Glass classification and identification based on $mRMR$-random forest algorithm

Nian Li, Zhuoyuan Chen *, Yalei Jian

School of Mechanical Engineering and Automation, Dalian Polytechnic University, Liaoning, China, 116034

* Corresponding Author Email: 3392837972@qq.com

Abstract. With the increasing importance of conservation and identification of antiquities, there are numerous methods for the identification of antiquities. In order to realise the analysis and identification of the components of ancient glass objects, this paper establishes a random forest identification model based on the original data, random forest feature extraction and the maximum correlation minimum redundancy algorithm (mRMR), and compares and analyses its effectiveness on various datasets. The mrmr algorithm-based random forest discrimination model achieved a combined accuracy and ROC curve coverage area (AUC value) of 1 on the unweathered dataset, which was close to a perfect model with high confidence in predicting the discrimination results; while the AUC value of the repeatedly trained model was not stable enough on the weathered dataset. Thus, this paper proposes a mrmr-based random forest-based discrimination model, which achieves an optimal AUC value of 0.9393 when the number of decision trees is 30 and the maximum number of splits is 10.

Keywords: Random Forest, Minimum redundancy, algorithm Feature Extraction.

1. Introduction

Ancient glass was a new material synthesised for the first time in the true sense of the word by man, and because the raw glass was so different from the finished glass, the glass process carried an extremely high technical and social significance. Our earliest glassware began at the end of the Spring and Autumn Period and the beginning of the Warring States period. The glass of that time was of a single variety and extremely similar in appearance, but the chemical composition of it was very different [1]. The analysis and identification of the composition of ancient glass objects is particularly important, and the analysis of their composition can be used to obtain the relationship between various variables, which can be used to obtain the relationship between the various basic information of glass artefacts.

2. Modelling glass classification

2.1. Random Forest Model Fundamentals

In 2001 Brieman and Cutler borrowed from the stochastic decision forest method to combine classification trees into a random forest, the Random Forest integrated learning algorithm (RF). The specific steps are as follows [2]. The random forest flow chart is shown in Figure 1:

a) Randomly select k variables from the collected dataset, a total of m variables (where k is less than or equal to m), and then build a decision tree based on these k variables.

b) Repeat the above process n times so that n different decision trees are constructed from these k variables in different random combinations.

c) For each decision tree predict the outcome using random variables and record all the predicted outcomes, which will result in n outcomes from the n decision trees.

d) Calculate the number of votes each predicted outcome receives and then choose the mode, i.e. the predicted outcome with the highest number of votes is taken as the final predicted outcome of the random forest algorithm [3][4].

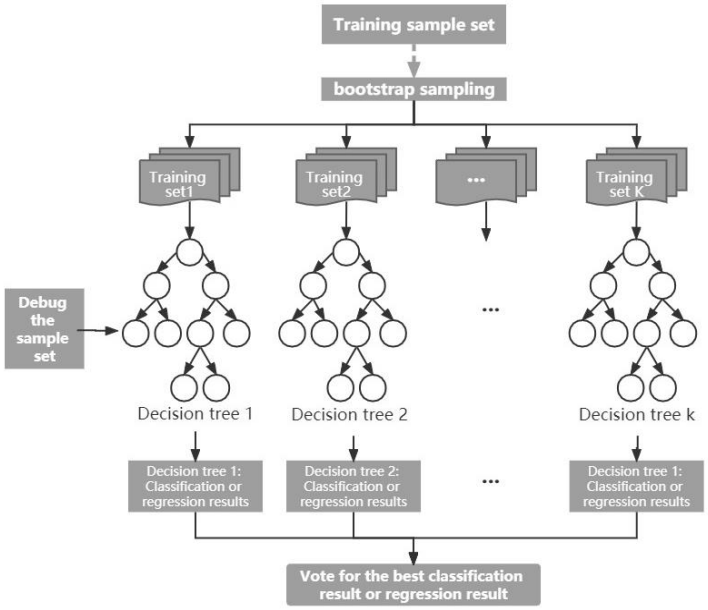


Figure 1. Random Forest Flowchart

2.2. Model evaluation indicators

Evaluation metrics are the key to measuring how good a model is, and the evaluation metrics used in this paper are Accuracy, Micro F1 score and AUC value [5].

