Artificial Intelligence Technologies and Applications
C. Chen (Ed.)

doi:10.3233/FAIA231369

Improved Glass Composition Analysis and Identification of Cultural Heritage with Limited Data Using Data Augmentation and CatBoost

Xianmei HU[a1], Jinyu ZHANG[a2], Ziyang WANG[a3], Jiahui ZHAO[b4]
[a]*School of Computer Science and Technology, Jilin University, Changchun, China*
[b]*Materials and Manufacturing Research Group, James Watt School of Engineering, University of Glasgow, Glasgow, UK*
Both Xianmei HU and Jinyu ZHANG are the first authors

Abstract. Glass artifacts play a significant role in cultural heritage, offering valuable insights into ancient craftsmanship and cultural exchange. However, accurately analyzing and identifying ancient glass objects presents challenges due to limited data. This study aims to enhance the analysis and identification of glass compositions in cultural heritage by employing data augmentation techniques and the CatBoost prediction model. Firstly, data augmentation techniques are applied to expand the limited dataset, increasing sample quantity and diversity to improve the model's generalization capability. The TOPSIS method is employed to comprehensively evaluate different augmentation factors and select the most suitable ones. Subsequently, the CatBoost prediction model is utilized, and the model parameters are optimized using a random search method to further enhance predictive performance. Experimental research on ancient glass artifacts validates the effectiveness and feasibility of the proposed methods. The final model demonstrates high predictive performance and a good fit on the training set, cross-validation set, and test set. For example, when predicting the sodium oxide content before weathering in glass artifacts, the average R-squared(R^2) reaches 0.998, and the Mean Squared Error(MSE) is 0.003. These results signify the accurate prediction of glass artifact compositions and the model's stable predictive capabilities across different datasets. Utilizing the predicted chemical composition, the identification of glass artifacts achieves a classification accuracy of 100%, indicating the excellence of the model. In conclusion, this study presents an improved approach for analyzing and identifying glass compositions by overcoming the limitations posed by limited data through data augmentation and the CatBoost model. These advancements provide valuable tools and methods for preserving and researching cultural heritage, contributing to the progress of ancient civilization studies and technological development.

Keywords. CatBoost; TOPSIS; data augmentation; Glass artifacts; machine learning

[1] Corresponding Author: Xianmei HU, School of Computer Science and Technology, Jilin University, Changchun, China; Email: yami.hu061@gmail.com
[2] Jinyu ZHANG, School of Computer Science and Technology, Jilin University, Changchun, China; Email: jyzhang2121@mails.jlu.edu.cn
[3] Ziyang WANG, School of Computer Science and Technology, Jilin University, Changchun, China; Email: 3123208367@qq.com
[4] Jiahui ZHAO, Materials and Manufacturing Research Group, James Watt School of Engineering, University of Glasgow, Glasgow, UK; Email: 2740166z@student.gla.ac.uk

1. Introduction

As a valuable heritage of ancient civilizations, glass artifacts play a significant role worldwide. The history of glass production can be traced back to over 2000 years ago in ancient Egypt and ancient Rome, where people melted sand and other materials and employed techniques such as blowing and casting to create exquisite glass artworks and containers[1]. Over time, glass craftsmanship has evolved and been passed down through different cultures, carrying rich historical and cultural information. Therefore, in-depth research and analysis of glass artifacts hold great significance in understanding ancient civilizations, including their manufacturing techniques, technological inheritance, and cultural exchanges.

With the advancement of archaeological studies and technological progress, computational methods like machine learning have gained increasing attention in the field of archaeology. These methods provide archaeologists with new tools and approaches to explore cultural heritage more comprehensively[2]. However, glass artifacts are prone to weathering, leading to changes in their chemical compositions. During the weathering process, internal elements of glass undergo significant exchange with environmental elements, resulting in alterations of composition ratios and posing challenges in predicting the chemical compositions of glass artifacts before weathering[3].

Traditional weathering prediction models encounter certain issues in forecasting and evaluating the weathering process of glass artifacts. Firstly, traditional regression models struggle to handle complex non-linear relationships and high-dimensional data. The weathering process of glass artifacts is influenced by multiple factors, and the relationships between these factors are often non-linear, involving a large number of chemical components and structural parameters. Traditional regression models often fail to capture these intricate relationships adequately, resulting in lower prediction accuracy. Secondly, the collection of glass artifact data is a challenging task. Researchers often face small-scale datasets due to the scarcity of artifacts, protective requirements, and limitations in data acquisition. Machine learning algorithms typically require substantial amounts of data for training and modeling, which limits their effectiveness when applied to small datasets[4][5][6].

