Preliminary Results for the Explanation of Neural Network-based Handwriting Identification in Historical Manuscripts

Riccardo De Cesaris[1], Valerio Caravani[2], Arianna Pastorini[3], Serena Ammirati[4], Paolo Merialdo[5]
[1] Università degli Studi "Roma Tre", Italy – riccardo.decesaris@uniroma3.it
[2] Università degli Studi "Roma Tre", Italy – valerio.caravani@uniroma3.it
[3] Università di Bologna, Italy – arianna.pastorini2@unibo.it
[4] Università degli Studi "Roma Tre", Italy – serena.ammirati@uniroma3.it
[5] Università degli Studi "Roma Tre", Italy – paolo.merialdo@uniroma3.it

ABSTRACT (ENGLISH)
This work tackles the "black-box" problem in Digital Paleography—specifically in Handwriting Identification—by introducing a tailored explainability approach designed to enhance the interpretability of AI-based systems in manuscript analysis. As a novel contribution, we apply Explainable Artificial Intelligence (XAI) techniques to the case study of "Vat.lat.653", an 11th-century manuscript preserved in the Vatican Apostolic Library. Leveraging a Convolutional Neural Network (CNN)-based classifier, we adapt LIME, a well-established explanation method, to identify the most influential features on each page with respect to the model's decisions. The present study represents one of the first attempts to systematically integrate XAI methods into paleographic analysis of historical handwritten texts, providing tools and methodology to assess the reliability and consistency of automated systems, helping promote their use in this domain. Although our findings provide insights into the potential of XAI methods for paleographic analysis, they remain preliminary and require further validation in order to fully assess the effectiveness and generalizability of the proposed approach.

Keywords: Explainable AI; Handwriting Identification; Digital Paleography

ABSTRACT (ITALIANO)
Risultati Preliminari per l'interpretazione dell'Identificazione della Grafia effettuata con Reti Neurali su Manoscritti Storici.

Il presente lavoro affronta il problema "black-box" nella Paleografia Digitale—in particolare nell'Identificazione della Grafia—introducendo un approccio progettato su misura per migliorare l'interpretabilità dei sistemi basati su IA nell'analisi dei manoscritti. Come contributo originale, applichiamo tecniche di Explainable Artificial Intelligence (XAI) al caso di studio del manoscritto "Vat.lat.653", un documento dell'XI secolo conservato nella Biblioteca Apostolica Vaticana. Utilizzando un classificatore basato su Reti Neurali Convoluzionali (CNN), adattiamo LIME, un metodo consolidato per la generazione di spiegazioni, al fine di identificare le caratteristiche più influenti in ogni pagina rispetto alle decisioni del modello. Questo studio rappresenta uno dei primi tentativi di integrare sistematicamente metodi XAI nell'analisi paleografica di manoscritti storici, fornendo strumenti e metodologie per valutare l'affidabilità e la coerenza dei sistemi automatici, promuovendone l'adozione in questo ambito. Sebbene i risultati offrano spunti sul potenziale delle tecniche XAI applicate all'analisi paleografica, rimangono preliminari e richiedono ulteriori validazioni per una valutazione completa dell'efficacia e della generalizzabilità dell'approccio proposto.

Parole Chiave: Intelligenza Artificiale Interpretabile; Identificazione della Grafia; Paleografia Digitale

1. INTRODUCTION
In recent years, the exploitation of Deep Learning techniques for the automated analysis of both historical and modern manuscripts has gathered increasing interest, offering the potential of both simplifying and significantly improving the work of paleographers and domain experts (Hu, Wang, Li, & Wang, 2021; Nieddu, Firmani, Merialdo, & Maiorino, 2021). Among the various tasks that can be addressed using such approaches there is "Handwriting Identification", which consists in partitioning texts into parts which can be attributed to different scribes, each one characterized by his own writing style.

However, the adoption of Deep Learning models raises the critical challenge known as the "black-box problem" (Stokes, 2015), one of the most debated issues in the deployment of AI systems in real-world applications. Briefly, this problem concerns the lack of transparency in model decision-making, where

outcomes are typically expressed only through performance metrics, limiting their interpretability and usefulness for non-technical users. Addressing this limitation requires developing explanation strategies capable of making model predictions both understandable and meaningful to experts in the target domain.

