ISSN 2616-5805 Vol. 4, Issue 5: 1-8, DOI: 10.25236/AJMS.2023.040501

# Multilayer Perceptron Classification Model Based on Weighted Majority Voting for Classifying Ancient Glass Products

**Xinyang Li[1,a,*], Donghao Li[1,b]**

[1]*Department of Mathematics, Henan University of Technology, Zhengzhou, China*
[a]*1849709379@qq.com*, [b]*jiehao1021@163.com*
[*]*Corresponding author*

***Abstract:*** *This study explores the application of weighted majority voting combined with machine learning and artificial neural networks to the categorization of ancient glass artifacts. The study developed an optimized Multilayer Perceptron (MLP) classification model based on the cyclically optimized results from various well-performing classification models through weighted majority voting. The model demonstrated strong performance in cross-validation tests, achieving a prediction accuracy of 92.75%, demonstrating high stability and precision. This research provides a novel approach and methodology for the classification of ancient glass artifacts, potentially contributing to further advancement in this field.*

***Keywords:*** *Multilayer Perceptron, Ancient Glass Artifacts, Machine Learning, Weighted Majority Voting*

## 1. Introduction

China boasts a 2500-year history of utilizing glass, which not only served as a highly sought-after commodity in trade but also acted as a cultural envoy [1]. However, ancient glass was extremely vulnerable to environmental changes when buried and underwent weathering, resulting in significant element exchange between the glass and its surroundings. This process altered their composition ratio and affected accurate typological judgment. Chemical composition of the ancient glasses unearthed from Neimenggu area and Boshan was determined quantitatively by the external-beam PIXE technique in [2]. The research team led by Cui utilized laser ablation inductively coupled plasma mass spectrometry (LA-ICP-MS) to analyze the chemical composition of 11 glass bi disks and 2 glass eyeballs excavated from Chu tombs in the Hunan River Basin. Multivariate statistical analysis revealed that these glass artifacts all belong to the lead-barium-silicate glass system[3]. Fu. utilized a portable energy-dispersive X-ray fluorescence spectrometer (pXRF) and a laser Raman spectrometer (LRS) to conduct chemical composition analysis and micro-area phase analysis of 21 silica-based artifacts from the late Warring States period to the Han dynasty, discovered in Baofeng and Xinzheng, Henan Province[4]. And Huang et.al. conducted non- destructive analysis on 11 decorative silicate beads from the Gansu M4 burial site of the Warring States period using X-ray fluorescence spectroscopy, Raman spectroscopy, and X-ray diffraction analysis[5]. Notably, some researchers recently proposed optimized random forest models that were applied to predict the compressive strength of concrete with good outcomes[6]. Zheng et al. [7] combined the random Forest-multilayer sensing algorithm with laser- induced breakdown spectroscopy (LIBS) to accurately distinguish ginseng from different origins. Lu[8] integrated feature selection methods and classification algorithms from machine learning into the research on chemical composition analysis and category identification of ancient glass artifacts. She constructed an integrated feature selection model for selecting the chemical composition of ancient glass artifacts and a random forest model for identification and classification, with accuracy and AUC as performance metrics for classification.

Based on this, this paper proposes an MLP classification model optimized by the weighted majority voting method. This process involves employing principal component analysis for dimension reduction of variables, followed by the establishment of multiple classifiers. The precision of each classification model is enhanced through k-fold cross-validation. By comparing their performance on existing test and training sets, the Random Forest model emerges as the most prominent one. Ultimately, the categorization of ancient glass artifacts with unknown categories is accomplished using the Random Forest model validated by 5-fold cross-validation. This research holds significant reference and practical implications

ISSN 2616-5805 Vol. 4, Issue 5: 1-8, DOI: 10.25236/AJMS.2023.040501

for ancient glass artifact categorization and specific component analysis.