Accuracy: The percentage of samples that are correctly classified for the entire sample space.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{1}$$

F1 Score: A measure in classification problems that weighs Precision and Recall and is defined as the summed average of Precision and Recall.

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

$$F1 = 2 * \frac{Precision*Recall}{Precision+Recall} \tag{4}$$

Micro-F1: the metric used to evaluate multi-category problems is the result of the F1 merge.

$$Precision_{micro} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i+\sum_{i=1}^{n} FP_i} \tag{5}$$

$$Recall_{micro} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i+\sum_{i=1}^{n} FP_i} \tag{6}$$

$$\mathrm{F1}_{\mathrm{micro}} = 2 * \frac{\mathrm{Precision}_{\mathrm{micro}}*\mathrm{Recall}_{\mathrm{micro}}}{\mathrm{Precision}_{\mathrm{micro}}+\mathrm{Recall}_{\mathrm{micro}}} \tag{7}$$

AUC value: the area covered by the ROC curve, which can take into account a combination of Recall, Precision and Accuracy metrics.

3. Modelling glass identification

3.1. Random forest based on raw data

As this question only needs to explore the classification pattern of high potassium glass and lead-barium glass, in line with the principle of controlling variables, we first split the data of each type of glass chemical composition in Annex Form 2 into two data sets according to the two types of weathering and unweathering, and then trained all 14 chemical compositions as sample data to

explore the effect of the random forest model identification. The ROC curves of the model for the weathered and unweathered datasets are shown in Figures 2 and 3 below

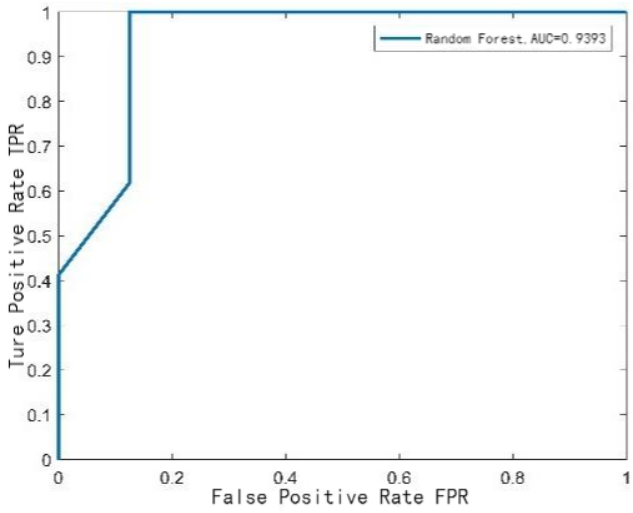


Figure 2. Weathering data set

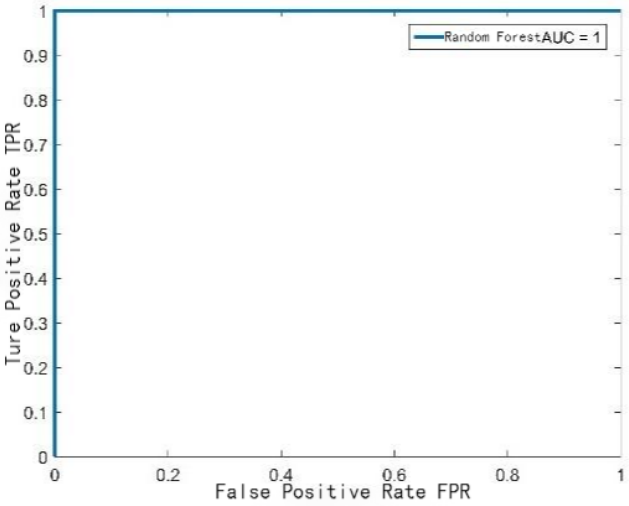


Figure 3. Unweathered data set

Meanwhile, the corresponding AUC and MicroF1 scores for the two datasets are shown in the table 1 below:

Table 1. AUC and MicroF1 scores corresponding to the two datasets

	Weathering data set	Unweathered data set
$Micro_f1_score$	0.9524	1
AUC	0.9393	1

It can be seen that the random forest performs perfectly on classification, especially for the weathered data set, and is very close to a perfect classifier, indicating that this integrated learning algorithm has excellent performance on this data set.

