Neural Network-Based Classification of XRF Profiles Of Pottery Shards Using Synthetic Data[1]

Ankita Nandy
Independent Researcher,
Hooghly, West Bengal, India

DOI:10.37648/ijrst.v13i03.007

Received: 17 June 2023; Accepted: 20 August 2023; Published: 20 August 2023

ABSTRACT

Across all archaeological sites, pottery fragments have been abundant and insightful resources for understanding the societies which crafted, traded and/or used them. Their chemical profiles generated using X-ray fluorescence (XRF) techniques are used to characterize the raw materials and manufacturing techniques, and thus map them to their potential origins. Numerous researchers have published their XRF analysis results for pottery shards across the world, which exist as isolated small datasets. The collation of these datasets and subsequent usage in training classifiers aids in studying potential migration routes and trade and diplomatic relations which have bridged civilizations. Artificial Neural Networks (ANNs) have gained popularity across all domains considering their versatility in learning non-linear, complex patterns. However, training them requires large datasets which is less common in the field of archaeology. To solve the problem posed by data availability, the classifier ANN is trained with synthetically generated data. The 297 XRF records, when added to synthetically generated data, swelled to over 13k records, and could classify pottery shards across 11 geographies with an accuracy of over 80 per cent.

Keywords: *pottery shards; XRF; synthetic data; classification; neural networks*

INTRODUCTION

The analysis of the raw materials and technology used in building earthenware or ceramics serves as geographical tracers to their place of origin. The clay used, the decorative elements, kiln temperatures et cetera tell tales of the bygone era, such as the socio-economic hierarchies and the trade and diplomacy with other regions. X-ray fluorescence (XRF) is a non-intrusive and non-destructive method in geochemical analysis which has been around for decades. On radiating a sample with X-rays, emissions of different wavelengths are observed; these emissions can be mapped to certain elements. This unravels their detailed composition. XRF is just one of the numerous techniques in practice for such analysis and is often complemented by or supplemented with results from other methods, such as using the radioactive isotopes ratios of the pottery material used in [7].

A. Neural Networks for Classification

The segregation of these XRF records is a supervised classification problem. Artificial neural networks and deep learning networks have gained wide popularity for their versatility across domains. ANNs are designed based on the structure and functioning of the human brain. The basic structural unit is called a neuron. The output of a neuron is the response of an activation function to a weighted stimulus and a bias. Several neurons are interconnected and arranged in consecutive layers. The training algorithm iteratively determines the weights and biases which can generate desired outputs. The activation functions of the component neurons and the overall architecture determines the capabilities of the neural network. For binary classification, the sigmoid activation function, and for multi-class classification, softmax function is used in the ultimate layer. Theoretically, the ANNs can be trained to learn any mathematical function. But these algorithms need a lot of labeled data for training. Unlike the volume or veracity of e-commerce transactional data, as mentioned in [2], the field of archeological studies does not generate such data and obtaining adequate labeled data can be a challenge. The samples in one site may outnumber those gathered in another. One of the popular techniques to solve the problems of data adequacy and data imbalance is the generation of synthetic data.

[1] How to cite the article: Nandy A. (July 2023); Neural Network-Based Classification of XRF Profiles Of Pottery Shards Using Synthetic Data; *International Journal of Research in Science and Technology*, Vol 13, Issue 3, 65-71, DOI: http://doi.org/10.37648/ijrst.v13i03.007

