

## DEEP LEARNING APPROACHES FOR HISTORICAL MANUSCRIPT DIGITIZATION: A SCOPING REVIEW OF EMERGING TRENDS AND RESEARCH GAPS

Muhammad Iqbal Ripo Putra\*, Muhammad Zikri Wiguna  
IKIP PGRI Pontianak, Jalan Ampera no 88, Indonesia  
\*correspondence email: [ripoputra87@gmail.com](mailto:ripoputra87@gmail.com)

Received: 27th November 2023

Accepted: 20th January 2024

Published: 30th January 2024

---

### Abstract

Historical manuscript digitization has been transformed by deep learning, replacing rule-based approaches in binarization, layout analysis, handwritten text recognition, writer identification, and restoration. However, recent evidence remains fragmented across workflows and manuscript traditions. This scoping review aims to characterize deep-learning approaches in manuscript digitization, map the geographic and linguistic distribution of studies, identify methodological trends, and reveal research gaps. Following PRISMA-ScR guidelines and the Population–Concept–Context framework, literature from six databases and three repositories published between 2015 and February 2026 was reviewed. Findings reveal seven major themes: image binarization, layout analysis, handwritten text recognition, writer identification, restoration and virtual unwrapping, OCR-free document understanding, and low-resource transfer learning. Research is dominated by Latin, Arabic, and Chinese manuscripts, while Southeast Asian and other underrepresented scripts receive limited attention. Deep learning has substantially improved digitization workflows, yet challenges remain in annotated corpora, evaluation consistency, reproducibility, and interdisciplinary integration with paleographic scholarship.

**Keywords:** historical manuscripts, deep learning, manuscript digitization

---

### INTRODUCTION

Historical manuscripts are non-renewable cultural assets that constitute much of what is recoverable about the textual past. National libraries, monastic archives, university collections, and private estates hold tens of millions of folios in scripts ranging from Latin and Greek through Arabic, Hebrew, and Syriac to Devanagari, Chinese, and the indigenous scripts of Africa, Mesoamerica, and Southeast Asia. Digitization programmes pursued by institutions such as the Vatican Apostolic Library, the British Library, Gallica at the Bibliothèque nationale de France, and a growing constellation of regional and community-led archives have over the past two decades converted a substantial proportion of these holdings into high-resolution digital images. The bottleneck has shifted: imaging is no longer the binding constraint, but transcription, transliteration, and structural enrichment remain prohibitively expensive when performed manually.

Deep learning has emerged as the dominant computational paradigm for closing this last-mile gap. Convolutional neural networks (CNNs) replaced handcrafted features for character classification in the mid-2010s. Convolutional recurrent neural networks (CRNNs) with Connectionist Temporal Classification loss made end-to-end handwritten text recognition (HTR) practical at the line level. Encoder-decoder transformer models such as TrOCR pushed accuracy further on diverse handwriting styles. Most recently, large vision-language models (VLMs) and OCR-free document understanding architectures such as Donut have begun to reframe digitization as a multimodal reasoning task rather than a pixel-to-character mapping. Specialized branches of research have addressed binarization of degraded substrates, layout analysis for complex page geometries, paleographic dating, writer identification, and even the virtual unwrapping of carbonized scrolls demonstrated by the Vesuvius Challenge in 2024.

Despite this proliferation of methods, the field lacks a recent, methodologically transparent synthesis that maps deep-learning approaches across the full manuscript digitization workflow rather than focusing on a single task. Existing surveys are typically scoped to HTR alone, to a single language family, or to a methodological subfield such as binarization. A scoping review is therefore appropriate: it can map the breadth of activity, characterize emerging methodological trends, and identify the research gaps that prevent generalization from well-resourced scripts to under-represented ones.

