The Feasibility of AI Models in Assisting Archaeological Research

Minglang Huang
Computer Science and Technology
Southern University of Science and Technology
Shenzhen, China
12011828@mail.sustech.edu.cn

Tong Zhang
Computer Science and Technology
Southern University of Science and Technology
Shenzhen, China
11911831@mail.sustech.edu.cn

***Abstract*—With the rapid development of Machine Learning, the research field about using AI models to assisting archaeology research become more and more vibrant. In ancient jade identification, more and more AI models are being used to promote the development of relevant research work.**

***Index Terms*—Artificial Intelligence, Machine Learning, Jade, Archaeology**

I. Introduction

In recent years, artificial intelligence has been developed and popularized in many fields. It has also been rapidly popularized and applied in archaeology, promoting the efficiency of archaeological work in some relevant fields, such as using pattern recognition cultural relics restoration, using computer vision technology in archaeological exploration, inferencing ancient characters with deep learning and so on. In this journal, we will focus on trying to use several AI models to identify different kinds of jade according to their unique feature like spectral, mineral, crystal characteristics.

II. Applicability of Jade Identification with AI

A. Background Survey

Since the 1980s, with some Breakthroughs in Chinese archaeology research, a great amount of jade which was excavated need to be accurately classified and detected with some jade's details. With the requirement of relevant knowledge and experience, this kind of jobs can be very overwhelming and challenging for archaeology researchers.

Also in these day, many provinces in China have witnessed an upsurge in the manufacture of fake antiques, and the manufacture of fake ancient jade is one of these phenomenon. By using the current modern production technology, the imitation of ancient jade can be extremely high, even people can not tell them accurately. Antique jade made in different regions can be imitated in various ways. The manufacture of modern antique jade is closely related to the development of archaeological discoveries and ancient jade identification, and many archaeological research achievements have been used for reference.

Over the last decade, artificial intelligence (AI) has become widespread across science and technology. With the development of Machine Learning, nowadays it can be convenient and efficiently to apply several AI models not only to help archaeology researchers by working jade identification, but also to help people identify different kinds of jade without being cheated.

B. Application Cases

When it comes to the topic of cultural relics identification, we can found there exists many AI models applications in these relevant fields, such as pottery identification, rock geological identification and jade identification. To extend our ideas on this topic, we do some research on relevant cultural relics identification's topic. And We will discuss some common cultural relics identification methods by using some cases we research as below.

- Automatic Recognition of Pottery with DNN
  The ArchAIDE project raised by italian archaeology researchers mainly use two different deep neural networks to dedicate image recognition and shape recognition. This project basically use pipeline framework to implement the application, using two kinds of classifiers: Appaerence-based recognition classifier and Shaped-based recognition classifier to handle the input image information. Once these images can be identified into correct type, the result can be linked to the photographed sherd and stored in databases which can be accessed online. For studying this case, we find a new idea to identify jade: classification with DNN.

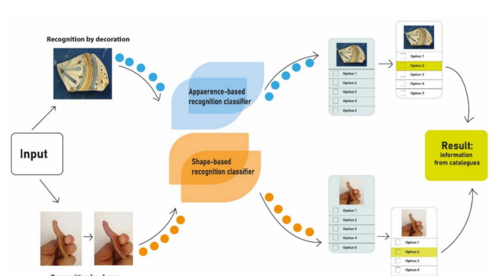

Fig. 1. ArchAIDE's Double workflow for appearance-based and shape-based recognition from an input image to results.

- Intelligent Mineral Identification with Spectral Analysis
  Considering ancient jade is a special kind of minerals, hence we focus on researching some study about mineral

analysis. In the field of Mineral Identification, American scientist Lopez-Reyes established three mathematical models of multivariate analysis based on principal component analysis (PCA), partial least squares (PLS) and artificial neural networks (ANNs), and propose a similarity learning method based on Siamese network, which classification accuracy reaches 0.94 in Raman Minerals Test. Its main idea is to analyze the spectral data of minerals under different light and the size of mineral suspended particles, which help us realize how to accurately identify the types of minerals including jade.