B. Synthetic Data Generation

Synthetic data comes with a host of benefits, as enumerated in [9]. It allows for data sharing without compromising on the privacy of the subjects. It creates additional data where the availability of real data might be limited, such as in the study of rare diseases, or in staff training programs with hypothetical customers. It is especially useful for generating audio-visual responses for interactive digital assistants. [13] synthesizes the spectra of common elements used in artists' pigments using the Fundamental Parameters model, later used to train a Convolutional Neural Network tasked with the classification of pigments on paintings through their XRF profiles. One very common method used to generate synthetic data for an imbalanced dataset, is the Synthetic Minority Over-sampling Technique (SMOTE), presented in [4]. In this, for any point in the minority class, a noisy replica is created which is closer than the true nearest neighbor(s). [12] employs SMOTE for synthetic data generation for a comparative evaluation of different machine learning algorithms for classification. Generative Adversarial Networks is a two-network architecture, one called the generator, tasked with creating fake data samples similar in properties to the real ones such that the other network, called discriminator or critic, cannot detect the fake from real, as described in [1]. As deep learning networks are gaining popularity in synthetic data generation applications [9], GANs have also been enhanced into several variants, which can generate not just numeric records, but audio clips and images. For generating tabular records while preserving mutual relationships between different columns and to mitigate the problem of class imbalance, [24][25] train the Generator conditional to the discrete/categorical column(s). [5] combine the simplicity of SMOTE with deep learning to generate synthetic high-quality images. Synthetic Data Vault (SDV) presented in [17] is a project maintaining several data generation techniques for standalone and relational tables. The tools are available as a Python library which has made it an easy choice to be used in this work.

This work collates XRF profiles of samples from various published sources, corresponding to different locations, generates synthetic data using CTGAN synthesizer from the SDV and trains an ANN for multi-class classification.

RESULT AND ANALYSIS

A. Data

Data for this work has been gathered by consolidating the compositions of samples across publications listed as follows. The assimilation of data from past publications facilitates the reuse of their analyses. Some of the papers which had published their XRF analysis in tabular format were gathered for this work. The published tables were converted to data files using the Tesseract OCR library [22]. Though papers had published the percentage compositions of several elements, information on the oxides of Silicon, Aluminum, Iron, Titanium, Potassium and Calcium were retained. [3] obtained their samples from sites in northeastern Tamil Nadu, India. These samples have been marked as A. [21] analyzes samples from Dahan-e Ghulaman in Iran, believed to be the capital of Sistan in the heights of its glory. [17] studies pottery fragments from Shahr-e-Sokhta, another important archeological site in Sistan. Of the samples, all but five are found to be locally made. [20] also studies shards from sites in the Sistan area. The XRF profiles of these samples have been marked as B. [19] uses the samples from archaeological sites in present day Gujarat and Maharashtra in India to compare them with some samples in the Emirate of Sharjah. These XRF profiles have been marked as C. Ancient pottery shards from the archaeological sites of Conjunto Vilas and São João in Amazonia have been studied [15], which have been marked as D. [11] studies celadon porcelain unearthed from Longquan and Jingdezhen sites in China. The body profiles of the samples from Longquan have been used, which have been marked as E. In [6], excavated samples were spread out along the northern coast of the Black Sea. These have been marked as F. [23] obtains the XF profiles of pottery samples from the Lijiaba site, in the Three Gorges reservoir in Chongqing, China, have been gathered. These samples have been marked as G. [18] studies votive tablets from Chawas cave in Hulu Kelantan in Malaysia. The tablets are associated with Mahayana Buddhism. As reported, these are not locally sourced and could have been transported from other regions where the faith was prevalent, however for this work they have been marked as category H. [8] study samples from neolithic sites of Esh Shaheinab, Kadaro and Jebel Umm Marahi, all located in present day Sudan. Some of these sites were inhabited as early as the 8^{th} millennium BCE. The XRF profiles of these samples have been marked as location I. Termez, an archeological site in present day Uzbekistan, was located on the Silk Road. [10] analyze pottery shards from this site dating back to 9^{th} to 12^{th} centuries CE. These XRF profiles have been marked as J. [16] study six pottery samples unearthed from a site in present day Udayagiri, in the state of Odisha, India. The site was bustling city attracting Buddhist scholars and pilgrims. The XRF profiles of these samples have been marked as category K. 279 samples are collected which are unevenly distributed across 11 categories which reflect the geographical locations, they were excavated in. However, as some of these sites were active trading centers, the categories may not correspond to their origins. The labeling has been summarized in Table 1.

B. Preprocessing

The comparative compositions are visualized using violin plots. The first two components obtained through principal components analysis (PCA) on the dataset are used to plot the data points, and observe the similarities and overlaps across locations. The data is scaled using max-min scaling.

