Based on Machine Learning Classification Correlation Study and Species Classification of Glass Artifact Under the Influence of Weathering

Jingyi Wang [1,*], Xueqing Li [2]

[1] College of Information and Computer, Taiyuan University of Technology, Taiyuan, Shanxi, 030600

[2.]College of Economics, Tianjin Normal University, Tianjin, 300387

* Corresponding Author Email: kexi_jyw@163.com

Abstract. The five thousand years of Chinese culture has congealed countless crystallization of wisdom, and glass is also one of the products. However, because in ancient times the technology of the China glass production failed to reach the current level, it is highly susceptible to be weathered by the burial environment. The paper provides theoretical support for the archaeological study of ancient glass by correlating the relationship between surface weathering and the features of glass artifacts and by exploring the classification rules of ancient glass of known glass varieties (lead-barium glass and high-potassium glass). For the relationship between color and surface weathering, the most suitable AHP model is selected in the paper, while the importance scale of AHP is quantified by combining the physical model. Secondly, the Kendall coefficient consistency test model was used to analyze the correlation between glass varieties and surface weathering. Finally,the chi-square test model was used to analyze the correlation between ornament and surface weathering. The paper draws a conclusion that the surface weathering of glass relics has more correlation with its color, the deeper the color, the greater the possibility of surface weathering; the correlation between lead-barium and weathering is high; and the degree of correlation between glass ornamentation and weathering is very low in agreement.In view of the classification problem, the paper uses linear discriminant analysis (LDA) to establish a linear classifier and quantify the classification rule, in order that data with low frequency are not reduced, so as not to truly reflect the classification law, and finally the paper obtains a linear classifier with accurate classification.

Keywords: Analytic Hierarchy Process (AHP), Kendall coefficient consistency test model, chi-square test, linear discriminant analysis (LDA).

1. Introduction

Glass, as one of the earliest man-made materials invented by mankind, has a history of more than two thousand years in China. It carries rich cultural and technical information, and is an important carrier for the development of Chinese civilization. Therefore, the accurate identification of the chemical composition, variety and color and so on of Chinese ancient glass can promote and enrich the exploration and research of scientific and technological archaeology history, the protection of cultural relics and human civilization [1].

However, the ancient glass is easily weathered by the burial environment. In the process of weathering, the chemical elements of glass products exchange with the external environment elements, resulting in the chemical composition, variety and color and so on cannot be accurately judged. As two main varieties of Chinese ancient glass —— lead-barium glass and high potassium glass, the former is usually considered to be China's own invention of glass varieties, the latter is a very popular variety around Lingnan, China, and changes in chemical composition before and after weathering can make it difficult to distinguish between these two varieties. The rising cost of testing inevitably increases the difficulty of preserving and restoring cultural objects.

It is undeniable that there are many monographs, topics and papers on the restoration of cultural relics in academia. However, there are few papers that really look at the effect of weathering on various properties of cultural objects and use this effect to obtain a classification rule and topics of

DHPAC research in China possessed inconspicuous orientation and consistency. [2]. The paper integrates different types of data, analysis data into the meaningful archaeology and anthropology research framework, and combined with similar fine typology and technology research [3], obtaining reliable rules for classifying different varieties of glass (lead-barium, high potassium) makes it easier to precisely target heritage conservation and restoration work.

2. Materials and Methods

2.1. data

2.1.1 brief introduction

In the paper, 58 groups of cultural relics samples have been accurately detected, including ornamentation, varieties (lead-barium, high potassium), color and surface weathering. At the same time, the paper obtained the proportion of the main chemical composition of the 58 groups of partially weathered/unweathered sites, with a total of 69 data.

2.1.2 data preprocessing

(1) Data cleaning：

For the color data missing value processing, the paper adopts the deletion method. The goal can be achieved by simply deleting a small number of samples; On the contrary, the interpolation method can affect the process of statistical analysis and easily interfere with the results of the study, making the results biased and even drawing wrong conclusions. 54 valid data were counted.