## 2. Preliminaries

### *2.1. K-fold cross-validation*

The technique of "cross-validation" [9] involves partitioning a dataset into K equally sized and mutually exclusive subsets using stratified sampling. Each subset is used as a test set, while the union of the remaining subsets is used as the training set. This process is repeated K times, and the average of the K test results is obtained. The structure of K-fold cross validation is shown in Fig.1:

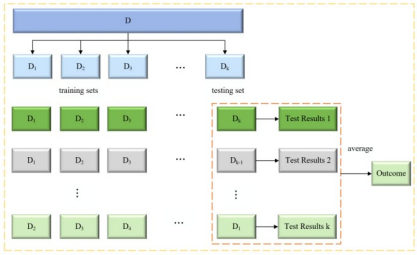


*Figure 1: K-fold cross-validation structure*

### *2.2. Majority Voting*

Majority Voting[10] is an ensemble learning strategy that abides by the principle of majority rules, aiming for lower variance and higher robustness by integrating multiple models to attain high accuracy. It primarily consists of hard voting and soft voting approaches. Unlike hard voting's simple 'majority rules' principle, soft voting considers extra information such as predictive probabilities, which are averaged. In a binary classification problem, if the probability exceeds 50%, the prediction is Category A, or otherwise Category B. This paper adopts the binary weighted soft voting method. The predicted probability of the *i*-th (*i*=1, 2) base learner is represented by matrix $M_i$,

$$M_i = \begin{bmatrix} P(y_i = 1) \\ P(y_i = 2) \end{bmatrix}$$

where $P(y_i = k), k = 1, 2$ refers to the predicted probability of the *k*-th class by the *i*-th base learner. The weight $W_i$ of the *i*-th base learner is $W_i = A_j \Big/ \sum A_j$ ,where $A_j$ is the $F_1$ score of the *j*-th base learner's prediction. The final prediction is

$$\hat{y} = \sum_{i=1}^{m} W_i M_i. \tag{1}$$

Category A if the combined probability $\hat{y}$ exceeds 0.5, or otherwise Category B.

### *2.3. Random Forest Model*

Random Forest[11] has gained widespread attention over the past decade due to its robust classification ability, scalability, and ease of use. The process involves bootstrap sampling to select n samples randomly for training. A decision tree Di is then constructed, and the process is repeated multiple times. The categories predicted by each tree are summarized through majority voting. And the flow chart is shown in Fig 2 below.

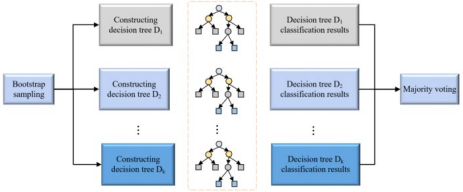


*Figure 2: Flow chart of random forest*

ISSN 2616-5805 Vol. 4, Issue 5: 1-8, DOI: 10.25236/AJMS.2023.040501

### *2.4. AdaBoost Model*

AdaBoost[12] is a quintessential boosting algorithm that adjusts the distribution of training samples by increasing the weights of samples with high prediction errors. It combines all the base learners in a weighted manner, increasing the weights of base learners with good prediction performance and reducing those with inferior accuracy. And the flow chart is shown in Fig 3 below.

### *2.5. XGBoost Model*

Extreme Gradient Boosting (XGBoost)[13] is also a boosting algorithm and an improvement of GBDT. By expanding the GBDT loss function to a second-order Taylor series and adding regularization penalties to the objective function of GBDT, XGBoost optimizes and improves the performance, effectively preventing overfitting. It also takes advantage of random forest algorithms to support dataset sampling. The XGBoost algorithm incrementally adds a tree in each iteration, constructing a linear combination of K trees

$$\hat{y}_i^{(t)} = \sum_{k=1}^{K} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i), \qquad f_k \in F,$$

where the function space F encompasses all trees, while $f_k(x_i)$ denotes the weight assigned to the *i*-th sample classified into the leaf node in the *k*-th tree.

### *2.6. KNN Model*

The KNN (K-Nearest Neighbor)[14] algorithm posits that if most samples in a feature space belong to a certain category, then the k most similar samples in this space belong to the same category. By selecting k neighbors and a distance measurement method, the *k* closest neighbors to the sample to be classified are found, followed by majority voting based on the labels.

