2025 IMEKO TC26 International Conference on
Metrology for Archaeology and Cultural Heritage
Bergamo, Italy, October 15-17, 2025

Exploring the Application of Interpretable Neural Networks for the Petrographic Classification of Ceramic Samples from the Levant

Capriotti Sara[1], Devoto Alessio[2], Genovese Donatella[2], Mignardi Silvano[1], Scardapane Simone[3], Medeghini Laura[1]

[1]*Department of Earth Sciences, Sapienza University of Rome, P.le Aldo Moro 5, 00185 Rome, Italy, {sara.capriotti,silvano.mignardi,laura.medeghini}@uniroma1.it*
[2]*Department of Computer, Control and Management Engineering, Sapienza University of Rome, Via Ariosto 25, 00185 Rome, Italy, {alessio.devoto, donatella.genovese}@uniroma1.it*
[3]*Department of Information Engineering, Electronics and Telecommunications, Sapienza University of Rome, Via Eudossiana 18, 00184 Rome, Italy, simone.scardapane@uniroma1.it*

***Abstract* – The archaeological context of the Levantine region is both rich and complex, particularly during the transition from the Late Chalcolithic to the Early Bronze Age, a period marked by urban development, craft specialization, and interregional trade. This study explores the use of Artificial Intelligence techniques to classify Levantine ceramic thin sections based on their petrographic *fabrics*. Deep learning models, including Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), were applied to a large dataset of thin section images from ceramics dating to the Uruk period, Bronze Age, and Iron Age, collected from various archaeological sites across the Levant. To improve model transparency, explainable AI methods such as Guided Grad-CAM and attention maps were applied to identify key features and interpret latent representations. The results show that deep learning can achieve high accuracy in automated ceramic classification and provide important insights into ancient ceramic technologies and cultural interactions.**

I. INTRODUCTION

This study applies Artificial Intelligence (AI) methodologies to the grouping and classification of Levantine ceramics based on their petrographic *fabrics*. The Early Bronze Age represents a critical phase in the archaeological history of the Levant, characterized by increasing complexity in settlement patterns, production systems, and trade networks. These developments are particularly evident in pottery production, where ceramics are considered key indicators for understanding technological practices and patterns of cultural interaction across time and space [1]. The Levantine area played a central role in this process, acting as a hub of ceramic innovation and craft specialization. From the Early Bronze Age onward, the region experienced the consolidation of urban life, the expansion of interregional exchange, and the emergence of political centralization. These transformations reflect a path to centralized social organization, with a simplification of material culture and the restructuring of ceramic production systems [2, 3]. Although scholarly interest in this region has grown significantly, the study of ancient exchange systems and the reconstruction of trade patterns remains a major challenge. The lack of sufficient comparative datasets limits the ability to study long-distance interactions and trade connections.

Artificial intelligence is a brunch of computer science that is increasingly being applied to archaeological research. Over the past two decades, AI-based classification tools, particularly those involving machine learning (ML) and deep learning (DL), have been used as complementary methods in archaeology and archaeometry [4-7]. In ceramic studies, the ML and DL models have been used to classify pottery based on typological, decorative, and morphological characteristics, as well as for provenance analysis [8-11]. More recently, DL approaches based on convolutional neural networks (CNNs) have been applied to the classification of ceramic thin sections according to their petrographic *fabrics* [12, 13].

Building on this progress and on our previous research [14], which successfully implemented deep learning models for the classification of ceramic petrographic *fabrics*, the present study extends this approach by incorporating a significantly expanded dataset. Although the same architectures of convolutional neural networks (CNNs) and vision transformers (ViTs) are adopted, this study introduces new ceramic samples from additional archaeological sites and includes new petrographic classes. This broader dataset enables a more comprehensive and representative analysis of Levantine pottery variability. In addition to model training and evaluation, the study also applied interpretability through the use of Explainable Artificial In-

telligence (XAI) techniques. In fact, although DL methods are known for their high accuracy, their internal mechanisms are often difficult to interpret, making them described as "black boxes" due to lack of transparency in their decision-making processes [15, 16]. XAI seeks to overcome this limitation by making these processes more accessible and understandable. Visual explanation tools such as Guided Grad-CAM (guided gradient-based Class Activation Maps) and attention maps are used to highlight the features that contribute the most to classification decisions. This approach supports a more transparent and informed application of AI in archaeological research.