C. Data Synthesis and Classifier Trainings

The scaled data is split into training and testing subsets. The training data is used to train the synthesizer and generate additional samples. These samples, together with the training data, are used to train a neural network classifier with 1500 neurons in the input layer, one hidden layer with 300 neurons, and 11 neurons in the ultimate layer, using the Adaptive Moment Estimation (ADAM) optimizer.

D. Results

Figure 1 shows the category wise compositions for each component for the original data. Categories B, E and J demonstrate large variations in the Calcium Oxide contents. Samples from A are lowest in Silica content. Figure 2 presents the scatterplot of the pottery samples for the two major principal components. The categories F and I are far removed from others, as is obvious from its being celadon, while others are earthenware. The categories A, B, E and J are spread out, indicating the variety in their compositions. D, G, H and K are densely packed and overlap. These similarities can be attributed to similar clay compositions and potential trades or to mere chance. Figure 3 shows the category wise compositions for each component for the synthetic data. The similarities with Figure 1 are obvious. This indicates the CTGAN synthesizer could capture the peculiarities of the data very well. A comparison of the real and synthetic data shows a similarity of 80 per cent. Out of the 112 samples set aside for testing, 99 were correctly classified, registering a 88.4 percent accuracy. The misclassified instances might have been samples found in a certain location but with origins elsewhere, as reported by some authors in the original publications. As observed with samples marked J and B in Figure 2, the sites being on or near the Silk Road might have samples manufactured in some other city along the road, or brought in by nomadic groups.

Table I. Categories of Pottery Samples with Sources and Counts

Label	***Locations***	***Count***
A	Melpadi, Paiyampalli, Thirumani, Karivadu, Arcot, Vallimalai, Kaveripakkam, PanchaPandavar Malai, Walajah, Udayandiram in Vellore, Tamil Nadu [3]	10
B	Sistan,Iran [21][14][20]	54
C	Dwarka, Prabhas Patan, Padri, Nevasa, Junnar and Nasik [19]	11
D	Conjunto Vilas and São João in Amazonia [15]	10
E	Dykyj Sad, Subotiv, Glinjeni II-La Șanț, Kartal and Nemyriv in the northern coast of Black Sea [5]	35
F	Longquan, China [11]	44
G	Lijiaba, Chongqing, China [23]	53
H	Chawas Cave, Hulu Kelantan, Malaysia [18]	10
I	Esh Shaheinab, Kadaro and Jebel Umm Marahi, Sudan [8]	7
J	Termez, Uzbekistan [10]	39
K	Udaygiri, Odisha, India [16]	6

