

Hands in the Dark: Semantically Rich Explanations for Open-Set Scribal Hand Discovery in Historical Manuscripts

Serena Ammirati¹, Riccardo De Cesaris², Paolo Merialdo^{2,*}, Antonella Poggi³,
Riccardo Torlone² and Alessio Vernarelli³

¹Department of Humanities – Roma Tre University, Italy

²Department of Engineering – Roma Tre University, Italy

³Department of Computer, Control and Management Engineering – Sapienza University of Rome, Italy

Abstract

Historical paleography relies on the qualitative, manual inspection of scribal hands, yet automated handwriting identification has so far remained confined to closed-set classification, producing predictions without inspectable evidence. *Hands in the Dark* is a research programme that aims to shift AI-assisted paleography from closed-set prediction to an explainable, open-set discovery process, in which artificial intelligence surfaces evidence expressed in the paleographer’s own conceptual vocabulary rather than as raw visual saliency. The programme is grounded in a recently developed and validated framework that generates faithful, class-conditional explanations for neural handwriting classifiers, assessed along four complementary axes and against expert paleographic judgement. Building on this foundation, *Hands in the Dark* pursues three research objectives: extending the framework to open-set scribal discovery at the scale of the *Codices Latini Antiquiores*; developing a three-level, ontology-grounded explainability pipeline that moves from segment-based saliency to graphic signatures to natural-language paleographic description; and supporting human-guided historical interpretation of the discovered hands.

Keywords

Explainable AI, Handwriting Identification, Digital Paleography, Historical Manuscripts, Open-Set Discovery, Codices Latini Antiquiores, Ontologies

1. Introduction

Deep neural networks, particularly computer vision models, are increasingly supporting research in the Digital Humanities [1]. *Historical Handwriting Identification* (HI), which aims to attribute a manuscript page (or part of it) to the scribe who produced it, is a particularly relevant example. In traditional paleographic practice, such attribution relies on the manual inspection of recurrent graphic traits—including letterform morphology, ligatures, allographic choices, and stroke execution—interpreted together with codicological and historical evidence [2]. By learning discriminative representations directly from manuscript images, deep learning models can support analyses that would otherwise require highly specialized and extremely time-consuming expert inspection [3, 4].

Their decision processes, however, remain largely opaque [5, 6]. This limitation is particularly critical in paleography, where AI is expected to support rather than replace scholarly judgement. Every attribution must therefore be inspectable, evaluated, and, when necessary, challenged in the light of paleographic knowledge. This need is amplified by the characteristics of medieval manuscripts: visual evidence is often compromised by material degradation, different scribes may deliberately conform to common graphic conventions, and reliable reference attributions are frequently sparse, uncertain, or disputed [7, 8, 9].

4th Italian Workshop on Artificial Intelligence for Cultural Heritage (IAI4CH 2026, <https://ai4ch.di.unito.it/>), co-located with the 24th International Conference of the Italian Association for Artificial Intelligence (AIIA 2026). 6–9 October 2026, Perugia, Italy.

*Corresponding author.

✉ serena.ammirati@uniroma3.it (S. Ammirati); riccardo.decesaris@uniroma3.it (R. De Cesaris);
paolo.merialdo@uniroma3.it (P. Merialdo); antonella.poggi@uniroma1.it (A. Poggi); riccardo.torlone@uniroma3.it
(R. Torlone); vernarelli.1946128@studenti.uniroma1.it (A. Vernarelli)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

To address this challenge, we recently developed a model-independent framework that generates faithful, class-conditional explanations for neural handwriting classifiers and evaluates them through a rigorous validation protocol [10]. Rather than treating explanations as qualitative visual artefacts, the framework assesses them along four complementary dimensions (faithfulness, stability, cross-model agreement, and computational efficiency) and compares them with independent paleographic analyses. These results demonstrate that explainability can provide grounded evidence that scholars can inspect and critically assess, establishing a solid methodological foundation for AI-assisted paleography.