C. Limitation of Jade Identification with AI models

- Huge Gap between Ancient Jade Recognition and AI
  The meaning of related minerals in ancient jade recognition can be qualitatively and clearly described, while machine learning is based on data, and the recognition process and standards are often difficult to explain directly. How to overcome the differences between these two aspects has become the biggest obstacle for researchers of different majors to conduct cross research.
- Lack of Jade Identification Data Sets
  There exists a shortage of ancient jade related data sets. Benchmark data sets play an important role in the research and innovation of ancient jade identification. Benchmark datasets based on high-quality and big data can build better discriminant models and support strict performance comparisons.
- Early Stage In Archaeology Application with AI
  Machine learning method is still in the early stage of exploration in ancient jade recognition, especially in mineral intelligent recognition, and lacks mature methods that are widely recognized, highly accurate and fast. The existing methods are usually used for scattered datasets of different objects and methods. The author sets his own standards to evaluate the performance in the absence of standard answers, which makes it difficult to accurately compare the detection performance, advantages and disadvantages of each method, thus hindering the application and change of the latest detection methods;
- High Development Cost and Uncertain Quality
  On the one hand, AI Archaeology Application's development cost can be high, as well as the AI model development's quality can be uncertain, which means the high development cost of the ancient jade identification project with the low efficiency, leading to difficulty to carry out large-scale application. On the other hand, due to the specificity of cultural relics, it is difficult for such solutions to generate a universal solution, which may require developers to frequently adjust the relevant parameters of the model to reach the standard of the actual situation.

III. Machine Learning in Jade Identification

To test what we have learnt in a more directed approach, we implement several AI models in Jade identification to explain the theoretical ideas in practical example.

We choose jade as the research object for the following reasons. Firstly, jade has been a long-lasting symbol for ancient China. There is a famous jade story dates back to the Warring States Period, when the Qin king offered 15 towns to the state of Zhao in exchange of the rare treasure 'He Shi Bi'. We could easily tell how important it is for ancient people as a connection bond and socializing symbol. The inheritance of jade has been passed down to this day. Jade often represents as a family heirloom to witness the rise and fall of a family, the family members' spirit and emotion, cohesion of the family strength. The inheritance of jade often carries the ancestors' expectations for future generations, hoping that future generations can become a person with faith, pursuit and gentle as jade stands for. While the specific way of using jade and its social significance vary greatly in different cultures, social circumstance and historical stages. The awareness, recognition, use and respectfulness of ancestors for jade are directly related to the physical properties of jade and the social value as well as cultural connotation endowed by the physical properties. Nephrite's unique high toughness may be used to embody and construct the character and social relationship with life vigor, lasting, eternal and unbreakable. The high toughness determines that jade can be used for a long life, so a piece of jade can be passed down from generation to generation, while its physical properties remain basically unchanged. This is also an important reason for nephrite to be studied in archaeology. The toughness of nephrite is determined by the properties of the fiber bundles themselves and the intersection between fiber bundles. The smaller single crystals that compose fiber bundles, the greater degree of dislocation between the crystals, and hence greater strength of fiber bundles. To put it in a nut shell, the smaller the fiber bundle, the higher the degree of interweaving, the higher the toughness of nephrite which leads to a higher density and better stability. It is the particularity of fiber structure of nephrite that determines the special high toughness and density of nephrite, and then the special warm luster of nephrite, which together constitute the special texture of nephrite.

The structural composition of nephrite is tremolite-actinolite, that is, mineral type with chemical formula $Ca2(Mg,Fe)5Si8O22(OH)2$. In order to analyze the structural components of jade on the basis of protecting the cultural relics themselves, we adopt the non-destructive mineralogy analysis method of jade, that is, to scan the jade in an all-round way through the spectral scanner, and obtain a set of spectral data of the jade.

Different bands contain different chemical element information. We can identify whether an artifact is nephrite or not, and whether it has been incinerated or not by judging the characteristics of different bands.

The data used for processing is a group of spectral data files in.mat format. There are two dimensions in the file: the jade number corresponding to each spectral data, and the spectral data obtained at the frequency of 1300nm-2500nm at the sampling rate of 2nm, which is represented by the reflectance size at the corresponding frequency.

In addition, we manually marked the jade data through the jade characteristic labels prepared by the archaeologists provided by the Humanities and Social Sciences Center. Among them, the mineral composition is analyzed by near infrared spectroscopy: Np-i-nephrite Type I, NP-II-nephrite Type II, Np III nephrite Type III, AT-leaf serpentine, CT-chlorite, IT-illite, DC-Dicarkite, QZ-chalcedony, cryptocrystalline quartz, UN-unknown mineral or rock. The other characteristics, degree of sensitivity, and texture classification of objects include: A - objects whose mineral composition is Np I; The mineral composition of B, B1, B2 - is similar to that of A, is Np I, but contains trace vermiculite, B - has no or weak exposure or secondary changes (not including type A type I jade pipe), B1 - moderate or strong exposure or secondary changes (not including type A type I jade pipe), B2-A type I jade tube, no or weak qinling or secondary changes; C-A type I jade pipe, mineral composition is Np II; D - objects with mineral composition of Np III; S - Minerals or rocks not jade.