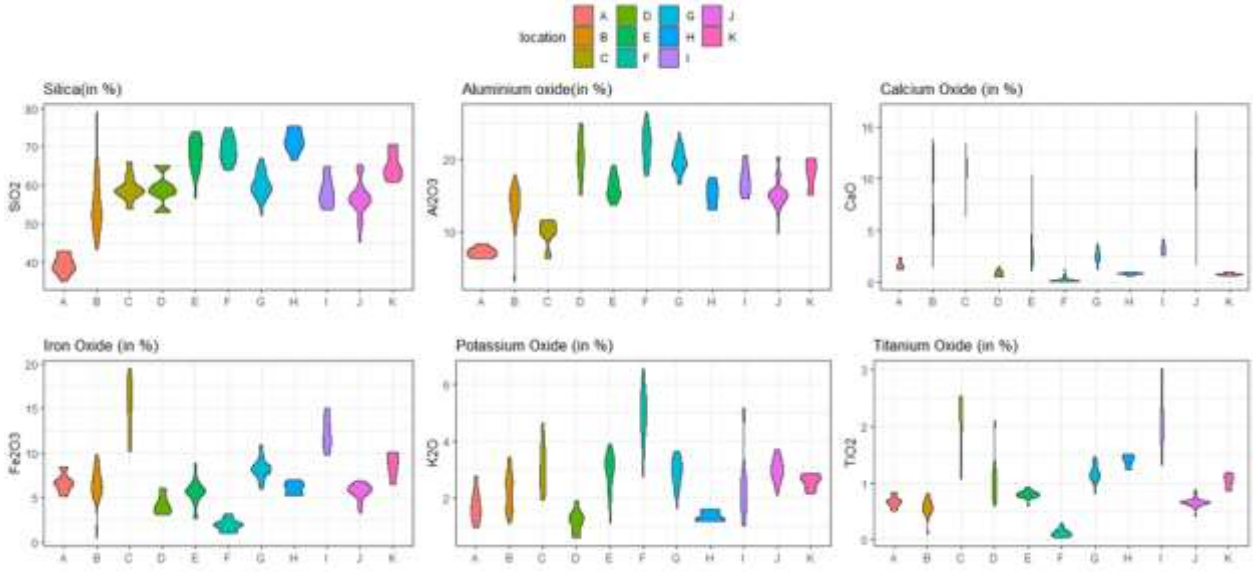

Figure 1. Violin Plots: Compositions of pottery shards across locations and elements.

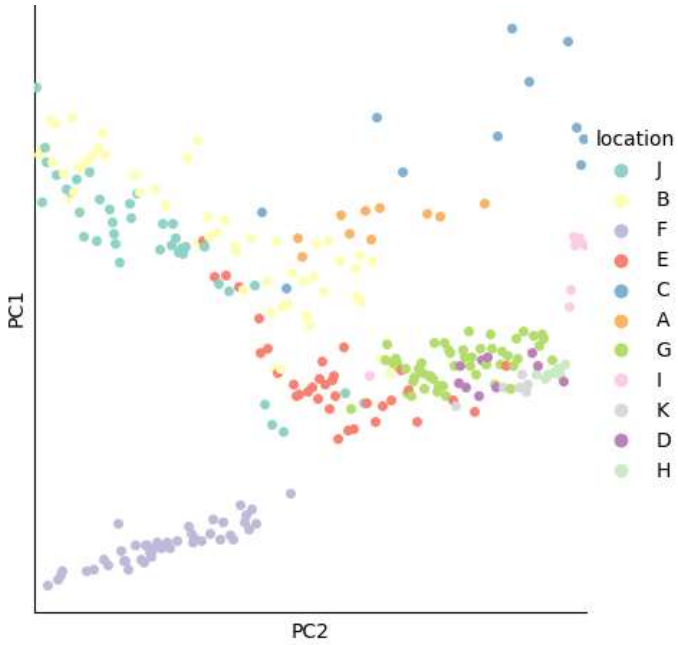

Figure 2. Scatterplot: Samples plotted by location across first two principal components.