The objectives of this scoping review are to characterize the range and depth of deep-learning approaches applied to historical manuscript digitization, to map the geographic, linguistic, chronological, and substrate-related coverage of the evidence base, to identify methodological trends that have emerged in the past three years, and to articulate the research gaps that constrain progress on under-represented scripts and damage profiles. These objectives are operationalized through four research questions:

1. Which deep-learning architectures and training paradigms are currently applied to the principal sub-tasks of historical manuscript digitization, including image enhancement, layout analysis, handwritten text recognition, writer identification, and restoration?
2. How is the empirical evidence distributed across scripts, languages, geographies, time periods, and material substrates, and which categories are most clearly under-represented?
3. What methodological trends have emerged between 2023 and 2026, particularly with respect to vision-language models, diffusion-based data augmentation, self-supervised pretraining, and parameter-efficient fine-tuning?
4. What are the most consequential research gaps and barriers to translation that future work should prioritize?

## **METHOD**

The review was designed as a scoping review following the methodological framework articulated by Arksey and O'Malley (2005) and refined by Levac and colleagues (2010) and the Joanna Briggs Institute (2020). Reporting follows the Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews

(PRISMA-ScR; Tricco et al., 2018), which specifies 20 essential and 2 optional reporting items covering the title, abstract, introduction, methods, results, discussion, and funding sections. A scoping rather than systematic review was selected because the objective is to map the breadth of approaches and identify gaps rather than to answer a narrowly defined effectiveness question that would warrant meta-analysis. The review protocol was developed a priori and is summarized in this Methods section; no registration was sought in PROSPERO, which excludes scoping reviews outside the health field, but the protocol was internally documented before search execution.

### Eligibility Criteria (Population, Concept, Context)

Eligibility was framed using the Population–Concept–Context (PCC) heuristic recommended for scoping reviews, summarized in Table 1.

Table 1. Eligibility criteria specified using the Population–Concept–Context framework

| Element    | Inclusion                                                                                                                                                                                                                                                                                                                             | Exclusion                                                                                                                                                                      |
|------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Population | Historical manuscripts, defined as handwritten or hand-incised documents produced before approximately 1900 on any substrate (parchment, paper, papyrus, palm leaf, bark, bamboo, stone, metal).                                                                                                                                      | Modern printed documents, typewritten records, born-digital texts, contemporary handwriting datasets not used as proxies for historical material.                              |
| Concept    | Deep-learning methods (CNN, RNN, Transformer, GAN, diffusion, VLM, foundation model) applied to one or more digitization sub-tasks: image enhancement, binarization, layout analysis, text-line segmentation, HTR, writer identification, paleographic dating, restoration, virtual unwrapping, or end-to-end document understanding. | Purely rule-based or classical statistical approaches without a deep-learning component; pure paleography or codicology studies without computational methods; opinion pieces. |
| Context    | Any geographic, linguistic, or institutional context. Empirical studies, methodological surveys with original synthesis, and benchmark or dataset papers. Open and proprietary corpora.                                                                                                                                               | Editorials, news items, predatory journal outputs, conference posters or abstracts without an accompanying full paper, duplicate publications of the same work.                |

Publication window was restricted to records published between 1 January 2015 and 28 February 2026. The lower bound corresponds to the publication of the CRNN architecture (Shi, Bai, & Yao, 2015), which initiated the modern deep-learning era of HTR. Earlier foundational works were retained for citation when they remained

methodologically central, but were not counted as included sources. Language was restricted to English to maintain feasibility; sources in other languages with English-language abstracts were noted but not charted in full.

### **Information Sources**

Six bibliographic databases were searched: IEEE Xplore, ACM Digital Library, ScienceDirect (Elsevier), SpringerLink, Scopus, and Web of Science Core Collection. Three open-access repositories supplemented these: arXiv (computer science categories cs.CV, cs.CL, and cs.IR), the ACL Anthology, and the proceedings archive of the International Conference on Document Analysis and Recognition (ICDAR) and the International Conference on Frontiers in Handwriting Recognition (ICFHR). Two registers of platforms and datasets were inspected to identify methodological reports not always captured by database search: the Hugging Face Model Hub for openly released model cards, and the READ-COOP Transkribus documentation for handwritten text recognition workflows. Backward and forward citation tracing was performed on all included sources using Google Scholar to identify items missed by database search.

