \pm \frac{n_2(n_2+1)}{2} - R_2 \tag{2}$$

U_1、U_2respectively, for the two sets of sample U values;n_1、n_2 are the sample sizes of the two sets of samples, respectively;R_1、R_2 are the rank sums of the two sets of samples, respectively.

2.1.3 Calculating statistical variables

The two groups of samples involved in the statistics were combined to compile and calculate the rank sum of each group of samples. In turn, the Mann-Whitney U test was then calculated for the U_1 and U_2 values. Finally, the U_1 and U_2 The smaller value of the Mann-Whitney U test was selected as the value of the Mann-Whitney U test. U .

2.1.4 Determining critical values and deriving test results

The critical value P was determined based on the sample size and the significance derived from the one-sided or bilateral test.U is less than or equal to the critical value P, then the two independent samples are different.

2.2. SOM Cluster Analysis Model Construction

In order to accurately subclassify each class according to its chemical composition. In this paper, a neural network-based SOM clustering analysis model is used to obtain the desired subclass classification results by selecting 60 sets of data in the dataset as the training set for machine learning and 7 sets of data as the test set for testing its reasonableness. The specific process is as follows.

2.2.1 Data initialization processing

The data obtained from the training set and the test set are normalized separately. Firstly, the smaller random values of the data set are initialized and used as weights. And then the relevant data are initialized.

$$X' = \frac{X}{\|X\|} \tag{3}$$

$$\omega_i' = \frac{\omega_i}{\|\omega_i\|} \quad 1 \le i \le m \tag{4}$$

$\|x\|$and$\|\omega_i\|$ denote the Euclidean parametrization of the sample vector and the weight vector, respectively.

2.2.2 Acquisition of winning neurons

The Euclidean distance between the sample vector and the weight vector is calculated, and the neuron with the smallest distance is the neuron that wins the competition and is called the winning neuron.

2.2.3 Updating the weights

Neurons within the topological neighborhood of the winning neuron are updated and normalized after machine learning. After normalizing the learning rateϑ and topological neighborhood N are updated, the convergence of the operation results is determined computationally.

$$\omega(t+1) = \omega(t) + \vartheta(t,n) * \big(x - \omega(t)\big) \tag{5}$$

$$\vartheta(t,n) = \vartheta(t)e^{-n} \tag{6}$$

ϑ:For the study rate

2.2.4 Model Testing

The test set data were processed in the same way as the training set data and the data from the training and test sets were compared to verify the robustness and accuracy of the SOM clustering analysis model.

2.3. Construction of Random Forest Model

A random forest algorithm model is developed to accurately identify the chemical classes of several glass artifacts based on existing data. The model is trained by using existing data, and then it has the ability to accurately identify unknown glass artifacts. The specific process is as follows:

2.3.1 Data processing

First set the number of samples and divide the training and test sets: the root set the number of samples to 50 and divide the sample data into two categories: training set and test set. According to the different categories of data, 70% of the sample data are taken out as the training set using a loop and the remaining 20% of the sample data are the test set.

The data are then disordered and normalized in turn: the training and test sets are disordered to prevent the recurrence of the same data set, which in turn avoids any impact on the operation of the relevant programs. The data are then normalized so that all the data are compressed in the range [0, 1] and the data are kept in the same numerical unit. The calculation formula is:

$$Z = \frac{X-Min}{Max-Min} \tag{7}$$

When a data is exactly the minimum value, it is normalized to 0. If the data is exactly the maximum value, it is normalized to 1.

2.3.2 Decision tree construction

First based on the training dataset$S$ regression tree is constructed [12]$f(x)$ and secondly, based on the Gini index

$$Gini(S,A) = \frac{S_1}{S}Gini(S_1) + \frac{S_2}{S}Gini(S_2) \tag{8}$$

Select the optimal feature A of the classification tree to generate the CAT tree; then use the CAT pruning algorithm to find the optimal decision tree in the sequence of decision subtreesT_0、$T_1 \cdots\cdots T_n$ and then use CAT pruning algorithm to find the optimal decision tree in the sequence of decision subtreesT_α Finally, a total of 50 decision tree models are constructed in this paper.

2.3.3 Simulation test and performance evaluation

Firstly, in this paper, we conduct simulation tests on the training and test sets. Using simulation tests can help us to better plan the relevant models and accurately locate the test points to verify the effect of the final chosen model. Then, we compare the predicted values with the real values to verify the performance of the model.

