npj | heritage science

Article

https://doi.org/10.1038/s40494-026-02370-5

Ancient chinese glass heritage classification based on compositional data and machine learning

Check for updates

Pengxiang Tang [1], Xiaoting Gan [2] ✉ & Jiade Tang [2] ✉

Ancient Chinese glass resembles foreign glass in appearance but differs in chemical composition, which is further altered by weathering, thereby complicating the classification of artefacts. According to classification information and compositional proportion data for a set of ancient glass samples, we applied compositional data analysis based on the centered log-ratio (CLR) transformation, combined with chi-square and Fisher's exact tests, to investigate the relationships between surface weathering, glass type, emblazonry and color. Summary statistics, box plots, normality tests and two-sample *t* tests were used to compare chemical compositions before and after weathering and to estimate pre-weathering compositions from median ratios. Decision trees, logistic regression, support vector machines and random forests were then used to classify high-potassium and lead-barium glass, and ANOVA, significance tests and K-means clustering were used to divide their compositional sub-categories. The resulting models show robust classification performance and provide a reproducible, data-driven framework for the classification of ancient Chinese glass.

The ancient Silk Road served as a crucial conduit for material and cultural exchange between China and the West, with glass artifacts representing one of the hallmark trade goods during the early stages. Initially, early Chinese glass objects were introduced from Western Asia and Egypt, primarily in the form of glass beads[1]. After assimilating foreign glass-making techniques, Chinese artisans began to produce glass locally using autochthonous raw materials. Consequently, while domestically produced glass in China often resembled imported glass in appearance, its chemical composition differed markedly[2] from that of other materials. The primary component of ancient Chinese glass was quartz sand (SiO_2), with limestone added as a stabilizer that, upon calcination, was converted into calcium oxide (CaO). Owing to the high melting point of pure quartz, various flux agents were required during the refining process to lower the melting temperature. In ancient China, common flux agents included lead ore, plant ash, natural natron, and saltpeter, each contributing to distinct compositional characteristics. As a result, ancient Chinese glass can be broadly classified into two main types: lead-barium glass and high-potassium glass. The former, which is produced using lead ore as the flux, typically contains elevated levels of lead oxide (PbO) and barium oxide (BaO), whereas the latter, which is refined with plant ash, is characterized by a higher content of potassium oxide (K_2O). As an important category of archeological material, ancient Chinese glass not only reflects the technological evolution of domestic glassmaking but also provides essential clues regarding raw material sourcing, manufacturing techniques, and Sino-foreign exchange along the Silk Road. However, over time, glass undergoes weathering because of environmental exposure, during which time internal elements may be exchanged with external elements, altering the original chemical composition. As such, identifying glass types solely based on the present concentrations of PbO, BaO, or K_2O can lead to inaccurate conclusions. Currently, the classification of ancient Chinese glass relies heavily on microscopic observation, compositional analysis, and expert interpretation. Cultural heritage specialists often base their judgments on surface characteristics such as emblazonry, color, and degree of weathering, along with their own experience. This approach, however, is highly subjective, labor intensive, and prone to error. More importantly, it tends to overlook deeper structural patterns embedded in the compositional data.

In recent years, the interdisciplinary integration of mathematics with archeology and materials science has opened new avenues and provided unique insights for the study of ancient glass artifacts. Traditional analytical methods in archaeometry rely primarily on high-resolution structural imaging under microscopy and techniques such as scanning electron microscopy with energy dispersive X-ray spectroscopy (SEM-EDS)[3] and inductively coupled plasma-mass spectrometry (ICP-MS)[4] to examine material composition. While these techniques are effective for characterizing archeological materials, the final interpretation still depends heavily on manual comparison and expert judgment. When faced with a large volume

[1]The Chinese University of Hong Kong (Shenzhen), School of Medicine, Shenzhen, China. [2]Chuxiong Normal University, School of Mathematics and Statistics, Chuxiong, China. ✉e-mail: gxt@cxtc.edu.cn; tangjd@cxtc.edu.cn

of excavated materials or time-sensitive identification tasks, such approaches become increasingly limited in terms of their efficiency. Moreover, inconsistencies in classification criteria and interpretative perspectives across different institutions and experts often make it difficult to reach a consensus or extract deeper patterns. Currently, the application of statistical methods in archeology remains relatively limited and often relies on basic compositional comparisons, which are insufficient for analyzing the high-dimensional data inherent in archeological materials. With the rapid expansion of excavation datasets, there is a pressing need for a mature, replicable analytical framework that can handle both complexity and scale. Mathematical modeling offers a powerful tool for abstracting real-world problems into mathematical frameworks, enabling structured analysis and deeper understanding. In various fields, modeling has proven instrumental in providing practical solutions and critical insights. For instance, in agricultural production, researchers have employed mathematical models to optimize drying conditions for Garcinia fruit, ensuring better preservation without compromising nutritional quality[5]. In aerospace engineering, Xu Yeshou et al.[6] developed a mathematical model to describe the relationship between the macroscopic damping performance of dampers and the microstructure of viscoelastic materials, effectively predicting dynamic behavior under different testing conditions. In chemistry, mathematical models have been used to determine the optimal reaction conditions and yields for the ethanol coupling synthesis of C4 olefins[7]. As a data-driven extension of modeling, machine learning enables computers to automatically learn and improve performance without being explicitly programmed. It addresses the limitations of traditional statistical methods and has been successfully applied in various domains, such as wastewater treatment[8], biosensor design[9], and cardiovascular risk prediction[10]. In the field of archeology, researchers have begun to apply machine learning and mathematical modeling techniques[11], primarily in ceramic analysis, demonstrating the potential of data-driven approaches in material classification and interpretation[12].

Based on experimental data derived from archeologists' analyses of a collection of ancient Chinese glass samples, this study conducts compositional data processing to explore the statistical characteristics of chemical weathering in ancient glass. By applying mathematical modeling and machine learning techniques, it identifies the classification patterns of high-potassium glass and lead-barium glass, ultimately aiming to achieve efficient and accurate classification of previously unidentified glass artifacts. In the domain of ancient glass research, Li et al.[13] proposed a joint machine learning algorithm (JMLA) that integrates Daen-LR, ARIMA-LSTM, and MLR for the classification of unknown types of glass. Guo et al.[14] suggested using the slime mold algorithm to optimize the parameters of a support vector machine (SVM) and analyzed a dataset comprising 69 groups of glass chemical compositions. Their research demonstrated that the SVM algorithm, combined with the slime mold strategy, provides a reliable classification reference for future glass artifacts. Zou[15] systematically studied the changes in the composition of ancient glass due to weathering. They analyzed the surface weathering of glass relics and its correlation with three properties and established a multivariable time series model to predict the chemical composition content before weathering. Compared with these existing studies, our work is more systematic, comprehensive, and deeply analytical. First, the chemical composition data of ancient Chinese glass are preprocessed using central log-ratio (CLR) transformation to account for the closed nature of the compositional data. Subsequently, the correlations between multiple compositional features and the classification of lead-barium and high-potassium glass are investigated. Several statistical methods are then applied to examine the underlying statistical characteristics of chemical components within the classified groups. After confirming the statistical characteristics of the component content in ancient glass, we provide a novel prediction of the chemical component content in ancient glass before weathering. We subsequently employed multiple machine learning methods to explore and compare the classification rules of ancient Chinese glass, making it possible to identify unknown glass types. To ensure the robustness and practical applicability of the classification system, a sensitivity analysis is performed on the models. Ultimately, this study proposes a specific classification scheme for ancient Chinese glass and demonstrates its reliability and interpretative value for archeological and materials science research.

Methods

Sources of data

The original experimental data used in this paper are from the attachment of Question C of the 2022 Higher Education Community Cup National Mathematical Contest in Modeling for College Students[16]. All the datasets used in this study are provided in Supplementary Tables S1–S3, which collectively document the provenance, typology, and chemical composition of the analyzed ancient glass samples. The descriptive metadata, including the cultural relic number, decorative motif (emblazonry), glass type (such as high-potassium or lead-barium), color, and surface-weathering condition (before or after weathering), are listed in Supplementary Table S1. These categorical attributes serve to contextualize each artifact and provide a qualitative framework for subsequent compositional interpretation. Supplementary Table S2 presents the quantitative elemental compositions (wt%) of the samples used to construct the classification model, encompassing major, minor, and trace oxides—SiO_2, Na_2O, K_2O, CaO, MgO, Al_2O_3, Fe_2O_3, CuO, PbO, BaO, P_2O_5, SrO, SnO_2, and SO_2. This dataset represents the analytical foundation for distinguishing technological groups and glass-making traditions. Finally, Supplementary Table S3 includes compositional measurements of unclassified or newly examined glass specimens, recorded with the same chemical parameters and contextual information. These data were employed to test and validate the performance of the classification model, thereby bridging the analytical dataset with its archeological application.

Symbol description

Handling missing values in compositional glass data. The original dataset comprises various feature data, including elemental compositions, derived from ancient Chinese glass samples and obtained through instrumental analysis or manual inspection. Owing to potential inaccuracies caused by measurement errors, environmental weathering, and sample degradation, data preprocessing is essential to ensure the reliability and consistency of subsequent mathematical modeling and machine learning analyses.

In Supplementary Table S1, the color of samples 19, 40, 48, and 58 is missing. The after-weathering lead-barium glasses numbered 19 and 48 belong to emblazonry A, and samples 40 and 58 belong to emblazonry C. The modes of the colors of these two types of samples in Supplementary Table S1 are 'light blue', and the missing color is filled with 'light blue'. The values of some chemical components in Supplementary Tables S2, S3 are missing; the reason may be that the content of the component is low, and the instruments, therefore cannot detect it. It is also possible that there is no chemical component in the sample, although this case corresponds to two different situations. Hence, for the convenience of analysis, it is regarded as the same class. Moreover, because it is necessary to take the logarithm of the chemical composition data and then perform the central log-ratio Transformation, the missing null value of the component data is 0, or the component that is not detected is replaced by a smaller positive number, 0.0001.