### *2.7. SVM Model*

The Support Vector Machine (SVM)[15] is a data mining method rooted in statistical learning theory that can effectively handle regression problems, pattern recognition, and other tasks. SVM finds an optimal classification hyperplane that best meets the classification requirements while maximizing the empty areas on both sides of the hyperplane. The optimal classification function is

$$f(x) = \operatorname{sgn}\left\{(w^* \cdot x) + b^*\right\} = \operatorname{sgn}\left\{\left(\sum_{j=1}^{l} a_j^* y_j (x_j \cdot x_i)\right) + b^*\right\}, \qquad x \in R^n,$$

where w* is the optimal weight vector and b* is the optimal bias.

### *2.8. MLP Model*

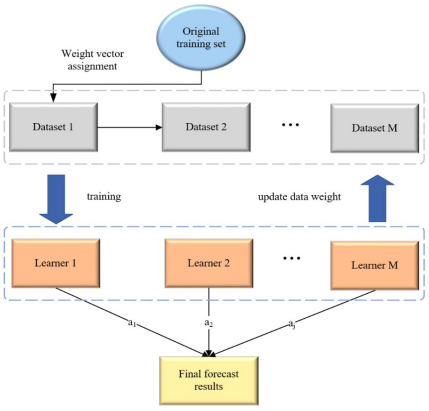


*Figure 3: Flowchart of AdaBoost algorithm*

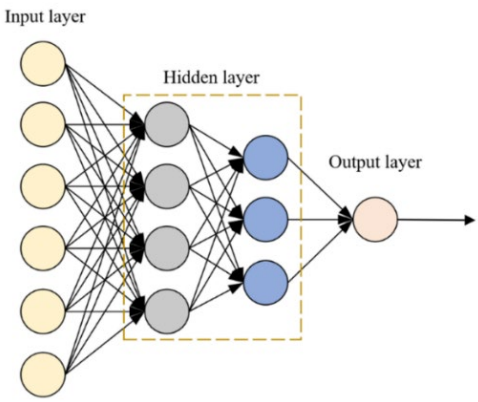


*Figure 4: The structure diagram of MLP model*

The MLP (Multilayer Perceptron)[16] network comprises three layers: the Sensory (S), Association (A), and Response (R). Each layer consists of neurons of the same type. The S layer forms the input layer

ISSN 2616-5805 Vol. 4, Issue 5: 1-8, DOI: 10.25236/AJMS.2023.040501

for feature vector input, the A layer is the hidden layer, and the R layer is the output layer. The relationship between the S and A layers forms an association matrix for processing objects. The connection between A layer units and R layer units forms the decision matrix for processing objects. Through training and organization, the network forms an ordered and stable structure with decision-making capabilities. The structure diagram is shown Fig. 4.

## 3. Optimized MLP Classification Model

This paper proposes an optimized MLP classification model based on multi-model weighted majority voting. The main components of the model include: model construction, weighted majority voting, extraction of differential results, multiple model training, and result output.

1) Model Construction: Initially, the available data is divided into a 7:3 ratio for training and testing sets. Next, the MLP, Random Forest, AdaBoost, and XGBoost models are trained using the training set, and their results are denoted as mlp_result, random_forest_result, AdaBoost_result, and XGBoost_result, respectively.

2) Weighted Majority Voting: The results of the three models mentioned above are combined using weighted majority voting. The classification results with votes greater than 2 are considered as the final outcome, labeled as vote_result.

3) Extraction of Differential Results: The mlp_result array is compared with the vote_result array, and items with differences form a new array called difference_result.

4) Multiple Model Training: The difference_result array is fed back into the MLP classification model that was trained using the training set. The dataset is predicted again, and the results are recorded as second_result. This process is repeated by comparing second_result with vote_result (step 3) until there are no differential items.