2.3.4 Sensitivity test

At the end of the experiment, we perform a sensitivity check on the model to ensure that the developed model has high robustness and accuracy.

The specific process of model creation is as follows.

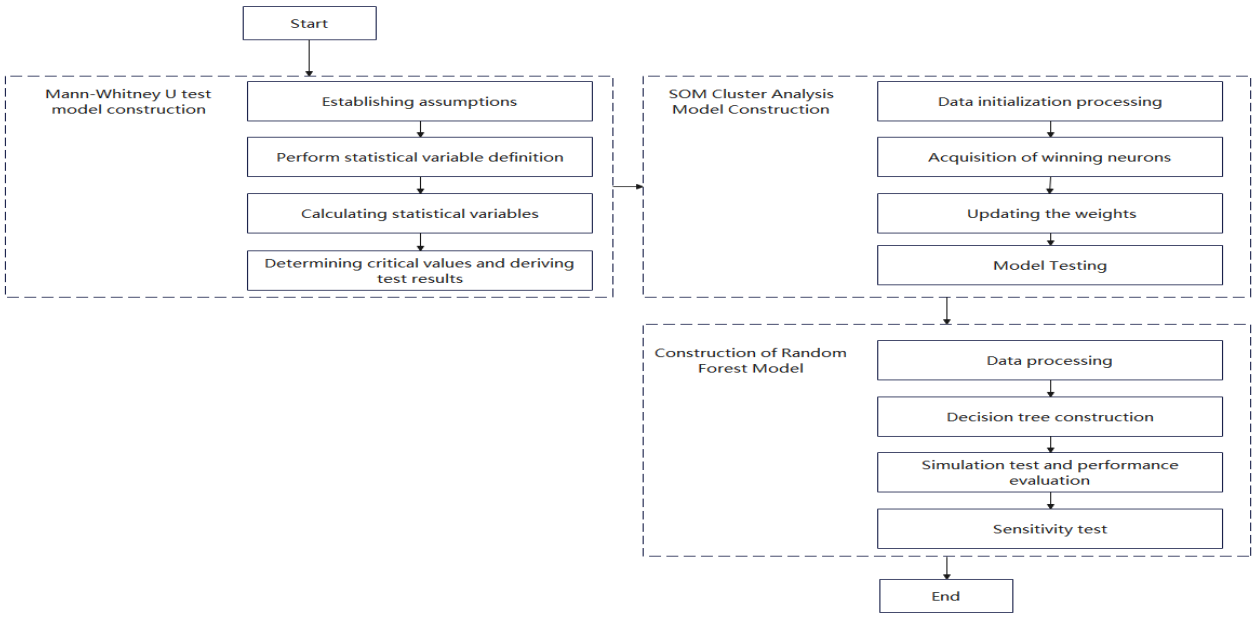

Figure 1. The specific process of model creation

3. Analysis of results

3.1. Acquisition and processing of sample data

The data used in this paper were obtained from the organizing committee of the 2022 National Student Mathematical Modeling Contest, which was tested and processed by professionals with high authority. In this paper, the data with the cumulative sum of component proportions between 85% and 100% are considered as valid data in the subsequent processing.~ Therefore, after calculating the summation, the data sets that are not in this range are excluded and are not used for the subsequent analysis and research. In addition, the vacant values in the existing data were interpolated with 0 to facilitate the subsequent study.

3.2. Experimental procedure and analysis of its results

In this paper, SPSS data analysis software and Matlab programming software were applied to analyze and process the obtained data. In the data processing process, Mann-Whitney U analysis processing, SOM clustering analysis processing, and random forest integrated learning analysis processing were performed in turn, and the desired analysis results were finally obtained.

3.2.1 Mann-Whitney U analytical treatment

Table 1 shows the obtained data for the Mann-Whitney U analysis processing results. Based on the analysis of the resultant data, it can be seen that there are five components with high significance P-values and Cohen's d values: silica, strontium oxide, barium oxide, lead oxide and potassium oxide. However, two chemical components, silica and strontium oxide, need to be excluded in the subsequent classification. The specific reasons are as follows: since the main chemical composition of the glass is silica, it cannot be used as a reference basis; meanwhile, the chemical composition of the added fluxes all have a high content of strontium oxide, so the final distinction between high potassium glass and lead-barium glass is made by the three chemical compositions of barium oxide, lead oxide and potassium oxide. Through further analysis of the chemical composition of the two glass artifacts, the high potassium oxide content in the high potassium glass and the higher content of barium oxide and lead oxide in the lead-barium glass are classified according to the above classification pattern.