Validation and screening of compositional glass data

When the sum of the chemical components of ancient glass is between 85% and 105%, it is regarded as valid data. The sum of the components of samples 15 and 17 is calculated to be 79.47% and 71.89%, respectively, which are classified as invalid data. Therefore, these two sets of data were deleted from Supplementary Tables S1, S2 and not considered for the next processing steps.

Limiting conditions for compositional analysis of ancient glass

The chemical composition data of ancient glass are compositional, where the total amount of all oxides theoretically is 100%. However, incomplete

detection of some minor elements often causes the total to deviate from 100%. Instead of forcing normalization, which may distort the relative proportions among detected oxides, only variables with acceptable completeness rates were retained for multivariate analysis. Following the common practice in archaeometric studies, components detected in more than 75–85% of the samples were considered reliable. In accordance with this criterion, SiO_2, CaO, Al_2O_3, CuO, and P_2O_5, to a lesser extent, PbO and BaO, as shown in Supplementary Tables S2, S3, were selected as the representative variables for subsequent statistical modeling. When using variables such as PbO or others close to the detection limit, a univariate method is considered to process them.

Central Log-Ratio Transformation (CLR) of compositional data

When machine learning and mathematical modeling techniques are applied in archeology—particularly in studies involving the elemental variation of archeological materials—it is essential to first perform a proper central log-ratio (CLR) transformation on the compositional data. The reason is that the elemental data in archeological samples typically represent closed compositions, where the sum of all components equals 100% (i.e., mass percentages). Such data exhibit compositional constraints, meaning that the components are not independent but are proportionally interdependent, rendering the use of conventional Euclidean-based statistical methods inappropriate. Owing to these constant-sum constraints, compositional variables are inherently collinear, making traditional multivariate analyses susceptible to spurious correlations and misleading results. To address this issue, British statistician John Aitchison[17–20] introduced log-ratio transformation as a methodology for analyzing compositional data. This approach is grounded in the principle that the ratios between components are not influenced by the closure constraint and that the logarithms of these ratios often approximate a normal distribution. As a result, conventional statistical and machine learning techniques can be reliably applied to the transformed data, enabling valid inference and robust classification in compositional domains such as archeological materials analysis. Owing to the fixed sum constraint, the variables of the composition data have obvious collinearity, which makes the traditional statistical analysis method invalid. It is necessary to solve such problems through appropriate transformations, such as the central log-ratio transformation (CLR), asymmetric logarithmic ratio transformation, equidistant logarithmic ratio transformation, etc.

Statistical analysis of categorical associations

To assess whether macroscopic attributes correlate with surface weathering, we modeled the relationships among weathering state, glass type, emblazonry, and color using contingency-table methods. First, we constructed contingency tables between surface weathering and each categorical variable (glass type, emblazonry, and color) using the 56 artifacts listed in Supplementary Table S1. Before applying the chi-square test, we verified that the total sample size was greater than 40, that no expected cell count was less than 1, and that at least 80% of the expected cell counts were greater than or equal to 5, ensuring the validity of the chi-square approximation. For surface weathering vs. glass type, these conditions were satisfied and the chi-square test was applied to test independence at the 5% significance level. For cross-tabulations involving low-frequency categories (emblazonry and color), the chi-square assumptions were violated. In these cases, we employed Fisher's exact test to obtain exact p-values based on the hypergeometric distribution. Because emblazonry has three levels (A, B, C), we performed pairwise Fisher tests on the derived AB, AC, and BC groups to probe potential differences in weathering distributions between motif pairs. Fisher's exact test was also used to evaluate the relationship between weathering state and color.

Statistical analysis of chemical compositions by weathering state

To characterize how weathering affects chemical composition within each glass type, we conducted univariate analyses on CLR-transformed oxide contents. Using the five groups T_1–T_5, we first visualized distributions of SiO_2 and other selected oxides via boxplots to obtain an overview of intergroup variability. For formal hypothesis testing, we examined the distributional assumptions of SiO_2 (and other key oxides as needed) in each group. Normality was assessed using Q–Q plots together with the Lilliefors test, which is suitable for small samples with unknown mean and variance. For pairs of groups representing the same glass type before and after weathering (e.g., T_3 vs. T_4 for lead-barium), we applied two-sample t tests to evaluate whether the mean CLR-transformed oxide content differed significantly between states. To mitigate the influence of potential outliers, we examined the coefficient of variation (CV) for each group.

Table 1 | Symbols and definitions of the parameters

Parameter	Explanation
ID	Cultural relic number
i	Chemical composition number
j	Glass type
$T1$	Before-weathering high-potassium glass
$T2$	After-weathering high-potassium glass
$T3$	Before-weathering lead-barium glass
$T4$	After-weathering lead-barium glass
$T5$	Severe-weathering lead-barium glass
μ	Sample average
σ	Sample variance
μ_c	95% confidence interval of sample average
σ_c	95% confidence interval of sample variance
CV	Coefficient of variation

Estimation of pre-weathering compositions

To obtain approximate pre-weathering compositions for weathered samples, we used group-level statistics from the glass type–weathering categories defined above. For each oxide and for each glass type (high-potassium and lead-barium), we calculated the central tendency (median or mean, depending on the dispersion) in the before-weathering groups and in the corresponding after-weathering groups, and took their ratios. These type-specific ratios were then used as multiplicative correction factors, applied to the measured compositions of after-weathering samples to estimate their pre-weathering values. This procedure was only used for glass types for which both before- and after-weathering data were available, and its performance was evaluated using artifacts that have both interior (before-weathering) and surface (after-weathering) measurements (Table 1).

Supervised classification of the high-potassium glass and lead-barium glass

To investigate the internal compositional heterogeneity within each major glass type, we applied unsupervised clustering to the CLR-transformed data. High-potassium and lead-barium glasses were analyzed separately. *K-means* clustering was used to partition samples into compositional groups based on their oxide contents. For each glass type, we first explored clustering solutions using the full set of available oxides and then performed one-way ANOVA on individual variables to identify those that contributed most strongly to between-cluster differences. Reduced-variable clustering was then carried out on these more discriminative oxides. The quality of alternative clustering schemes (different numbers of clusters and variable subsets) was compared using standard internal validity indices, including the silhouette coefficient, Davies–Bouldin index and Calinski–Harabasz index.

Unsupervised clustering for subclassification within glass types

To investigate the internal compositional heterogeneity within each major glass type, we applied unsupervised clustering to the CLR-transformed data. High-potassium and lead-barium glasses were analyzed separately. *K-means* clustering was used to partition samples into compositional groups

Table 2 | Sample data preprocessing and CLR summary table (part)

ID	Type	Surface weathering	SiO_2	Na_2O	K_2O	CaO	MgO
01	High-potassium	Before-weathering	6.76	−6.69	4.82	4.37	2.38
02	Lead-barium	After-weathering	5.57	−7.23	2.03	2.83	2.14
03	High-potassium	Before-weathering	8.63	−5.05	5.81	4.86	−5.05
03	High-potassium	Before-weathering	5.29	−8.05	3.68	2.93	1.27
04	High-potassium	Before-weathering	6.69	−6.71	4.78	4.47	2.95
05	High-potassium	Before-weathering	6.07	−7.27	4.34	3.94	2.52
06	High-potassium	Before-weathering	6.20	−7.22	3.98	−7.22	2.67
06	High-potassium	Before-weathering	5.24	−8.06	3.19	2.84	1.70
07	High-potassium	After-weathering	9.49	−4.25	−4.25	5.03	−4.25
08	Lead-barium	Severe-weathering	3.75	−6.99	−6.99	3.38	−6.99
08	Lead-barium	After-weathering	5.29	−6.93	-6.93	2.68	−6.93
09	High-potassium	After-weathering	9.01	−4.75	3.93	3.98	−4.75
10	High-potassium	After-weathering	9.75	−4.03	5.09	3.62	−4.03

All the data analysis in this paper is based on the data in Supplementary Table S1.

based on their oxide contents. For each glass type, we first explored clustering solutions using the full set of available oxides and then performed *one-way ANOVA* on individual variables to identify those that contributed most strongly to between-cluster differences. Reduced-variable clustering was then carried out on these more discriminative oxides. The quality of alternative clustering schemes (different numbers of clusters and variable subsets) was compared using standard internal validity indices, including the silhouette coefficient, *Davies–Bouldin* index and *Calinski–Harabasz* index. The final subclassification schemes and their compositional characteristics are presented and interpreted in the Results and Discussion sections.

Results

CLR of Chinese ancient glass compositional data

After the sample data (Supplementary Tables S1, S2) were processed using the CLR, the data in Table 2 were obtained. All the calculated data can be found in Supplementary Table S4.

Analysis of the relationships between the surface weathering of glass relics and the glass type, emblazonry, and color

Studying the relationships between the surface weathering of ancient glass artifacts and their glass type, decorative style, and color is highly important. On the one hand, it facilitates an understanding of the aging and degradation patterns of glass with different characteristics from a materials science perspective; on the other hand, it helps to exclude the potential interference of certain features in subsequent analytical procedures. The weathering, emblazonry, color, and type of sample cultural relics in Supplementary Table S1 are fixed class variables, and the chi-square test[21,22] can be used to analyze the relationship between them. To investigate whether the expected frequency meets the requirements, the weathering condition is used as the grouping variable, with emblazonry, type, and color as the variables. The contingency table (Supplementary Table 3) was calculated by SPSS.