To address these issues, this study aims to improve the prediction and identification of chemical compositions of glass artifacts before weathering in cultural heritage. We propose an innovative approach that combines data augmentation (DA) techniques with the CatBoost prediction model to enhance the analytical capabilities and identification accuracy of ancient glass artifacts. Data augmentation techniques allow us to expand and modify the original dataset, increasing sample size and introducing diversity, thereby improving the model's generalization ability. The CatBoost classification model, known for its adaptive learning rate adjustment mechanism and optimization techniques, is well-suited for small datasets, as it effectively adapts to data distribution and features, leading to enhanced training and modeling performance.

Predicting the chemical compositions of glass artifacts before weathering holds significant importance in the fields of cultural heritage preservation, historical research, research method development, as well as identification and counterfeit detection. It not only contributes to the preservation and restoration of cultural heritage but also provides valuable insights into ancient manufacturing techniques, technological inheritance, and cultural exchanges. Accurate prediction of chemical compositions facilitates a deeper understanding of ancient glass artifact production methods and techniques, thus

advancing historical research and cultural heritage conservation. Moreover, precise identification and counterfeit detection of ancient glass artifacts address crucial challenges in the field of cultural heritage, while providing valuable auxiliary information for authenticating artifacts and ensuring the healthy development of the artifact market. The following sections will provide a detailed description of the methods and experimental design employed in this study, present the experimental results, and conduct discussions.

2. Research Process

2.1. Data Preparation and Preprocessing

This study selected a batch of ancient glass artifacts and collected relevant data. Archaeologists have classified these artifacts into two types: high-potassium glass and lead-barium glass, based on their chemical compositions and other detection methods. A total of 67 sample data points were included, which provided information on the type, weathering, color, and chemical composition proportions of the artifacts. To ensure the accuracy, completeness, and consistency of the data, the study used the mode-filling method to handle missing values. Considering that the detection methods and other factors may result in the sum of chemical composition proportions of some sample data not being equal to 100%, this study treated data with a sum of proportions between 85% and 105% as valid data and excluded those that did not meet this criterion. During the data augmentation process, the original data was expanded by a certain factor, and random noise ranging from -0.1 to 0.1 was added. Finally, the data was divided into a 70% training set and a 30% test set.

2.2. Research Method

2.2.1. Exploring Weathering Patterns Based on Statistical Tests

In this section, the potential weathering patterns of glass artifacts were explored through various data mining techniques. The following methods were employed:

a) Frequency analysis: The data was analyzed to determine the occurrence and distribution of different weathering patterns in the glass artifacts.

b) Categorical aggregation: The glass artifacts were categorized based on their weathering patterns to facilitate further analysis.

c) Cross-tabulation analysis: A cross-tabulation was performed to examine the relationship between weathering patterns and other variables, such as pattern type, color, etc.

d) Correlation analysis: In this study, we utilized the Spearman correlation coefficient to examine the correlation between weathering patterns and relevant factors, aiming to identify potential relationships or dependencies. Unlike the Pearson correlation coefficient, the Spearman correlation coefficient measures the strength and direction of the monotonic relationship between two variables using ranked data points, without assuming a linear relationship. The coefficient (rho) ranges from -1 to 1, where -1 indicates a perfect negative monotonic relationship, 1 indicates a perfect positive monotonic relationship, and 0 indicates no monotonic relationship. It is a valuable tool to assess the ordinal relationship between variables based on the comparison of ranked

data points[7]. The formula for calculating the Spearman correlation coefficient is as follows:

$$\rho = 1 - (6\sum d_i^2)/(n(n^2 - 1)) \quad (1)$$

Where ρ represents the Spearman correlation coefficient, $\sum d_i^2$represents the sum of squared differences, and n represents the sample size.

Through the application of these methods, we were able to gain preliminary insights into the underlying patterns of weathering in glass artifacts and uncover factors and patterns influencing their weathering processes. This has laid the foundation for our subsequent research and analysis.

2.2.2. CatBoost Prediction Model

This study employs data augmentation and CatBoost regression to predict changes in the chemical composition of glass artifacts before and after weathering. Data augmentation expands and adds noise to the original data, enhancing sample diversity and variability. This approach improves the model's ability to capture different patterns and variations. Through data augmentation, we generated additional samples, introducing real-world uncertainties and variations in the weathering process, creating a more comprehensive and representative dataset for training. This enables capturing a wider range of data samples, improving the model's generalization and prediction accuracy. To select the optimal expansion factor, we used the TOPSIS method based on five metrics. Furthermore, a random search was performed to optimize the CatBoost regression model by tuning parameters and improving its performance.