The overall algorithm is depicted as follows:

| **Algorithm 1 Optimized MLP model** |
|---|
| Input: Model classification results: random_forest_result, AdaBoost_result,<br>XGBoost_result, mlp_result,<br>MLP model trained on the training set: MLP(x),<br>Original dataset: M |
| Initialization: Weighted majority voting result array: vote_result,<br>Difference corresponding sequence values: difference_index |
| Procedure: |
| 1:for t in len(random_forest_result) do |
| 2: if random_forest_result[t] + AdaBoost_result[t]+XGBoost_result[t] ≥ 2 then |
| 3: Add vote_result[t] = 1 |
| 4: else |
| 5: Add vote_result[t] = 0 |
| 6: end if |
| 7:end for |
| 8:for m in len(vote_result) do |
| 9: if vote_result[m]= mlp_result[m] then |
| Entry its sequence value into difference_ index |
| 10: end if |
| 11:end for |
| 12:for j in difference_index do |
| 13: Get the subset M[j] from M, forming difference_result |
| 14: Predict MLP(difference_result) once again, and the result is used as the new mlp_result |
| 15:end for |
| 16: The prediction results repeat the above process again until there is no difference between vote_result and mlp_result |

The structure description is shown in Figure 5:

ISSN 2616-5805 Vol. 4, Issue 5: 1-8, DOI: 10.25236/AJMS.2023.040501

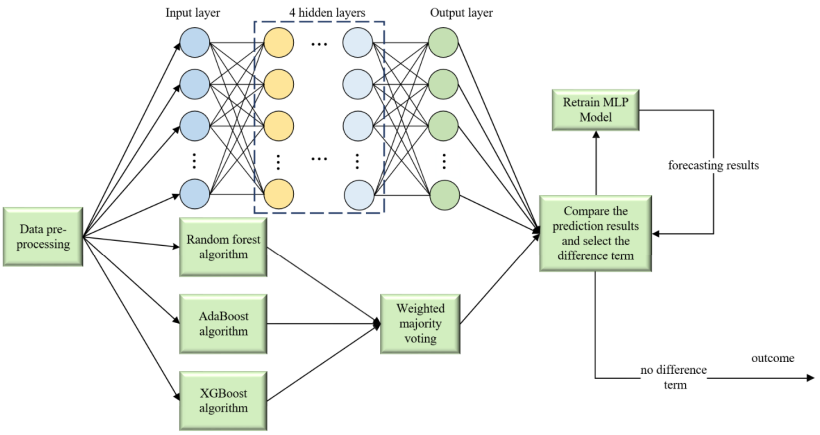


*Figure 5: The flow chart of MLP model optimization*

The flowchart illustrates that the proposed optimization method involves weighted majority voting on the prediction results of three well-performing classification models. It continually compares the weighted majority voting results with MLP's prediction results, extracts differential items, and re-trains the MLP model with these items, gradually reducing their difference to 0. This process ultimately improves the accuracy of the MLP model's results.

## 4. Experiment

### *4.1. Data Source and Preprocessing*

The data for this study was obtained from the official website of the 2022 National College Student Mathematical Modeling Competition, providing relevant data on ancient glass products in China. Archaeologists classified the samples into high-potassium glass and lead-barium glass based on their chemical composition and other detection methods. However, due to reasons such as detection methods, the sum of the component proportions is not always 100%. In this study, two groups of data with incomplete sums were removed, leaving 69 valid data sets for research.

### *4.2. Experimental Analysis*

Before model establishment, as the dataset has a high-dimensional space, this paper first conducted dimensionality reduction using principal component analysis and variable selection based on correlation coefficient.

1) Principal Component Analysis (PCA): SPSS software was used to perform principal component analysis on the dataset, calculating eigenvalues and cumulative variance explained by each variable. The important variables for modeling were selected based on their contribution rates to achieve dimensionality reduction. Fig. 6 shows the cumulative contribution rate curve of the ancient glass dataset, with components 1 to 14.