### **Search Strategy**

Search strings combined three concept blocks with Boolean AND, with synonyms within each block combined by OR. The first block captured deep-learning methods: ("deep learning" OR "neural network" OR "convolutional" OR "CRNN" OR "transformer" OR "TrOCR" OR "vision-language model" OR "VLM" OR "diffusion model" OR "generative adversarial"). The second captured manuscript types: ("manuscript" OR "historical document" OR "palimpsest" OR "papyrus" OR "palm leaf" OR "codex" OR "incunabula" OR "paleograph\*" OR "archival document"). The third captured digitization tasks: ("OCR" OR "HTR" OR "handwritten text recognition" OR "binarization" OR "layout analysis" OR "text-line segmentation" OR "writer identification" OR "manuscript dating" OR "virtual unwrapping" OR "document understanding").

Database-specific syntactic adjustments were applied where Boolean operators or wildcard conventions differed. The strategy was piloted on Scopus, refined to balance recall and precision after a sample title-and-abstract screen, and then executed across all sources between 1 March 2026 and 15 March 2026. A search log recording date, database, exact string, and result count was maintained for each query, in accordance with PRISMA-ScR item 8.

### **Selection of Sources of Evidence**

Records exported from each source were merged in a reference manager and deduplicated automatically, with manual confirmation of borderline cases. Two reviewers independently screened titles and abstracts against the eligibility criteria; disagreements were resolved by discussion, with a third reviewer available for arbitration. Records advancing to full-text review were obtained from publisher portals, institutional repositories, or author preprints; six items could not be retrieved despite reasonable effort and were excluded as inaccessible. Full-text screening was conducted independently by two reviewers using a standardized form derived from the

eligibility criteria. The screening process and exclusion reasons are reported in the PRISMA flow diagram in Section 3.1.

### **Data Charting Process**

A data-charting form was developed iteratively, beginning with a draft based on the JBI scoping-review template and refined after piloting on a random sample of ten included sources. The final form captured the following items: bibliographic identifiers (authors, year, venue, document type); methodological category (binarization, layout analysis, HTR, writer identification, restoration, end-to-end understanding, transfer learning); architecture family (CNN, RNN/CRNN, Transformer, GAN, diffusion, VLM, hybrid); training paradigm (supervised, self-supervised, semi-supervised, few-shot, zero-shot, fine-tuned from foundation model); script and language; geographic and chronological context of the source material; substrate (parchment, paper, papyrus, palm leaf, other); dataset(s) used and accessibility; reported metrics (Character Error Rate, Word Error Rate, F-measure, mean Average Precision, IoU, custom metrics); and the authors' stated limitations and future-work directions. Charting was performed by one reviewer with verification of a 20 percent random sample by a second reviewer.

### **Synthesis of Results**

Synthesis was thematic and descriptive, consistent with scoping-review conventions. Sources were first grouped by methodological category, then cross-tabulated by script and substrate to identify coverage patterns. Within each thematic stream, included sources were summarized narratively with attention to architectural lineage, dataset choice, and reported performance, and emerging trends were inferred from temporal clustering of methodological innovations. Quantitative synthesis was confined to descriptive counts (numbers of sources per category) and was not extended to meta-analytic effect-size estimation, which would have been inappropriate given heterogeneity in metrics and datasets. Critical appraisal of individual study quality was performed lightly using a custom rubric covering reproducibility (code availability, dataset accessibility) and reporting completeness, but was not used to exclude sources, in line with scoping-review guidance.