Although this framework demonstrates that faithful explanations are achievable, it also highlights the limitations of the current paradigm. Most existing computational approaches to handwriting identification, including our own previous work [10, 11], formulate the problem as *closed-set* classification, in which the set of candidate scribes is fixed and known in advance. While appropriate for verification tasks, this assumption rarely reflects the way paleographic research is actually conducted. Scholars often begin the analysis of a manuscript without knowing how many scribes contributed to it, where scribal changes occur, or whether the same hand appears across different codices. The problem is therefore not simply to assign pages to predefined identities, but to discover previously unknown scribal groups from visual evidence before historical interpretation can even begin.

A second limitation concerns the form in which explanations are communicated. Current XAI methods typically express explanations as relevance maps over pixels, patches, or image regions [11, 12]. These visualizations indicate *where* a model focuses its attention, but they do not explain *what* graphical characteristics distinguish one scribal hand from another in terms that paleographers naturally use. As AI moves from closed-set classification towards open-set discovery, this semantic gap becomes even more critical: newly discovered groups cannot be validated through class labels alone, but require explanations expressed in a conceptual language that experts can interpret, discuss, and relate to established paleographic knowledge.

Hands in the Dark addresses these two limitations by recasting AI-based handwriting identification as an explainable process of historical discovery. Rather than assuming a predefined set of scribes, the project aims to discover scribal hands at the scale of the *Codices Latini Antiquiores* [13] (CLA), while progressively transforming low-level visual evidence into semantically grounded explanations expressed through a paleographic ontology and natural language. In this perspective, explainability is no longer an accessory to classification, but the mechanism through which AI becomes a credible instrument for scholarly reasoning. This paper presents:

- The vision and research agenda of *Hands in the Dark*, a programme that shifts AI-assisted paleography from closed-set prediction to explainable open-set discovery;
- The validated explainability framework that provides the project’s methodological starting point;
- Three research objectives—open-set scribal discovery, semantically grounded explainability, and human-guided historical interpretation—together with a three-year research roadmap.

2. A Validated Foundation for Explainable Handwriting Identification

We developed a model-independent framework that generates page-level explanations for neural handwriting classifiers. Given a manuscript page and a target scribe, the framework produces a signed relevance map (Figure 1, left) highlighting the visual evidence that supports or counteracts that attribution, computed through perturbation-based analysis (*GLIME-Binomial* [14]) over spatial patches. We focus on this explainer because it is model-agnostic and estimates the contribution of individual input regions, providing a natural point of contact with paleographic practice. Rather than treating explanations as qualitative visual artefacts, the framework subjects them to a multi-dimensional validation protocol comprising quantitative evaluation—*faithfulness* [15], *stability*, *cross-model agreement*, and *computational efficiency*—together with an independent qualitative assessment by expert paleographers. The framework was validated using two architecturally distinct classifiers (a ResNet-18 [16] and

a Swin Transformer Tiny [17]) on four datasets derived from digitized Latin manuscripts, summarized in Table 1: three codices from the Biblioteca Apostolica Vaticana and the *Chelles* dataset, whose ten scribal hands are documented through the copyists’ own subscriptions [18].

Table 1

Datasets used for empirical validation.

Dataset	Classes	Pages	Period / Script
Vat.lat.653	4	208	11th c., <i>romanesca</i>
Vat.lat.5951	3	310	early 9th c., pre-Caroline
Vat.lat.4221	4	280	mid-12th c., <i>carolina reformata</i>
Chelles	10	1,585	c. 800, documented hands

The empirical evaluation demonstrates that the framework provides explanations that are both behaviourally grounded and suitable for scholarly inspection. In particular, the quantitative evaluation and the expert assessment lead to the following conclusions:

- **Faithfulness.** Masking the regions identified as most relevant degrades classification accuracy substantially more than masking randomly selected regions, consistently across datasets and architectures, indicating that the explanations capture evidence genuinely used by the models.
- **Stability.** Repeated executions of the stochastic explainer produce highly reproducible attribution maps under fixed conditions, supporting the reliability of the generated explanations.
- **Cross-model agreement.** Architecturally different classifiers reaching the same prediction often rely on different spatial evidence, suggesting that explanations should be regarded as model-specific rather than model-independent.
- **Expert assessment.** Compared with an independent paleographic study of the Chelles corpus [18], the framework’s most relevant regions partially, though not systematically, coincide with documented diagnostic traits. For example, for the scribe Vera, the highlighted regions frequently recover her characteristic cursive *e-r* ligatures (Figure 1, right).