II. MATERIALS AND METHODS

The dataset was built by acquiring 4,768 images from 596 ceramic thin sections, selected from a broad range of archaeological sites across the Levantine region. These include Bethlehem, Jericho, Tell el-Far'ah (North), Al Jib and Jerusalem (West Bank); Khirbat Iskandar, Khirbat al-Batrawy, Tell Balata, and Deir Alla' (Jordan); Tell Qasile (Israel), Ebla, Tell Nebi Mend, Jebel Aruda and Tell Hadidi (Syria). The geographic and chronological diversity of these sites contributes to the representativeness and variability of the dataset (see table 1 for more information about the sites).

Table 1. Summary of the selected ceramic samples.

Archaeological sites	Age	N. of Samples
Tell el-Far'ah (North)	Early Bronze	58
Khirbat al-Batrawy	Early Bronze	46
Khirbat Iskandar	Early Bronze	33
Bethlehem	Bronze and Iron Ages	25
Ebla	Early and Middle Bronze	56
Jericho	Early Bronze	11
Tell Balata	Iron Age	6
Deir Alla'	Bronze and Iron Ages	204
Al Jib	Iron Age	6
Tell Hadidi	Early Bronze	13
Tell Qasile	Iron Age	6
Tell Nebi Mend	Middle Bronze	59
Jerusalem	Bronze and Iron Ages	47
Jebel Aruda	Late Uruk	26

To reflect realistic acquisition conditions, the images were intentionally captured using different microscope and camera systems. This introduces natural variability in

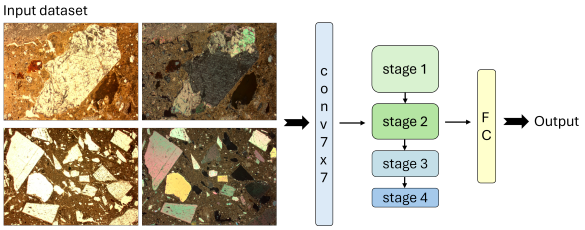

Fig. 1. Scheme of the ResNet18 architecture.

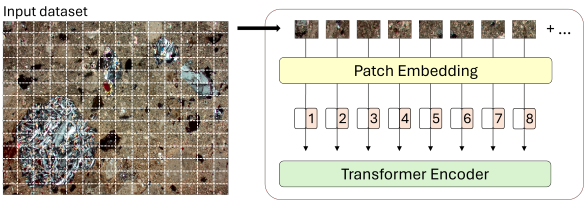

Fig. 2. Scheme of the ViT architecture.

resolution, colour calibration, and lighting; factors that, while adding heterogeneity to the dataset, also improve the realism and generalizability of the trained models, making them better suited for real-world archaeological applications where imaging conditions may vary. For each section, multiple images were taken under both plane-polarized light (PPL) and cross-polarized light (XPL) and at different magnifications varying from 2.5X, 4X and 10X. A total of eight images have been acquired per thin section. All images were grouped into 20 petrographic *fabrics* based on comparative petrographic criteria. Unlike the previous study [14], which used a smaller dataset and applied only a binary split (training/test with a ratio of 80:20), the present work benefits from a significantly expanded dataset and adopts a more robust three-way split: 70% for training, 15% for validation, and 15% for testing. This configuration not only supports more rigorous model evaluation but also allows for the selection and fine-tuning of model hyperparameters using the validation set.

Two types of deep learning models were selected for this study: A CNN using the ResNet18 [17] architecture and the Deit small Vision Transformer [18]. CNNs process images through multiple convolutional layers that detect increasingly complex features; specifically, the ResNet18 architecture consists of 18 layers organized in a total of 5 stages, including convolutional layers, batch normalization, and ReLU activations. The network ends in a pooling and fully connected layers for classification (fig. 1).