3.2. Random forest based on feature selection

Although the random forest model based on the original data has achieved better results, there are 14 indicators in the data set due to the presence of only 70 sample sizes. On the one hand, the number of indicators is too large and the distribution of samples in space may show sparsity, resulting in dimensional disasters [6]; on the other hand, too many indicators will increase the difficulty of the learning task and increase the training time. Thus, this paper proposes a random forest identification

model based on random forest feature extraction and feature selection based on the maximum correlation minimum redundancy algorithm (mRMR), and compares and analyses the advantages and disadvantages of the two feature selection methods.

3.3. Random Forest based on Random Forest feature extraction

An important feature of random forests is the ability to evaluate the importance of a feature variable [7]. The basic idea is that when a feature variable is added to a noise value, a significant decrease in the prediction accuracy (Gini value) of RF indicates that the feature variable is more important, and the specific calculation lesson is described as follows.

(1) The performance of the generated random forest is first tested with data on the chemical composition content of the glass to obtain a Gini value.

(2) A noise value is then artificially added to that data for a characteristic variable v. The performance of the random forest is then tested with the added noise data to obtain a new Gini value.

(3) The difference between the original Gini value and the noise added Gini value is used as the importance measure for the corresponding feature variable v. [8]

With the introduction of random forest feature extraction, the selection of features for the weathered and unweathered datasets is shown in Figure 4 below:

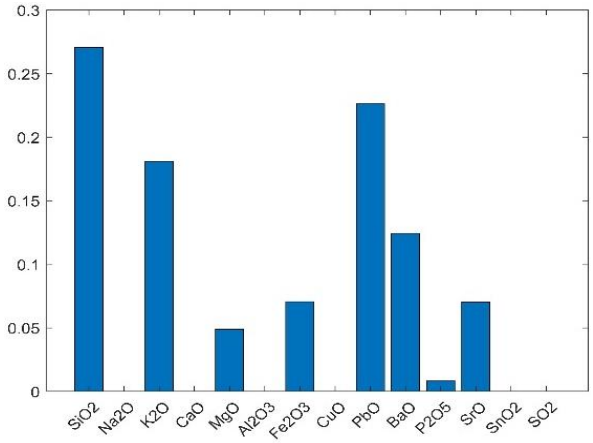


Figure 4. Random Forest based feature extraction

It can be seen that $SiO_2.K_2O.PbO$ and BaO These four chemical components have the highest importance to the random forest model, so, after excluding the data of the remaining chemical components, we performed random forest integration learning on the new data set. The ROC curves for the weathered and unweathered datasets were obtained as shown in Fig. 5.6 below.

