Analysis of the chemical composition of ancient glass products and identification of the types to which they belong

Haotian Lin [*, †], Jiannan Lin [†], Jianze Lin [†]

School of Ocean Information Engineering Jimei University Xiamen, China

* Corresponding Author Email: linhaotian@jmu.edu.cn

[†]These authors contributed equally.

Abstract. The Silk Road is a land artery connecting China and the West, an important passage for East-West trade, and glass products are one of the treasures of early exchanges. Due to different production techniques, the composition and performance of Eastern and Western glass are often very different, but the identification of glass has always been a problem in ancient times. Aiming at the analysis of the composition and source of ancient glass products, this paper designs a mathematical model to avoid duplication of work. Based on the data, this paper calculates the classification rules of the main components of high potassium glass, lead-barium glass; and by selecting the appropriate chemical components as the subcomponents for each category, the specific division methods and division results are given, and the rationality and sensitivity of the classification results are analyzed. And the chemical composition of the unknown category of glass artifacts is identified by the model to identify the type they belong to, and the sensitivity of the classification results is analyzed.

Keywords: Ancient glassware; DBI index; gradient ascent tree; random forest.

1. Introduction

The ancient Silk Road has rich historical and cultural value, and has made important contributions to the exchange of material and spiritual culture between China and the West [1]. Among them, the most emblematic of cultural differences are the glass products of the Silk Roads. After glass was introduced into China, it quickly obtained localized technical support in China, and differentiated different processes with the production area, which has high artistic value [2]. The chemical composition of glass made by different processes is often different, and glass is easily weathered in an easy burial environment, and weathering will cause its internal components to change, so the identification of glass products has become a hot research problem in archaeology.

Based on the methods of chemistry, mathematics and information science, this paper theoretically discusses the model of this problem, discusses the correlation of the weathering degree, appearance, category and glass category of glass, and further quantifies and subdivides the glass components on the basis of this preliminary exploration, and provides intuitive classification criteria for newly unearthed unclassified glassware to identify their types and perform sensitivity analysis on the classification results. , which provides a new direction for archaeological research to speculate on the process of glass [3].

2. Model Assumptions and Notation

2.1. Assumptions [4]

Assume that the type of weathering has no effect on the results of artifact identification

Assume that the content of all undetected components is 0

Assume that different parts of the artifacts are made with the same type of glass

Assume that there is no error in the classification of artifacts in all valid data

It is assumed that the weathering process is not subject to much environmental pollution and does not cause the production of new chemical components.

2.2. Notations

Important notations used in this paper are listed in TABLE 1.

Table 1. Notations.

Symbols	Description
R^2	Goodness of fit
MAPE	Absolute percentage error
MSE	Mean square error
K	Percentage of corresponding elements in oxide species
M(R)	Molar mass of a substance
W	Percentage of corresponding oxides in the glass component
S_i	The average distance between the point inside the class and the center of mass of the class
M_{ij}	Distance between two class centers
R_{ij}	Similarity index

3. Model construction and solving

3.1. Determination of the main classification principle.

For the main classification principle, which has been clearly determined based on the content of potassium, barium and lead, first, it is necessary to determine the specific determined content of each element in the artifact by first finding the percentage of that element in the oxide.

$$\mathrm{K} = \frac{nM(R)}{M(R_nO)} \tag{1}$$

Then, its share in the overall artifact can be known.

$$\frac{W_k K_k}{W_{P_b} K_{Pb} + W_{BQ} K_{BQ}} \tag{2}$$

Where K is the percentage of elements in the oxide and W is the percentage of the corresponding oxide in the composition of the glass.

Subsequently, the absolute numerical differences and multiples were calculated for each of the three factors, and then the numerical values were sorted to count the number of data where the classification deviated from the numerical ordering

In this paper, the multiplicity was chosen as the numerical principle for the determination of the main classification, which is also more in line with the general laws of statistics. It is also noted that there are three values deviating from the classification layer, but they all have very obvious characteristics, two data do not contain potassium, but also do not contain the elements of lead and barium at the same time, while one data contains lead and barium, but the amount of detection is very small, which may be an experimental error, so a principle is proposed.