In contrast, Vision Transformers work with a different approach. Input images are divided into fixed-size patches (196 patches of 16×16 pixels in our case, fig. 2), which are linearly embedded with positional information and then processed through a series of transformer encoder blocks

based on multi-head self-attention mechanisms. Our ViT implementation employs 12 transformer encoder blocks, each consisting of 6 attention heads and a hidden dimension of 384.

Both models were trained to classify images into their respective petrographic *fabrics* by minimizing a cross-entropy loss function. Transfer learning was applied by fine-tuning models pre-trained on ImageNet, a large dataset that contains more than one million images from 1,000 classes [19]. This approach allows the model to build upon previously learned general visual features, such as edges and textures, by retaining backbone weights and retraining only the final classification layers to adapt to our specific task. Training was performed for 25 epochs using the AdamW optimizer. For ResNet18, a learning rate of $1x10^{-4}$, a weight decay of 1×10^{-4}, and a batch size of 32 were used, while for ViT, a learning rate of 1×10^{-5}, a weight decay of 3×10^{-4}, and a batch size of 64 were applied. To improve model generalization and artificially increase the training data size, data augmentation techniques including random axis flipping, color jittering, and random cropping, were applied consistently during the training phase to reduce overfitting [20]. The performance of the two deep learning models was compared in terms of both classification accuracy and generalization capabilities. Their predictive performance was evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. In addition, a confusion matrix was generated for each model to visualize the performance and identify misclassification patterns. To further assess generalization ability, we conducted additional experiments in which data from specific archaeological sites (Jericho, Al Jib and Tell Hadidi) were entirely excluded from the training, validation and initial test sets and used solely as independent and out of domain new test sets. This approach simulates real-world scenarios in which the models must analyze previously unseen data from new excavation contexts.

III. RESULTS AND DISCUSSION

This study extends the work of Capriotti et al. (2025). In the earlier study, the dataset was relatively small and split into training and testing sets. With the improved and expanded dataset now available, we revisited the same models (ResNet18 and DeiT small), training them on the updated collection, which comprises 4,768 images and 20 petrographic *fabrics* as classes. The results were promising, with both models achieving strong performance across all evaluation metrics. Figure 3 presents the confusion matrices of ResNet18 and DeiT small, showing correct and incorrect predictions relative to the true labels. Accuracy, precision, recall, and F1-score values are reported in Table 2, confirming the overall robustness of both models. Specifically, all ResNet18 metrics exceeded 86%, with an accuracy of 86.45%, while all DeiT small metrics reached more 93%, with an accuracy of 93.58%.