Given the data in Table 3, it can be concluded that only the expected frequency between surface weathering and glass type meets the requirements of the chi-square test, and the chi-square test can be used; the expected frequency between surface weathering and emblazonry, surface weathering and color is equal to 0, and the chi-square test cannot be used.

The chi-square test (Table 4) is used to test the independence of surface weathering and glass type. The null hypothesis is that the glass surface weathering and glass type are independent, and the alternative hypothesis is that the glass surface weathering and glass type are not independent.

If the significance level is 0.05 and the *P* value is 0.02, which is lower than the significance level, at the 95% confidence level, the original hypothesis is rejected, and glass surface weathering is not independent of the glass type; that is, the glass type affects surface weathering.

Table 3 | Contingency results for surface weathering and emblazonry, color and type of sample cultural relics

Subject	Category	Surface weathering		Total
		Before-weathering	After-weathering	
Emblazonry	A	11 (50.00%)	11 (50.00%)	22
	B	0 (0.00%)	6 (100.00%)	6
	C	11 (39.30%)	17 (60.70%)	28
Total		22	34	56
Type	Lead-barium	12 (30.00%)	28 (70.00%)	40
	High-potassium	10 (62.50%)	6 (37.50%)	16
Total		22	34	56
Color	Light green	2 (66.70%)	1 (33.30%)	3
	Light blue	6 (27.30%)	16 (72.70%)	22
	Dark green	3 (42.90%)	4 (57.10%)	7
	Dark blue	2 (100.00%)	0 (0.00%)	2
	Purple	2 (50.000%)	2 (50.000%)	4
	Green	1 (100.00%)	0 (0.00%)	1
	Bluish green	6 (40.00%)	9 (60.00%)	15
	Black	0 (0.00%)	2 (100.00%)	2
Total		22	34	56

Fisher's exact test can be used for the independent analysis of surface weathering and emblazonry, surface weathering and color[23,24]. Compared with the chi-square test, Fisher's exact test is more accurate when dealing with low-frequency observations (such as fewer than 5 observations). The basic principle of Fisher's exact test is to use the hypergeometric distribution to calculate the probability of the observation data. By comparing the difference between the observed data and the randomly distributed data, it is possible to determine whether there is a significant correlation between the two categorical variables. Because the emblazonry in Supplementary Table S1 has three levels, A, B, and C, it cannot be directly used for Fisher's exact test. First, emblazonry A and B in Supplementary Table S1 are selected as a group, A and C, as a group, and B and C, as a group, which are called the

Table 4 | Independent chi-square test results for surface weathering and glass type

Subject	Name	Surface weathering		Total	Test method	χ^2	P
		Before-weathering	After-weathering				
Type	Lead-barium	12	28	40	Pearson chi-square test	5.06	0.02
	High-potassium	10	6	16			
Total		22	34	56			

Table 5 | Fisher's exact test results for surface weathering and the emblazonry AB group

Subject	Title	Surface weathering		Total	P
		Before-weathering	After-weathering		
Emblazonry	A	11	11	22	0.06
	B	0	6	6	
Total		11	17	28	

Table 6 | Valid data classification for the CLR data

Category	Before-weathering	After-weathering	Severe-weathering
High-potassium	12	6	0
Lead-barium	23	23	3

emblazonry AB group, emblazonry AC group, and emblazonry BC group, respectively; then, Fisher's exact test is used.

If the significance level is 0.05, the significant *P* value in Table 5 is 0.06, which is greater than the significance level, and the original hypothesis cannot be rejected. Therefore, no significant difference was observed between the surface weathering and emblazonry AB group. Using the same method, Fisher's exact test *P* values of the surface weathering and the emblazonry AC and BC groups were 0.57 and 0.15, respectively, which are greater than 0.05. There was no significant difference between the surface weathering and emblazonry AC and BC groups. Using the same test method, it can also be concluded that there is no significant difference between surface weathering and color. Through the above discussion, it can be concluded that only the type of glass has a significant effect on surface weathering and that the color and emblazonry of glass have no significant effect on surface weathering.

Analysis of the contents of the corresponding chemical components in ancient glass before or after weathering

Statistical analysis of the chemical composition of different glass types. The 67 valid data points in Table 2 are divided into three categories, before-weathering, after-weathering, and severe-weathering, according to the weathering of the sampling points, and then divided into lead-barium glass and high-potassium glass according to the type of glass. Because there are no sampling data for severe-weathering high-potassium glass, the data in Table 2 can be divided into five categories: before-weathering high-potassium, after-weathering high-potassium, before-weathering lead-barium, after-weathering lead-barium, and severe-weathering lead-barium. These five categories are represented by the letters T1, T2, T3, T4, and T5, respectively. Note that the surfaces of the glass cultural relic samples numbered 49, 50 and 53 are after-weathering, but there is a before-weathering sampling point, and the glass weathering type should be classified as before-weathering at this sampling point. The classification of all the valid data is shown in Table 6.

To analyze the variation in chemical compositions of these five glass types (T1, T2, T3, T4, and T5), SiO_2 was taken as an example to construct a box diagram of SiO_2.

As shown in Fig. 1, the content of SiO_2 in high-potassium glass is greater than that in lead-barium glass. After weathering, the content of SiO_2 in high-potassium glass is the highest, with an average value of 9.09, and the content of SiO_2 in severe-weathering lead-barium glass is the lowest, with an average value of 4.44. Under the conditions of known high-potassium glass or lead-bismuth glass, in order to determine whether there is a significant difference in the content of SiO2 before and after weathering. Taking lead-barium glass as an example, whether the content of SiO_2 obeys a normal distribution at T3 and T4 is tested first, and a *QQ* diagram is constructed.

The data points in Figs. 2 and 3 are approximately distributed near a straight line, which can be considered to obey a normal distribution. In general, the Jarque–Bera test[25,26] is also needed to test normality. The *Jarque–Bera* test evaluates whether the null hypothesis that the sample obeys a normal distribution with unknown mean and variance is true, but the Jarque–Bera test cannot be used for small sample tests. When the sample data volume is small, the Lilliefors test[27,28] can be used. With respect to this problem, the classification sample size of the sample of cultural relics is small, and the Lilliefors test can be used. After MATLAB calculation, $p = 0.13$ and $p = 0.26$ are obtained, which are greater than the significance level of 0.05, and the null hypothesis cannot be rejected. That is, the content of SiO_2 under T3 and T4 obeys a normal distribution, which also verifies the theory that the component data in the introduction often obey a normal distribution after the central log-ratio transformation (CLR).

The two-sample *t* test[29,30] is used to test whether the mean values of the two are equal. The null hypothesis is that there is no significant difference in the mean value. Using MATLAB calculations, $p = 0.001$, which is less than the significance level of 0.05. Thus, the null hypothesis is rejected; that is, there is a significant difference in the mean value of SiO_2 in lead-barium glass before weathering (T3) and after weathering (T4). The sample mean μ, variance σ, median, 95% confidence interval μ_c, σ_c, and coefficient of variation (CV) of SiO_2 in five glass types (T1, T2, T3, T4, and T5) were calculated (Table 7).

The coefficient of variation *CV* of T2 and T3 was less than 0.15, indicating that there was a small probability of outlier values in the current data and that the average value could be used for statistical analysis. The coefficient of variation *(CV)* of T1, T4, and T5 was greater than 0.15, indicating that there might be outliers in the data. The medians of 6.59, 5.25 and 4.10 were used for statistical analysis. Using the same analysis method, the mean (or median) of other chemical components under different types of glass can be calculated.

From Table 8, it can be concluded that the mean values (medians) of SiO_2, K_2O, PbO, and CuO in the high-potassium glass before and after weathering are quite different; in the lead-barium glass, the mean values (median) of Na_2O, K_2O, Fe_2O_3, and P_2O_5 before and after weathering are quite different. The mean values (medians) of SiO_2, Fe_2O_3, PbO, BaO, and SO_2 in the severe-weathering and after-weathering glass types are quite different.

According to the after-weathering point data, the chemical composition content before weathering is predicted. As shown in Table 2, the cultural relics of samples 49 and 50 are lead-barium glass, which has sampling data for both before-weathering points and after-weathering points. Chemical composition histograms of samples 49 (Fig. 4) and 50 (Fig. 5) before and after weathering, respectively, are drawn below.

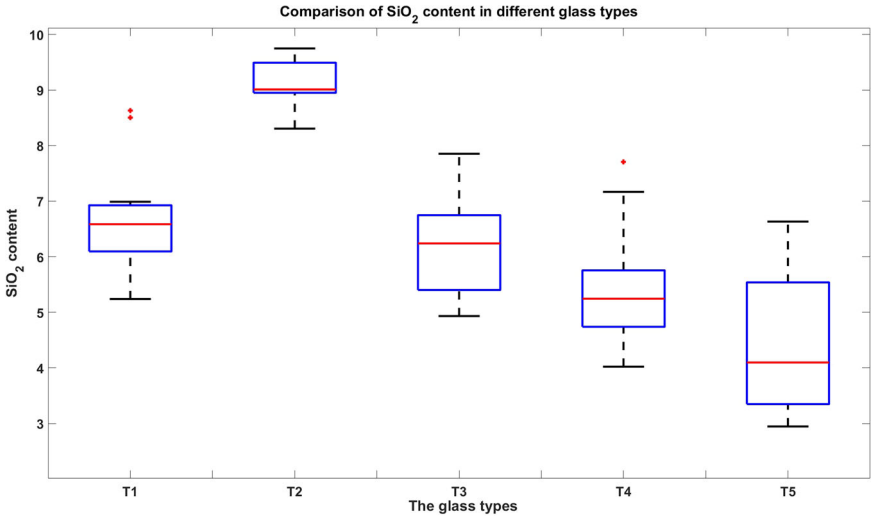

Fig. 1 | Boxplot of SiO_2 content in T1, T2, T3, T4, and T5.