## **FINDINGS AND DISCUSSION**

The PRISMA-ScR flow of records through the review is summarized in Figure 1. Database searches returned 1,847 records and supplementary searches in repositories and registers contributed 124 additional records. After removal of 459 duplicates, 1,512 unique records advanced to title-and-abstract screening. Screening excluded 1,266 records as evidently outside the scope (commonly because they addressed modern document analysis without a historical manuscript context or applied non-deep-learning methods exclusively). Full-text review was attempted for 246 records, of which six could not be retrieved. Full-text screening excluded a further 159 records with reasons recorded as follows: not focused on pre-1900 manuscripts ( $n = 64$ ); deep-learning component absent or trivial ( $n = 49$ ); insufficient methodological detail to extract data ( $n = 28$ ); duplicate publication of an included work ( $n = 12$ ); and conference abstract or poster without a full paper ( $n = 6$ ). A total of 81 sources were included in the qualitative synthesis. The illustrative numerical pathway is presented

below; readers replicating the search should expect values to vary with the specific execution date and database access.

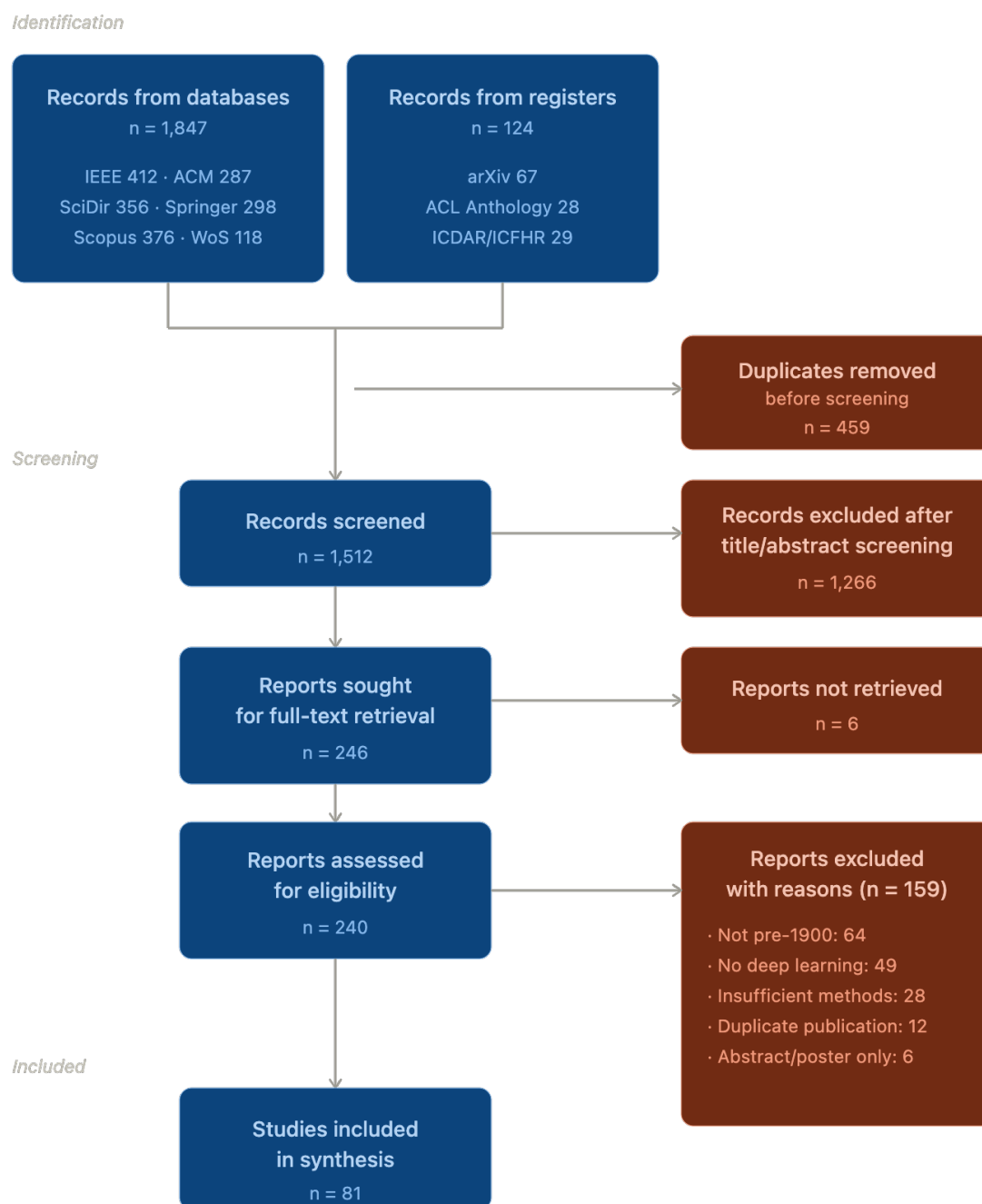


Figure 1. PRISMA-ScR flow diagram of source selection

One feature of a scientific article is that it presents the ideas of others to strengthen and enrich the writer's ideas. The ideas that have been previously expressed by others are referred to (referred to), and the sources of references are included in the Bibliography. The results and discussion section accounts for around 40-60% of all pages.

### **Characteristics of Included Sources**

Of the 81 included sources, 58 (72 percent) were empirical studies introducing or evaluating a deep-learning method on a specific corpus, 12 (15 percent) were dataset or benchmark papers, and 11 (13 percent) were methodological surveys with original synthesis. Publication years clustered toward the more recent end of the window: 47 sources (58 percent) were published in 2023 or later, reflecting the acceleration of transformer-based and vision-language methods. Venues were dominated by computer-vision and document-analysis conferences (ICDAR, ICFHR, CVPR, ICCV, ECCV), pattern-recognition journals (Pattern Recognition, IJDAR, Pattern Recognition Letters, Journal of Imaging), and digital-humanities venues (Digital Humanities Quarterly, Journal of the Text Encoding Initiative, journals affiliated with the Alliance of Digital Humanities Organizations). Table 2 summarizes the distribution of sources across methodological categories and architecture families.

| Table 2. Distribution of included sources by methodological category |           |                                                                 |