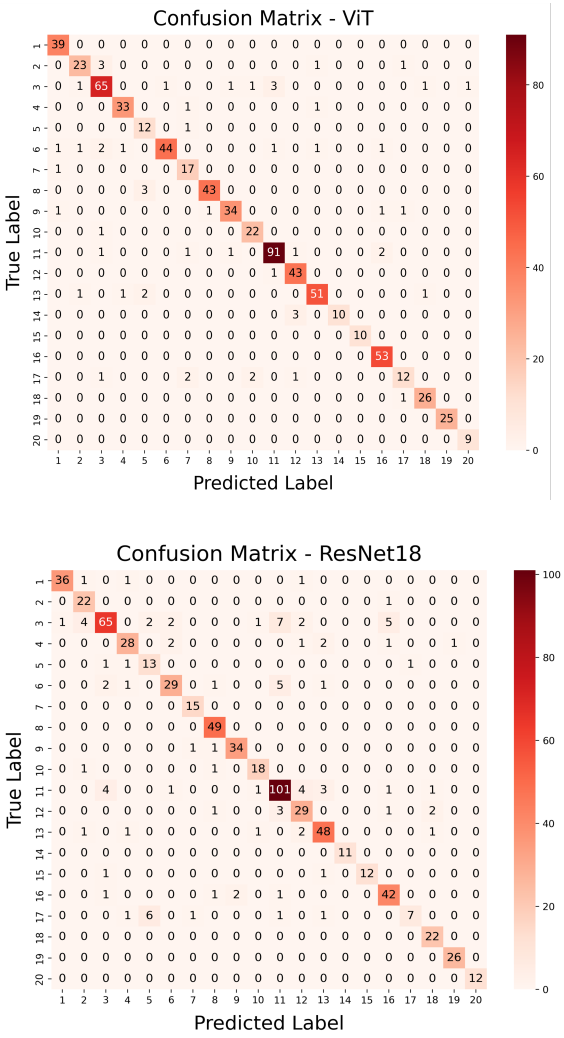

Fig. 3. Confusion Matrices of ViT small and ResNet18.

Moreover, to evaluate the generalization ability of ResNet18 and Deit small, we decided to exclude all the samples from Jericho, Al Jib and Tell Hadidi from the original dataset and to use them as a separate domain for the final testing. It should be noted that this new dataset includes only a subset of the classes present in the original collection, making the evaluation more challenging. Both models were first trained, validated, and tested on the remaining dataset, and then evaluated on this new domain to simulate a "real-world" scenario. On this unseen dataset, results are reported in table 3. ResNet18 achieved a macro-F1 of 0.767, with a micro- and weighted F1 of 0.81, and a confidence interval of precision 95% of 67.4% - 79.0%. DeiT small showed a very similar performance, with a macro-F1 of 0.76, a micro- and weighted F1 of 0.80, and an accuracy

Table 2. Accuracy, precision, recall (True positive rate) and F1-score (harmonic mean of precision and recall) metrics of the CNN and ViT.

Model Metrics	ViT (small)	ResNet18
Accuracy	93,58%	86,45%
Precision	93,88%	87,43%
Recall	93,41%	87,69%
F1-score	93,57%	86,85%

Table 3. Macro-F1, Micro-F1, Weighted-F1 and Accuracy (95% CI) of ResNet18 and ViT (small) on the new domain (Jericho, Al Jib, Hadidi Tomb).

Model Metrics	ViT (small)	ResNet18
Macro-F1	76,0%	76,7%
Micro-F1	80,0%	81,0%
Weighted-F1	80,0%	81,0%
Accuracy 95% CI	67,0–78,6%	67,4–79,0%

95% confidence interval of 67% – 78.57%. In both models, the performance was higher on more frequent classes (e.g., class 16), while smaller classes (such as 2, 4, 11) exhibited more variability, reflecting the challenges posed by class imbalance. For clarity, we report primarily the F1-scores, as they provide a more reliable evaluation metric than accuracy, precision, and recall in this context, given the class imbalance and the reduced number of classes present in the new domain.

A possible explanation for the performance drop of the DeiT small model in the out-of-domain evaluation, compared to its in-domain accuracy of around 93%, could depend on the intrinsic differences between convolutional neural networks (CNNs) and vision transformers (ViTs). While transformers are capable of capturing global dependencies across image patches, they lack the strong spatial inductive biases of CNNs, such as locality and translation equivariance. As a result, ViTs typically require larger and more diverse datasets to learn robust and transferable representations, and they may overfit more easily to the distribution of the training data. In contrast, ResNet18, with its convolutional architecture, tends to extract more stable local features that generalize better when exposed to new domains with different distributions. This effect becomes particularly evident in our case, where the unseen dataset is smaller, class-imbalanced, and only partially overlaps with the original label space. Under these conditions, DeiT small shows reduced robustness and higher variability across less represented classes, while ResNet18 maintains slightly more consistent performance.