To handle the missing value of chemical composition content data, the fixed value (0) has been adopted to fill the missing value, because the blank in the table indicates the abnormal value of chemical composition content, the component proportion and the data between 85% and 105% as valid data, thus the composition content of each sample, select the invalid data (No.15 and 17), a total of 67 valid data.

(2) Data transformation: the classification indicators are transformed into quantitative indicators

The attribute values of glass variety, ornament or color are composed of character type, which is very inconvenient in the modeling process and will take up more computer resources.Therefore, the paper uses the whole data to express the meaning of the original attribute value:The glass varieties "lead-barium" and "high potassium" are replaced by {0, 1}; the patterns "A", "B" and "C" are replaced by {1, 2, 3}; the colors "light blue", "blue-green", "dark blue", "light green", "dark green", "green", "purple" and "black" are replaced by {1, 2, 3, 4, 5, 6, 7, 8}, and "weathered" and "unweathered" are replaced by {1, 0}.

(3) Data statute

Both the two relics of "black" in color are weathered, so the paper speculated that "black" color is easy to weathering, and "black" data does not participate in AHP modeling; meanwhile, "purple" color has four relics, two relics are not weathered and have no statistical significance, so they do not participate in AHP modeling.

2.1.3 signalment

The normality test was conducted for ornamentation, variety, color, and surface weathering, and as can be seen from Table 1, the sample sizes of the data studied were all greater than 50, so the K-S test was used. Specifically, all of them showed significance ($p<0.05$), which means that the original hypothesis (original hypothesis: normal distribution of data) was rejected and all of them did not have the quality of normality.

Table 1. Analysis results of normality test

name	Size of sample	average value	standard deviation	skewness	kurtosis	The Kolmogorov-Smirnov test		Shapiro-Wilk test	
						Statistics values D	*p*	Statistical amount value W	*p*
ornamentation	54	2.148	0.940	-0.306	-1.837	0.336	0.000**	0.695	0.000**
variety	54	1.333	0.476	0.727	-1.529	0.425	0.000**	0.595	0.000**
color	54	4.056	2.334	0.383	-1.572	0.237	0.000**	0.816	0.000**
Surface weathering	54	1.556	0.502	-0.230	-2.023	0.368	0.000**	0.632	0.000**

* $p<0.05$ ** $p<0.01$

2.2. Method introduction

For the study of the correlation between the surface weathering and glass relic properties, the widely applicable correlation coefficient model include Pearson correlation coefficient and Spearman's rank correlation coefficient. However, the Pearson's correlation coefficient requires two variables to follow the normal distribution or the amount of data is relatively large, the Spearman's rank correlation coefficient needs to determine the level of the variables at each point (period), and the two sets of data are required to have no repeated observations or less observations. According to the characteristic description results, the two models are not suitable. Also, unfortunately, p-XRF has several drawbacks for the investigation of ancient glass. The most evident is a significant limitation in providing accurate information for “light” elements, which may constitute a large part of the total mass fraction in the glass matrix[4].

According to the relationship between single color variable and surface weathering, the paper selects AHP model: Hierarchical analysis(AHP) is a decision making method that decomposes the elements related to decision making into levels such as objectives, criteria, and options, and performs qualitative and quantitative analysis based on the above. Quantitative models are based on a set of variables that vary over a specific domain, while quantitative and causal relationships have been defined between these variables[5].Due to the subjectivity of the AHP model, the paper combines the physical model as a quantitative model and the objective facts, which greatly reduces the subjective ambiguity of the model and makes it clearer.

According to the relationship between single variety variable and surface weathering of glass variety, the paper selects the consistency test model of Kendall coefficient: Kendall rank correlation coefficient is a rank statistical parameter with non-parametric property (unrelated to the distribution), which conforms to the characteristic that the "variety" attribute does not have the normal characteristic.