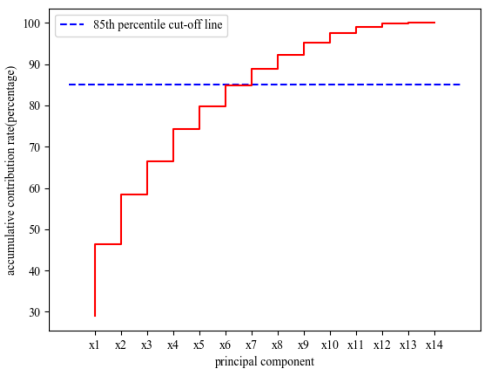


*Figure 6: Ladder diagram of cumulative variance*

ISSN 2616-5805 Vol. 4, Issue 5: 1-8, DOI: 10.25236/AJMS.2023.040501


*Figure 7: Heat map of correlation coefficient contribution rate of each principal component*

The 85% threshold line indicates that the first 8 principal components account for more than 85% of the cumulative contribution rate and are considered important for modeling.

2) Variable Selection Based on Correlation Coefficients: The Python "corr" function was used to calculate the Pearson correlation coefficients between variables. Variables with strong correlations were filtered, and the remaining variables were used for modeling. Fig. 7 is a heatmap showing the correlation coefficients between variables in the ancient glass dataset.

The color intensity represents the magnitude of the correlation coefficient. Based on the heatmap, the paper chose to discard four variables (potassium oxide, lead oxide, copper oxide, and sulfur dioxide) due to their strong correlations with other variables, using the remaining variables as the basis for modeling.

An MLP model with four hidden layers was established using the MLP Classifier module from the sklearn package in Python. Each hidden layer had 100 neurons, and the activation function for all layers was set to Sigmoid. The model was trained and tested using the data processed through the two dimensionality reduction methods and the original data.

The results of 15 predictions using the two dimensionality reduction methods and the raw data were compared in terms of Accuracy values, as shown in Fig. 8.

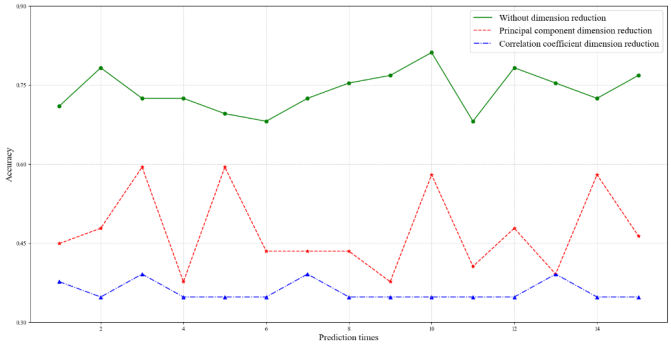


*Figure 8: Comparison of Accuracy values of dimensionality reduced and non-dimensionality reduced models*

The experimental results indicated that the models obtained through the two dimensionality reduction methods had higher prediction errors compared to the direct modeling analysis. Therefore, the dimensionality reduction process was removed, and the optimization of model establishment was carried out directly.

Using the sklearn package in Python, five models (Random Forest, AdaBoost, XGBoost, SVM, KNN) were implemented, and then an MLP model with four hidden layers and a Sigmoid activation function was established. The dataset was split into a 7:3 ratio for training and testing the models. All models achieved 100% prediction accuracy on the training set. The average performance of the models on the testing set was recorded and presented in Table 1.

The experimental data showed that the MLP model among all models had weaker performance on the testing set and required optimization for classification prediction. However, all classification models exhibited overfitting tendencies, so 5-fold cross-validation was used in subsequent experiments to reduce overfitting risk.