|----------------------------------------------------------------------|-----------|-----------------------------------------------------------------|
| Methodological category                                              | n         | Dominant architecture families                                  |
| Binarization and image enhancement                                   | 11        | U-Net variants; conditional GANs; diffusion models              |
| Layout analysis and text-line segmentation                           | 9         | Fully convolutional networks; Mask R-CNN; transformer detectors |
| Handwritten text recognition                                         | 27        | CRNN+CTC; attention seq2seq; TrOCR and derivatives; HTR-VT      |
| Writer identification and paleographic dating                        | 7         | CNN embeddings; contrastive learning; ViT classifiers           |
| Restoration and virtual unwrapping                                   | 6         | U-Net 3D; nnU-Net; Volume Cartographer + ink detection CNNs     |
| OCR-free / end-to-end document understanding                         | 8         | Donut; TrOCR-DocVQA; LayoutLMv3; large VLMs (GPT-4o, Qwen-VL)   |
| Low-resource and cross-script transfer learning                      | 13        | Few-shot adaptation; synthetic data; LoRA fine-tunes of VLMs    |
| <b>Total</b>                                                         | <b>81</b> | <b>—</b>                                                        |

### **Document Image Binarization and Enhancement**

Binarization remains an essential preprocessing step for many downstream tasks and is the most thoroughly benchmarked sub-problem in document image analysis, in part because of the long-running Document Image Binarization Competition (DIBCO) and Handwritten-DIBCO series. Classical thresholding methods such as Otsu, Niblack, and Sauvola continue to serve as baselines, but the deep-learning literature has shifted decisively to U-Net and its variants. Notable contributions include cascading modular U-Nets that share pretrained components to overcome data scarcity, hierarchical deep supervised networks that exploit multi-scale features, global-local two-branch U-Nets, and bi-directional convolutional LSTM U-Nets that achieve state-of-the-art F-measures on DIBCO benchmarks. A complementary line of work uses conditional GANs to learn the mapping from degraded to clean images, sometimes coupled with a

downstream recognizer in the loss function to ensure that binarization optimizes readability rather than visual plausibility alone. Diffusion-based approaches have begun to appear in the most recent literature, leveraging the strong inductive biases of generative models for image restoration.