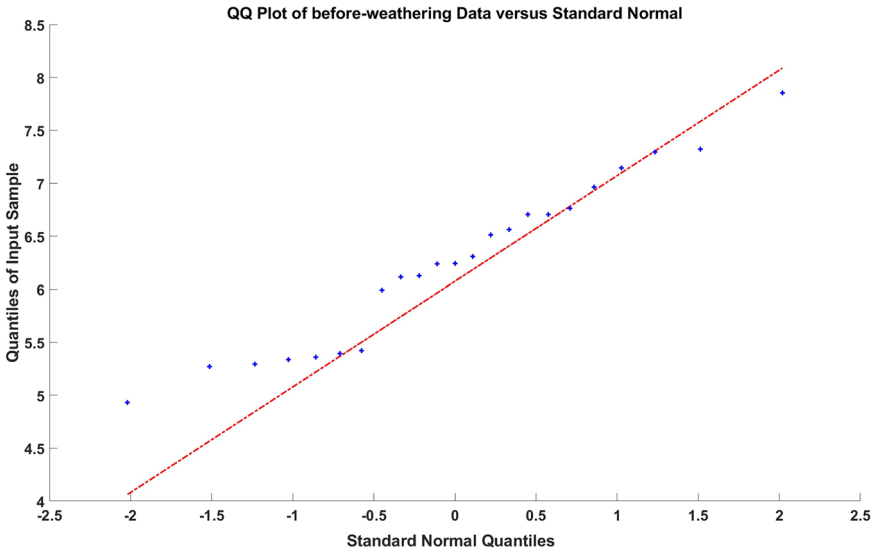

Fig. 2 | qq plot of the SiO_2 content in T3.

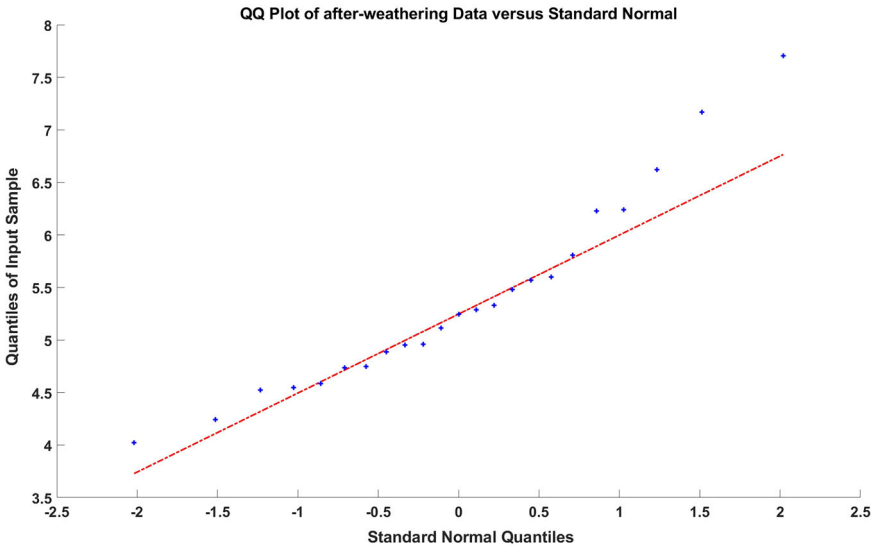

Fig. 3 | Content qq plot of SiO_2 in T4.

The proportion of each chemical component in the two samples is generally essentially the same, as shown in Fig. 5, and the specific values are slightly different. Unfortunately, only these two sets of matching data before-weathering and after-weathering have a small amount of data and lack high-potassium glass matching data. Therefore, according to the after-weathering point data, predicting the chemical composition content before weathering through regression, neural networks, and other models is inappropriate.

From the discussion in Section "Analysis of the relationships between the surface weathering of glass relics and the glass type, emblazonry, and color", it is evident that only the type of glass has a significant effect on surface weathering and that the color and emblazonry of the glass have no significant effect on surface weathering. Therefore, when predicting the chemical composition content before weathering, only the type of glass needs to be considered, and the color and emblazonry do not need to be considered. Through the comparison of Fig. 4 and Fig. 5, the numerical changes in each chemical component in different samples before and after weathering are different, but the overall 'appearance' is consistent; that is, the proportion of each chemical component is nearly consistent; therefore, it is speculated that under different glass types, the ratio of the mean value (median) of chemical composition before weathering to the mean value (median) after weathering can be used to describe the change rule between them. From the compositional statistics summarized in Table 8, the median (or mean) concentration of each oxide for the different glass types was obtained. Based on these values, the proportional change in each chemical component before and after weathering was computed according to Eq. (1), yielding the ratio parameter $k_{ij}(i = 1, 2, \ldots, 14;\ j = 1, 2)$, where i represents the fourteen analyzed oxides and j corresponds to the high-potassium ($j = 1$) and lead-barium ($j = 2$) glass groups.

$$k_{i1} = \frac{T1(i)}{T2(i)}, k_{i2} = \frac{T3(i)}{T4(i)} \tag{1}$$

The chemical composition content before weathering can be predicted by multiplying k_{ij} by the detection data of the after-weathering points. To verify the feasibility of this prediction method, the data of cultural relics No. 49 and No. 50 are used for verification. First, the value of k_{i2} is multiplied by the chemical composition data of the after-weathering points to obtain the prediction data before weathering, after which the absolute error between the prediction data and the measured data is calculated. The complete table content can be found in Supplementary Table S5.

From Table 9, it can be concluded that the absolute error of the prediction of 9 chemical components, such as SiO_2, CaO, MgO, Al_2O_3, and CuO, is small, and the absolute error of the prediction of 4 chemical components, such as Na_2O, K_2O and Fe_2O_3, is large. The reason for this finding may be that although the sample cultural relics are divided into three categories—before-weathering, after-weathering, and severe-weathering—the situation of the same type of individuals is different.

Table 7 | Sample mean, variance and 95% confidence interval of SiO_2

SiO_2	High-potassium		Lead-barium		
	T1	*T2*	*T3*	*T4*	*T5*
μ	6.65	9.09	6.26	5.37	4.44
σ	1.06	0.50	0.79	0.92	1.94
Median	6.59	9.01	6.24	5.25	4.10
μ_c	[5.98, 7.32]	[8.56, 9.61]	[5.91, 6.60]	[4.98, 5.77]	[-0.37, 9.26]
σ_c	[0.75, 1.79]	[0.31, 1.22]	[0.61, 1.12]	[0.71, 1.30]	[1.01, 12.19]
CV	0.16	0.06	0.13	0.17	0.36

Table 8 | The mean (or median) of each chemical composition for different glass types

i	High-potassium		Lead-barium		
	T1	*T2*	*T3*	*T4*	*T5*
SiO_2	6.59	9.09	6.26	5.25	4.10
Na_2O	−4.22	−4.67	−2.40	−5.44	−6.66
K_2O	4.64	3.71	−1.81	−4.27	−3.90
CaO	4.09	4.24	2.00	2.83	2.73
MgO	2.04	−1.77	−0.97	−1.47	−3.56
Al_2O_3	4.28	5.09	3.62	2.72	2.33
Fe_2O_3	2.88	3.19	−2.89	−0.92	−6.66
CuO	3.02	4.93	1.83	2.07	3.37
PbO	−1.85	−4.67	5.36	5.64	5.70
BaO	−3.53	−4.67	4.28	4.19	5.20
P_2O_5	2.64	2.98	−0.58	3.31	4.25
SrO	−3.44	−4.67	−0.97	0.72	1.59
SnO_2	−5.93	−4.67	−5.87	−6.30	−6.66
SO_2	−4.69	−4.67	−6.48	−6.21	1.31

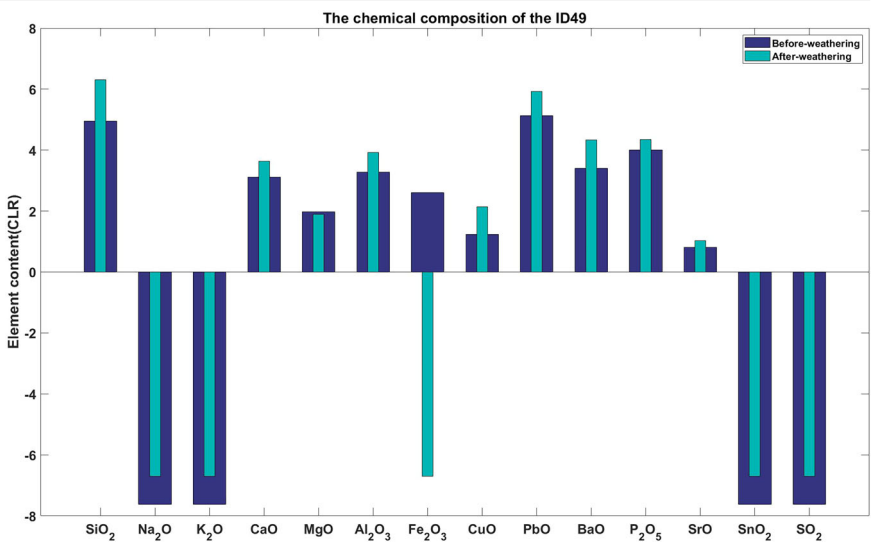

Fig. 4 | Histogram of before-weathering and after-weathering chemical components for sample number 49.