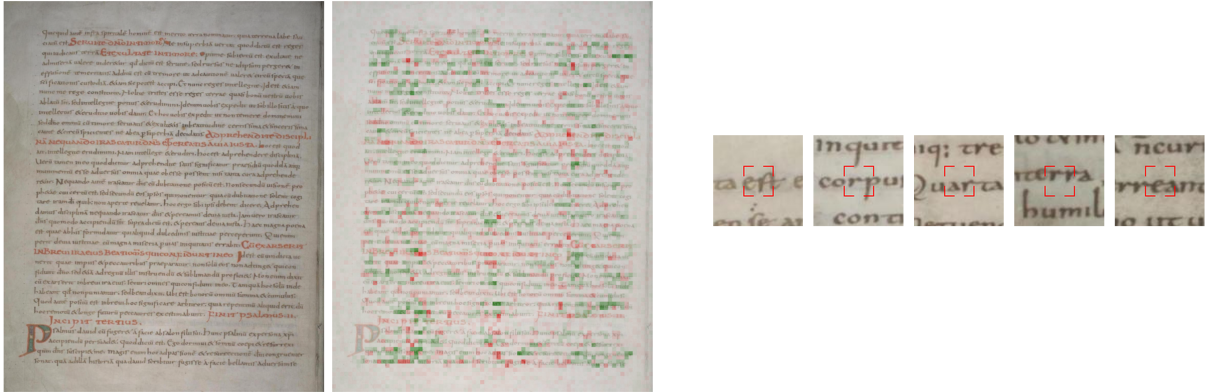


Figure 1: (Left): Grid-based explanation: manuscript page (f. 4v, Köln 63, Chelles) and rectangular-patch attribution map, green/red marking positive/negative relevance. (Right): Stroke-level detail: cursive *e-r* ligatures for scribe Vera (Scribe 9), matching a trait documented in the paleographic literature.

These results establish that the framework produces explanations that are behaviourally grounded, reproducible, and paleographically meaningful, providing the methodological foundation on which *Hands in the Dark* is built.

3. Hands in the Dark

The validated framework presented in Section 2 demonstrates that faithful explanations can be generated and rigorously assessed for supervised handwriting identification. *Hands in the Dark* builds

on this foundation to address a substantially broader challenge: transforming AI-assisted handwriting identification from a closed-set classification task into an explainable process of historical discovery.

Achieving this goal requires extending the current framework along two complementary directions. First, handwriting identification must move beyond supervised classification towards *open-set discovery*, where neither the number nor the identity of scribal hands is known beforehand. Second, explanations must evolve from low-level visual saliency into semantically meaningful descriptions expressed through the concepts and terminology of paleography. These advances enable AI not merely to classify manuscript pages, but to support the formulation, validation, and refinement of historical hypotheses.

The project focuses on the *Codices Latini Antiquiores* (CLA), the most comprehensive census of Latin manuscripts produced before 800 CE [13]. This corpus comprises more than 2,000 manuscripts, for which the number of scribal hands is frequently unknown or disputed, reference annotations are sparse, and substantial visual variability coexists with strong similarities arising from common scribal traditions. In this setting, the problem is no longer assigning pages to predefined classes, but discovering previously unknown scribal groups and providing interpretable evidence supporting such discoveries.

The overall research workflow is summarized in Figure 2. Starting from the unlabeled CLA corpus, the project progressively transforms visual evidence into increasingly richer representations, culminating in ontology-grounded descriptions that can be directly inspected and discussed by paleographers.