CatBoost is a Gradient Boosting Decision Tree(GBDT) framework based on symmetric decision trees. It is primarily designed to address challenges related to handling categorical features, gradient biases, and prediction offsets, to improve algorithm accuracy and generalization capability[8][9].While CatBoost shares the overall algorithm framework with GBDT, it introduces significant improvements in handling categorical features, boosting techniques, and decision tree scoring. These enhancements enable CatBoost to effectively handle string features and achieve faster model fitting speed compared to XGBoost and LightGBM for the same dataset.

a) Handling categorical features

CatBoost adopts an innovative approach for processing categorical features, moving away from simple greedy objective-based statistics for node splitting. Instead, CatBoost introduces a prior distribution term that considers the specific nature of categorical features when computing node gains. This approach effectively mitigates the influence of low-frequency features and noise on decision tree generation.

$$x_{i,k} = \frac{\sum_{j=1}^{p-1}[x_{\sigma j,k}=x_{\sigma p,k}]\cdot Y_j + a\cdot p}{\sum_{j=1}^{p-1}[x_{\sigma j,k}=x_{\sigma p,k}]+a} \quad (2)$$

In the formula, σj represents the j-th data sample, xi,k denotes the k-th discrete feature of the i-th row in the training set, a is a prior weight, and p is the prior distribution term. For regression problems, the prior term is usually set as the mean of the predicted labels in the training set, while for binary classification problems, it corresponds to the prior probability of positive instances. The square brackets [] denote

an indicator function that outputs 1 if the internal condition is satisfied, and 0 otherwise. Through the improved Target Encoding with Symmetry (TS) method, CatBoost can convert categorical features into numerical values while minimizing information loss.

b) Ordered boosting

Traditional GBDT models adopt a method without row or column sampling, where all the base learners, usually represented by Classification and Regression Trees(CART) decision trees, are trained on the complete dataset using gradient boosting. In each iteration, the negative gradient of the previous round's trees is used for training. However, this approach can lead to the accumulation of prediction bias and overfitting. To mitigate the overfitting effect, XGBoost, and Microsoft's LightGBM introduced row and column sampling as well as regularization techniques. CatBoost goes a step further and proposes the Ordered Boosting method. The pseudocode of the algorithm is shown in Figure 1.

Algorithm: Ordered boosting
Input$\{x_k, y_k\}_{k=1}^{n}$ ordered according to σ ,the numbers of trees I
σ←randompermutationof[1,n]
M ← Ofori=1ton
for iter ← 1 to I do
for i←1 to n do
for j ← 1 to i-1 do
$g_i \leftarrow \frac{d}{da} Loss(y_i, a)|_{a=M_j(X_j)}$
M ← learn a tree(X_j, g_j)
$M_i \leftarrow M_i + M$
return $M_1, M_2, \cdots, M_n$

Figure 1. Pseudocode of Ordered Boosting

Among them, σ represents the number of times the training set is randomly shuffled, and I represent the number of symmetric decision trees to be generated, which is equivalent to the number of learners. For all n samples, initialize M_i as 0. Then, through sampling on the random sequence and obtaining gradients based on it, the purpose of σ permutations is to enhance the robustness of the algorithm and effectively avoid overfitting. These permutations are the same as those used to calculate the improved TS. For each random permutation σ, train the n different models $M_i$as shown above. Then, sequentially calculate the gradients g_j of the loss function (Loss) for the first i-1 data points, and use the i-1 g_j to construct a residual tree in the symmetric tree. Update the initial model Mi from $M_i(X_1)$ to $M_i(X_i)$. The purpose of this process is to remove the influence of $X_i$on the model's prediction for X_i, thereby reducing the interference of noise on the model. For each permutation in s permutations, we build n models M_i, resulting in an overall complexity of approximately $O(s \cdot n^2)$. To accelerate the algorithm, when updating Mi, CatBoost does not store and update $O(n^2)$ models $M_i(X_i)$, but instead uses $M_i'(X_j)$, where i=1 to log2(n) and j<2i+1. $M_i'(X_j)$ is an approximation based on the same j of the previous 2i samples. Finally, the prediction complexity of $M_i'(X_j)$ will not exceed $\sum_{0 \le i \le log2(n)} 2^{i+1} < 4n$.

c) Fast Scoring

CatBoost uses Oblivious Decision Trees (ODT) as base learners, which have the following structure shown in the diagram below. Unlike general decision trees, ODTs have identical feature selection and threshold for internal nodes at the same depth. Therefore, ODTs can be transformed into decision tables with 2^d entries, where d represents the depth of the decision tree. This tree structure is more balanced and

features faster processing speed compared to typical decision trees. Additionally, by uniformly treating floating-point features, statistical information (such as user IDs), and one-hot encoded features as binary, the model greatly reduces the need for hyperparameter tuning. The Figure 2 below illustrates the structure of an Oblivious Decision Tree.