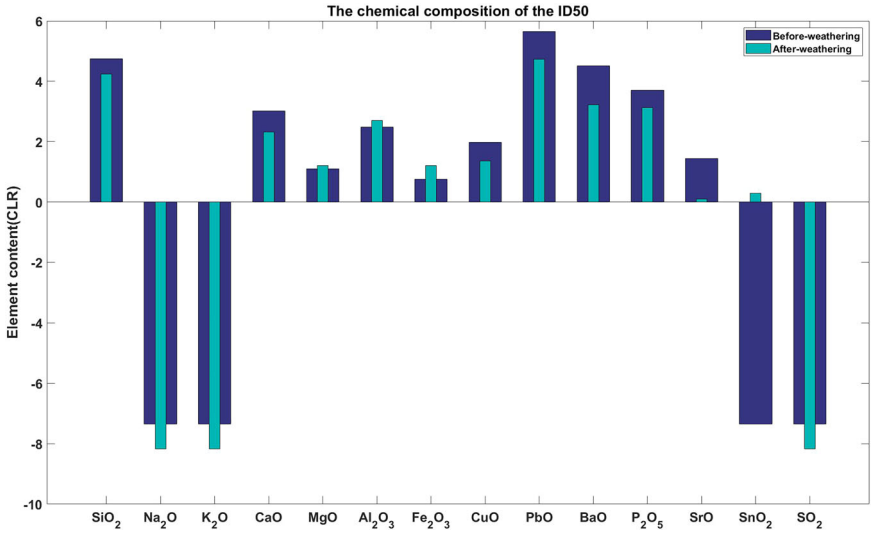

Fig. 5 | Histogram of before-weathering and after-weathering chemical components for sample number 50.

Table 9 | Numbers 49 and 50 The predicted values and errors before the weathering of cultural relics (part)

ID49	SiO_2	Na_2O	K_2O	CaO	MgO	Al_2O_3	Fe_2O_3	CuO
Measured data	4.95	−7.62	−7.62	3.11	1.98	3.27	2.60	1.23
Forecast data	5.90	−3.37	−3.22	2.20	1.31	4.36	8.15	1.09
Absolute error	0.63	4.57	3.29	0.20	0.15	1.22	6.65	0.62
ID50	**SiO_2**	**Na_2O**	**K_2O**	**CaO**	**MgO**	**Al_2O_3**	**Fe_2O_3**	**CuO**
Measured data	4.24	−8.17	−8.17	2.31	1.21	2.70	1.21	1.35
Forecast data	5.06	−3.61	−3.45	1.64	0.80	3.59	3.81	1.20
Absolute error	0.31	3.74	3.90	1.38	0.30	1.11	3.06	0.78

Analysis of the classification law of high-potassium glass and lead-barium glass

For a cultural relic worker, fast and accurate classification of glass relics is necessary, which is a supervised classification problem. In Section "Analysis of the contents of the corresponding chemical components in ancient glass before or after weathering", the mean (median) of each chemical composition in different glass types is calculated, and Table 8 is obtained. From Table 8, it can be concluded that the values of SiO_2, K_2O, PbO, and BaO in high-potassium glass and lead-barium glass are quite different. The key to classifying ancient Chinese glass relics is the selection of one or more chemical components to categorize and distinguish different types of glass. In the following, several types of classification models commonly used in machine learning are employed for analysis.

Decision trees. In accordance with the data in Table 2, MATLAB is used to construct Fig. 6.

The results revealed that the classification results were related only to the chemical composition of No. 9, that is, the value of PbO. The threshold value of the decision tree output was 3.04; when the content of PbO was less than 3.04, the sample was classified as high-potassium; otherwise, it was classified as lead-barium. The evaluation index of a classification method can be described by the importance of the chemical composition of the classification, the accuracy of the training set and the test set, and the F1 score. The above indices of the decision tree were calculated. The results indicate that the characteristic importance of PbO is 100%, and the accuracy and F1 score are both 1; that is, the use of PbO can completely divide the sample cultural relics into two categories, and the accuracy of each sample classification is 100%.

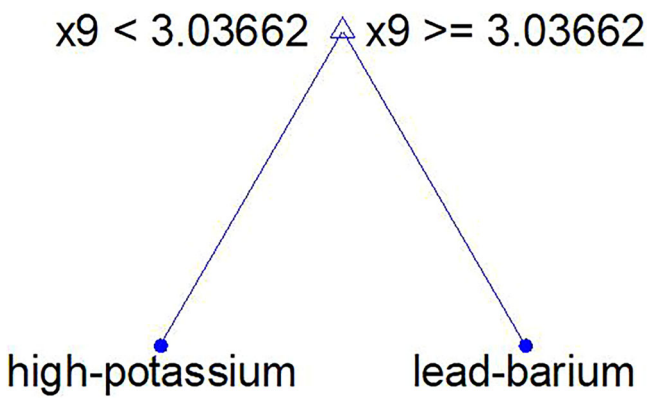

Fig. 6 | Decision tree for high-potassium and lead-barium glass.

To further explore the use of decision trees to classify cultural relics, according to the weathering of cultural relics, they are divided into two categories. Decision tree models are established for before-weathering cultural relics and after-weathering cultural relics, and the branches of the decision trees are compared.

It can be concluded from Fig. 7 that the classification of before-weathering cultural relics is consistent with the decision tree of the overall classification; that is, the classification result is related only to the size of the PbO. When the content of PbO is less than 3.04, it is classified as a high-potassium class; otherwise, it is classified as a lead-barium class. In after-weathering cultural relics, the classification index is SiO_2. When the content of SiO_2 is less than 8.006, the after-weathering

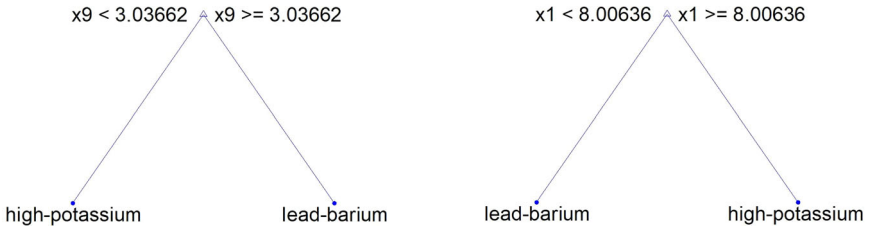

Fig. 7 | Before-weathering and after-weathering decision trees.

Table 10 | Type identification of unknown cultural relics under different classification models

Relic number	Decision tree	After-weathering decision tree	Binary logit	Support Vector Machine (SVM)	Random forest
A1	High-potassium	High-potassium	High-potassium	High-potassium	High-potassium
A2	Lead-barium	*High-potassium*	Lead-barium	Lead-barium	*High-potassium*
A3	Lead-barium	Lead-barium	Lead-barium	Lead-barium	Lead-barium
A4	Lead-barium	Lead-barium	Lead-barium	Lead-barium	Lead-barium
A5	Lead-barium	Lead-barium	Lead-barium	Lead-barium	Lead-barium
A6	High-potassium	High-potassium	High-potassium	High-potassium	High-potassium
A7	High-potassium	High-potassium	High-potassium	High-potassium	High-potassium
A8	Lead-barium	Lead-barium	Lead-barium	Lead-barium	Lead-barium

cultural relics are classified as lead-barium; otherwise, they are classified as high-potassium.

Logit regression. Logit regression[31,32], in this classification problem, because the sample data in Table 2 have been transformed by the central logarithm ratio, the collinearity between the chemical components is eliminated. The glass type (high-potassium, lead-barium) is regarded as the two-class classification data, and the chemical components are quantitative data. Using the binary logit model, F1 = 1 in the training set and test set can be obtained by SPSS calculation. The logit regression confusion matrix heatmap shows that logit regression correctly classifies the sample's cultural relics.

Support Vector Machine (SVM). A support vector machine (SVM)[33,34] can solve the problem of machine learning in the case of small samples and simplify common classification and regression problems. When SPSS is used for calculations, the model results indicate that both the accuracy and the F1 score are 100%. All the lead-bismuth glasses can be classified correctly. It can be concluded that the classification effect of support vector machines is also good.

Random forest. In machine learning, random forest[35,36] is a classifier containing multiple decision trees, and the output category is determined by the mode of the category output by individual trees. Leo Breiman[37,38] and Adele Cutler[39] developed the random forest algorithm. The results of the random forest can be obtained by SPSS calculations, for which F1 = 0.973; the importance of the characteristics involves many chemical components: PbO accounts for 27.3%, BaO accounts for 15%, and SrO accounts for 10.6%. The results show that one sample was misclassified.

Identification and sensitivity analysis of unknown types of glass relics

According to the evaluation results of the above four commonly used classification models, the F1 values for the decision tree, binary logit regression and support vector machine (SVM) are all 1, and the classification effect is the best. The F1 value for random forest is 0.976, and the classification effect is better.

To analyze the sensitivity of the above five methods, the data in Supplementary Table S3 (after normalization and central log-ratio transformation) are supplemented with a certain proportion of noise and then classified by the above five methods. After adding 5% noise to the data in Supplementary Table S3, the A6 category obtained by the after-weathering decision tree changed; after adding 10% noise to the data in Supplementary Table S3, the discrimination of two cultural relics by the after-weathering decision tree changed, the discrimination of one cultural relic by random forest changed, and the discrimination of cultural relics by the decision tree, binary logit and support vector machine (SVM) did not change. Therefore, the decision tree, binary logit, and support vector machine (SVM) methods demonstrated better anti-interference effects.