When the sum of lead and barium elements in the overall composition of strictly less than 4%, regardless of how much potassium is present, are high potassium glass, otherwise, when and only when the ratio of potassium / (barium + lead elements) is greater than or equal to 0.3, the glass is considered to be high potassium glass, otherwise, lead-barium glass.

3.2. Clustering of non-principal component variables

A dependent variable-autonomous variable model is needed to first cluster the data simply and assign labels; otherwise, supervised learning will be difficult. Again, two models are considered, a clustering approach based on systematic clustering and a clustering approach based on K-means++ clustering [5].

For the two clustering approaches, which are only briefly described in this paper.

Firstly, potassium oxide, lead oxide and barium oxide, which contain the main components, are deleted. For the remaining parameters, based on the common sense of geology, silica is greatly affected by weathering and is not suitable as a subclass classification indicator, so they are also removed, and the remaining 10 parameters are used to cluster the samples.

First of all, the system clustering, the specific number of class clusters of the system clustering selection, need to be based on the elbow rule [6]to make a judgment, give the image of the elbow rule as follows Fig. 1.

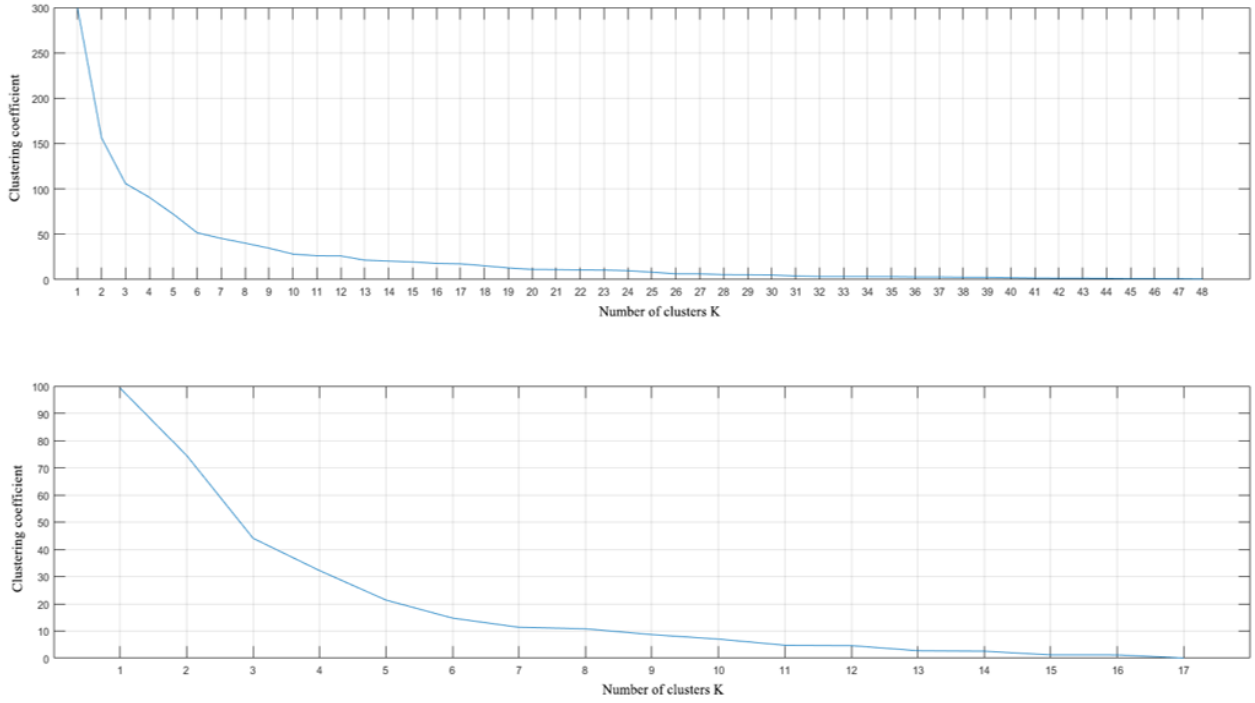

Fig 1. Elbow Law.