For the relationship between the single variable of ornamentation and surface weathering, the paper selects the chi-square test model: the chi-square test is the deviation degree between the actual observed value of the statistical sample and the theoretical inferred value, the deviation degree between the actual observed value and the theoretical inferred value determines the size of the value of the chi-square, if the larger the square value, the greater the deviation of the two; on the contrary, the smaller the deviation between the two.

To explore the classification rules of ancient glass with known glass varieties (lead-barium, high potassium), the paper adopts the machine learning method, using the known historical data to "train" the machine and then "learn" a certain law, and establishes a model to predict the future results. Machine learning is divided into supervised learning methods and unsupervised learning methods (Figure 1). Supervised learning methods mainly include classification and regression, which can realize the prediction of the identification of new data. The classification method mainly targets discrete data, and the regression method mainly targets continuous data. The unsupervised learning

method mainly includes clustering, but has no predictive effect. Therefore, the paper selects the classification method of supervised learning, and corrects the accuracy of the model by selecting some known data as test data.

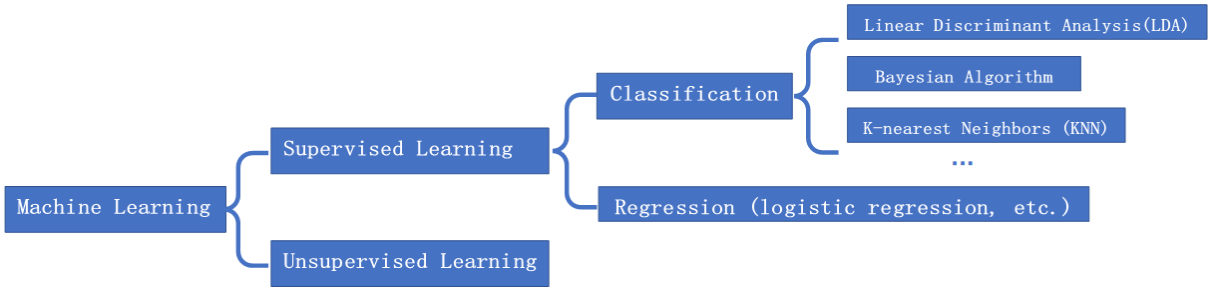

Figure 1. Classification of machine learning

For naive Bayes classification (NaiveBayes, NB) and K-nearest neighbor classification (KNN) such algorithms using Bayes theorem for classification, the process is simple and fast. However, for a category variable feature in the test set, if not seen in the training set, the result probability is 0, and the prediction function is invalid.

The paper selects the given training sample set of linear discriminant analysis (LDA), and tries to project the sample to a straight line, so that the projection point of the similar sample is as close as possible, and the projection points of the different sample are as far away as possible; when classifying the new sample, it is projected to the same line, and then the category of the new sample is determined according to the position of the projection point. Linear discriminant analysis (LDA) quantifies the classification rule, which avoids the dimension reduction of the data with low frequency, so as not to truly reflect the classification law. Also, linear discriminant analysis (LDA) gives acceptable accuracy in the analysis of the high-dimensionality data[6].

3. Model building and solution

3.1. Model of correlation between surface weathering and attributes of glass relics

3.1.1The AHP model combined with the physical model: the relationship between the color single variable and the surface weathering

Ancient glass is greatly affected by the light, and the light will cause fading and structural damage [7]. According to the physical equation $C = H \times \lambda$, $E = h \times H$ it is known that the smaller the wavelength of light, the greater the frequency, the greater the energy. Where E is the energy in eV (electron volt), h is Planck's constant, the value of 6.63 × 10-34J.s (Joule. sec), k is the constant, the value is 1.6×10-19J/eV, C is the speed of light, the value is 3×108m/s, λ is the wavelength in nm(nanometers), H is the frequency in Hz(hertz). Reviewing the relevant data [8] provides the following data in Table 2:

Table 2. Wavelengths of different color light

type	Light blue	blue-green	dark Blue	light green	dark green	green	purple
wavelength (nm)	475	500	440	535	515	525	400