In this work, we present an original contribution to the field of Digital Paleography by integrating Explainable Artificial Intelligence (XAI) techniques into the automated analysis of historical manuscripts. Among the numerous approaches (Pradhan, Lahiri, Galhotra, & Salimi, 2022) for interpreting learning-based models, our methodology leverages LIME (Ribeiro, Singh, & Guestrin, 2016), a well-known "feature-based" explainability technique. LIME's straightforward strategy succeeds in explaining a prediction provided by the original classifier leveraging an inherently interpretable model (with a known decision function) whose purpose is to approximate the behavior of the original classifier within a local region around the instance to be explained itself. For this reason, LIME can be defined as "model-agnostic", meaning it can be applied to any type of classifier, regardless of its complexity or architectural details.

Through the adaptation of LIME to the Handwriting Identification problem, we aim to identify the portions of the manuscript that most influence a CNN-based classifier (Lastilla, Ammirati, Firmani, Komodakis, Merialdo, & Scardapane, 2022) trained to attribute pages to individual scribes: in doing so, we are able to provide concrete and visual explanations for model's decisions. While the proposed approach shows promising results—with the generated explanations effectively reflecting the model's decision-making process for the task of interest—its broader effectiveness and generalizability need further validation across additional experiments and case studies.

2. CASE STUDY

Our case study focuses on "Vat.lat.653"[1], an 11th-century manuscript preserved in the Vatican Apostolic Library. The manuscript comprises 538 pages (or 269 folios) and was selected from the collection of manuscripts employed in (Lastilla, Ammirati, Firmani, Komodakis, Merialdo, & Scardapane, 2022), primarily due to the high classification performance achieved on its scribes, which forms a strong basis for a reliable explainability analysis. "Vat.lat.653" contains the *Expositio in Epistulas Pauli* by Aimon of Auxerre and was copied in the so-called "minuscola romanesca", a variant of Carolingian minuscule typical of Rome and its surroundings from the 10th to the 12th century; it was produced in the Monastery of St. Scholastic in Subiaco. In 522 out of the 538 pages, paleographers identified four distinct scribal hands contributing to the manuscript's composition. Assuming each page was written by a single scribe, the page distribution is reported below:

- **Scribe 0**, 42 pages: from folio 1 *recto* (r) to folio 21 *verso* (v);
- **Scribe 1**, 58 pages: from folio 237 *recto* (r) to folio 265 *verso* (v);
- **Scribe 2**, 202 pages: from folio 22 *recto* (r) to folio 121 *verso* (v);
- **Scribe 3**, 220 pages: from folio 122 *recto* (r) to folio 231 *verso* (v);

Each scanned page measures 902x1279 pixels, with the text organized into two columns per page, each containing approximately thirty lines.

3. METHODOLOGY

The original classifier (Lastilla, Ammirati, Firmani, Komodakis, Merialdo, & Scardapane, 2022) involved in our experiments consists of a ResNet18 (He, Zhang, Ren, & Sun, 2016) backbone encoder followed by two fully-connected linear layers. The model has been fine-tuned[2] on 380x380-pixel crops extracted from the "Vat.lat.653" pages with the objective of performing a multi-label classification task. The system achieved an overall accuracy greater than 86% on the Test Set, providing a sufficiently reliable foundation for the subsequent explainability analysis.

As outlined in the previous sections, the aim of the proposed explainability strategy is to evaluate the relevance of page sub-regions that the model relies on at inference time. In LIME terminology, these sub-

[1] Available Online at: http://www.mss.vatlib.it/guii/console?service=present&term=@5Vat.lat.653_ms&item=1&add=0&search=1&filter=&relation=3&operator=&attribute=3040

[2] The initial weights of the backbone encoder are the result of the experiments conducted in (Lastilla, Ammirati, Firmani, Komodakis, Merialdo, & Scardapane, 2022)

regions are referred to as "super-pixels": in our implementation, each super-pixel measures 50x50 pixels and enables the analysis of feature associated with short text sequences (2-3 consecutive characters) as well as features spanning across two consecutive lines of text. Given an instance to be explained, LIME approximates the behavior of the original classifier within a local region around the instance by training an inherently interpretable model.