It can be seen that the choice of class clusters of three (the derivatives change significantly) or four (the image tends to be smooth) are better choices. Since the main class classification has been performed, the sub classification should not be too cumbersome for practical work, and three is chosen as the final number of clusters. The results of K-means++ clustering, the amount of data in the class clusters is more average, so this paper uses the results of K-means++ clustering as the benchmark for subsequent modeling.

In this paper, a feature engineering based parametric analysis approach is proposed. Feature engineering considers that the attributes that are valuable for the current learning task are relevant features, while the irrelevant attributes are irrelevant features, and among the relevant features, the impact of different features on the final result is obviously different, and the magnitude of the impact of feature changes on the result is called feature importance.

Consider a decision tree, in the process of making decisions from top to bottom, each decision is a feature selection, and the feature selected first has a higher feature importance, and this decision tree is not always a standard tree, and each feature may be linked and become a directed graph, at this time, the new features that can be obtained by other features are redundant features.

For the problem analysis of feature engineering, it is necessary to firstly select suitable features, secondly select suitable feature algorithms, and finally check the features.

In this paper, we propose three models that can be used for feature engineering to solve the feature importance.

3.3. Gradient ascent tree (GBDT)

The principle of GBDT [7] is to obtain the predicted value by summing the results of all weak classifiers and then fitting the gradient/residual of the error function to the predicted value with the

next weak classifier (this gradient/residual is the error between the predicted value and the true value), GBDT can approximate the residual with a negative gradient. The representation of its weak classifier is the individual trees. As shown in the Fig. 2: Y=Y1+Y2+Y3

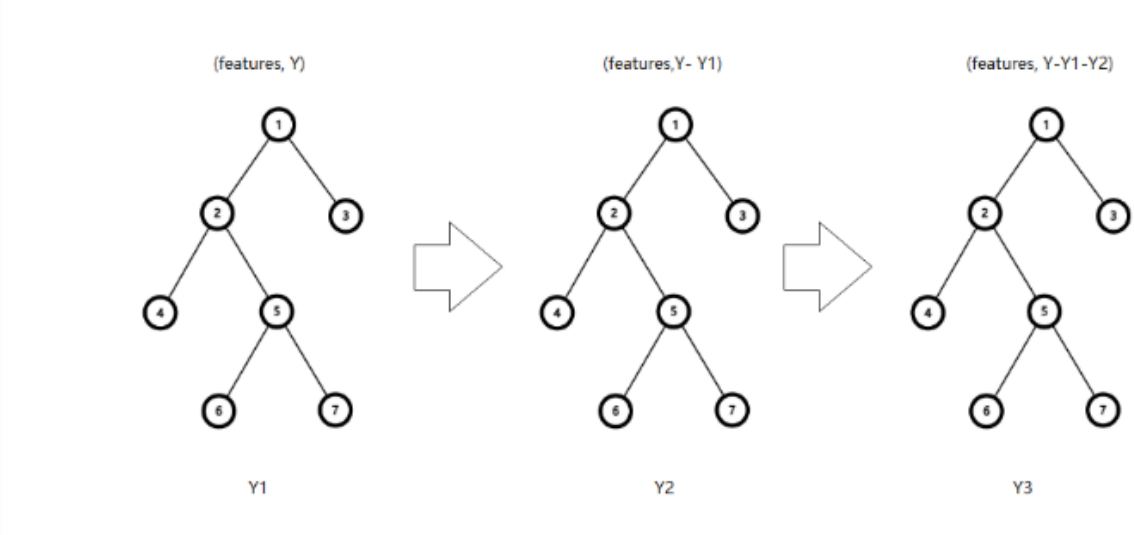

Fig 2. Gradient ascent tree schematic.

Under the regression task, GBDT has a predicted value for each sample at each iteration of the round, and the loss function at this point is the mean squared loss function:

$$l(y_i, y^i) = \frac{1}{2}(y_i - y^i)^2 \tag{3}$$

In this case, the negative gradient is calculated as follows:

$$-[\frac{\partial l(y_i, y^i)}{\partial y^i}] = y_i - y^i \tag{4}$$

Therefore, when the loss function is chosen as the mean-square loss function, the value of each fit is (the true value - the value predicted by the current model), i.e., the residual. At this point, the variable y^i, which is the value of the current prediction model, is the negative gradient of it.