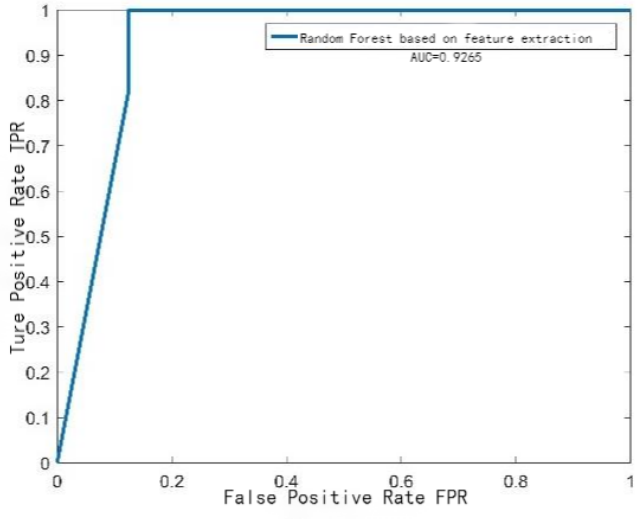


Figure 5. Weathering dataset ROC curves

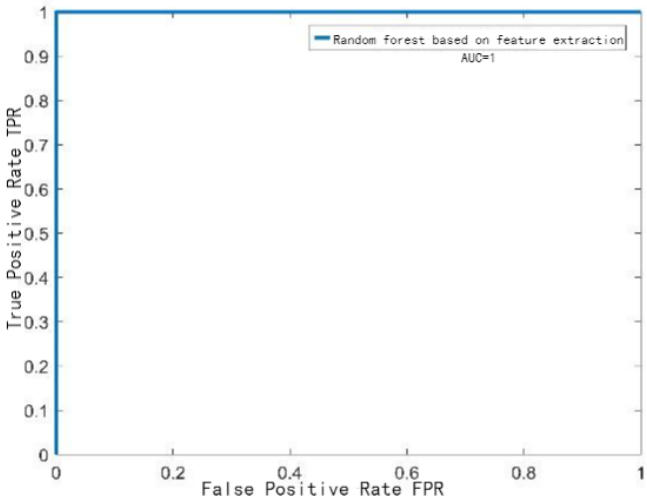


Figure 6. ROC curves for the unweathered dataset

Their classification results are compared in the following table 2.

Table 2. Classification results

	Weathering Original	Unweathering Original	Weathering Feature extraction	Uneathering Feature extraction
$Micro_f1_score$	0.9524	1	0.9762	1
AUC	0.9393	1	0.9265	1

Compared with the random forest based on the original weathered data, although the AUC value of the model based on random forest feature extraction slightly decreases, the micro F1 score of the model has a larger improvement, while fewer feature metrics can make the integrated learning more efficient, proving that the optimization idea of feature extraction is feasible.

3.3.1 Random forest based on mRMR feature selection-Principle of the mRMR algorithm

The maximum correlation minimum redundancy (mRMR) algorithm is a mutual information based heuristic feature selection algorithm that calculates the correlation between features and attributes based on an evaluation function, ranks the original features and obtains a feature set with high correlation and few redundant features [9]. mRMR algorithm's ideas can be summarised in two points.

(1) The first idea of the mRMR algorithm is to extract a subset S of features with m features that are highly correlated with the category that will be output. The maximum correlation ($Max-Relevance$) is achieved by using the average of the mutual information of all features $\{x_t\}$ in the same category. The mutual information MI is defined as shown in Eq. 8.

$$I(x,y)=\iint p(x,y)log\frac{p(x,y)}{p(x)p(y)}dxdy \tag{8}$$

For the variables, the value of m is calculated as in Eq. 9 shown in.

$$max_{t\epsilon S}D(S)\ ,D=\frac{1}{|S|}\sum_{t\epsilon S}I(Y,x_t) \tag{9}$$

where $Y=[y^{(1)},y^{(2)},\dots,y^{(n)}]^T$ is denoted as the column vector of label features in the set of instances and $y^{(i)}$ is the label that makes the ith instance in the set of instances. Searching for features that satisfy the criterion of having maximum relevance, the larger the value of $Max-Relevance$, the higher the sign of relevance.

(2) The second idea of the $mRMR$ algorithm is that the criterion of maximum relevance of features for class variables should be complemented by minimum redundancy ($Min-Redundancy$) between features [10]. If only correlation is achieved, there is a high risk of increasing the dependencies between features. Therefore, the minimum redundancy should be achieved by

appropriately removing a particular redundant feature that has little impact on the classification result without affecting its relevance. The correlation is calculated as shown in Eq.10, where $I(x_k, x_t)$ denotes the mutual information between feature k and feature t.Combining max and min to form the mrmr algorithm, as shown in Eq. 11.

$$min_{t\epsilon S} R(S), R = \frac{1}{|S|^2} \sum_{k.t\epsilon S} I(x_t, y_t) \tag{10}$$

$$max\varphi(D, R), \varphi = D - R \tag{11}$$

3.3.2 Random forest based on $mRMR$ feature selection-Develop a model

This paper uses the $mRMR$ algorithm for feature selection on both weathered and unweathered glass datasets. The results are shown in Figures 7 and 8 for the weathered data set:

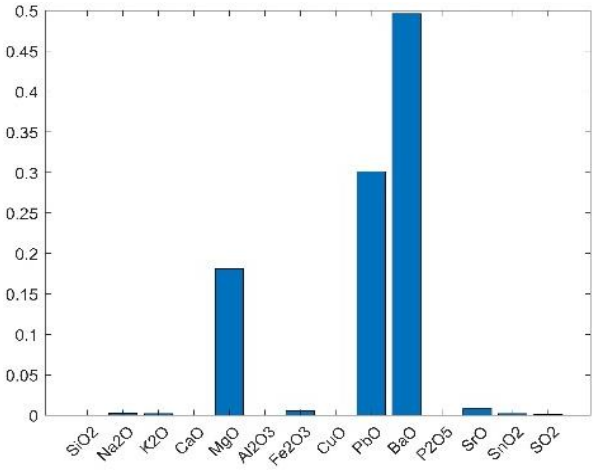


Figure 7. Weathering-$mRMR$

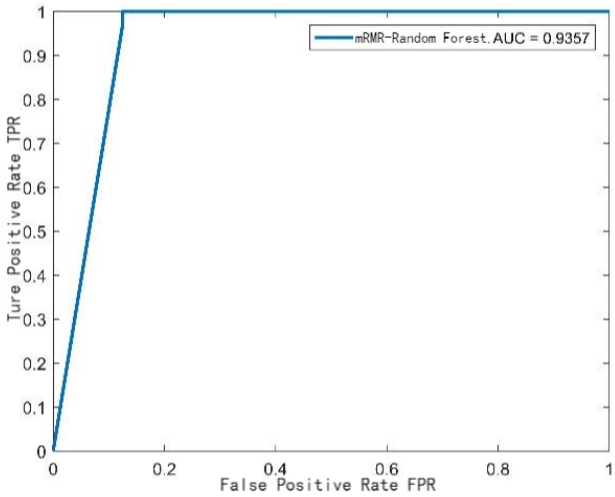


Figure 8. Weathering $mRMR$-random forest ROC curve

It can be seen that for the weathered glass data set,PbO and BaOthese two chemical components have the highest importance. So, after eliminating the rest of the chemical composition data, we perform random forest integration learning on the new data set.

The features of the unweathered data set were selected as the most important feature indicators only using the algorithm as above. The results are shown in Figures 9 and 10 for the unweathered data set:

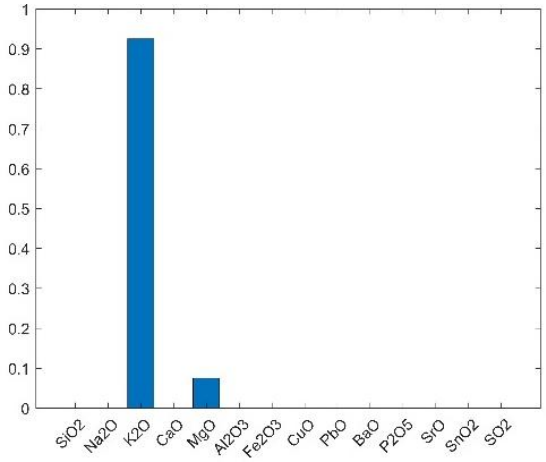


Figure 9. Unweathering-$mRMR$

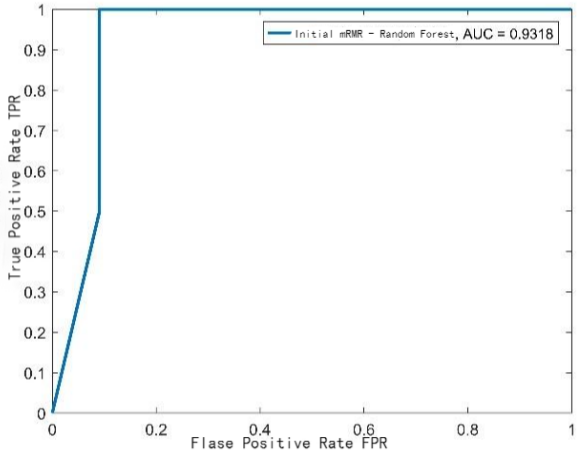


Figure 10. Unweathering $mRMR$-random forest ROC curve

A comparison of their classification results is shown in Table 3 below:

Table 3. Classification results

	Weathering Feature extraction	Unweathering Feature extraction	Weathering $mRMR$	Unweathering $mRMR$
$Micro_f1_score$	0.9762	1	0.9762	0.9600
AUC	0.9265	1	0.9357	0.9318

By comparing the random forest models based on feature extraction, the $mRMR$-based random forest had higher AUC values and better models on the weathered data set of glass; however, the $mRMR$-based random forest AUC values decreased much lower on the unweathered data set of glass instead, indicating that a part of the classification of the original data set was lost when only K_2O was extracted as the data set for integrated learning features. So, based on this, we combined the $mRMR$ feature selection of the glass weathered unweathered data set and combined the data for the three chemical components $PbO.BaO$ and K_2O into a new data set, again using Random Forest for integration learning. The ROC curve is obtained in Figure 11:

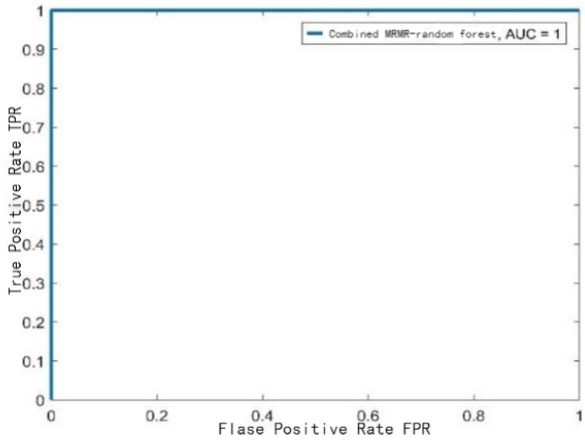


Figure 11. Unweathered - combined with $mRMR$ -Random forest ROC curve

It can be seen that the random forest model combined with $mRMR$ feature selection achieves an AUC value of 1 on the unweathered data set, which is very close to a perfect model, and the feature metrics are more refined compared to those extracted based on random forest features, and the model is more efficiently integrated for learning.

3.4. Random forest based on grid search for hyperparameter optimization

Although the random forest model was optimised using feature selection, the random forest model based on mrmr feature selection achieved even better results in terms of accuracy and efficiency. However, the AUC value of the random forest model based on mrmr feature selection was still slightly lower than the original model on the weathered dataset of glass [11]. Thus, after optimising the feature metrics, we focused on optimising the hyperparameters of the random forest.

3.4.1 Grid search method

The grid search method is an exhaustive search algorithm for parameter optimisation that is general and simple to use [12]. The essence of the method is to divide the parameter space into a number of grids and to optimise the model to be trained by traversing all the combinations of parameters at the intersections of the grid, while calculating the root mean square error or accuracy of the corresponding model. Only by traversing all the nodes in the grid plane can the combination of parameters with the lowest root mean square error or highest accuracy be obtained, i.e. the optimal combination of parameters for the model.[13]

3.4.2 Improving random forests based on grid search

In order to optimize the AUC value of the random forest model based on $mRMR$ feature selection, so that it can find the optimal solution of the model more quickly, the $K-fold$ cross validation and grid search method are combined to improve the random forest model, and the average root mean square error of the model is minimized as the goal of the optimization parameters, and the two parameters of the random forest, the number of decision trees k and the maximum number of splits m, which are more important, are adjusted and optimized. The model is optimised for the more important parameters of the random forest, the number of decision trees k and the maximum number of splits m [14]. The resulting is guaranteed to be the best combination of parameters over the full range of the grid.The specific steps are shown in Figure 12 below[15].