We used Extra Tree Classifier, Support Vector Machine (SVM), Gaussian Naive Bayes and K Neighbors to distinguish the spectral data features of jade and non-jade.

A. Extra Tree Classifier

Extra Tree Classifier is a tree-based ensemble machine learning algorithm that is used for classification tasks. It is an extension of the decision tree algorithm and is often used as a substitute for random forests, which are another type of tree-based ensemble algorithm. It works by building a set of decision trees on a subset of the training data and then combining the predictions of these trees to make a final prediction. Each tree in the ensemble is trained on a different subset of the training data, and the predictions of all the trees are combined using a majority vote or an averaging method. And it is known for its fast training speed and high accuracy, especially for datasets with a large number of features. It is also resistant to overfitting, which means that it is able to generalize well to unseen data. However, it is not as robust to noisy or imbalanced data as some other algorithms.

B. Support Vector Machines (SVMs)

Support Vector Machines (SVMs) are a type of supervised machine learning algorithm that is used for classification and regression tasks. They work by finding a hyperplane in a high-dimensional space that maximally separates different classes. They are known for high accuracy and ability to handle high-dimensional data. They are particularly useful for datasets with a large number of features or when the relationships between features are complex. SVMs are also resistant to overfitting, which means that they are able to generalize well to unseen data. However, SVMs can be sensitive to the choice of kernel function and require careful tuning of hyperparameters to achieve good performance. They can also be computationally expensive to train, especially for large datasets.

C. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a type of instance-based learning algorithm that is used for classification and regression tasks. It works by storing all the training data and then predicting the label or value for a new sample by finding the K nearest neighbors in the training data and using their labels or values to make a prediction. It is a simple and intuitive algorithm that is easy to understand and implement. It requires no training and can be used for both classification and regression tasks. However, KNN can be computationally expensive to use, especially for large datasets, and it is sensitive to the choice of K and the distance metric used to measure similarity between samples.

D. Naive Bayes

Naive Bayes is a type of probabilistic machine learning algorithm that is used for classification tasks. It is based on the Bayes theorem, which states that the probability of an event occurring is equal to the prior probability of the event multiplied by the likelihood of the event given some evidence. In the context of classification, Naive Bayes works by using the probabilities of different classes and the probabilities of different features given each class to predict the class for a new sample. It is called "naive" because it makes the assumption that all the features in the data are independent, which is often not the case in real-world data. Despite this assumption, Naive Bayes is often surprisingly effective and is widely used in a variety of applications. It is simple to implement and requires relatively little training data. However, it can be sensitive to the presence of correlated features and may not perform as well as other algorithms for more complex datasets.

E. Experiments

The following is the confusion matrix of the results of mineral composition classification:

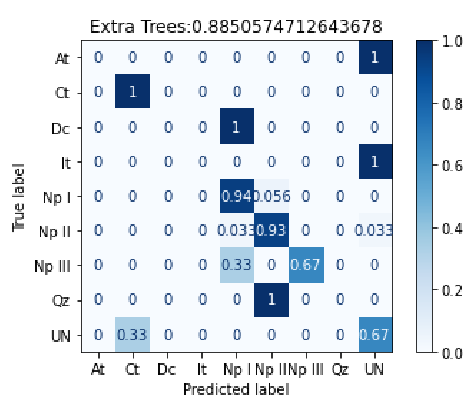

Fig. 2. Extra Tree Classifier

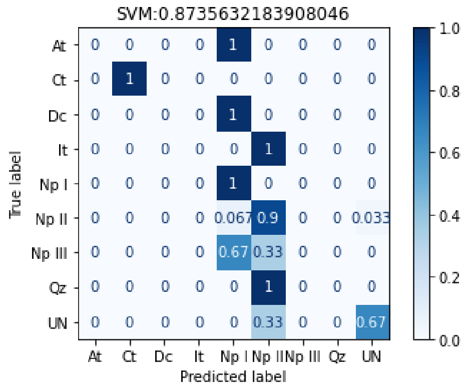

Fig. 3. SVM

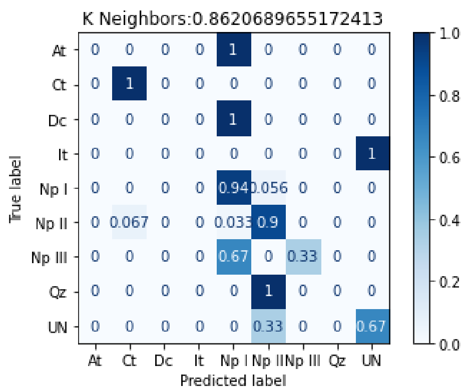

Fig. 4. K Neighbors

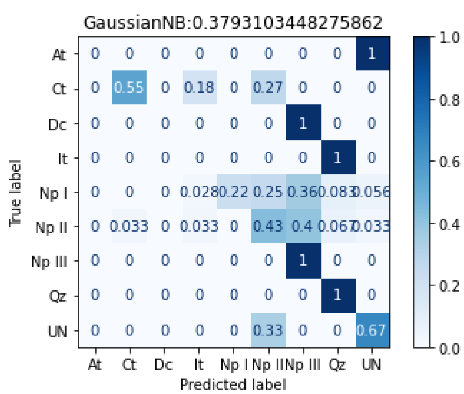

Fig. 5. Gaussian Naive Bayes

According to the confusion matrix and score of the supervised learning classification results of the above minerals, SVM can give better classification results in this case, while the Extra Tree Classifier is unstable, with the dominant result exceeding 0.88 and sometimes lower than 0.85, and K Neighbors slightly lower than the first two. However, Gaussian NB has poor classification effect, this is due to the limits when using Gaussian NB, it requires the datasets' features are independent as we explained above. However, the data we obtained from Jade has 600 features, even though we use feature extraction to reduce it to 50 features, they are still not independent. We can tell from the figure below, for a feature -OH, it is represented by 1380nm-1420nm. Hence, even though we reduced features in jade datasets, the features are still not independent which leads to bad performance of Gaussian NB.

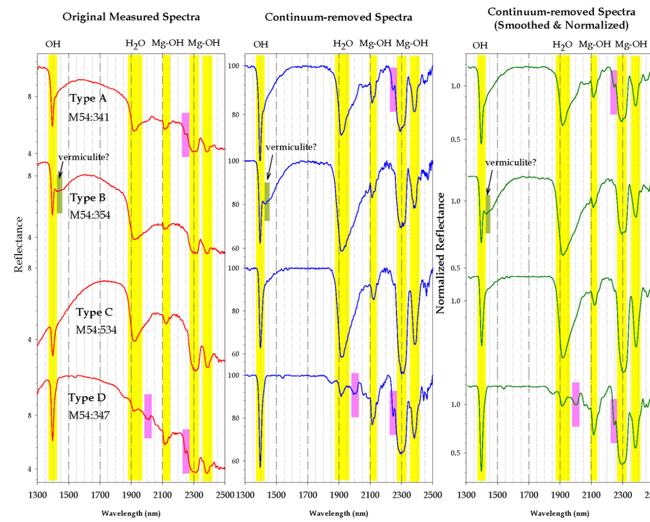

Fig. 6. Frequency and Chemical Bonds

Acknowledgment

Thanks to the introduction of Artificial Intelligence taught at CS103, we have the chance to be involved in such interesting multi-disciplinary research fields. Special thanks to prof. Jimmy Liu, prof. Zhichun Jing and all the class tutors who instruct us together in completing the survey as well as answer our technical and domain knowledge problems.

References

[1] Meng Lei. (2013). Near infrared spectroscopy analysis of coal quality based on machine learning (doctoral dissertation, China University of Mining and Technology)
[2] Guang Wen, Zhichun Jing. (1997). Research on Archaeology of Ancient Jade Geology in China. Institute of Geology, Chinese Academy of Geological Sciences. (eds.) Collected Works of Institute of Geology, Chinese Academy of Geological Sciences (29-30) (pp. 292-307). Geological Publishing House
[3] Guang Wen. (1990). Research on Chinese Ancient Jade Building material geology (02), 2-10+49-50+2
[4] Gualandi 2021 An Open System for Collection and Automatic Recognition of Pottery through Neural Network Algorithms
[5] Batsaki 2021 AI and Cultural Heritage Image Collections-Opportunities and challenges
[6] Landscape Archaeology and Artificial Intelligence 2021
[7] Kazimi 2022 Deep Learning for Archaeological Object Detection
[8] Pawlowicz 2021 Applications of deep learning to decorated ceramic typology and classification 2021
[9] LYONS 2021a Ceramic Fabric Classification of Petrographic Thin Sections with Deep Learning
[10] ABITBOL 2021 Machine Learning Based Assembly of Fragments of Ancient Papyrus 2021
[11] A tutorial on automatic hyperparameter tuning of deep spectral modeling for regression and classification tasks 2022
[12] Computer Vision in the Infrared Spectrum 2022

[13] Geological Archaeological Research on M54 Unearthed Jade Articles 2008
[14] Middleton 2006 Jade – Geology and Mineralogy
[15] Wen Guang.Evaluation criteria for jade. 1900
[16] Wenying Li, Bin Cao, Chunshui Cao, Yongzhen Huang. A bronze inscription recognition method based on deep learning [J]. Journal of Automation, 2018, 44 (11): 2023-2030.