To achieve this, LIME generates a set of random samples (referred to as "neighboring points") by applying random modifications to the features of the original input instance. These perturbed samples are fed into the original classifier to obtain their corresponding outputs. Finally, using the generated neighboring points and their associated responses from the original black-box model, LIME trains an interpretable model that serves as a local approximation of the original one. In our context, LIME locally approximates the original classifier's behavior for 380x380-pixel crops extracted from the instance to be explained, assigning an importance score to each super-pixel. The scores are normalized within the range [-1, 1]: for each super-pixel, a positive value indicates that the corresponding super-pixel contributes positively to the correct classification, while negative values suggest that its information misleads the model's prediction.

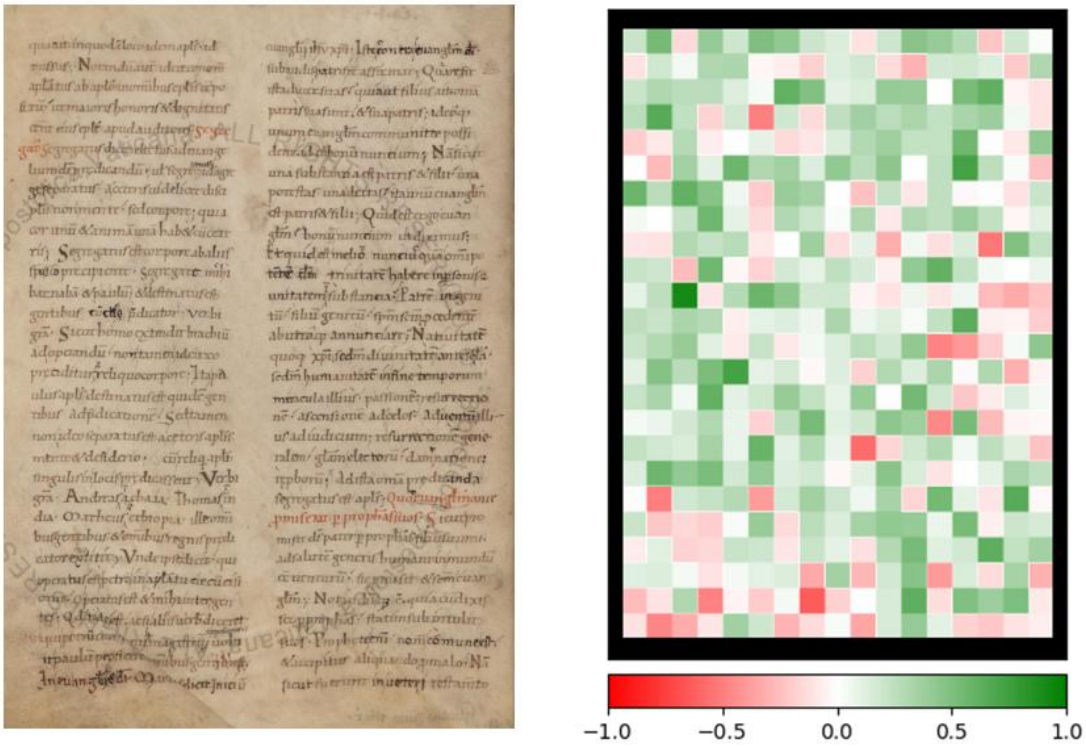

Figure 1. (i) Folio 1 *verso* – (ii) Its attribution Map

To identify the most important super-pixel at the level of a single page, we collected a set of overlapping 380x380 crops covering the entire page using a sliding window mechanism.
For the instance to be explained, its "attribution map" (Figure 1) is constructed by aggregating the attribution scores of its "super-pixels": specifically, for each "super-pixel" present on the page, its attribution score is calculated as the average of the scores obtained from all crops that include the corresponding sub-region. To reduce possible fluctuations (Zhang, Song, Sun, Tan, & Udell, 2019; Tan, Tian, & Li, 2024) in the importance score values caused by the random perturbation process, the "attribution map" is computed by averaging the results over 3 iterations.

In the attribution maps, green regions correspond to super-pixels deemed relevant by LIME, red regions indicate misleading super-pixels whose content negatively affects the prediction, and white areas represent information that appears to have no significant impact on the model's decision.