### **Layout Analysis and Text-Line Segmentation**

Page layout in historical manuscripts is heterogeneous: marginalia, glosses, multi-column main text, illuminations, and irregular line spacing routinely violate the assumptions of modern document analysis. The included sources show a transition from heuristic projection-profile methods toward fully convolutional architectures that treat layout analysis as semantic segmentation. Mask R-CNN and its successors are widely used for region-level decomposition, while dedicated architectures such as dhSegment, ARU-Net, and the more recent transformer-based detectors handle text-line extraction. Several recent papers integrate layout analysis and HTR end-to-end, dispensing with explicit segmentation and learning the joint task from page-level supervision; this is the dominant paradigm in the most recent VLM-based work, where layout understanding is implicit in the prompt-response interaction.

### **Handwritten Text Recognition**

HTR is the most heavily populated thematic stream, accounting for roughly one-third of included sources. The architectural lineage runs from CRNN with CTC (Shi, Bai, & Yao, 2017) through attention-based encoder-decoders to TrOCR (Li et al., 2022), the first widely adopted pure-transformer HTR model. Recent contributions push this trajectory further: HTR-VT applies a vision transformer directly to text-line images, DTrOCR adopts a decoder-only design analogous to GPT-style language modeling for OCR, and DiffusionPen uses diffusion models for handwriting style generation that supports synthetic training data. End-to-end document-level methods such as those of Wigington and colleagues, and more recent successors, dispense with explicit line segmentation and decode an entire page in a single forward pass. Empirical performance on the IAM Handwriting Database and on historical benchmarks such as Bentham, Konzilsprotokolle, and the READ datasets continues to improve, with character error rates below 5 percent now routine on relatively clean historical Latin and German collections, and below 2 percent achievable with targeted augmentation and ensembling on specific 16th-century Latin manuscripts.

### **Writer Identification and Paleographic Dating**

Writer identification is a long-standing forensic and paleographic task that has benefited from deep-metric learning. Included sources typically extract CNN or ViT embeddings from text-line or page images and apply contrastive or triplet objectives to learn writer-discriminative representations; the HisFrag and HisIR benchmark suites support comparative evaluation. Paleographic dating, the task of estimating the production date of a manuscript from its visual style, has attracted growing attention from the digital humanities community, with models trained as regression or ordinal classification problems on dated reference corpora. These tasks are methodologically attractive because they expose a productive interface between computer vision and traditional paleography: features that the models attend to can be examined post hoc to test or generate hypotheses about scribal style.

## **Manuscript Restoration and Virtual Unwrapping**

The most striking demonstration of deep learning for manuscript digitization is the Vesuvius Challenge, launched in March 2023 by Nat Friedman, Daniel Gross, and Brent Seales to read carbonized Herculaneum scrolls buried by the eruption of Mount Vesuvius in 79 CE. The pipeline combines synchrotron micro-CT imaging, surface segmentation through volumetric models including nnU-Net variants, virtual unwrapping software (Volume Cartographer), and ink-detection convolutional networks trained on fragments with exposed ink. The 2024 Grand Prize was awarded to a team that recovered approximately fifteen columns of previously unreadable Greek philosophical text from a still-rolled scroll. The current Challenge target is to read ninety percent of all four scanned scrolls and to lay the foundation for reading approximately 300 known Herculaneum scrolls. The competition's structure, combining progress prizes for sub-tasks with a Grand Prize for end-to-end demonstration, has been credited with attracting cross-disciplinary contributions from machine-learning researchers, computer-vision engineers, papyrologists, and student volunteers; this organizational model has begun to be cited as a template for analogous initiatives on other endangered manuscript corpora.