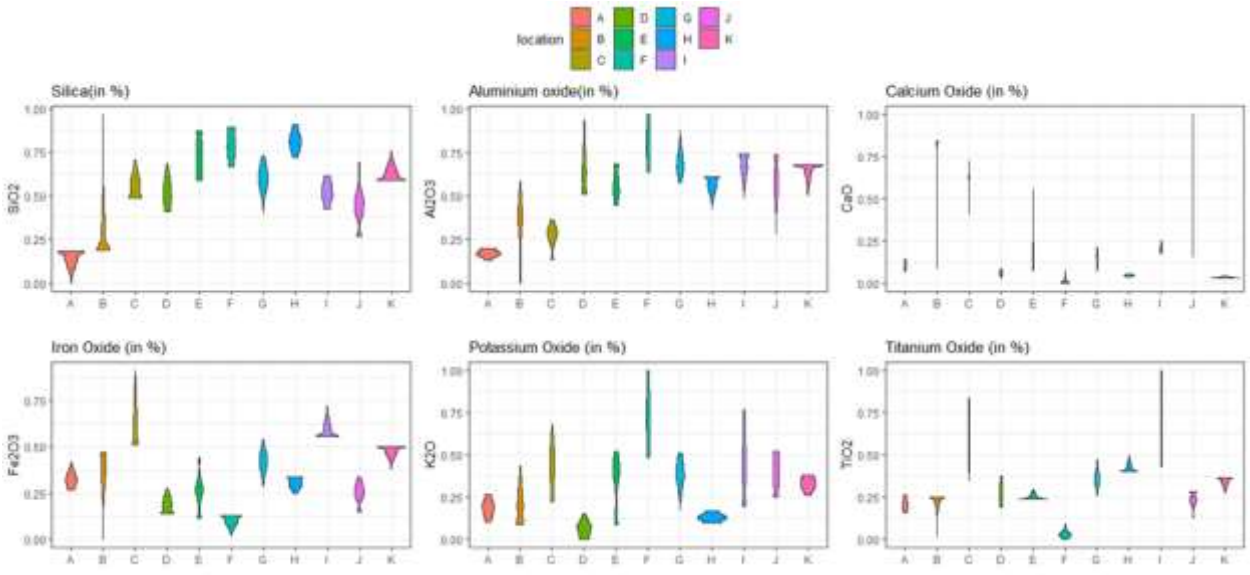

Figure 3. Violin Plots: Compositions of pottery shards using synthetic data.

Table 2. ANN Classifier Confusion Matrix

		PREDICTED										
		A	***B***	***C***	***D***	***E***	***F***	***G***	***H***	***I***	***J***	***K***
A	***A***	4										
C	***B***		18			1					3	
T	***C***			4								
U	***D***				3			1				
A	***E***					13					1	
L	***F***						18					
	G							21				
	H								4			
	I									3		

	J		4			2					10	
	K					1						1

FUTURE WORK

The sources used for this dataset admit some samples to be different from the others, but this discrimination could not be made and the samples were mapped to their location of excavation. The compositions corresponding to just six elements were used, the inclusion of additional components might have been useful in enhancing the classification. Refinement of the labeling and inclusion of additional features can be attempted in the future phases of this work.

CONCLUSION

The work employs open-source tools in R and Python, and achieves a high classification accuracy. This gives the application of neural networks and related architectures in the field of archeology a positive push. Human history is a shared resource, linked by exchanges through trades, conquests, and migration. Automated classification of samples unearthed at new locations can provide rudimentary insights, which can later be refined/corrected by experts, therefore speeding up the process, and potentially assist in unraveling the missing links in the journey of civilization.

REFERENCES

1. Alqahtani, H., Kavakli-Thorne, M., & Kumar, G. (2021). Applications of generative adversarial networks (GANs): An updated review. *Archives of Computational Methods in Engineering*, *28*, 525-552.
2. Bickler, S. H. (2021). Machine learning arrives in archaeology. *Advances in Archaeological Practice*, *9*(2), 186-191.
3. Chandrasekaran, A., Naseerutheen, A., & Ravisankar, R. (2017). Dataset on elemental concentration and group identification of ancient potteries from Tamil Nadu, India. *Data in brief*, *10*, 215-220.
4. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.
5. Dablain, D., Krawczyk, B., & Chawla, N. V. (2022). DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*.
6. Daszkiewicz, M., Gavrylyuk, N., Hellström, K., Kaiser, E., Kashuba, M., Kulkova, M., ... & Winger, K. (2020). Possibilities and limitations of pXRF as a tool for analysing ancient pottery: a case study of Late Bronze and Early Iron Age pottery (1100–600 BC) from the northern Black Sea region. *Praehistorische Zeitschrift*, *95*(1), 238-266.
7. Długosz-Lisiecka, M., Sikora, J., Krystek, M., Płaza, D., & Kittel, P. (2022). Novel method of ancient pottery analysis based on radioactive isotope ratios: a pilot study. *Heritage Science*, *10*(1), 1-18.
8. Elbashir Siddig, F., Elbashir, A. A., Lepper, V., & Hussein, A. (2018). Spectroscopic approach for characterization of archaeological potsherds excavated from some Neolithic sites from Sudan. *International journal of experimental spectroscopic techniques*, *3*(2), 1-11.
9. Figueira, A., & Vaz, B. (2022). Survey on synthetic data generation, evaluation methods and GANs. *Mathematics*, *10*(15), 2733.
10. Fusaro, A., Martínez Ferreras, V., Gurt Esparraguera, J. M., Angourakis, A., Pidaev, S. R., & Baratova, L. (2019). Islamic pottery from ancient Termez (Uzbekistan): new archaeological and archaeometric data. *ArcheoSciences. Revue d'archéométrie*, (43-2), 249-264.
11. He, Z., Zhang, M., & Zhang, H. (2016). Data-driven research on chemical features of Jingdezhen and Longquan celadon by energy dispersive X-ray fluorescence. *Ceramics International*, *42*(4), 5123-5129.
12. Heyburn, R., Bond, R. R., Black, M., Mulvenna, M., Wallace, J., Rankin, D., & Cleland, B. (2018). Machine learning using synthetic and real data: similarity of evaluation metrics for different healthcare datasets and for different algorithms. In *Data Science and Knowledge Engineering for Sensing Decision Support: Proceedings of the 13th International FLINS Conference (FLINS 2018)* (pp. 1281-1291).