①The structure of the index system is constructed, and the results are shown in Table 3

Table 3. Index architecture table

Evaluation target	Evaluation of first-level indicators	Evaluation of secondary indicators
Surface weathering M	Blue A_1	Dark blue B_1
		Light blue B_2
		Blue green B_3
	Green A_2	Blue green B_3
		Dark green B_4
		Light green B_5

②Construct the judgment (pairwise comparison) matrix, hierarchical single ranking and its consistency test

I.Establish the importance degree scale table

The importance between the two elements is quantified to establish the importance scale table,

In order to compare their advantages and disadvantages and their degree.(Table 4.)

Table 4. Important Degree Scale Table

Factor ai than factor bj	Equally important	Slightly important	Strongly important	Intensely important	Extremely important	The median value of two adjacent levels
X_{ij}	1	3	5	7	9	2，4，6，8

II. Establish the judgment matrix, hierarchy single ranking and its consistency test

The stochastic consistency index RI is related to the order of the judgment matrix. Generally, the larger the order of the matrix, the greater the possibility of a random deviation of consistency, and the corresponding relationship is as follows in Table 5 below:

Table 5. Average random consistency index RI standard values

Matrix order	1	2	3	4	5	6	7	8	9	10
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49

Since the deviation of consistency may be caused by random reasons, it is also necessary to compare the CI and the random consistency index RI to obtain the test coefficient CR, so as to test whether the matrix has satisfactory consistency.

The following results are obtained with the matlab run:

1)Evaluation objectives, i. e., surface weathering judgment matrix, Table 6 below: (comparison of relative importance between each evaluation level relative to evaluation objectives)

Table 6. Surface weathering judgment matrix

M	A_1	A_2	W_i	W	$\lambda_{max}=2$
A_1	1	5	0.9806	0.8333	CI=0
A_2	1/5	1	0.1961	0.1667	RI=0
					CR=0

The second-order judgment matrix itself is fully consistent, and the maximum feature root $\lambda_{max}=2=n$ meets the objective condition $\lambda_{max}=2$, CI=0, CI=0, RI=0, CR=0<0.1, through the consistency test.

2)Evaluation first-level index, namely the color judgment matrix: (the comparison of the relative importance of the evaluation second-level index relative to the evaluation first-level index)

Table 7. Blue series judgment matrix

A_1	B_1	B_2	B_3	W_i	W	λ_{max}=3.0092
B_1	1	2	3	0.8468	0.5396	CI=0.0046
B_2	1/2	1	2	0.4660	0.2970	RI=0.58
B_3	1/3	1/2	1	0.2565	0.1634	CR=0.0088

According to Table 7 above, λ_{max}=3.0092, CI=0.0046, RI=0.58, CR=0.0088 <0.81

Table 8. Green series of judgment matrix

A_2	B_3	B_4	B_5	W_i	W	λ_{max}=3.0092
B_3	1	2	3	0.8468	0.5396	CI=0.0046
B_4	1/2	1	2	0.4660	0.2970	RI=0.58
B_5	1/3	1/2	1	0.2565	0.1634	CR=0.0088

According to Table 8 above, λ_{max}=3.0092, CI=0.0046, RI=0.58, CR=0.0088 <0.1, which were tested for consistency.

III. Overall hierarchical ranking and its consistency test

Using the results of the single sorting of all levels in the same level, you can calculate the weight of the importance of all factors at this level in Table 9:

Table 9. Weights of the importance of all factors

administrative levels A / administrative levels B	A_1	A_2	Weight total ordering of hierarchy B relative to hierarchy A
	0.8333	0.1667	
B_1	0.5396		0.4496
B_2	0.2970		0.2475
B_3	0.1634		0.1362
B_3		0.5396	0.0900
B_4		0.2970	0.0495
B_5		0.1634	0.0272
			Total：1.0000

3.1.2 Kendall Coefficient consistency test model: the relation between a single variable and surface weathering of the glass variety