In the traditional gradient ascent tree, only first-order derivatives are generally used for the residuals, but the gradient ascent tree used in this paper additionally considers the second-order derivatives and adds regularization to prevent overfitting.

3.4. Neural network-based fitting

Fully connected network (MLP) [8] is a modeling approach based on interconnecting neurons in each layer and fitting the output results by different weights and activations on the link path, which is divided into two stages: forward propagation and backward propagation.

(1) MLP forward propagation

Forward propagation is actually the prediction process, which is the process from input to output.

(2) MLP back propagation

The backpropagation of MLP over is the training process for the neural network. Here, the training is the weights on each of the previous edges.

(3) Total error

$$E_{total} = \sum \frac{1}{2}(target - output)^2 \tag{5}$$

Target is the correct value of this sample and output is the value predicted in this round.

$$E_{ol} = \frac{1}{2}(target - output)^2 \tag{6}$$

Here there are two outputs, so the summation is done for all outputs and the final calculation of E_total

$$E_{total} = E_{o1} + E_{o2} \tag{7}$$

Extending to N outputs (classifications), the error of the outputs (generally classification probabilities) in each of the N classifications is found separately and finally summed.

The total error here in the following application depends mainly on how many errors are accepted (if only one error is accepted, then only one error is used).

(4) Output layer parameter update

Take the weight parameter w5 as an example, if you want to know how much influence w5 has on the overall error, you can use the overall error to find the partial derivative of w5: (chain rule of derivation)

$$\frac{\partial E_{total}}{\partial w_5} = \frac{\partial E_{total}}{\partial out_{o1}} \times \frac{\partial out_{o1}}{\partial net_{o1}} \times \frac{\partial net_{o1}}{\partial w_5} \tag{8}$$

$$\frac{\partial net_{o1}}{\partial w_5} \times \frac{\partial out_{o1}}{\partial net_{o1}} \times \frac{\partial E_{total}}{\partial out_{o1}} = \frac{\partial E_{total}}{\partial w_5} \tag{9}$$

(5) Implicit layer parameter update

$$\frac{\partial E_{total}}{\partial w_1} = \frac{\partial E_{total}}{\partial out_{\mathrm{h1}}} \times \frac{\partial out_{\mathrm{h1}}}{\partial net_{\mathrm{h1}}} \times \frac{\partial net_{\mathrm{h1}}}{\partial w_1} \tag{10}$$

$$\frac{\partial E_{total}}{\partial out_{\mathrm{h1}}} = \frac{\partial E_{o1}}{\partial out_{\mathrm{h1}}} + \frac{\partial E_{o2}}{\partial out_{\mathrm{h1}}} \tag{11}$$

(6) The neural network used in this paper

The β-mish activation function is a new activation function proposed in 2022 [9], which is the predecessor of the mish activation function [10], and has certain advantages over the Swish activation function proposed by Google Brain, and the advantages are even more obvious when compared to the commonly used Relu activation function, whose formula and derivatives are

$$f(x) = x \cdot \tanh\left(\ln\left(1 + e^{\frac{ax}{\sqrt{\beta+x^2}}}\right)\right) \tag{12}$$

$$f'(x = \frac{\frac{a\beta x\sqrt{\beta+x^2}e^{\frac{ax}{\sqrt{\beta+x^2}}}}{\cosh^2(softplus(x))} + \left(x^2+\beta\right)^2\left(1+e^{\frac{ax}{\sqrt{\beta+x^2}}}\right)\tanh\left(softplus(x)\right)}{(x^2+\beta)^2\left(1+e^{\frac{ax}{\sqrt{\beta+x^2}}}\right)} \tag{13}$$

Where the value of α over β is 5. In this paper, we take β as 150

Then the expression of each layer of nodes is

$$Output = p - mish\left(\sum_{i=1}^{n} power_i \times lastlayeroutput_i\right) \tag{14}$$

The Alpha Drought technique, on the other hand, is based on the optimization of the Drought technique, which is based on the principle of making some neurons in the training network deactivate during each training session, thus making the fit not overfitting.