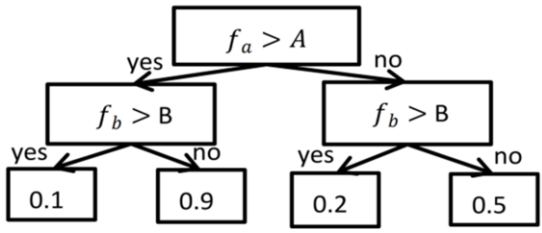

Figure 2. Oblivious Decision Tree (ODT) structure

d) Feature Importance Ranking

CatBoost not only achieves high prediction accuracy but also allows us to identify the relative contributions of different influencing factors (i.e., the features used for prediction) to the prediction results. The relative contribution of a feature in a single decision tree is measured using the following formula:

$$J_j^2 = \frac{1}{M\sum_{m=1}^{M} J_j^2(T_m)} \tag{3}$$

Where M represents the number of iterations (number of trees) and J_j^2 represents the global importance of feature j.

$$J_j^2(T) = \sum_{t=1}^{L-1} i_t^2 I(v_t = j) \tag{4}$$

In the formula, L represents the number of leaf nodes in the tree, L-1 represents the number of non-leaf nodes, v_t is the feature associated with node t, and i_t^2 represents the squared loss reduction after the split at node t. A higher value of i_t^2 indicates a greater benefit from the split, indicating a higher feature importance for the corresponding node.

2.2.3. Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS)

TOPSIS is a multi-criteria decision analysis method used to evaluate and rank multiple alternative solutions. It is based on the principle of comparing each solution to both the ideal and anti-ideal solutions and determining their relative superiority by calculating distances. The basic idea of TOPSIS is to construct a normalized matrix by normalizing the original data with the same trend, and then measure the differences between the evaluation objects and the ideal and anti-ideal vectors to assess their differences[10]. Assuming there are n evaluation objects and m criteria, the basic steps of TOPSIS are as follows:

Step 1: Same-trend normalization of the original data:

Differentiate the categories of indicators in the criteria system (higher is better or lower is better) and perform forward transformation according to different formulas for

different types of indicators. Construct an n x m matrix X_{ij}, where X represents the value of the j-th criterion for the i-th object.

Step 2: Construct the normalized matrix using the following formula.

$$Z_{ij} = \frac{X_{ij}}{\sqrt{\sum_{k=1}^{n} {X_{ij}}^2}} \tag{5}$$

Step 3: Calculate the distances between each evaluation criterion and the ideal and anti-ideal vectors using the following formula, where j represents a specific evaluation criterion, m represents the number of evaluation criteria, w_j is the weight of the j-th criterion, Z_j^+represents the data for the ideal solution of the j-th criterion,z_{ij} represents the normalized data of the j-th criterion for a specific evaluation object i, and Z_j^-represents the data for the anti-ideal solution of the j-th criterion.

$$\begin{cases} D_i^+ = \sqrt{\sum_{j=1}^{m} w_j (Z_j^+ - z_{ij})^2} \\ D_i^- = \sqrt{\sum_{j=1}^{m} w_j (Z_j^- - z_{ij})^2} \end{cases} \tag{6}$$

Step 4: Measure the proximity of the evaluation objects to the ideal solution using the following formula, where D_i^+ and D_i^- represent the positive and negative ideal distances of the i-th object, and a larger value of Ci indicates a more optimal evaluation object.

$$\mathrm{C_i} = D_i^- / D_i^+ + D_i^- \tag{7}$$

3. Statistical Result

3.1. Analysis of Weathering Patterns

The Table 1 below shows the results of the cross-tabulation analysis with surface weathering as the grouping variable and pattern, type, and color as the analysis variables. It includes variables, frequencies, and percentages.

Table 1 Table of Cross-tabulation Analysis

	Name	Surface Weathering		Total
		No weathering	*weathering*	
Pattern	***A***	14(50.000%)	14(50.000%)	28
	B	0(0.000%)	6(100.000%)	6
	C	11(33.300%)	22(66.700%)	33
Total		25	42	67
Type	***Lead-Barium***	13(26.500%)	36(73.500%)	49
	High Potassium	12(66.700%)	6(33.300%)	18
Total		25	42	67
Color	***Light Green***	2(66.700%)	1(33.300%)	3
	Light Blue	6(23.100%)	20(76.900%)	26
	Dark Green	3(42.900%)	4(57.100%)	7
	Dark Blue	3(100.000%)	0(0.000%)	3
	Purple	2(33.300%)	4(66.700%)	6

	Name	Surface Weathering		Total
		No weathering	*weathering*	
	Green	1(100.000%)	0(0.000%)	1
	Blue-Green	8(47.100%)	9(52.900%)	17
	Black	0(0.000%)	4(100.000%)	4
Total	25		42	67

Pattern B of the glass artifacts has undergone complete weathering, indicating that it provides less protection against external factors such as humidity, light, and chemicals compared to other patterns. The simplicity and smaller coverage area of Pattern B contributes to its lower level of protection and higher susceptibility to weathering. In contrast, Pattern A is considered the most complex, offering a higher level of protection, followed by Pattern C. However, overall, more than 60% of the samples in all patterns have experienced weathering, suggesting that patterns have limited effectiveness in protecting glass artifacts.