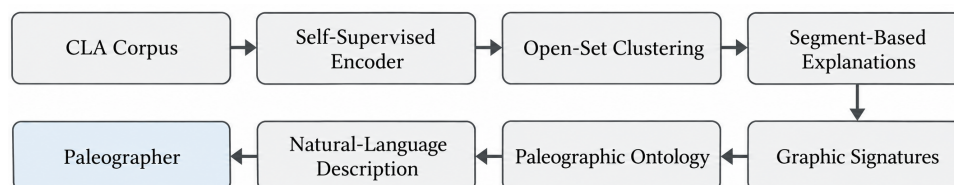


Figure 2: The *Hands in the Dark* workflow: from the unlabeled CLA corpus to open-set discovery, progressively richer explanations, and human interpretation.

The programme is organized around three closely connected research objectives.

3.1. Research Objective 1: From Closed-Set Classification to Open-Set Discovery

The first objective is to overcome the closed-set assumption underlying current handwriting-identification systems. While the framework presented in Section 2 explains the predictions of a supervised classifier, paleographic research rarely starts from a predefined set of candidate scribes. Instead, scholars must determine how many hands contributed to a manuscript, identify transitions between them, and recognize recurring hands across different manuscripts.

To address this challenge, we will extend the self-supervised representation learning strategy introduced by Lastilla et al. [4] from individual manuscripts to the entire CLA corpus. Corpus-specific pretraining on the unlabeled collection will be combined with open-set clustering techniques, including deep-clustering and density-based approaches, capable of automatically determining the number of scribal groups while coping with severe class imbalance and manuscripts represented by only a few pages. Whenever available, expert annotations from projects such as CLASSICS¹ and ELMSS² will provide weak supervisory signals for cluster validation and refinement.

3.2. Research Objective 2: From Visual Saliency to Semantically Grounded Explanations

Moving from supervised classification to open-set discovery substantially changes the role of explainability. In the closed-set setting, explanations justify why a page has been assigned to one known class rather than another. In the open-set setting, explanations must instead justify why previously

¹<https://studiumanistici.uniroma3.it/ricerca/progetti-di-ricerca/classics-codices-latini-antiquiores-supplements-scripts-in-depth-through-computer-sciences/>

²<https://elmss.nuigalway.ie/about>

ligatures, abbreviations, writing conventions, and material characteristics. The ontology will be formalized in OWL⁵/RDF⁶, enabling interoperability with existing cultural-heritage resources while supporting reasoning over the generated descriptions and facilitating integration with projects such as CLASSICS and ELMSS.

3.3. Research Objective 3: From Explainability to Historical Interpretation

The ultimate objective of *Hands in the Dark* is not simply to discover clusters of visually similar manuscript pages, but to support historical reasoning.

The graphic signatures and ontology-grounded descriptions generated by the previous stages will be systematically compared with established paleographic knowledge, leading to three complementary outcomes. First, they may provide *confirmation* by recovering diagnostic features already recognized by paleographers. Second, they may enable *discovery* by revealing previously undocumented graphical regularities shared by groups of manuscripts. Finally, they may support *failure analysis*, where discrepancies between computational evidence and expert interpretation reveal limitations of either the computational models or current scholarly understanding. Through this iterative interaction between AI and domain expertise, computational explanations become the medium through which visual evidence contributes to the reconstruction of scribal practices, writing-centre organization, and manuscript transmission.

4. Related Work

Handwriting identification. Computational handwriting identification has evolved from hand-crafted descriptors [21] processed through methods such as k -NN and SVMs to deep-learning approaches for representation learning [22], classification of pre-extracted features [12], and end-to-end recognition [3, 4]. Our own prior work [4] trains a ResNet-18 [16] on medieval manuscripts via self-supervised Siamese pre-training and triplet-loss fine-tuning, and is the classification backbone extended here.