In our previous study [12], we explored in-domain

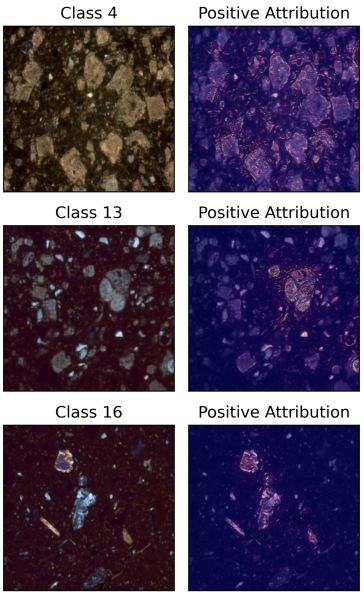

Fig. 4. Guided Grad-CAM results for samples GE-4, AJ-4, and HT-12 belonging to class 4 (Dolomite A fabric), class 13 (Microfossils fabric), and class 16 (Biotite fabric).

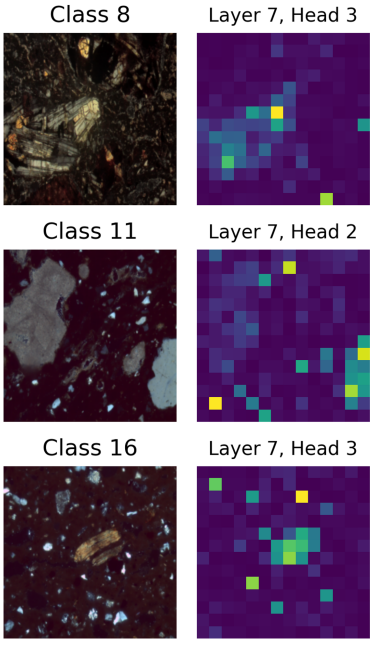

Fig. 5. Attention maps results for samples GE-6, GE-50, and HT-15 (Basalt A fabric), class 11 (Carbonate + Quartz fabric), class 16 (Biotite fabric).

visual explainability methods, applying Guided Grad-CAM and transformer attention maps to ResNet18 and ViT small, respectively, and highlighted petrographically

meaningful characteristics of the micrographs. Building on this evidence, in the present work we extended the explainability analyses to the out-of-domain dataset (Jericho, Al Jib, Tell Hadidi ceramic samples).

As in the earlier study, for ResNet18 we applied Guided Grad-CAM, whereas for DeiT small we analyzed its transformer attention maps. Fig 4 shows the Guided Grad-CAM results for samples GE-4, AJ-4, and HT-12 belonging to class 4 (Dolomite A *fabric*), class 13 (Microfossils *fabric*) and class 16 (Biotite *fabric*), respectively. Fig. 5 presents the DeiT small attention maps for samples GE-6, GE-50, and HT-15 corresponding to class 8 (Basalt A *fabric*), class 11 (Carbonate + Quartz *fabric*), and again class 16 (Biotite *fabric*). The visual explainability maps obtained on the out-of-domain dataset are consistent with those observed in the in-domain setting. This consistency indicates that both models are able to capture and emphasize the petrographic and textural characteristics that are most relevant for the correct classification of the ceramic samples in their petrographic *fabrics*, reinforcing the reliability of their predictions across different domains.

IV. CONCLUSIONS

Overall, both ResNet18 and DeiT small demonstrated comparable performance in the out-of-domain evaluation. However, while DeiT small performed better than ResNet18 in the in-domain setting, its relative drop in performance was more pronounced when exposed to unseen data. This suggests that ResNet18, despite achieving a lower accuracy, may provide greater robustness across domains. These results confirm that models can generalize to previously unseen samples, although targeted strategies to improve predictions on less represented classes could further improve performance. More broadly, the findings highlight the importance of training on datasets with balanced class distributions and sufficient variability. At the same time, they underscore the potential of deep learning for the automated classification of ceramic thin sections. In this context, the integration of explainable AI techniques can provide insights into model behavior, improve transparency and reliability, and support the real-world applicability of these methods in archaeometric research.

V. ACKNOWLEDGMENTS

The authors acknowledge Dr. Kamal Badreshany (Durham University, UK), Dr. Dennis Braekmans (Leiden University, The Netherlands), and The Levantine Ceramic Project group for their assistance in expanding the dataset by providing access to their samples. Their contributions were essential to the completion of this study.