This approach enables domain experts to visually identify the sequences considered most influential by the classifier, assess their paleographic relevance, investigate potential parallels between automated and human methodologies, and refine the process of identifying different writing styles within a document.

Furthermore, the identified features can be aggregated across multiple pages to evaluate the presence of character sequences that consistently influence the prediction of a specific scribe.

4. EXPERIMENTAL RESULTS & VALIDATION

The proposed methodology was validated through two orthogonal procedures:

- **Training Set Masking & Re-Training**: starting from the explanations computed for the Training Set instances, the "super-pixels" estimated as the most relevant were removed. The system was then re-trained using the "masked" version of the Training Set instances, and the system's accuracy on the Test Set was finally re-computed.
- **Test Set Masking & Accuracy Re-Computation**: starting from the explanations computed for the Test Set instances, the "super-pixels" estimated as the most relevant were removed. The system accuracy on the Test Set was then re-computed.

Training Set Masking & Re-Training: for each Training Set instance, we incrementally removed the most relevant crops, masking the top 5%, 10% and 20% of the total area of a given instance.
After performing this "masking process", we re-trained the original classifier and subsequently measured again the overall accuracy on the Test Set. The results obtained (Table 1) are reported below:

Masked Area	Overall Accuracy on Test Set
0% (baseline)	86,36%
5%	78,41%
10%	73,86%
20%	64,77%

Table 1. Accuracy Measures for baseline and re-trainings

Test Set Masking & Accuracy Re-Computation: for each Test Set instance, we incrementally removed the most relevant crops (Atanasova, Simonsen, Lioma, & Augenstein, 2020), masking from the top 1% to the top 20% (with a step size of 1%) of the total area of a given instance. After performing this "masking process", we provided the modified Test Set as input to the original classifier and subsequently re-measured the overall accuracy on these instances. Figure 2 highlights how the system's performance (Test Accuracy) degrades as the proportion of masked relevant information increases.