Explainable AI. Post-hoc feature-based explainability methods can be broadly grouped in gradient- and backpropagation-based techniques—such as Saliency Maps [23], Grad-CAM [24], Integrated Gradients [25], and transformer extensions like Attention Rollout [26], and LeGrad [27]—and model-agnostic perturbation-based techniques, including Occlusion [28], RISE [29], LIME [30], and SHAP [31].

XAI in Digital Humanities. Explainability for historical document analysis is emerging. Prior work on script-type identification and writer profiling (clustering over learned descriptors [32], bigram feature importance [33], template-similarity [12], heatmaps [34], saliency-based dating [35]) yields diagnostic artefacts valuable for hypothesis generation but rarely assessed for model faithfulness. Our work is among the first to jointly evaluate CNN- and transformer-based writer identification under multi-dimensional validation [10], and, next, the first to extend this to *open-set* discovery on a pre-Carolingian corpus, so far addressed, retrieval-only and without XAI, only by Monk (Groningen).⁷

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) for paraphrasing and rewording, improving writing style, grammar and spelling checking, and L^AT_EX formatting assistance. The tool was not used to generate original scientific content, figures, results, analyses, conclusions, or research ideas. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

⁵<https://www.w3.org/TR/2012/REC-owl2-primer-20121211/>

⁶<https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/>

⁷<https://www.ai.rug.nl/~lambert/Monk-collections-english.html>

References

- [1] E. Nieddu, D. Firmani, P. Merialdo, M. Maiorino, In codice ratio: A crowd-enabled solution for low resource machine transcription of the vatican registers, *Information Processing & Management* 58 (2021) 102606.
- [2] P. A. Stokes, Digital approaches to paleography and book history: Some challenges, present and future, *Frontiers in Digital Humanities* 2 (2015) 5. doi:10.3389/fdigh.2015.00005.
- [3] N. D. Cilia, C. De Stefano, F. Fontanella, C. Marrocco, M. Molinara, A. Scotto Di Freca, An end-to-end deep learning system for medieval writer identification, *Pattern Recognition Letters* 129 (2020) 137–143.
- [4] L. Lastilla, S. Ammirati, D. Firmani, N. Komodakis, P. Merialdo, S. Scardapane, Self-supervised learning for medieval handwriting identification: A case study from the vatican apostolic library, *Information Processing & Management* 59 (2022) 102875.
- [5] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, *arXiv preprint arXiv:1910.10045* 1 (2019).
- [6] R. Pradhan, A. Lahiri, S. Galhotra, B. Salimi, Explainable ai: Foundations, applications, opportunities for data management research, in: *Proceedings of the 2022 international conference on management of data*, 2022, pp. 2452–2457.
- [7] A. Derolez, *The palaeography of Gothic manuscript books: From the twelfth to the early sixteenth century*, Cambridge University Press, 2003.
- [8] M. Parkes, *Their Hands Before Our Eyes: A Closer Look at Scribes : the Lyell Lectures Delivered in the University of Oxford, 1999*, Lyell lectures in bibliography, Ashgate, 2008. URL: <https://books.google.it/books?id=NCQrYGWCAEUC>.
- [9] A. S.-B. de Port-Valais (Bouveret), S. Saint-Benoît de Port-Valais (Abbey : Bouveret, *Colophons de manuscrits occidentaux des origines au XVIe siècle*, number v. 4 in *Colophons de manuscrits occidentaux des origines au XVIe siècle*, Ed. universitaires, 1965. URL: <https://books.google.it/books?id=LLci0V5Zh5kC>.
- [10] R. De Cesaris, V. Caravani, P. Merialdo, S. Ammirati, Generating and assessing page-level explanations for neural historical handwriting identification, 2026. URL: <https://ssrn.com/abstract=7286720>. doi:10.2139/ssrn.7286720, available at SSRN.
- [11] R. De Cesaris, V. Caravani, A. Pastorini, S. Ammirati, P. Merialdo, et al., Preliminary results for the explanation of neural network-based handwriting identification in historical manuscripts, in: *Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale*, 2025, pp. 386–391.
- [12] S. Grieggs, C. Henderson, S. Sobacki, A. Gillespie, W. Scheirer, The paleographer's eye ex machina: Using computer vision to assist humanists in scribal hand identification, in: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 7177–7186.
- [13] E. A. Lowe, *Codices Latini Antiquiores: A Palaeographical Guide to Latin Manuscripts Prior to the Ninth Century*, Clarendon Press, Oxford, 1934–1971. 12 vols.
- [14] Z. Tan, Y. Tian, J. Li, Glime: general, stable and local lime explanation, *Advances in neural information processing systems* 36 (2023) 36250–36277.
- [15] P. Atanasova, J. G. Simonsen, C. Lioma, I. Augenstein, A diagnostic study of explainability techniques for text classification, in: *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, 2020, pp. 3256–3274.
- [16] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [17] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012–10022.
- [18] L. Pani, *Un lavoro di squadra: le copiste delle enarrationes in psalmos (köln, dombibliothek, 63,*