①First, draw the distribution map of glass varieties under weathered and unweathered conditions according to the data, and the results are shown in Figure 2:

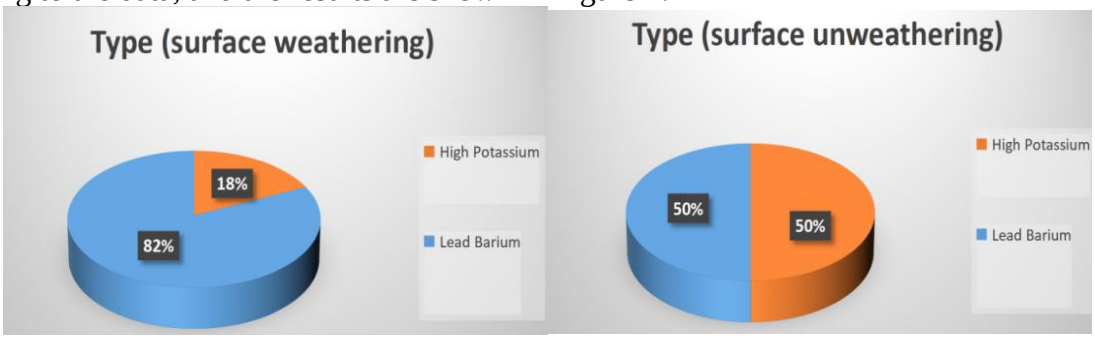

Figure 2. Distribution diagram of glass varieties under weathered and unweathered conditions

②The significant relationship of the statistics was tested by SPSS (Table 10) to determine whether the P-value was significant (P <0.05), if the data was consistent.

Table 10. Analysis of the results

name	Rank average	Results of the Kendall's W analysis median	The Kendall's W coefficient	X^2	P
variety	1.362	0	0.11	6.4	0.011
Surface weathering	1.638	1			

③ Analysis of the analysis results: Kendall coefficient consistency test is to analyze the correlation of the overall (all data): the significance of the overall data P 0.011, significant level significant, reject the null hypothesis, so the data consistency, while the Kendall coordination of the model is 0.11, so the degree of correlation is very low consistency[9].

3.1.3 Chi-square test model: relationship between single variable and surface weathering

①First, draw the distribution map of glass ornamentation under weathered and unweathered conditions according to the data, and the results are shown in Figure 3:

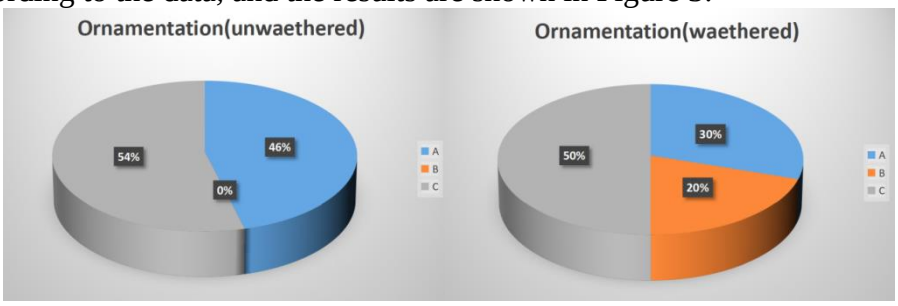

Figure 3. Distribution diagram of glass decoration in weathered and unweathered cases

②Variable X: {ornamentation}; variable Y: {surface weathering}, was calculated by matlab (see Table 11) to analyze whether the model is significant ($P<0.05$). If it is significant, the null hypothesis is rejected, then there are significant differences between the samples.

Table 11. Significance analysis of decorative model

subject	name	ornamentation C	ornamentation A	ornamentation B	total	X^2	P
Surface weathering	unweathered	13	11	0	24		
	weathered	17	11	6	34	4.9565	0.0839
Total		30	22	6	58		

③Analysis of the analysis results:based on the ornamentation and surface weathering, the significant P-value is 0.0839, which does not appear significant on the level of the analysis, and the original hypothesis is accepted, so there is no significant difference between the ornamentation and surface weathering data.