In the choice of optimizer, the most classical SGD optimizer is abandoned and the Nadam optimizer is used, which is an optimized version of the adam-based adaptive learning rate optimizer.

The training results are presented in Table 2 below.

Table 2. Training results.

TYPE	VALACC	ACC
HIGH	0.75	0.928
POTASSIUM	0.83	0.948

It can be seen that the fitting effect is poor because the data volume is small but the dimensionality is large, and the dispersion of the original data is large and the normality of the distribution is not good, so in the subsequent calculations, the neural network is no longer considered in this paper.

3.5. Random Forest

Random forests [11] belong to the Bagging (short for Bootstrap AGgregation) method in integrated learning. A random forest is composed of many decision trees, and there is no association between different decision trees. When a classification task is performed, new input samples enter and each decision tree in the forest is allowed to judge and classify them separately, and each decision tree gets a classification result of its own.

When constructing a decision tree, the current label can be considered as the entropy of the classification system, and the entropy of the system always decreases as the number deepens.

To make the decision tree most efficient, the entropy of the decision tree needs to be made to decrease the fastest, which is also based on the gradient fastest descent method in machine learning.

Random forest is to plant multiple decision trees in the same forest, each decision tree has different decisions, and randomly draw features from the training data, then learn the features and give different decision weights to each decision maker. Since decision trees are naturally interpretable, random forests also inherit this advantage to some extent, but they also have the disadvantage that they do not give a continuous output. When performing regression, random forests are not able to make predictions beyond the range of the training set data, which may lead to overfitting when modeling certain specific noisy data.

3.6. Calculation of feature importance

For the tree model feature importance, using gini coefficients, he is judging how much each feature has contributed to each tree, and then the practice of averaging. For the neural network, such a calculation is more difficult, so the attention returns again to the narrowband Gaussian noise, where the output of the neural network is not an exact classification, but the probability that corresponds to the possibility of the label, and it is only necessary to perturb each feature with noise, and the importance of each feature can be determined by the final probability change.

First, the importance of the features based on the gradient ascent tree is computed.

Finally, the importance of features based on random forest is calculated

Combining the results of the feature importance calculations for both models, the following conclusions can be drawn

For lead-barium glass, phosphorus should be used as a subclass classification criterion, while for high-potassium glass, calcium should be used as a classification criterion.

3.7. Characterization of the model

In order to evaluate the merit of clustering, it is necessary to consider its DBI[12,13], which takes into account the intra-class sample similarity and the inter-class sample difference , and the smaller its value is, the higher the effectiveness of clustering is characterized . Suppose there are m sequences, and these sequences are clustered into n classes by the algorithm, using the DBI clustering effectiveness evaluation method. The specific formula is as follows.

$$R_{ij} = \frac{Si+S_j}{M_{ij}} \tag{15}$$

Calculate Si

The DBI formula begins with the definition of the Si variable. Si calculates the average distance of the data within a class to the cluster center of mass, representing the dispersion of each time series in cluster class i, and is calculated as

$$Si = \left(\frac{1}{T_i} \sum_{j=1}^{T_i} |X_j - A_i|^p \right)^{1/p} \tag{16}$$

Calculation M_{ij}

The calculation formula is.

$$M_{ij} = \|A_i - A_j\|_2 = \left(\sum_{k=1}^{N} |a_{ki} - a_{kj}|^p \right) \tag{17}$$

Calculation R_{ij}

The calculation formula is.

$$R_{ij} = \frac{Si + S_j}{M_{ij}} \tag{18}$$

Calculation DBI

Based on the above formula, we make an n^2 nested loop based on the number of cluster classes n. For each cluster class i, calculate the maximum value ofR_{ij}:

$$D_{ij} = \max_{j \neq i} R_{ij} \tag{19}$$

That is, the maximum similarity value of cluster class i with other classes, that is, the worst result is taken out. Then the DBI index is obtained by taking the average value of the maximum similarity of all classes, which is calculated by the formula

$$DBI = \overline{D} = \frac{1}{N} \sum_{i=1}^{N} D_i \tag{20}$$

A comparison of the DBI indices of the two clustering approaches is given by calculating (Table 3).