13. Jones, C., Daly, N. S., Higgitt, C., & Rodrigues, M. R. (2022). Neural network-based classification of X-ray fluorescence spectra of artists' pigments: an approach leveraging a synthetic dataset created using the fundamental parameters method. *Heritage Science*, *10*(1), 1-14.
14. Moradi, H., Sarhaddi-Dadian, H., Ramli, Z. & Rahman, N. H. S. N.A. (2013). Compositional analysis of the pottery shards of Shahr-I Sokhta, South Eastern Iran. *Research Journal of Applied Sciences, Engineering and Technology*, *6*(4), 654-659.
15. Oliveira, L. S. S., Abreu, C. M., Ferreira, F. C. L., Lopes, R. C. A., Almeida, F. O., Tamanaha, E. K., ... & Souza, D. N. (2020). Archeometric study of pottery shards from Conjunto Vilas and São João, Amazon. *Radiation Physics and Chemistry*, *167*, 108303.
16. Panda, S. S., Jena, G. N., & Garnayak, D. B. (2019). Characterization of Representative Ancient Potteries: Chemical, Mineralogical and Morphological Studies. *International Journal of Conservation Science*, *10*(2), 317-326.
17. Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 399-410). IEEE.
18. Ramli, Z., Rahman, N. H. S. N. A., Samian, A. L., Razman, M. R., Zakaria, S. Z. S., Jusoh, A., ... & Dadian, H. S. (2014). X-Ray Diffraction (XRD) and X-Ray Fluorescence (XRF) analysis of proto-historic votive tablets from Chawas cave, Hulu Kelantan, Malaysia. *Research Journal of Applied Sciences, Engineering and Technology*, *7*(7), 1381-1387.
19. Reddy, A., Attaelmanan, A. G., & Mouton, M. (2012). Pots, plates and provenance: sourcing Indian coarse wares from Mleiha using X-ray fluorescence (XRF) spectrometry analysis. In *IOP Conference Series: Materials Science and Engineering* (Vol. 37, No. 1, p. 012010). IOP Publishing.
20. Sarhaddi-Dadian, H., Ramli, Z., Rahman, A., & Mehrafarin, R. (2015). X-ray diffraction and X-ray fluorescence analysis of pottery shards from new archaeological survey in south region of Sistan, Iran. *Mediterranean Archaeology and Archaeometry*, *15*(3), 45-56.
21. Sarhaddi-Dadian, H., Moradi, H., Zuliskandar, R., & Purzarghan, V. (2017). X-ray fluorescence analysis of the pottery shards from dahan-E ghulaman, the achaemenid site in Sistan, east of Iran. *Interdisciplinaria Archaeologica*, *8*(1), 35-41.
22. Smith, R. (2007, September). An overview of the Tesseract OCR engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)* (Vol. 2, pp. 629-633). IEEE.
23. Wu, Q. Q., Zhu, J. J., Liu, M. T., Zhou, Z., An, Z., Huang, W., ... & Zhao, D. Y. (2013). PIXE-RBS analysis on potteries unearthed from Lijiaba Site. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, *296*, 1-6.
24. Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. *Advances in neural information processing systems*, *32*.
25. Zhao, Z., Kunar, A., Birke, R., & Chen, L. Y. (2021, November). Ctab-gan: Effective table data synthesizing. In *Asian Conference on Machine Learning* (pp. 97-112). PMLR.