Table 1. Mann Whitney U test

Variable Name	Variable Value	Sample size	Median	Standard deviation	Statistical quantities	P	Median Value Difference	Cohen's d Value
Barium oxide (BaO)	High Potassium	18	0	0.842	34	0.000***	8.94	1.407
	Lead Barium	49	8.94	8.331				
	Total	67	6.65	8.425				
Lead oxide (PbO)	High Potassium	18	0	0.514	0	0.000***	31.9	2.574
	Lead Barium	49	31.9	14.947				
	Total	67	23.02	19.513				
Potassium oxide (K_2O)	High Potassium	18	7.525	5.308	763.5	0.000***	7.535	2.286
	Lead Barium	49	0	0.276				
	Total	67	0.2	3.879				

3.2.2 SOM clustering analysis processing

Based on the above analysis, the samples can be classified into high potassium glass and lead-barium glass based on the content and percentage of potassium oxide (K_2O), barium oxide (BaO), and lead oxide (PbO) in the chemical composition of each group of samples. The data of each category were set up in separate tables and it was found that there was some connection between the various properties of the different samples of a single category. In order to explore the linkages between the attributes and to further divide the individual categories, it was decided to learn and divide the data autonomously using the SOM cluster analysis [13] method in machine learning, considering the number of attribute values and the amount of data.

From this we construct two neuron network models [14] to carry out the problem solving. In the graph of the classification effect of the response neuron in Figure 2, the red line represents that the two neurons are connected, and the shade of color indicates the difference in the two neurons, and the darker color indicates the better classification effect. Therefore, it can be concluded that there is a connection between our classifications. And the red color between two neurons in Figure 2 indicates that the difference between two neurons is higher and the classification effect is better.

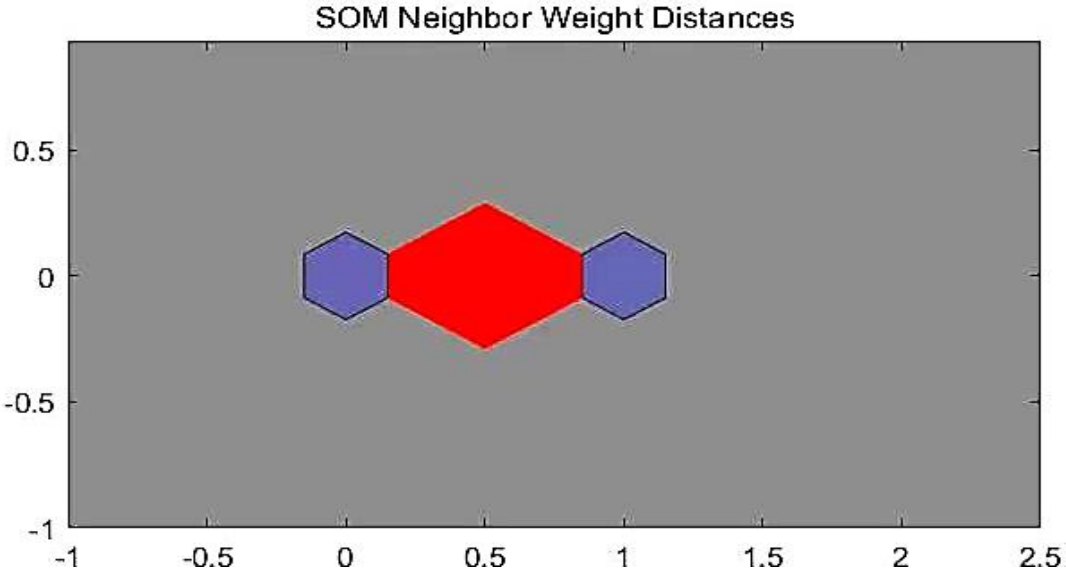

Figure 2. Effect of response neuron classification

In the response neuron effect figure 3, the darker the color in the neuron, the more weight it takes up. In Figure 3, it is clear that black has a greater weight, i.e., more artifacts in classification 1.