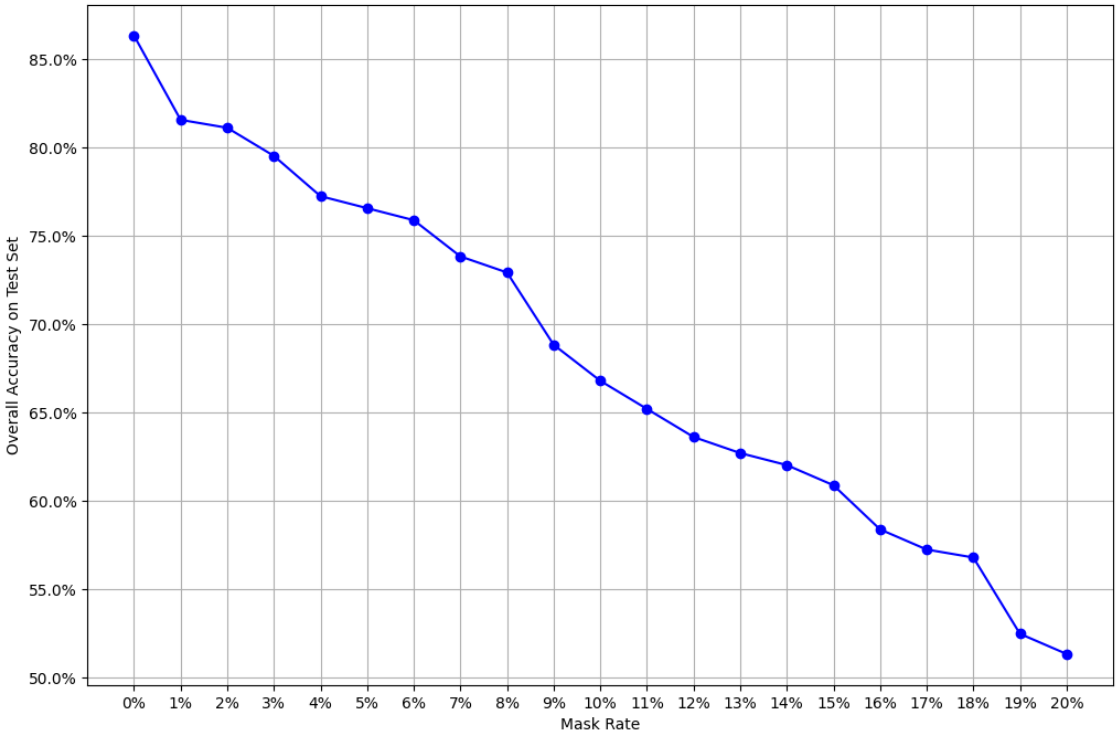

Figure 2. Relationship between Mask Rates and Test Accuracies (in percentage)

The experiments conducted and the corresponding outcomes highlight the effectiveness of the proposed explainability approach in identifying the features critical to the model decision-making process. More precisely, the observed performance degradation upon removing the information deemed relevant by LIME further confirms that the system genuinely relies on these features to make its predictions.

5. CONCLUSIONS AND FUTURE WORK

The approach proposed in this work tackles the "black-box" problem within the field of Digital Paleography, specifically concerning the Handwriting Identification task. Our technique enables the identification of a limited set of relevant features for each page under study, providing new insights into the interpretability of AI-based systems in manuscript analysis. By applying this method across multiple pages, it is also possible to uncover additional distinctive features that may have a greater impact on the classifier's prediction for a given scribe.

The results obtained show promising potential, yet they remain preliminary. As part of future work, further validation is needed to confirm the effectiveness and reliability of the proposed approach. The technique will be extended and tested on different models and datasets to evaluate its generalizability. Additionally, we aim to leverage the insights gained from the explainability process itself to improve the classifier's performance and refine its predictions in subsequent iterations.

REFERENCES

Atanasova, P., Simonsen, J.G., Lioma, C., Augenstein, I. (2020). A Diagnostic Study of Explainability Techniques for Text Classification. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). https://doi.org/10.18653/v1/2020.emnlp-main.263

He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR.2016.90

Hu, P., Wang, W., Li, Q., & Wang, T. (2021). Touching text line segmentation combined local baseline and connected component for Uchen Tibetan historical documents. Information Processing & Management, Volume 58, Issue 6. https://doi.org/10.1016/j.ipm.2021.102689

Lastilla, L., Ammirati, S., Firmani, D., Komodakis, N., Merialdo, P., Scardapane, S. (2022). Self-supervised learning for medieval handwriting identification: A case study from the Vatican Apostolic Library. Information Processing & Management, Volume 59, Issue 3. https://doi.org/10.1016/j.ipm.2022.102875

Nieddu, E., Firmani, D., Merialdo, P., Maiorino, M. (2021). In Codice Ratio: A crowd-enabled solution for low resource machine transcription of the Vatican Registers. Information Processing & Management, Volume 58, Issue 5. https://doi.org/10.1016/j.ipm.2021.102606

Pradhan, R., Lahiri, A., Galhotra, S., Salimi, B. (2022). Explainable AI: Foundations, Applications, Opportunities for Data Management Research. In Proceedings of the 2022 International Conference on Management of Data (SIGMOD '22). https://doi.org/10.1145/3514221.3522564

Ribeiro, M.T., Singh, S., Guestrin, C. (2016). "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 (KDD '16). https://doi.org/10.1145/2939672.2939778

Stokes, P.A. (2015). Digital Approaches to Paleography and Book History: Some Challenges, Present and Future. Frontiers in Digital Humanities 2, 5. https://doi.org/10.3389/fdigh.2015.00005

Tan, Z., Tian, Y., & Li, J. (2024). GLIME: general, stable and local LIME explanation. Advances in Neural Information Processing Systems, *36*. https://proceedings.neurips.cc/paper_files/paper/2023/hash/71ed042903ed67c7f6355e5dd0539eec-Abstract-Conference.html

Vatican Apostolic Library: Website of the Vatican Apostolic Library. https://www.vaticanlibrary.va/en/

Zhang, Y., Song, K., Sun, Y., Tan, S., & Udell, M. (2019). "Why should you trust my explanation?" Understanding Uncertainty in LIME explanations. arXiv preprint arXiv:1904.12991. https://arxiv.org/abs/1904.12991