To evaluate the effectiveness of the proposed mathematical model, we applied it to classify ancient glass artifacts of unknown types (Supplementary Table S3) and further assessed its robustness under data perturbations. The data for each chemical component in Supplementary Table S3 were normalized, and the central log-ratio transformation (CLR) was performed. The above decision tree, binary logit regression, support vector machine (SVM) and random forest methods were used for classification. The results of classifying unknown types of ancient glass using machine learning models are shown in Table 10.

The cultural relic numbers A1, A3, A4, A5, A6, A7, and A8 and the five classification methods are consistent. The cultural relic number A2, decision tree, binary logit, and support vector machine (SVM) are classified as lead-barium, and the other two methods are classified as high-potassium. According to the discussion of the first five classification methods, A2 should be classified as lead-barium. All the unknown types of glass artifacts were successfully classified, and the results are presented in Table 11.

To evaluate the robustness of the five classification methods described above, we introduced controlled noise into the dataset (Suplementary Table S3), which had been preprocessed through central log-ratio (CLR) transformation. Specifically, noise was added at two levels (5% and 10%) to the composition data of the unknown-type glass artifacts, and the classification performance of the five models was reevaluated. When 5% noise was added, the classification result of one artifact (A6) changed under the weathering decision tree model. At the 10% noise level, the weathering decision tree misclassified two artifacts, and the random forest model misclassified one artifact. In contrast, the standard decision tree, binary logistic regression (logit), and support vector machine (SVM) models maintained consistent classification results across both noise levels. These findings indicate that the decision tree, binary logit, and SVM classifiers exhibit strong resistance to data perturbation and demonstrate superior robustness.

Table 11 | Classification table of unknown types of ancient glass

Relic number	Select the model	The prediction results are _Y	Probability of prediction results for lead- barium	Probability of prediction results for high-potassium
A1	Decision tree binary logit SVM	High-potassium	0	1
A2		Lead-barium	1	0
A3		Lead-barium	1	0
A4		Lead-barium	1	0
A5		Lead barium	1	0
A6		High-potassium	0	1
A7		High-potassium	0	1
A8		Lead-barium	1	0

Table 12 | K-means clustering statistical table of high-potassium glass

Element	Cluster category (mean ± standard deviation)		*F*	*P*
	Category 2 (*n* = 13)	Category 1 (*n* = 5)		
SiO_2	7.867 ± 1.364	6.418 ± 1.348	4.092	0.060*
Na_2O	−3.345 ± 3.974	−7.02 ± 1.23	3.991	0.063*
K_2O	3.277 ± 3.513	2.107 ± 5.067	0.316	0.582
CaO	4.356 ± 0.431	−0.493 ± 5.245	12.103	0.003***
SnO_2	−5.702 ± 1.176	−5.007 ± 5.617	0.196	0.664
SO_2	−3.786 ± 3.118	−7.02 ± 1.23	4.928	0.041**
Al_2O_3	4.724 ± 0.587	4.062 ± 0.92	3.366	0.085*
SrO	−4.749 ± 1.982	−1.51 ± 3.073	7.14	0.017**
MgO	−0.839 ± 3.946	2.582 ± 1.268	3.498	0.080*
P_2O_5	1.462 ± 3.301	2.786 ± 1.331	0.734	0.404
Fe_2O_3	2.306 ± 2.342	1.208 ± 3.555	0.598	0.451
BaO	−5.702 ± 1.176	0.752 ± 3.343	39.252	0.000***
PbO	−3.816 ± 3.397	−0.115 ± 2.938	4.573	0.048**
CuO	3.947 ± 1.307	1.251 ± 3.596	5.818	0.028**

***, **, and * represent the significance levels of 1%, 5% and 10%, respectively.

Table 13 | K-means clustering evaluation index of high-potassium glass (*K* = 2)

Contour Coefficient	DBI	CH
0.246	1.597	4.746

Table 14 | K-means clustering evaluation index of selected chemical components of high-potassium glass (*K* = 2)

Contour coefficient	DBI	CH
0.411	1.118	9.878

Table 15 | K-means clustering variance table of selected chemical components of high-potassium glass (*K* = 2)

Element	Cluster category (mean ± standard deviation)		*F*	*P*
	Category 1 (*n* = 13)	Category 2 (*n* = 5)		
SO_2	−3.786 ± 3.118	−7.02 ± 1.23	4.928	0.041**
CaO	4.356 ± 0.431	−0.493 ± 5.245	12.103	0.003***
SrO	−4.749 ± 1.982	−1.51 ± 3.073	7.14	0.017**
PbO	−3.816 ± 3.397	−0.115 ± 2.938	4.573	0.048**
BaO	−5.702 ± 1.176	0.752 ± 3.343	39.252	0.000***
CuO	3.947 ± 1.307	1.251 ± 3.596	5.818	0.028**

*** and ** represent the significance levels of 1% and 5%, respectively.

Subclass division of high-potassium glass and lead-barium glass

In Part 3.4, the classification rules of high-potassium glass and lead-barium glass are discussed. Sometimes, it is necessary to classify these two types of glass into subcategories. This is an unsupervised classification problem. The high-potassium glass data in Table 2 are subjected to K-means clustering, $K = 2$.

It can be concluded from Table 12 that the *P* values of CaO, SO_2, SrO, PbO, BaO, and CuO are all less than 0.05, which is significant at this level. The null hypothesis is rejected, indicating that there are significant differences between the above chemical components in the categories classified by cluster analysis.

The contour coefficient is the average of the contour coefficients of all the samples. The value range of the contour coefficient is [−1, 1]. The closer the distance of the samples in the same category is, the farther the distance of the samples in different categories is, the higher the score is, and the better the clustering effect is. The DBI (Davies–Bouldin) index is used to measure the ratio of the intracluster distance to the intercluster distance of any two clusters. The smaller the index is, the better the clustering effect is. The CH (Calinski–Harbasz score) index is obtained as the ratio of separation to tightness. The larger the CH is, the better the clustering effect is. As shown in Table 13, when the contour coefficient = 0.246, DBI = 1.597, and CH = 4.746, these indicators do not look too good. Taking $K = 3$ and $K = 4$, the contour coefficients can be calculated to be 0.251 and 0.22, respectively. With increasing K, the increase in the contour coefficient is not obvious.

The analysis of the data in Table 12 shows that the chemical components of CaO, SO_2, SrO, PbO, BaO, and CuO significantly differ among the categories classified by cluster analysis, and these six chemical components were selected for clustering again.

The contour coefficient and CH greatly improved, and the clustering effect significantly improved (Table 14).

From Table 15, it can be concluded that high-potassium glass can be divided into two subcategories according to its chemical composition. The contents of SO_2, CaO, and CuO in the first subcategory are relatively high, while the contents of SrO, PbO, and BaO in the second subcategory are relatively high. The first subcategory is named high-potassium-CaO-CuO glass, and the second subclass is named high-potassium-BaO-PbO glass. The subclassification of lead-barium glass is carried out according to the same logic.

The calculation results (Tables 16 and 17) of the contour coefficient and CH are good. The lead-barium glass can be clustered into three subclasses by using Na_2O, MgO, SO_2, P_2O_5, CuO, BaO and Fe_2O_3. The contents of MgO and Fe_2O_3 in the first subclass are high, the content of Na_2O in the second subclass is high, and the contents of BaO and CuO in the third

Table 16 | K-means clustering variance table ($K = 3$) of selected chemical components of lead-barium glass

	Cluster category (mean ± standard deviation)			F	P
	Category 2 ($n = 27$)	Category 3 ($n = 13$)	Category 1 ($n = 9$)		
Na_2O	−6.766 ± 1.637	3.083 ± 1.137	−6.406 ± 0.954	226.426	0.000***
MgO	0.41 ± 3.285	−1.563 ± 4.384	−6.406 ± 0.954	13.944	0.000***
SO_2	−7.099 ± 0.869	−7.156 ± 0.713	−0.356 ± 4.925	35.166	0.000***
P_2O_5	2.328 ± 3.062	−2.264 ± 3.971	2.418 ± 2.87	9.379	0.000***
CuO	0.629 ± 3.603	1.989 ± 1.112	3.932 ± 0.902	4.897	0.012**
BaO	2.838 ± 3.401	4.144 ± 1.012	5.813 ± 0.896	4.52	0.016**
Fe_2O_3	0.045 ± 3.729	−3.946 ± 3.913	−6.406 ± 0.954	13.942	0.000***

*** and ** represent the significance levels of 1% and 5%, respectively.

Table 17 | K-means clustering evaluation index of selected chemical components of lead-barium glass ($K = 3$)

Contour coefficient	DBI	CH
0.335	1.119	20.107

subclass are high. These subclasses can be called lead-barium-MgO-Fe_2O_3, lead-barium-Na_2O, and lead-barium-BaO-CuO glass, respectively.

Discussion

In this study, we present a data-driven classification framework for ancient glass, moving beyond traditional typological approaches by applying multivariate clustering to compositional data. We demonstrate that within the broadly classified high-potassium and lead-barium glass groups, distinct compositional subcategories emerge—namely, a high-potassium–CaO–CuO class versus a high-potassium–BaO–PbO class and three lead-barium subclasses enriched in MgO–Fe_2O_3, Na_2O, and BaO–CuO. By linking categorical variables (surface-weathering state, glass type, emblazonry, and color) via χ^2 and Fisher's exact tests, we found that only the glass type correlates significantly with the weathering state, whereas emblazonry and color do not. These findings align with those of mechanistic studies showing that glass durability is strongly governed by composition and ionic mobility rather than serving merely as decorative or esthetic attributes[40]. Specifically, alkali-rich networks (such as high-potassium glass) are known to leach modifier cations more rapidly, thus accelerating surface degradation, whereas glasses containing lead and barium may form more chemically resistant structures, thereby reducing weathering[41,42].