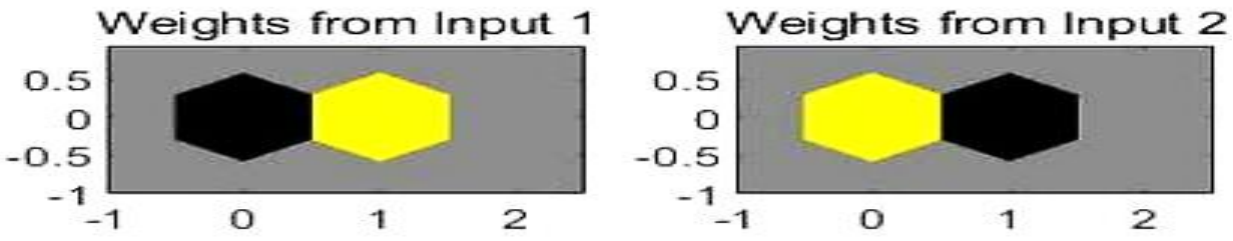

Figure 3. Response neuron effect diagram

After SOM cluster analysis, we can get that the high potassium class glass artifacts can be divided into two subclasses, and the lead-barium class glass artifacts can be divided into three subclasses. The specific results of subclass division are shown in Table 2 and Table 3 below.

Table 2. Subclassification results of lead-barium glass artifacts

Sub-classification results for lead-barium glass artifacts								
Artifact Number	20	56	57	23	43	42	49	08
Subcategories	1	1	1	1	1	1	1	1
Artifact Number	31	40	19	58	24	44	48	54
Subcategories	1	1	1	1	1	1	1	1
Artifact Number	50	51	43	11	49	32	28	30
Subcategories	2	2	2	2	2	3	3	3
Artifact Number	41	52	51	35	55	54	50	25
Subcategories	3	3	3	3	3	3	3	3
Artifact Number	30	53	45	38	46			
Subcategories	3	3	3	3	3			

Table 3. Subclassification results of high potassium class glass artifacts

Subclassification results for high potassium glass artifacts								
Artifact Number	5	18	1	13	16	21	27	6
Subcategories	1	1	1	1	1	1	1	1
Artifact Number	3	7	9	10	22	3	14	12
Subcategories	1	2	2	2	2	2	2	2

After deriving the SOM clustering analysis results, we conducted a sensitivity test on the model and obtained a new graph of SOM clustering analysis results, and the classification results were in full agreement with the above classification results, which proved the accuracy of the model. The specific classification results of the sensitivity test are shown in Figure 4 and Figure 5 below, respectively.

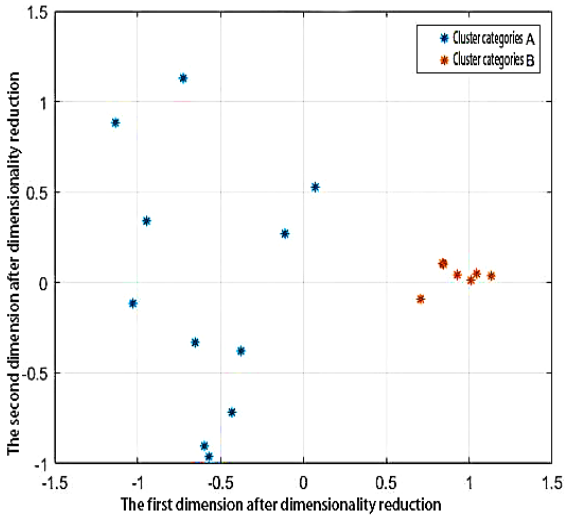

Figure 4. high potassium class glass artifact classification results

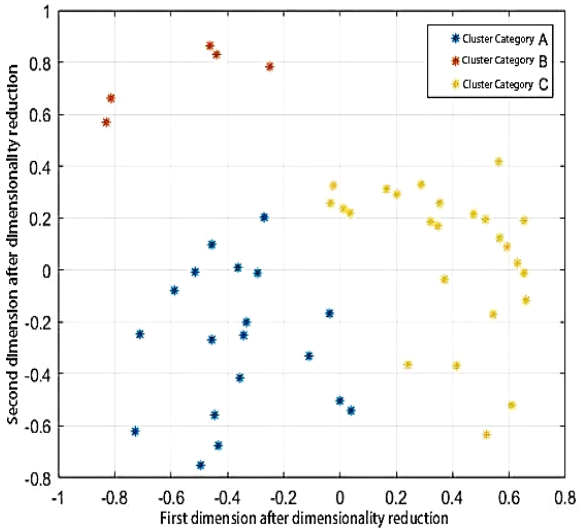

Figure 5. lead barium glass class artifact classification results

3.2.3 Prediction results of random forest model

According to the feature selection algorithm of random forest [15] and from the importance analysis graph of random forest model features, we can see that among the features selected in the decision tree, the 9th feature has the highest importance, so it has the greatest impact on the output of our results; the 4th feature has the lowest importance, so it has the least impact on the output of our results. The specific results are shown in Figure 6.