ISSN 2616-5805 Vol. 4, Issue 5: 1-8, DOI: 10.25236/AJMS.2023.040501

*Table 1: Comparison of evaluation index of preliminary classification results of each model*

| Model | Accuracy | Recall | Precision | F1 |
|---|---|---|---|---|
| SVM | 0.63 | 0.75 | 0.65 | 0.70 |
| MLP | 0.65 | 0.61 | 0.86 | 0.71 |
| KNN | 0.74 | 0.63 | 1.00 | 0.78 |
| AdaBoost | 0.71 | 0.61 | **1.00** | 0.76 |
| XGBoost | **0.80** | 0.75 | 0.88 | 0.81 |
| Random forest | 0.71 | **1.00** | 0.71 | **0.83** |

Based on the performance of each model on the testing set, three models (Random Forest, AdaBoost, XGBoost) with better performance were selected for optimizing the MLP model.

The predict_proba function of each model was used to predict class probabilities for the dataset. Then, the Accuracy values of each model were calculated, and the weighted weights of the three models were determined. The weighted majority voting prediction results were obtained using the formula (1), and these results were compared with the MLP model's prediction results. The differential items were selected for retraining the MLP model until there were no differential items in the prediction results of both models. The $F_1$ values after each optimization cycle of the MLP model were shown in Fig. 9.

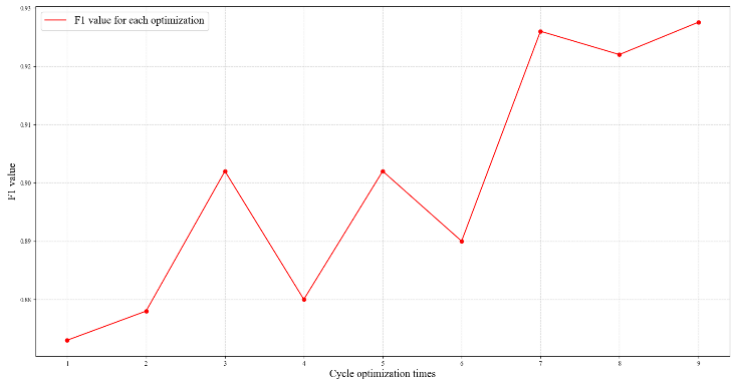


*Figure 9: F1 value of loop optimization*

After multiple cycles of training optimization, the original dataset was mixed with unknown data, and the MLP model was used to predict multiple times, and the average value was taken. The predicted results were compared with the official results announced by the competition to verify the accuracy and practical significance of the model. The comparison of classification results' errors between the optimized and unoptimized MLP models is shown in Fig. 10.

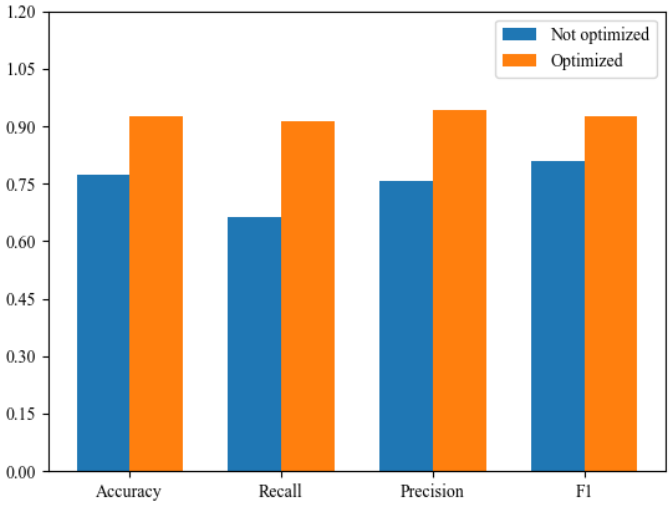


*Figure 10: Comparison of MLP model performance before and after optimization*

The experimental results showed that the optimized MLP model had significant improvements in all evaluation metrics, achieving a prediction accuracy of 92.75%.

ISSN 2616-5805 Vol. 4, Issue 5: 1-8, DOI: 10.25236/AJMS.2023.040501

## 5. Conclusion

Based on existing data on ancient glassware, this study first conducted principal component analysis and variable selection through correlation coefficients to reduce dimensionality, but it was found to have a reverse effect on the results. Secondly, by comparing the performance of various classifiers on a 5-fold cross-validation set, it was found that the MLP performed poorly in all evaluation metrics. Therefore, multiple classification models with better experimental results were used for weighted majority voting to iteratively optimize the initial MLP model. The final result is an optimized MLP model based on weighted majority voting. This research has practical significance and effective value for the study of the composition of ancient glassware materials and the identification of their quality categories in China.