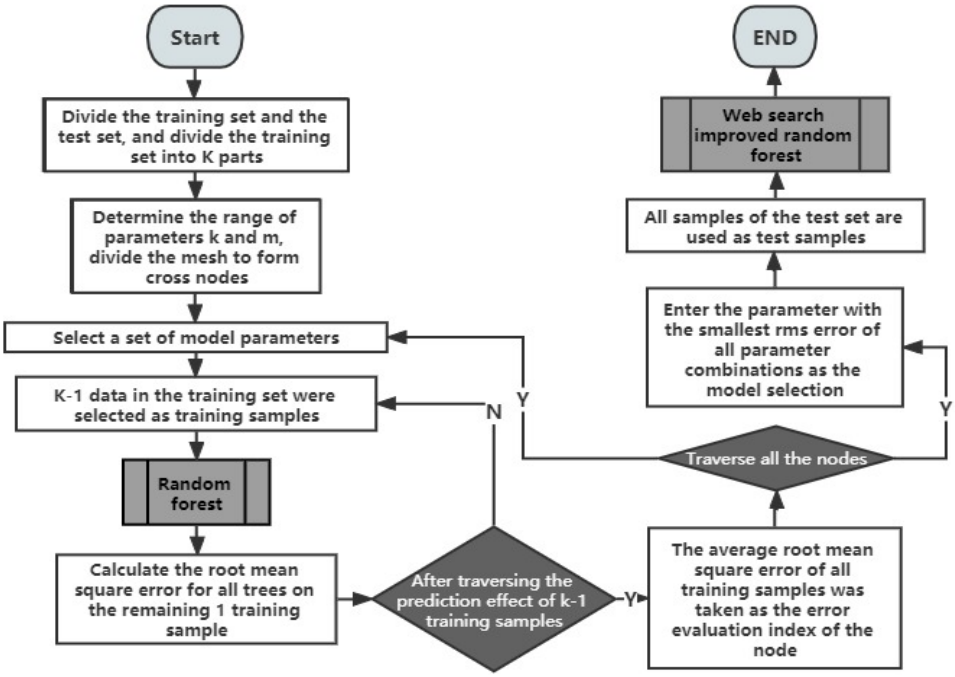


Figure 12. Flow chart of improved random forest based on grid search

Following the above steps, on the weathering data set of glass, We used grid search to refine the - based hyperparameters of the random forest model and performed feature selection to obtain AUC heat maps for different parameter combinations, see Figure 13.

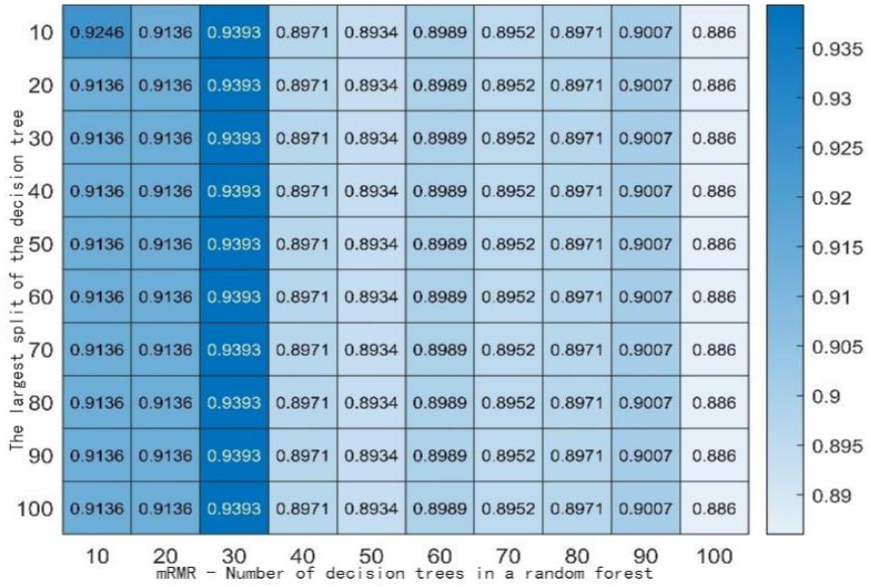


Figure 13. AUC heat map for different combinations of hyperparameters

4. Results

It can be seen that the AUC reaches the optimal value of 0.9393 when the maximum number of splitting of decision trees is chosen as 10 and the number of decision trees is chosen as 30. Therefore, on the weathered data set of glass, we use the $mRMR$ random forest model with 30 decision trees and the maximum number of splitting of each decision tree as 10. Based on this it is possible to optimise the hyperparameters of $mRMR$ based on network search Feature extraction random forest prediction yields artefact categories. The final identification results obtained are shown in Table 4 below

Table 4. Categories for which heritage predictions were obtained

Heritage number	Forecast Category	Heritage number	Forecast Category
A1	High Potassium Glass	A5	Barium lead glass
A2	Barium lead glass	A6	High Potassium Glass
A3	Barium lead glass	A7	High Potassium Glass
A4	Barium lead glass	A8	Barium lead glass

5. Conclusions

In this paper, we compare two algorithms based on random forest feature extraction and the maximum correlation minimum redundancy algorithm ($mRMR$) based algorithm and their effectiveness on each dataset. The combined accuracy and ROC curve coverage area (AUC value) of the random forest discrimination model based on the $mRMR$ algorithm reached 1 on the unweathered dataset, which is close to the perfect model, and the confidence of the prediction discrimination results is high. This paper proposes a random forest discrimination model based on mrmr, in order to solve the situation that the AUC value of the multiple training model is not stable enough on the weathered dataset. The final result is achieved: the model achieves an optimal AUC value of 0.9393 when the number of decision trees is 30 and the maximum number of splits is 10.