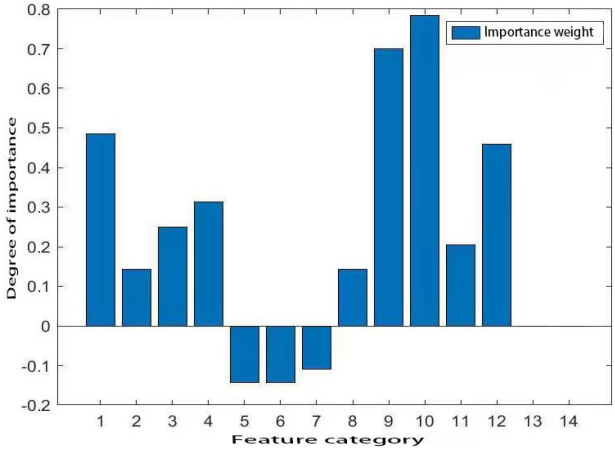

Figure 6. Results of importance analysis of model features

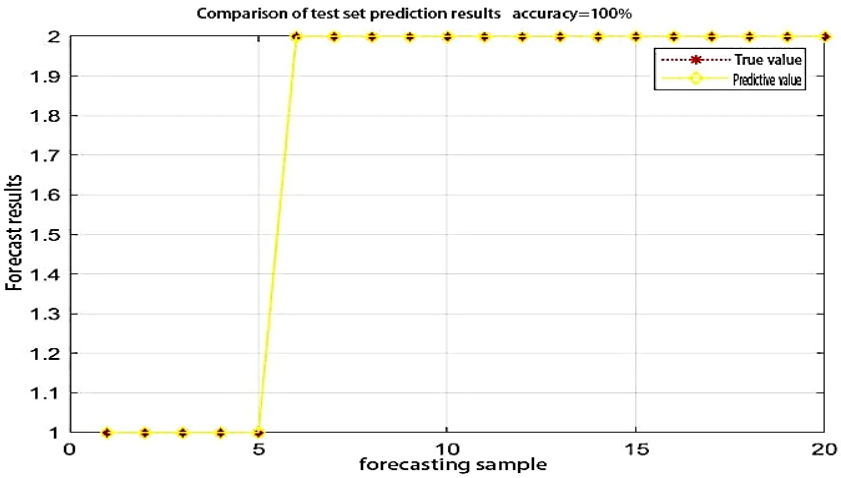

Figure 7. Results of sensitivity test analysis

The test set data of the model was subjected to sensitivity testing, and some of the data of the original make sample was modified before testing. The modifications were made in the interval of [-5,5] for the increase or decrease of the original data, and the modified data samples were run out of the prediction of the modified data using the original program to test its sensitivity. The specific modifications are shown in Table 4 below.

Table 4. Comparison table of sensitivity detection data

Sampling Points	01	14	06 Part 2	03 Part 1	51 Part 2	11
Element Type	SiO2	CaO	MgO	Al2O3	CuO	BaO
Original data	69.33	8.23	1.73	4.06	0.75	14.61
New Data	72	10	0.73	3	2.75	12.61

According to the sensitivity detection results, it can be seen that the accuracy of the predicted value of the test set compared with the real value of the results is 100%, which is basically consistent with the results of the sample data, proving that the program has high robustness and sensitivity. The specific results are shown in Figure 7.

Finally, the model was used to predict the unknown categories in the original data samples, and the high potassium glass artifacts and lead-barium glass artifacts were quantified as 1 and 2, respectively, before using the model for prediction. the final prediction results are shown in Table 5 below.