3.2. Classification rule model of ancient glass of known glass varieties

Prior to the ML classification, the preprocessing program extracts features with critical information, such as principal component analysis (PCA), independent component analysis (ICA), linear discriminant analysis (LDA) [10], etc.The basic idea of LDA classification model: assuming

that the sample data of each category meet the Gaussian distribution, the projection with LDA can calculate the mean and variance of the projection data of each category by maximum likelihood estimation, and then obtain the probability density and function of the Gaussian distribution of the category. When a new sample arrives, it can be projected, and then the projected sample features can be brought into the Gaussian distributed probability density function of each category to calculate the probability of belonging to this category. The paper uses positive and negative of the dependent variable of the discriminant function to distinguish lead-barium and high potassium

First, the matlab is used to obtain the W, lambda:W=0.0202,lambda=$1.0376e^{-7}$;

Second, a linear discriminant analysis using spss yields the coefficients of the discriminant function, as shown in Table 12.

Table 12. Coefficients of the discriminant functions of the various classes

type	Non-high potassium
nodal increment; intercept	26.263
silicon dioxide(SiO_2)	-0.15
sodium oxide(Na_2O)	-1.709
Potassium oxide(K_2O)	1.167
Calcium oxide(CaO)	-0.349
Magnesium oxide(MgO)	-0.597
Aluminum oxide(Al_2O_3)	-1.339
Iron oxide(Fe_2O_3)	0.744
copper oxide(CuO)	2.019
Lead oxide(PbO)	-0.678
Barium oxide(BaO)	-1.314
Phosphorus pentoxide(P_2O_5)	-0.456
Strontium oxide(SrO)	-0.441
Tin oxide(SnO_2)	0.599
Sulfur dioxide(SO_2)	-2.827

At the same time, the proportion of sample size participating in the model training was set to 0.7, and the LDA model was evaluated using the test set (Table 13): the accuracy is the proportion of predicted correct samples in the total sample; recall rate: the proportion of predicted positive samples in the results of actual positive samples. The accuracy is the proportion of positive samples that are predicted to be positive. F1 is the harmonic average of the precision rate and the recall rate. The precision rate and the recall rate affect each other. If you need to consider both, then the F1 index can be used.(The larger the value, the better the model)

Table 13. Table of test-set evaluation coefficients for the LDA model

	accuracy rate	Recall Rate	Accuracy rate	F1
training set	1	1	1	1
test set	1	1	1	1

4. Results

According to the analysis, the paper shows that the weight value of the energy strength of different colors of light is the ranking: dark-blue> light-blue> blue-green> dark-green> light-green. Due to the too little data given in the dark blue in the attachment, the conclusions are not universal.The correlation between the color of dark blue relics and its surface weathering, but the remaining data are consistent with the data. Therefore, the surface weathering of glass cultural relics is related to its color. The darker the color, the greater the possibility of surface weathering. The proportion of lead-barium in the variety of weathered cultural relics is about four times that of high potassium varieties, so the glass varieties have a high correlation between lead-barium and weathering, and the correlation between glass varieties and weathering is very low, so the correlation between glass varieties and weathering is low.The influence order of ornamentation type on weathering degree is A> B> C. Meanwhile, there is no significant difference in the data of ornamentation and surface weathering, so there is a high correlation between ornamentation and weathering.