Building upon the compositional clustering framework described above, the present analysis highlights that the accurate classification of ancient glass cannot rely solely on macroscopic indicators such as surface weathering but must integrate the quantitative chemistry of the artifacts. The elemental composition of archeological glass is inherently compositional and constrained by the constant-sum rule, which renders conventional statistical methods inappropriate and results in calculation error[43,44]. To address this issue, the present work applies the centered log-ratio (CLR) transformation, a technique designed to preserve the relative structure of compositional data while enabling valid multivariate inference. According to the classification characteristics of the sample data, glass sample cultural relics are divided into five categories: before-weathering high-potassium, after-weathering high-potassium, before-weathering lead-barium, after-weathering lead-barium, and severe-weathering lead-barium. Taking SiO_2 as an example, the distribution of the chemical composition in these five types of glass was discussed. The difference in SiO_2 before and after weathering was tested by a two-sample t test, which revealed a significant difference in SiO_2 before and after weathering. The sample mean or median of each chemical composition in the five kinds of glass was calculated. The change rule of each chemical composition before and after weathering was reflected by their ratio so that the chemical composition before weathering could be estimated according to the chemical composition after weathering. Such a data-driven approach establishes a quantitative bridge between glass alteration chemistry and classification modeling, providing a unique perspective to restore the elemental content of cultural relics.

The sum of the chemical composition ratios of ancient glass products should be 100%, which is typical component data. Many negative correlations were observed in the ancient glass composition data after CLR conversion. The fundamental reason for this negative correlation is that the sum of the chemical composition of the glass is zero. Since the content of trace elements in the ancient glass sample is very low, the glass is weakly affected by the overall additive effect. The conversion of CLR results in an artificial additive effect, resulting in a large number of negative correlations between trace elements. ILR is an effective method for analyzing geochemical composition data because of its superior theoretical properties. However, when the sum of components is not equal to 1, caution needs to be taken[45]. The calculation process of ILR is based on the assumption that the sum of the components is equal to 1. Notably, ILR does not involve one-to-one transformation. The transformed data cannot be used for the study of single elements, but can be used only for those based on a correlation coefficient matrix or a covariance matrix. In the analysis of ancient glass chemical composition data, the proportion of trace elements is very small, and the influence of the main elements on their content is weak. Therefore, when only trace elements or the chemical composition data of ancient glass for which the sum of each component is not equal to 1 are used, the use of central log-ratio transformation (CLR) is the most appropriate method.

To further elucidate the classification patterns derived from compositional clustering, multiple machine-learning algorithms—including decision tree, logistic regression, support vector machine (SVM), and random forest—were implemented to model the relationships between oxide composition and glass type. Among these, the decision tree, logistic regression, and SVM methods exhibited strong predictive performance and robustness against noise, confirming the stability of compositional boundaries across analytical uncertainties. Notably, the decision tree model identified *PbO content* as the most decisive variable for distinguishing lead-barium glass subclasses, reflecting the pivotal technological role of lead oxide as both a flux and optical modifier in ancient Chinese glassmaking[46,47]. The consistent prominence of PbO in model splits supports previous archaeometric interpretations that lead addition marked a deliberate technological innovation for achieving higher refractive indices and enhanced workability[48]. Sensitivity analysis performed on unknown or weathered samples demonstrated that the trained classifiers could reliably infer glass type even when partial compositional data were missing, thus underscoring their potential for nondestructive provenance and authenticity assessments.

This study demonstrates that integrating compositional data analysis with machine learning provides a powerful framework for decoding technological patterns in ancient Chinese glass. Beyond its archeological implications, this approach exemplifies how quantitative data-mining

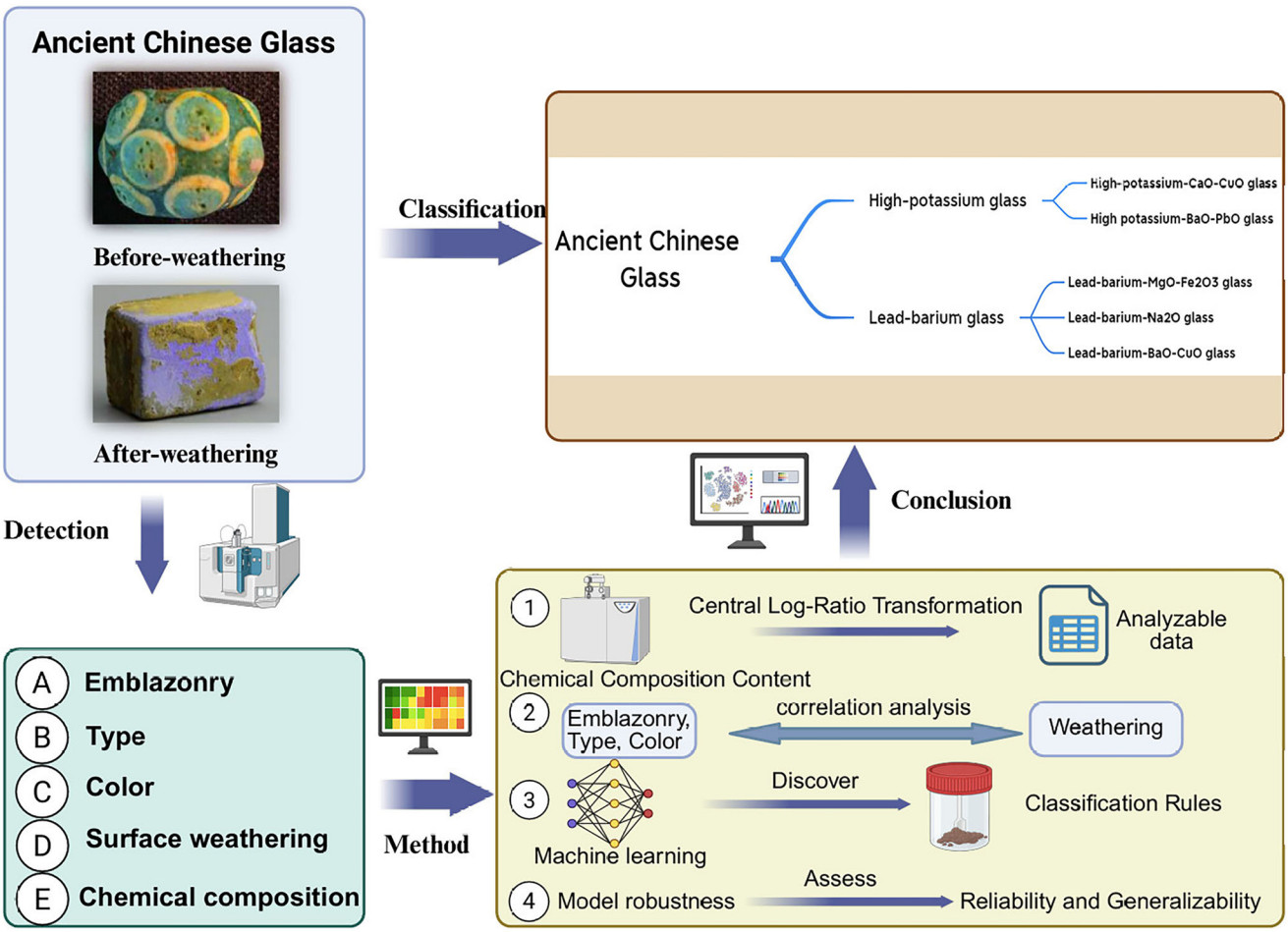

Fig. 8 | Schematic diagram of the research workflow.

techniques can enhance material characterization and authentication in conservation science. Future work should expand the dataset to include isotopic and microstructural parameters, allowing the integration of multimodal data into unified predictive frameworks[49,50]. Ultimately, the methodology presented here establishes a scalable model for transforming complex heritage datasets into interpretable technological narratives, bridging chemistry, archeology, and computational science within the broader discourse of cultural heritage research.

Data availability

The data in this paper come from Question C of the 2022 Chinese Contemporary Undergraduate Mathematical Contest in Modeling (CUMCM). All the data, programs and materials are available from the corresponding author upon request.