Table 5. Unknown category for prediction results

Artifact Number	A1	A2	A3	A4	A5	A6	A7	A8
Running results	1	2	2	2	2	1	1	2
Actual Type	High Potassium	Lead Barium	Lead Barium	High Potassium	Lead Barium	High Potassium	High Potassium	Lead Barium

4. Conclusion

Since ancient glass artifacts have been buried underground for a long time, they have been subjected to the corrosion of chemical substances in the soil for a long time, which increases the difficulty of determining the type to which they belong. In this paper, we analyzed the data related to the chemical composition of glass artifacts by using Mann-Whitney U analysis, SOM cluster analysis, and random forest algorithm with the help of computer software such as Matlab and Spass, respectively. The Mann-Whitney U analysis was performed by using the glass type as the variable value to identify the three main components with the greatest variability: potassium oxide, barium oxide, and lead oxide. Then, a neural network-based SOM clustering model was used to finally classify the high potassium glass into two subclasses and the lead-barium glass into three subclasses. Finally, using the random forest algorithm, combined with the classification results already obtained, a model for accurate identification of glass artifacts based on their chemical composition to the categories they belong to was established with the glass category as the target and the relationship between different chemical compositions as the internal nodes. The model is of great significance for the accurate identification of glass artifacts in archaeological excavations in China.

References

[1] Cheng Qian, Zhang Jianlin. Scientific analysis and research on glass beads excavated from the Xiaoling Tomb of Emperor Wu of the Northern Zhou Dynasty[J]. Archaeology and Cultural Properties, 2011(1):107-112

[2] Conradt, Reinhard (2019). Prospects and physical limits of processes and technologies in glass melting. Journal of Asian Ceramic Societies, (), 1–20.

[3] Liu, S., & Zhang, Y. (2022). Based on the analysis and identification of ancient glass products. Highlights in Science, Engineering and Technology, 21, 212–221.

[4] Ishii, Keizo (2019). PIXE and Its Applications to Elemental Analysis. Quantum Beam Science, 3(2), 12–.

[5] Sharma, V., Acharya, R., Bagla, H. K., & Pujari, P. K. (2021). Standardization of an external (in air) PIGE methodology using tantalum as a current normalizer in conjunction with INAA for rapid and non-destructive chemical characterization of "as-received" glass fragments towards forensic applications. Journal of Analytical Atomic Spectrometry, 36(3), 630–643.

[6] Li, Fei; Ge, Liangquan; Tang, Zhuoyao; Chen, Yi; Wang, Jing (2019). Recent developments on XRF spectra evaluation. Applied Spectroscopy Reviews, (), 1–25.

[7] K. P. Sinaga and M. -S. Yang, "Unsupervised K-Means Clustering Algorithm," in IEEE Access, vol. 8, pp. 80716-80727, 2020.

[8] Wall Emerson, R. (2023). Mann-Whitney U test and t-test. Journal of Visual Impairment & Blindness, 117(1), 99–100.

[9] C. Yang, S. Su, X. Ju and J. Song, "A Mobile Sensors Dispatch Scheme Based on Improved SOM Algorithm for Coverage Hole Healing," in IEEE Sensors Journal, vol. 21, no. 18, pp. 21080-21089, 15 Sept.15, 2021.

[10] Chen, Y., Zheng, W., Li, W., & Huang, Y. (2021). Large group activity security risk assessment and risk early warning based on random forest algorithm. Pattern Recognition Letters, 144, 1–5.

[11] Lin, T., Chen, T., Liu, J., & Tu, X. M. (2021). Extending the Mann-Whitney-Wilcoxon rank sum test to survey data for comparing mean ranks. Statistics in Medicine, 40(7), 1705–1717.

[12] J. -W. Baek and K. Chung, "Context Deep Neural Network Model for Predicting Depression Risk Using Multiple Regression," in IEEE Access, vol. 8, pp. 18171-18181, 2020.

[13] Ghadiri, S. M. E., & Mazlumi, K. (2020). Adaptive protection scheme for microgrids based on SOM clustering technique. Applied Soft Computing, 88, 106062.

[14] Zhang, Shuai; Xu, Jiayue; Li, Mengdi; Zhao. Simulation analysis of firing activity of transcranial magnetoacoustic electrical stimulation neural network based on cortical neuron model[J]. Journal of Electrical Engineering Technology,2021, (18):3851-3859.

[15] Zhou, Jinglei; Zhou, Zhi; Cui, Lin. Application of variational modal decomposition and random forest feature selection algorithm for anomalous sound classification of loudspeakers[J]. Vibration and Shock,2022,41(20):277-28