When comparing different types of glass artifacts, lead-barium glass exhibits a higher weathering rate of 73.500%, while potassium-rich glass has a lower rate of 33.300%. This preliminary analysis indicates that lead-barium glass is more prone to weathering compared to potassium-rich glass.

Table 2 Table of Spearman correlation coefficient

	Pattern	Type	Color	Surface Weathering
Pattern	1 (0.000***)	-0.432 (0.000***)	-0.402 (0.001***)	-0.004 (0.977)
Type	-0.432 (0.000***)	1 (0.000***)	0.569 (0.000***)	0.368 (0.002***)
Color	-0.402 (0.001***)	0.569 (0.000***)	1(0.000***)	-0.033 (0.790)
Surface Weathering	-0.004 (0.977)	0.368 (0.002***)	-0.033 (0.790)	1 (0.000***)

[a.] *** , ** , and * represent significance levels of 1%, 5%, and 10%, respectively.

Table 3 Table of Joint Cross-Analysis of Type and Decoration

	Name	Type		Total
		High Potassium	*Lead-Barium*	
Pattern	***A***	8(28.600%)	20(71.400%)	28
	B	6(100.000%)	0(0.000%)	6
	C	4(12.100%)	29(87.900%)	33
Total		18	49	67

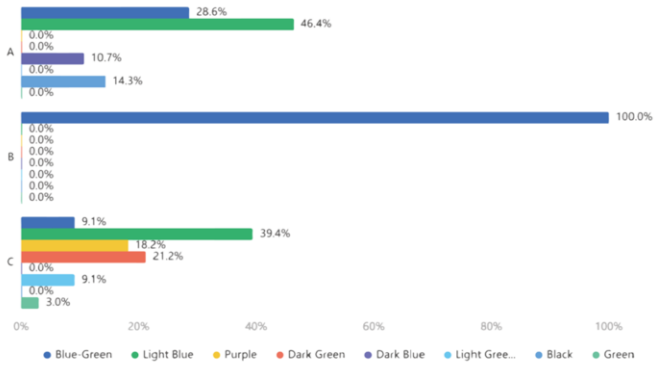

Figure 3. Cross-Chart of Decoration and Color.

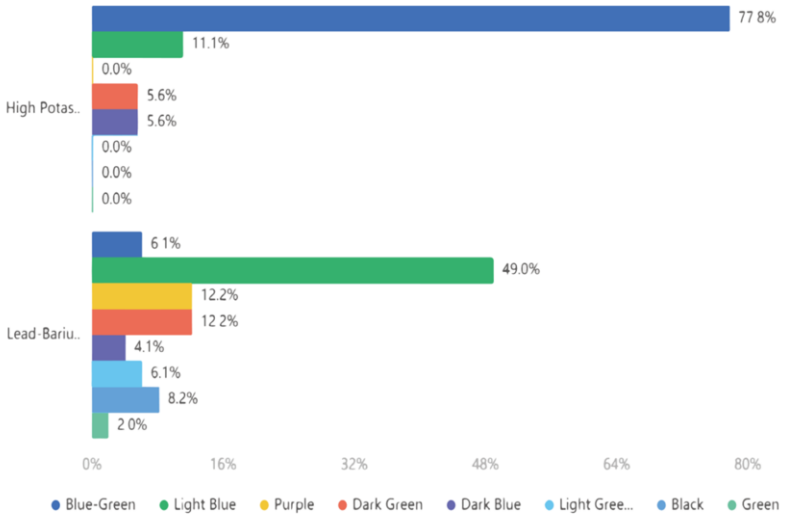

Figure 4. Cross-Chart of Type and Color.

Based on the Spearman correlation coefficients presented in Table 2 above, we can conduct the following analysis at the significance level of $\alpha = 0.05$:

- There is a significant negative correlation between the pattern and the type of glass artifacts, indicating a close association between the complexity of patterns and the type of glass artifacts. Further analysis in Table 3 reveals that high-potassium glass is predominantly associated with Pattern B, while lead-barium glass is mainly associated with Patterns A and C.
- There is a significant negative correlation between decoration and color, suggesting a certain association between the patterns and the colors of glass artifacts. Combining the analysis with Figure 3, Pattern A is mainly found in light blue and blue-green artifacts, Pattern B is predominantly observed in blue-green artifacts, and Pattern C is mainly found in light blue, purple, and deep green.
- There is a significant positive correlation between the type of glass artifacts and their color, indicating a relationship between the type of glass and its color. Analyzing Figure 4, high-potassium glass tends to have a blue-green color, while lead-barium glass exhibits a wider range of color variations, with light blue being the most common.
- There is a significant positive correlation between the type of glass artifacts and their surface weathering, suggesting that certain types of glass artifacts may be more susceptible to surface weathering, supporting the previous conclusion that lead-barium glass is more prone to weathering.