## References

*[1] Wang C., Tao Y. The weathering of silicate glasses. Journal of Chinese Ceramic Society, 2003, 31(1):78-85.*

*[2] Li, F., Li, Q., Gan, F., Zhang, B., Cheng, H. Chemical composition analysis for some ancient chinese glasses by proton induced X-ray emission technique. Journal of Chinese Ceramic Society, 2005, 33, 581-586.*

*[3] Cui J., Wu X., Tan Y., et al. Chemical analysis of ancient glass wares unearthed from chu cemeteries of the warring state period in the drainage area of the Yuanshui river, Hunan province. Journal of Chinese Ceramic Society, 2009, 37(11): 1909-1913+1918.*

*[4] Fu Q., Zhao H., Dong J., et.al. Nondestructive analysis of the silicate artifacts excavated from Baofeng and Xinzheng, Henan province. Spectroscopy and Spectral Analysis, 2014, 34(01): 257-262.*

*[5] Huang X., Yan J., Wang H. Analysis of the decorated silicate beads excavated from tomb M4 of the Ma-Jia-Yuan warring states cemetery, Gansu province. Spectroscopy and Spectral Analysis, 2015, 35(10): 2895-2900.*

*[6] Rajakarunakaran S., Lourdu A., Lourdu A., et.al. Prediction of strength and analysis in self-compacting concrete using machine learning based regression techniques. Advances in Engineering Software, 2022, 173: 103267.*

*[7] Zheng K., Li X., Song S., Gao X. Discrimination of ginseng origin by using laser-induced breakdown spectrum and machine learning algorithms. Microwave and Optical Technology Letters, 2022, 65(5): 1248-1254.*

*[8] Lu J. Classifying model of ancient glass products based on ensemble feature selection and random forest. Journal of Chinese Ceramic Society, 2023, 51(04): 1060-1065.*

*[9] Tang H., Xu Y., Lin A., et.al. Predicting green consumption behaviors of students using efficient firefly grey wolf-assisted K-nearest neighbor classifiers. IEEE Access 2020, 8, 35546-35562.*

*[10] Chandra, T., Verma K., Singh B. Coronavirus disease (COVID-19) detection in Chest X-Ray images using majority voting based classifier ensemble. Expert Systems with Applications, 2021, 165: 113909.*

*[11] Chen Y., Zheng W., Li W., Huang Y. Large group activity security risk assessment and risk early warning based on random forest algorithm. Pattern Recognition Letters. 2021, 144: 1-5.*

*[12] Wu Y., Ke Y., Chen, Z., Liang S., et.al. Application of alternating decision tree with AdaBoost and bagging ensembles for landslide susceptibility mapping. Catena, 2020, 187: 104396.*

*[13] M. Amjad, I. Ahmad, M. Ahmad, et al., Prediction of pile bearing capacity using XGBoost algorithm: modeling and performance evaluation, Appl. Sci. 2022, 12 (4): e01243.*

*[14] Gul A., Perperoglou A., Khan Z. et.al. Ensemble of a subset of kNN classifiers. Adv. Data Anal. Classif. 2018, 12, 827-840.*

*[15] Jiang Y., Xie J., Han Z. et.al. Immunomarker support vector machine classifier for prediction of gastric cancer survival and adjuvant chemotherapeutic benefit immunomarker SVM-based predictive classifier. Clinical Cancer Research. 2018, 24, 5574-5584.*

*[16] Mohammadi B., Guan, Y., Moazenzadeh R., Safari M. Implementation of hybrid particle swarm optimization-differential evolution algorithms coupled with multi-layer perceptron for suspended sediment load estimation. Catena, 2021, 198: 105024.*