Received: 24 August 2025; Accepted: 4 February 2026;
Published online: 27 February 2026

References

1. Lü, Q. Q., Henderson, J., Wang, Y. Q. & Wang, B. H. Natron glass beads reveal proto-Silk Road between the Mediterranean and China in the 1st millennium BCE. *Sci. Rep.* **11**, 13 (2021).
2. Gan, F. X., Cheng, H. S. & Li, Q. H. Origin of Chinese ancient glasses - study on the earliest Chinese ancient glasses. *Sci. China Ser. E Technol. Sci.* **49**, 701–713 (2006).
3. Davies, T. E. et al. Experimental methods in chemical engineering: Scanning electron microscopy and X-ray ultra-microscopy-SEM and XuM. *Can. J. Chem. Eng.* **100**, 3145–3159 (2022).
4. de León, C. A. P., Montes-Bayón, M. & Caruso, J. A. Elemental speciation by chromatographic separation with inductively coupled plasma mass spectrometry detection. *J. Chromatogr. A* **974**, 1–21 (2002).
5. Saniso, E. et al. Garcinia drying using mixed-mode solar dryer technique: Drying kinetics, mathematical modeling and quality characteristics. *Case Stud. Therm. Eng.* **66**, 13 (2025).
6. Xu, Y. S. et al. Investigation of the Dynamic Properties of Viscoelastic Dampers with Three-Chain Micromolecular Configurations and Tube Constraint Effects. *J. Aerosp. Eng.* **38**, 17 (2025).
7. Tang, P. X., Li, H. T., Zhang, X. M. & Sun, X. A mathematical model of catalyst combination design and temperature control in the preparation of C4 olefins through ethanol coupling. *Rsc Adv.* **13**, 10703–10714 (2023).
8. Lai, Y. Z. et al. Machine learning for membrane bioreactor research: principles, methods, applications, and a tutorial. *Front. Env. Sci. Eng.* **19**, 26 (2025).
9. Wekalao, J., Alsalman, O. & Patel, S. K. Graphene metasurface biosensor design for label-free peptide detection with machine learning optimization based on support vector regression with polynomial kernel. *Diam. Relat. Mat.* **153**, 16 (2025).
10. Mathur, P., Srivastava, S., Xu, X. W. & Mehta, J. L. Artificial Intelligence, Machine Learning, and Cardiovascular Disease. *Clin. Med. Insights Cardiol.* **14**, 9 (2020).
11. Bickler, S. H. Machine Learning Arrives in Archaeology. *Adv. Archaeol. Pract.* **9**, 186–191 (2021).
12. Ling, Z. Y., Delnevo, G., Salomoni, P. & Mirri, S. Findings on Machine Learning for Identification of Archaeological Ceramics: A Systematic Literature Review. *IEEE Access* **12**, 100167–100185 (2024).

13. Li, Z. X. et al. Analysis of the Composition of Ancient Glass and Its Identification Based on the Daen-LR, ARIMA-LSTM and MLR Combined Process. *Appl. Sci.* **13**, 24 (2023).
14. Guo, Y. H., Zhan, W. & Li, W. H. Application of Support Vector Machine Algorithm Incorporating Slime Mould Algorithm Strategy in Ancient Glass Classification. *Appl. Sci.* **13**, 14 (2023).
15. Zou, Y. Molecular-Composition Analysis of Glass Chemical Composition Based on Time-Series and Clustering Methods. *Molecules* **28**, 15 (2023).
16. China Undergraduate Mathematical Contest in Modeling. *Question C of the 2022 Chinese Contemporary Undergraduate Mathematical Contest in Modeling* (China Undergraduate Mathematical Contest in Modeling, 2022).
17. Aitchison, J. & Greenacre, M. Biplots of compositional data. *J. R. Stat. Soc. Ser. C Appl. Stat.* **51**, 375–392 (2002).
18. Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A. & Pawlowsky-Glahn, V. Logratio analysis and compositional distance. *Math. Geol.* **32**, 271–275 (2000).
19. Aitchison, J. Principal component analysis of compositional data. *Biometrika* **70**, 57–65 (1983).
20. Aitchison, J. Principles of Compositional Data Analysis. *Multivariate Analysis and Its Applications IMS Lecture Notes Monograph Series*, Vol. 24, 73–81 (Institute of Mathematical Statistics, 1994). https://doi.org/10.1214/lnms/1215463786
21. Goo, J., Sakhanenko, L. & Zhu, D. C. A chi-square type test for time-invariant fiber pathways of the brain. *Stat. Infer. Stoch. Proc.* **25**, 449–469 (2022).
22. Wu, Y. F., Ma, Y. R. & Jiang, Y. Double-state chi-square test based sparse grid quadrature filtering algorithm and its application in integrated navigation. *IET Contr. Theory Appl.* **17**, 1203–1213 (2023).
23. Sanchez, I., Keatley, D., Oklevski, S. & Speers, S. J. A large study evaluation of evidence types containing offender fingerprints from recorded crimes in North Macedonia from 2005 to 2015. *Sci. Justice* **62**, 43–49 (2022).
24. Sharif, O. et al. Analyzing the Impact of Demographic Variables on Spreading and Forecasting COVID-19. *J. Healthc. Inform. Res.* **6**, 72–90 (2022).
25. Gel, Y. R. & Gastwirth, J. L. A robust modification of the Jarque-Bera test of normality. *Econ. Lett.* **99**, 30–32 (2008).
26. Thadewald, T. & Büning, H. Jarque-Bera test and its competitors for testing normality -: A power comparison. *J. Appl. Stat.* **34**, 87–105 (2007).
27. Gonzalez, T., Sahni, S. & Franta, W. R. An Efficient Algorithm for the Kolmogorov-Smirnov and Lilliefors Tests. *ACM Trans. Math. Softw.* **3**, 60–64 (1977).
28. Blain, G. C. Revisiting the critical values of the Lilliefors test: towards the correct agrometeorological use of the Kolmogorov-Smirnov framework. *Bragantia* **73**, 192–202 (2014).
29. Kelter, R. A New Bayesian Two-Sample *t* Test and Solution to the Behrens-Fisher Problem Based on Gaussian Mixture Modelling with Known Allocations. *Stat. Biosci.* **14**, 380–412 (2022).
30. Al-Kassab, M. M. & Majeed, A. H. The use of two-sample t-test in the real data. *Adv. Appl. Stat.* **81**, 13–22 (2022).
31. Amin, M., Fatima, A., Akram, M. N. & Kamal, M. Influential observation detection in the logistic regression under different link functions: an application to urine calcium oxalate crystals data. *J. Stat. Comput. Simul.* **14**, 346–359 (2023).
32. Sopitpongstorn, N., Silvapulle, P., Gao, J. T. & Fenech, J. P. Local logit regression for loan recovery rate. *J. Bank Financ.* **126**, 14 (2021).
33. Dong, G. S. & Mu, X. W. A novel second-order cone programming support vector machine model for binary data classification. *J. Intell. Fuzzy Syst.* **39**, 4505–4513 (2020).
34. Qin, Z. F. & Li, Q. Q. An uncertain support vector machine with imprecise observations. *Fuzzy Optim. Decis. Mak.* **22**, 611–629 (2023).
35. Ghosh, D. & Cabrera, J. Enriched Random Forest for High Dimensional Genomic Data. *IEEE ACM Trans. Comput. Biol. Bioinform.* **19**, 2817–2828 (2022).
36. Rhodes, J. S., Cutler, A. & Moon, K. R. Geometry- and Accuracy-Preserving Random Forest Proximities. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 10947–10959 (2023).
37. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
38. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. Classification and Regression Trees. *Biometrics* **40**, 874 (1984).
39. Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A. & Hess, K. T. Random forests for classification in ecology. *Ecology* **88**, 2783–2792 (2007).
40. Zanini, R., Franceschin, G., Cattaruzza, E. & Traviglia, A. A review of glass corrosion: the unique contribution of studying ancient glass to validate glass alteration models. *NPJ Mater. Degrad.* **7**, 17 (2023).
41. Degryse, P. & Shortland, A. J. Interpreting elements and isotopes in glass: A review. *Archaeometry* **62**, 117–133 (2020).
42. Rodrigues, A., Fearn, S. & Vilarigues, M. Historic K-rich silicate glass surface alteration: Behaviour of high-silica content matrices. *Corrosion Sci.* **145**, 249–261 (2018).
43. Greenacre, M. Compositional data analysis. In *Annual Review of Statistics and Its Application*, Vol. 8 (ed. Reid, N.) 271–299 (Annual Reviews, 2021).
44. Vera Pawlowsky-Glahn, J. J. E. T.-D. Compositional data and their sample space. In *Modelling and Analysis of Compositional Data* 8–22 (Wiley, 2015).
45. Filzmoser, P., Hron, K. & Reimann, C. Principal component analysis for compositional data with outliers. *Environmetrics* **20**, 621–632 (2009).
46. Chen, W. & Chen, D. Research on the classification of ancient silicate glass artifacts based on machine learning. *Archaeometry* **67**, 72–86 (2025).
47. Ma, Q., Braekmans, D., Shortland, A. & Pollard, A. M. The Production and Composition of Chinese Lead-Barium Glass through Experimental Laboratory Replication. *J. Non Cryst. Solids* **551**, 8 (2021).
48. Henderson J. *Ancient Glass: An Interdisciplinary Exploration* (Cambridge University Press, 2013).
49. Yogurtcu, B., Cebi, N., Koçer, A. T. & Erarslan, A. A Review of Non-Destructive Raman Spectroscopy and Chemometric Techniques in the Analysis of Cultural Heritage. *Molecules* **29**, 19 (2024).
50. Zumpano, R., Simonetti, F., Genova, C., Mazzei, F. & Favero, G. Raman spectroscopy and SERS: Recent advances in cultural heritage diagnostics and the potential use of anisotropic metal nanostructures. *J. Cult. Herit.* **71**, 282–301 (2025).

Acknowledgements

The authors thank the Problem C Annex of the National College Students Mathematical Modeling Competition of the Higher Education Society Cup for experimental data support. Financial support from the National Natural Science Foundation of China (12461075) and The Special Basic Cooperative Research Programs of Yunnan Provincial Undergraduate Universities Association (No.202401BA070001-157) are acknowledged and appreciated. Figure 8 was Created in BioRender. Tang, P. (2025) https://BioRender.com/12j8m00. We would like to express gratitude to AJE for its language editing assistance on our manuscript. We are grateful to the editor and the reviewers for their professional comments and hard work. Our joint efforts have enhanced the quality of this research.

Author contributions

Conceptualization, methodology, software, validation, formal analysis, data curation, writing—original draft preparation, visualization, writing—review and editing, J.D., P.X.; writing—original draft, funding acquisition, project administration, X.T. All the authors have read and agreed to the published version of the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s40494-026-02370-5.

Correspondence and requests for materials should be addressed to Xiaoting Gan or Jiade Tang.