3.2. CatBoost Model for Predicting Changes in Chemical Composition before and after Weathering

In this section, we will use Na2O as an example to demonstrate in detail the process of improving the model through data expansion and parameter adjustment.

3.2.1. The Predictive Performance of the CatBoost Regression Model on the Original Dataset.

In this study, we assessed the predictive model using four evaluation metrics: MSE, RMSE, MAE, and MAPE. The MSE measures the squared difference between predicted and actual values, indicating prediction accuracy. The RMSE is the square root of MSE and quantifies the average difference between predicted and actual values, with a smaller value denoting better predictive performance. The MAE measures the average absolute difference, indicating accuracy, while the MAPE represents the average absolute percentage difference, reflecting prediction accuracy. As shown in Table 4, Our results indicate relatively accurate predictions on the training set, as evidenced by low MSE, RMSE, and MAE values. However, the larger MSE and MAE on the cross-validation and test sets suggest potential overfitting, and the relatively large MAPE values point to significant prediction errors.

Table 4 Table of CatBoost Regression Model Evaluation Results Based on Original Data

	MSE	RMSE	MAE	MAPE	R^2
Training Set	0.005	0.07	0.045	55.725	0.998
Cross-Validation Set	0.651	0.807	0.388	354.538	0.771
Test Set	0.56	0.748	0.463	289.538	0.755

In summary, the CatBoost regression model on the original data demonstrates good prediction accuracy on the training set. However, there are larger prediction errors on the cross-validation and test sets, indicating potential overfitting. Further steps such as optimizing model parameters, conducting feature engineering, and increasing the training data size may improve the model's generalization ability and prediction performance.

3.2.2. Data Augmentation Improves the Model.

In this section, we attempted to improve the performance of the model by using data augmentation techniques. We set different augmentation multipliers ranging from 1 to 10 and used the TOPSIS method to calculate scores based on the MSE, RMSE, MAE, MAPE, and R^2 metrics to select the most suitable augmentation multiplier. Tables 5 and 6 below show the results of the entropy weight method for weight calculation and the TOPSIS evaluation results.

Table 5 Table of Indicator Weight Calculation

Indicator	Entropy (e)	Information Utility (d)	Weight (%)
R^2	0.954	0.046	18.918
MSE	0.953	0.047	19.087
RMSE	0.95	0.05	20.585
MAE	0.951	0.049	20.181
MAPE	0.948	0.052	21.229

Table 6 Table of TOPSIS Evaluation Results

Multiplier	D+	D-	Comprehensive Score	Ranking
1	0.89708404	0.32725592	0.26729171	10
2	0.32410145	0.70190073	0.68411232	8
3	0.48276993	0.75694527	0.61057997	9
4	0.19928142	0.84657479	0.8094562	7
5	0.05569424	0.9516854	0.94471376	3
6	0.15781201	0.87312885	0.84692429	6
7	0.09350346	0.91674289	0.90744489	5

Multiplier	D+	D-	Comprehensive Score	Ranking
8	0.08395407	0.92540059	0.91682401	4
9	0.03015994	0.98224883	0.97020972	1
10	0.03107141	0.98542975	0.96943298	2

Choosing the highest-scoring augmentation factor may suggest that the model performs the best under the current data augmentation setting. However, it does not necessarily mean that it is the globally optimal choice. In the process of parameter tuning, we typically aim for the model to adapt better to different data and conditions, rather than just performing well under the current data augmentation setting. Therefore, we choose a good but not the highest augmentation factor to provide greater flexibility and exploration space for subsequent parameter tuning, to achieve a more stable and robust model configuration.

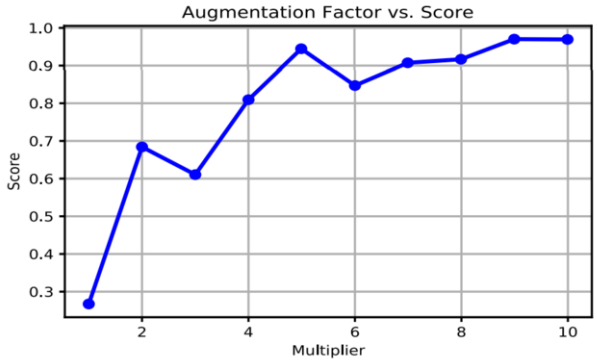

Figure 5. Scores at Different Multiples.

Figure 5 clearly shows the scores at different multiples, and it can be observed that the scores are already excellent from 4 to 8 multiples, fluctuating around 0.9. Here, we select 8 as our final augmentation multiple.