By determining the coefficient of the discriminant function, Finally, a linear classifier based on LDA model was established to represent the classification rules of high potassium glass and lead-barium glass: y=26.263-0.15 silica (SiO_2) -1.709 sodium oxide (Na_2O) + 1.167 potassium oxide (K_2O) -0.349 calcium oxide (CaO) -0.597 magnesium oxide (MgO) -1.339 aluminum oxide (Al_2O_3) + 0.744 iron oxide (Fe_2O_3) + 2.019 copper oxide (CuO) -0. 678 Lead oxide (PbO) -1.314 Barium oxide (BaO) -0.456 Phosphorus pentoxide (P_2O_5) -0.44 Strr 1 (SrO) + 0.599 Tin oxide (SnO_2) -2.827Sulfur dioxide (SO_2). When y is less than zero, the glass variety is lead-barium glass, and when y is greater than zero, the glass variety is high potassium glass.

5. Conclusions

The weathered surface of glass relics has a strong correlation with its color, which is consistent with the protection measures of ancient glass at the museum. Lead-barium glass artifact is more affected by weathering than high potassium glass artifact, which indicates that the identification of lead-barium glass should be more rigorous, so as to avoid incorrect identification because of weather and then accelerate the decay of glass artifacts. Combined with the data feature and the understanding of the advantages and disadvantages of relevant models, the paper puts forward a mathematical model with innovative points suitable for this research. At the same time, the paper makes a breakthrough in the problem of "finding rules" and transforms the qualitative model into a more rigorous and accurate quantitative model. However, due to the limited ability, the theoretical derivation process of the model is not rigorous and detailed.All sufficient conditions of models are not verified one by one, making the model results deviate slightly. For some problems with high dimensions,qualitative analysis should be used in the paper, and it should be combined with quantitative analysis to verify each other more effectively.

Acknowledgements

The authors would like to express gratitude to all those who helped them during the writing of this paper.

References

[1] Wiltshire, R. (2022). ‘WHAT SAID THIS RUDE ANTIQUE’: VICTORIAN RECEPTION OF ROMAN GLASS. Greece & Rome, 69(2), 267-284.

[2] Yun, B., Zhao, Y., & Yaolin, Z. (2022). Topic structure and evolution patterns of documentary heritage preservation and conservation research in china. Library Hi Tech, 40(3), 805-827.

[3] Rehren, T.; Freestone, I.C. Ancient glass: From kaleidoscope to crystal ball. J. Archaeol. Sci. 2015, 56, 233–241. [CrossRef].

[4] Yatsuk, O., Ferretti, M., Gorghinian, A., Fiocco, G., Malagodi, M., Agostino, A., & Gulmini, M. (2022). Data from multiple portable XRF units and their significance for ancient glass studies. Molecules (Basel, Switzerland), 27(18):1-3.

[5] Bertrand, J. W., & Fransoo, J. C. (2002). Operations management research methodologies using quantitative modeling. International Journal of Operations & Production Management, 22(2), 241-264.

[6] Noreldeen, H. A. A., He, S., Huang, K., Zhu, C., Zhou, Q., Peng, H., . . . Chen, W. (2022). Machine learning-based sensor array: Full and reduced fluorescence data for versatile analyte detection based on gold nanocluster as a single probe. Analytical and Bioanalytical Chemistry, 414(29-30), 8365-8378.

[7] Boonruang, C., Wonin, K., & Dararutana, P. (2021). SEM-EDS, PIXE and raman spectroscopies analysis of khlong thom ancient glass bead, southern thailand. Journal of Physics: Conference Series, 1963(1).

[8] Gaudfrin, K., Lopez, J., Gemini, L., Delaigue, M., Hönninger, C., Kling, R., & Duchateau, G. (2022). Fused silica ablation by double ultrashort laser pulses with dual wavelength and variable delays. Optics Express, 30(22), 40120-40135.

[9] Talukdar, S., Pal, S., Shahfahad, Naikoo, M. W., Parvez, A., & Rahman, A. (2023). Trend analysis and forecasting of streamflow using random forest in the punarbhaba river basin. Environmental Monitoring and Assessment, 195(1).

[10] Zhang, R., Han, X., Lei, Z., Jiang, C., Gul, I., Hu, Q., . . . Qin, P. (2022). RCMNet: A deep learning model assists CAR-T therapy for leukemia. Computers in Biology and Medicine, 150.