REFERENCES

[1] R. Greenberg, "The archaeology of the bronze age levant." Cambridge University Press, 2019.

[2] M. Steiner and A. E. Killebrew. The Oxford Handbook of the Archaeology of the Levant: c. 8000-332 BCE. Oxford Handbooks, 2014.

[3] R. Greenberg. Traveling in (world) time: transformation, commoditization, and the beginnings of urbanism in the Southern Levant. Oxbow Books, 2011.

[4] A. Guyot, M. Lennon, T. Lorho, and L. Hubert-Moy. Combined Detection and Segmentation of Archeological Structures from LiDAR Data Using a Deep Learning Approach. Journal of Computer Applications in Archaeology, 4(1):1, 2021.

[5] Ø. D. Trier, D. C. Cowley, and A. U. Waldeland. Using deep neural networks on airborne laser scanning data: Results from a case study of semi-automatic mapping of archaeological topography on Arran, Scotland. Archaeological Prospection, 26(2):165–175, 2019.

[6] M. Troiano, E. Nobile, F. Mangini, M. Mastrogiuseppe, C. Conati Barbaro, and F. Frezza. A comparative analysis of the bayesian regularization and levenberg–marquardt training algorithms in neural networks for small datasets: A metrics prediction of neolithic laminar artefacts. Information, 15(5), 2024.

[7] R. Zhang, Y. Cheng, J. Huang, Y. Zhang, and H. Yan. Unsupervised weathering identification of grottoes sandstone via statistical features of acoustic emission signals and graph neural network. Heritage Science, 12(1), 323, (2024).

[8] P. Navarro, C. Cintas, M. Lucena, J. M. Fuertes, C. Delrieux, and M. Molinos. Learning feature representation of Iberian ceramics with automatic classification models. Journal of Cultural Heritage, 48:65–73, 2021.

[9] G. Ruschioni, D. Malchiodi, A. M. Zanaboni, and L. Bonizzoni. Supervised learning algorithms as a tool for archaeology: Classification of ceramic samples described by chemical element concentrations. Journal of Archaeological Science, 49, 2023.

[10] A. Anglisano, L. Casas, I. Queralt, and R. Di Febo. Supervised Machine Learning Algorithms to Predict Provenance of Archaeological Pottery Fragments. Sustainability, 14, 2022.

[11] G. Barone, P. Mazzoleni, G. V. Spagnolo, and S. Raneri. Artificial neural network for the provenance study of archaeological ceramics using clay sediment database. Journal of Cultural Heritage, 38, 147-157, (2019).

[12] M. Lyons. Ceramic Fabric Classification of Petrographic Thin Sections with Deep Learning. Journal of computer applications in archaeology, 4(1):188–201, 2021.

[13] M. Lyons, F. Fecher, and M. Reindel. From LiDAR to deep learning: A case study of computer-assisted approaches to the archaeology of Guadalupe and northeast Honduras. it – Information Technology, 64(6):233–246, 2022.

[14] S. Capriotti, A. Devoto, S. Scardapane, S. Mignardi, and L. Medeghini, (2025). Interpretable Classification of Levantine Ceramic Thin Sections via Neural Networks. Machine Learning: Science and Technology.

[15] D. Castelvecchi. Can we open the black box of AI? Nature, 538(7623):20–23, 2016.

[16] R. Kashefi, L. Barekatain, M. Sabokrou, and F. Aghaeipoor. Explainability of Vision Transformers: A Comprehensive Review and New Perspectives. arXiv preprint arXiv:2311.06786, 2023.

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778., 2016.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, Zhai X., T. Unterthiner, M. Dehghani, M Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929, 2020.

[19] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li and L. Fei-Fei Imagenet: a large-scale hierarchical image database IEEE Conf. on Computer Vision and Pattern Recognition 248–55. 2009.

[20] S. Yang, W. Xiao, M. Zhang, S. Guo, J. Zhao, and F. Shen. Image Data Augmentation for Deep Learning: A Survey., 2023.