3.2.3. Parameter tuning further improves the model

After expanding the dataset using data augmentation techniques, we further optimized the model performance through parameter tuning. We employed the random search method to optimize the parameters of the CatBoost model. During the experiment, we considered the trade-off between time cost and model complexity by limiting the maximum number of iterations to 100 for CatBoost.

The parameters we tuned include:

- Learning Rate: Controls the step size for each iteration and influences the contribution of each tree. We explored different learning rates to find the optimal balance between convergence and learning speed.
- Depth: Determines the complexity and capacity of each tree in the CatBoost ensemble. By adjusting the depth parameter, we aimed to find the optimal tree depth level that maximizes model performance without overfitting the data.
- Regularization Parameters: CatBoost offers regularization options such as L1 and L2 regularization to prevent overfitting. These parameters control the degree of regularization in the model. By tuning the regularization parameters, we sought to strike a balance between model complexity and generalization ability.

We evaluated the performance of each parameter combination on the validation set using the mean squared error (MSE) as the evaluation metric. We set the initial parameter ranges based on typical values and iteratively narrowed down the parameter ranges in each round of optimization based on the results from the previous round, until convergence. The detailed optimization process is presented in Table 7.

Table 7 Table of Random Search Parameter Tuning Process

Iteration	Parameters	Random Count	Result（X1,X2,X3）	MSE
1	X1: [0.01, 0.1, 0.5, 1.0] X2: [3, 5, 7, 9] X3: [0.1, 0.5, 1.0, 3.0, 5.0]	50	(0.1,5,0.5)	0.0063
2	X1: [0.08,0.09,0.1,0.2,0.3] X2: [4, 5, 6] X3: [0.3,0.4, 0.5, 0.6]	50	(0.08,5,0.5)	0.0060
3	X1: [0.06,0.07,0.08,0.09] X2: [4, 5, 6] X3: [0.4, 0.5, 0.6]	20	(0.7,5,0.5)	0.0053

[b.] X1 represents the learning rate, X2 represents the depth, and X3 represents the l2_leaf_reg.

We used the final model to make predictions on the test set and training set and performed 5-fold cross-validation to calculate various evaluation metrics. The results are shown in Table 8.

Table 8 Table of Final Model Evaluation Results

	MSE	RMSE	MAE	MAPE	R^2
Training Set	0.001	0.038	0.026	26.823	0.999
Cross-Validation Set	0.005	0.073	0.042	98.718	0.998
Test Set	0.003	0.054	0.033	33.110	0.998

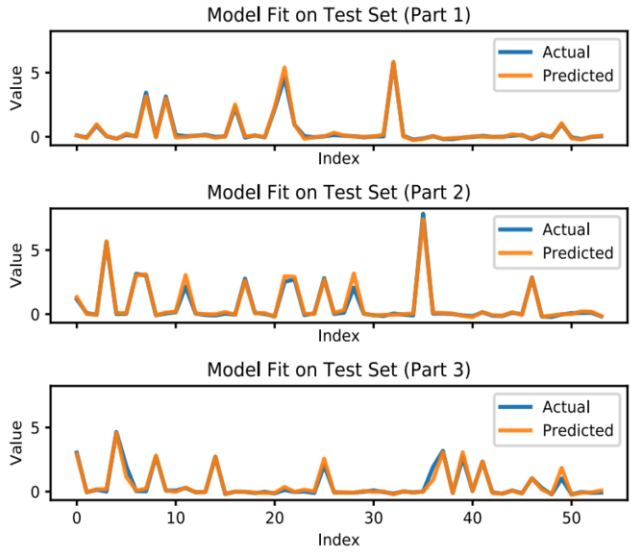

Figure 6. Fitment Effect Graph on the Test Set.

The comprehensive analysis of the table indicators indicates that the final model demonstrates high predictive performance and fitting ability on the training, cross-validation, and test sets. This implies that the model is capable of accurately

predicting the target variable and exhibits stable predictive capability across different datasets. This is crucial for practical prediction tasks as the consistency of the model across different datasets indicates good generalization ability and stability. These results further support the effectiveness and reliability of the model in predicting the target variable. The Figure 6 shows the fitting performance of the model on the test set.

3.3. The CatBoost Model for Classifying Glass Artifacts

Based on the previous discussions on predicting the chemical composition of glass artifacts before weathering, a CatBoost classification model was established. Table 9 below displays the evaluation results of the CatBoost classification model. The model achieved perfect scores in accuracy, recall, precision, and F1 score on all three datasets (training, cross-validation, and test). This indicates that the model performed exceptionally well in classifying glass artifacts, demonstrating high accuracy and effectiveness in identifying the glass artifact categories. This suggests that there are significant differences between the two categories of glass artifacts, and on small datasets, the CatBoost classification model is often able to effectively utilize the differences between categories to learn the feature patterns between them. Moreover, CatBoost is capable of leveraging the differences between categories to construct more powerful decision boundaries, thereby achieving better classification performance.