- 65, 67; berlin, staatsbibliothek, phillipps 1657), in: Dalla parte di lei. Scritti per Laura Casella dalle amiche del Dipartimento, Forum, 2025, pp. 91–105.
- [19] P. A. Stokes, Modeling medieval handwriting: A new approach to digital palaeography, in: ADHO 2012-Hamburg, 2012.
 - [20] S. Brookes, P. A. Stokes, M. Watson, D. M. De Matos, The digipal project for european scripts and decorations, *Essays and Studies* 68 (2015) 25–59.
 - [21] H. Bay, T. Tuytelaars, L. Van Gool, Surf: Speeded up robust features, in: European conference on computer vision, Springer, 2006, pp. 404–417.
 - [22] S. Fiel, R. Sablatnig, Writer identification and retrieval using a convolutional neural network, in: International conference on computer analysis of images and patterns, Springer, 2015, pp. 26–37.
 - [23] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, *arXiv preprint arXiv:1312.6034* (2013).
 - [24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: visual explanations from deep networks via gradient-based localization, *International journal of computer vision* 128 (2020) 336–359.
 - [25] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International conference on machine learning, PMLR, 2017, pp. 3319–3328.
 - [26] S. Abnar, W. Zuidema, Quantifying attention flow in transformers, in: Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 4190–4197.
 - [27] W. Bousselham, A. Boggust, S. Chayboudi, H. Strobelt, H. Kuehne, Legrad: An explainability method for vision transformers via feature formation sensitivity, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 20336–20345.
 - [28] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European conference on computer vision, Springer, 2014, pp. 818–833.
 - [29] V. Petsiuk, A. Das, K. Saenko, Rise: Randomized input sampling for explanation of black-box models, *arXiv preprint arXiv:1806.07421* (2018).
 - [30] M. T. Ribeiro, S. Singh, C. Guestrin, Why should i trust you? explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.
 - [31] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, *Advances in neural information processing systems* 30 (2017).
 - [32] M. Kestemont, V. Christlein, D. Stutzmann, Artificial paleography: computational approaches to identifying script types in medieval manuscripts, *Speculum* 92 (2017) S86–S109.
 - [33] C. Vandyck, W. Haverals, M. Kestemont, Making characters count. a computational approach to scribal profiling in 14th-century middle dutch manuscripts from the carthusian monastery of herne, *arXiv preprint arXiv:2509.00067* (2025).
 - [34] M. Popović, M. A. Dhali, L. Schomaker, Artificial intelligence based writer identification generates new evidence for the unknown scribes of the dead sea scrolls exemplified by the great isaiah scroll (1qisaa), *PloS one* 16 (2021) e0249769.
 - [35] J. Pavlopoulos, M. Konstantinidou, E. Perdiki, I. Marthot-Santaniello, H. Essler, G. Vardakas, A. Likas, Explainable dating of greek papyri images, *Machine Learning* 113 (2024) 6765–6786.