References

[1] Sichin Bilig, Li Qinghui, Gan Fuxi. Analysis of ancient Chinese potassium glass fractions by laser exfoliation-inductively coupled plasma-atomic emission spectrometry/mass spectrometry[J]. Analytical Chemistry,2013,41(09):1328-1333.

[2] WANG Jie, LI Mo, MA Qinglin, ZHANG Zhiguo, ZHANG Meifang, WANG Julin. A weathering study of an octagonal lead-barium glass vessel from the Warring States period[J]. Glass and enamel,2014,42(02): 6-13.DOI: 10.13588/j.cnki.g.e.1000-2871.2014.02.002

[3] Pei-Bei Zhang,Min Hu. Improved random forest based on grid search for customer satisfaction prediction[J]. Journal of Beijing University of Information Science and Technology (Natural Science Edition),2021,36(04):50-53+58.DOI: 10.16508/j.cnki.11-5866/n.2021.04.010.

[4] Li F. Research on Web user anomalous behavior analysis by mRMR-XGBoost two-layer model[D]. Liaoning University, 2021. doi: 10.27209/d.cnki.glniu.2021.000721.

[5] Yang ZL, Zhang D. Kmeans++-based adaptive superpixel drop point method[J]. Semiconductor optoelectronics,2022,43(03): 585-591.DOI:10.16818/j. issn1001-5868.2022011306.

[6] Ma Liangyu, Zhao Shangyu, Sun Jiaming, Yu Shilai. Abnormal state warning of wind turbine based on grid search optimization ERF model[J]. Thermal Power Engineering,2022,37(02): 160-166.DOI: 10.16146/j.cnki.rndlgc.2022.02.022.

[7] Fatih Özyurt. A fused CNN model for WBC detection with MRMR feature selection and extreme learning machine [J]. Soft Computing: A Fusion of Foundations, Methodologies and Applications,2020,24(9212).

[8] Cabrera Andrew, Bouterse Alexander, Nelson Michael, Razzouk Jacob, Ramos Omar, Chung David, Cheng Wayne, Danisa Olumide. Predicting short-term outcomes after posterior decompression and instrumented fusion for cervical spondylolisthesis using a random forest machine learning algorithm[J]. Journal of clinical neuroscience: the official journal of the Australasian Society for Neurosurgery, 2022.

[9] Gao Xuemei,Yang Xichao,Chen Berru,Lin Lin. Simulation of population spatialization in the Bohai Rim based on random forest model[J]. Journal of Geoinformation Science,2022,24(06):1150-1162.

[10] LI Cast, ZHANG Qingyong, KONG Linghua, LIAN Guofu, LI Peng. Identification of hardness of GCr15 steel based on laser-induced breakdown spectroscopy with random forest[J]. China Laser,2022,49(09):40-49.

[11] Cheng Xiaoyu, Zhou Aitao, Guo Yanzhen, Cheng Cheng, Li Debo. A method for predicting gas outflow from retrieval workings based on random forest and support vector machine[J]. Coal Mine Safety,2022,53(10): 205-211.DOI: 10.13347/j.cnki.mkaq.2022.10.028.

[12] Zhou Jinglei, Zhou Zhi, Cui Lin. Application of variational modal decomposition and random forest feature selection algorithm for anomalous sound classification of loudspeakers[J]. Vibration and Shock,2022,41(20): 277-283.DOI: 10.13465/j.cnki.jvs.2022.20.034.

[13] Hu W, Zhang XY, Li ZEN, Li Q, Wang H. Short-term load forecasting based on optimized VMD-mRMR-LSTM model[J]. Power System Protection and Control,2022,50(01): 88-97.DOI: 10.19783/j.cnki.pspc.210313.

[14] Li Shunyong, Wang Change. A new maximum correlation minimum redundancy feature selection algorithm[J]. Journal of Intelligent Systems,2021,16(04):649-661.

[15] Liu Jiaxing. Support vector regression based on grid search hyperparameter optimization[J]. Science and Technology Innovation,2022(13):71-74.