Table 9 Table of CatBoost Classification Model Evaluation Results Based on Original Data

	Accuracy	Recall	Precision	F1
Training Set	1	1	1	1
Cross-Validation Set	1	1	1	1
Test Set	1	1	1	1

4. Conclusion and Future Outlook

This study aimed to improve the prediction and identification of the pre-weathering chemical composition of glass artifacts in cultural heritage. By applying data augmentation techniques and combining them with the CatBoost prediction model, we successfully enhanced the analytical capabilities and identification accuracy of ancient glass artifacts. This research provided valuable support for a deeper understanding of ancient civilizations' manufacturing techniques, technological heritage, and cultural exchanges.

Through the evaluation of experimental results, we validated the effectiveness and feasibility of the proposed methods. The final model demonstrated high predictive performance and fitting on the training set, cross-validation set, and test set. This indicates that the model can accurately predict the target variables and exhibits stable predictive capabilities across different datasets. These findings are crucial for the preservation and restoration of cultural heritage, historical research, and identification and forgery detection.

However, this study still has some limitations that need to be addressed. Firstly, the size of the dataset remains relatively small, which may limit the model's generalization ability and applicability. Expanding the dataset is key to improving the model's performance, and we recommend collecting more diverse and extensive glass artifact data in future research. Secondly, this study focused solely on the prediction and

identification of the pre-weathering chemical composition of glass artifacts, without considering other relevant features and attributes. Future research can explore the introduction of additional features and attributes to enhance the predictive ability and identification accuracy of the model.

In terms of future outlook, we propose several areas for improvement. Firstly, further expanding the dataset by collecting more diverse and extensive glass artifact data will enhance the model's generalization ability and predictive performance. Secondly, incorporating additional features and attributes, such as structural parameters and physical properties of artifacts, will increase the model's diversity and generalization ability. Additionally, integrating other analytical techniques and methods, such as image processing and spectral analysis, can provide a more comprehensive and multi-dimensional analysis of glass artifacts.

In conclusion, through continuous research and improvement, we can further enhance the accuracy and reliability of predicting and identifying the pre-weathering chemical composition of glass artifacts. This will provide stronger support for the preservation and transmission of cultural heritage and promote advancements in the study of ancient civilizations and technological development.

References

[1] Brill, Robert H. "Ancient glass." Scientific American 209.5 (1963): 120-131.

[2] M. Singh, A. K. Rai, G. V. Krishna, Y. D. Kumar, R. Mishra and D. Kothandaraman, "Machine Learning and AI-based Robotic System for Archaeological Research," 2023 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 2023, pp. 48-52, doi: 10.1109/ICICT57646.2023.10134064.

[3] Brill, Robert H. "Crizzling–a problem in glass conservation." Studies in Conservation 20.sup1 (1975): 121-134.

[4] Kaplan, Maureen F. "Characterization of weathered glass by analyzing ancient artifacts." The scientific basis for nuclear waste management. Boston, MA: Springer US, 1980. 85-92.

[5] Yuan Y, Li S, Huang Y. Study the Chemical Composition Content of Ancient Glass Products before Weathering[J]. Analytical Chemistry: A Journal, 2022, 1(1): 54-60.

[6] W. Li, Y. Zhao, M. Dai, and J. Li, "Prediction and Classification of Ancient Glass Types Based on Logistic Regression Models," 2023 IEEE 3rd International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB), Taichung, Taiwan, 2023, pp. 331-336, doi: 10.1109/ICEIB57887.2023.10170035.

[7] Piantadosi, J.; Howlett, P.; Boland, J. (2007) "Matching the grade correlation coefficient using a copula with the maximum disorder", Journal of Industrial and Management Optimization, 3 (2), 305–312.

[8] Z. Li, Z. Li, J. Zhang, and J. Yao, "Prediction of 3D Cluster Energy Based on CatBoost Regression Model and Genetic Algorithm," 2022 International Conference on Applied Physics and Computing (ICAPC), Ottawa, ON, Canada, 2022, pp. 15-18, doi: 10.1109/ICAPC57304.2022.00009.

[9] C. Zhong, F. Geng, X. Zhang, Z. Zhang, Z. Wu, and Y. Jiang, "Shear Wave Velocity Prediction of Carbonate Reservoirs Based on CatBoost," 2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, 2021, pp. 622-626, doi: 10.1109/ICAIBD51990.2021.9459061.

[10] D. Datta, S. Biswas, and D. Datta, "An Innovative Technique for Intelligent Decision Making: Smart TOPSIS using Naïve Bayes Classification Algorithm," 2022 IEEE World Conference on Applied Intelligence and Computing (AIC), Sonbhadra, India, 2022, pp. 66-70, doi: 10.1109/AIC55036.2022.9848807.