Table 3. Comparison of the DBI indices of the two clustering methods.

High Potassium Kmeans	Lead Barium Kmeans	High potassium system	Lead Barium System
62%	0.1690	0.1106 0.1389	0.1106

It can be seen that the DBI index of both clustering methods are relatively small, clustering is more effective, while systematic clustering is better than K-means++, but because the results obtained by systematic clustering cluster data gap is too large, but then the lack of generalization ability in practical applications.

Next, the validity and robustness of the three models need to be verified. Verifying the validity cross-entropy has been calculated during the learning process, and the confusion matrix is used in this paper to perform another verification. When training the models, the original data has been randomly cut, and the cut test set can be used to verify the confusion matrix.

For the gradient ascent tree, the confusion matrices for the two major categories of high potassium and lead-barium are given in this paper as Fig. 3.

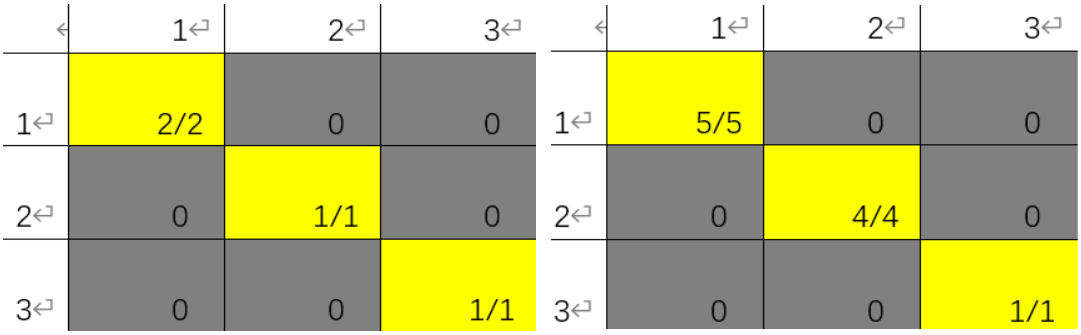

Fig 3. Confusion matrix for gradient ascent tree.

Similarly, for random forests, two corresponding confusion matrices are given in this paper as in Fig. 4.

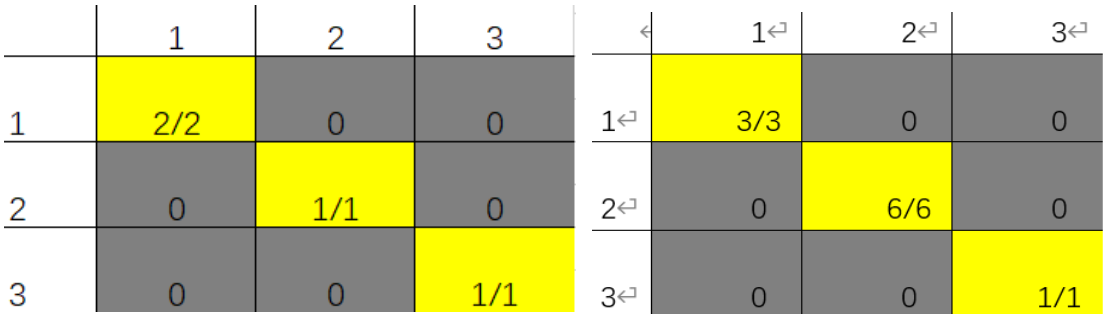

Fig 4. Confusion matrix for random forests.

It can be seen that both algorithms successfully pass the confusion matrix test.

Finally, the sensitivity of the model needs to be verified. For the classification model, the sensitivity of the model should not be too strong, and the result of the classification cannot be easily changed for slight perturbations, otherwise the generalization ability of the model will be greatly reduced, but it should not be too weak either. Therefore, narrowband Gaussian white noise is used, and each variable is finely perturbed to observe the change of the model prediction category

It can be seen that the robustness of the model is always, and the classification hardly changes when random noise is added, but as the noise is increased, a change in the classification gradually appears.

4. Conclusion

Because in practical applications, it is not always possible to use a computer to continue model model validation, it is necessary to propose a statistical classification scheme based on the model results.

It can be seen that for the classification of high potassium glasses, all those with calcium oxide percentage higher than 5% are clustered into the same category when the alumina content is lower than 10%, and those with calcium oxide lower than 5% are also clustered into the same category, so it is meaningful to choose calcium oxide as the classification criterion in this paper. In this paper, it is proposed that when the alumina content is lower than 10%, if calcium oxide accounts for more than 5% of the components, the glass is of high potassium and high calcium type, and vice versa, it is of high potassium and poor calcium type. In particular, if the alumina content is more than 10% and higher than the calcium content, it is considered as high potassium and high alumina glass (kaolinite glass).

For lead-barium glass, the same rule can be found, in which, when the phosphorus pentoxide content is higher than 5%, or phosphorus pentoxide content is higher than 3% and does not contain magnesium oxide, it is lead-barium high-phosphorus glass, otherwise it is lead-barium phosphorus-poor glass. Specially, if the glass has been severely weathered, it is judged to be severely weathered type.

References

[1] Liu Jing. Research on Sino Western Art Exchange from the Perspective of the Silk Road [J]. Grand View of Art, 2020 (02): 80-81

[2] Wang Yinghua. Study on the Transformation of the Artistic Value of "Dead Characters" on the Silk Road [J]. Overseas Chinese Garden, 2021 (07): 90-91

[3] Chen Liang, Li Xi, Deng Qian, Wei Hua Composition identification characteristics of gem grade natural glass [C]//. Proceedings of China International Jewelry and Jewelry Academic Exchange (2019). [Publisher unknown], 2019:283-287

[4] Yellow glow red Detection of mudstone lithology and weathering degree based on depth learning [D]. University of Electronic Science and Technology of China, 2022

[5] Tang Lin. Research on fast recognition algorithm of automatically picked target images based on K-means clustering algorithm [J]. Research on Agricultural Mechanization, 2023,45 (05): 32-36

[6] Lin Lei, Sun Jianmeng Research on well logging facies establishment method based on K-means clustering and elbow rule [C]// Proceedings of the 2020 China Geoscience Joint Academic Conference (XXVI) - Topic 76: Deep Mineral Resources Exploration Technology and Application, Topic 77: Petrophysics and the Frontier of Borehole Exploration, Topic 78: Borehole Geophysics and Deep Drilling and Surveying, 2020: 224-227

[7] Wang Yuxuan, Liu Xinggang, Li Wenyi, Luo Xiaochuan. Research on motor bearing fault diagnosis method based on Bayesian optimization limit gradient ascending tree method [J]. Large Motor Technology, 2022 (03): 33-36+54

[8] Zhang Yunfeng, Zhang Chao, Lv Zhao. Residual full connection network dynamic gesture recognition method based on key points [J]. Journal of Anhui University (Natural Science Edition), 2022,46 (02): 30-38

[9] Beta and Alpha Regularizers of Mish Activation Functions for Machine Learning Applications in Deep Neural Networks [J]. 2022(12): 10-11

[10] Mish: A Self-Regularized Non-Monotonic Activation Function [J]. 2020, 32(02):6-7.

[11] Ye Congzhou, Xiao Penglin, Qin Jun, Zhang Chengxiong, Chen Lie, Wang Ke. Research on prediction model of energy consumption load of super large buildings based on clustering and random forest regression [J]. Green Building, 2022-14 (05): 48-51+55

[12] Zhu Qiuzhen, Wu Qunyong, Yao Chengxin, Sun Haoyu. Fine division and recognition of traffic state based on DBI and sparse trajectory data [J]. Journal of Geoinformation Science, 2022, 24 (03): 458-468

[13] Wei Jiandong, Lu Jianfeng, Peng Furong. Research on a hierarchical initial clustering method with adaptive number of clusters [J]. Electronic Design Engineering, 2015, 23 (06): 5-8