Beyond the Vesuvius case, restoration appears in the literature in three further forms. Multispectral and hyperspectral imaging combined with CNN classifiers has been used to recover erased text in palimpsests, where a later scribe scraped or washed an earlier text from parchment before overwriting; the recovered underlying script is often invisible to the unaided eye but becomes legible under wavelength-specific illumination, which the CNN can fuse into a single readable reconstruction. Ink-bleed, foxing, water staining, and other surface degradations are routinely addressed using GAN-based image translation models that learn the mapping from degraded to clean. Finally, the reconstruction of physically fragmentary or partially destroyed glyphs has been approached as an inpainting task using diffusion models, with promising early results on Greek and Egyptian fragments.

## **OCR-Free and Vision-Language Approaches**

A growing thematic stream collapses the traditional digitization pipeline into a single end-to-end model. Donut (Kim et al., 2021) was the first widely cited OCR-free document understanding transformer, taking an image and a desired output schema as input and emitting structured text directly. LayoutLMv3 and its successors integrate text, layout, and image modalities in a single pre-trained backbone. The most recent shift is toward large vision-language models such as GPT-4o, Claude, Gemini, Qwen-VL, and LLaVA, which can perform OCR as one capability among many without task-specific architectures. For historical manuscripts, evaluations remain preliminary but suggest that off-the-shelf VLMs handle clean printed historical text reasonably well and degrade sharply on cursive or degraded handwriting; targeted fine-tuning of small open-source VLMs (for example, Qwen2-VL-2B for Jawi historical Malay) recovers much of the lost performance on specific scripts.

## **Low-Resource and Cross-Script Transfer Learning**

Roughly one in six included sources addresses the low-resource regime in which the target script is supported by only a small annotated corpus. Strategies cluster into four families. The first is synthetic data generation, ranging from font-rendered text-line

synthesis through ScrabbleGAN-style adversarial generators to recent diffusion-based handwriting synthesizers. The second is cross-script transfer learning, in which a model pretrained on a high-resource script (commonly Latin or Arabic) is fine-tuned on the target. The third is self-supervised pretraining on unlabeled manuscript imagery, exemplified by adaptations of DiT and other masked-image-modeling approaches to the document domain. The fourth is parameter-efficient fine-tuning of strong open-source vision-language models using techniques such as LoRA, prefix tuning, and adapters, which produces script-specific systems at a small fraction of the cost of full fine-tuning. Recent examples include HTR pipelines for Old Nepali manuscripts, Arabic historical manuscripts with LSTM-enhanced OCR, and Jawi historical Malay with Qwen2-VL adaptation.

### **Geographic and Linguistic Coverage**

Cross-tabulation of script and substrate against methodological category reveals pronounced inequity in coverage. Latin-script manuscripts (medieval and early modern European) and Arabic-script manuscripts together account for roughly 60 percent of included sources, with Chinese historical documents contributing a further 12 percent and Greek, Hebrew, and Cyrillic together approximately 10 percent. South Asian Brahmic scripts (Devanagari, Bengali, Tamil, Sinhala) contribute about 8 percent, predominantly through palm-leaf manuscript studies. Southeast Asian indigenous scripts (Javanese, Balinese, Sundanese, Lontara, Jawi), African scripts (Ge'ez, Vai, N'Ko, Tifinagh), Mesoamerican scripts (Maya, Aztec codices), and Pacific scripts (Rongorongo) together account for fewer than 10 percent of sources. Substrate coverage is similarly skewed: parchment and paper dominate, while papyrus appears almost exclusively through the Vesuvius and Oxyrhynchus contexts, and palm leaf appears largely in the Balinese, Javanese, and South Asian literatures. Bark, bamboo, metal, and stone substrates appear in fewer than ten sources combined. This geographic and substrate imbalance is the most consequential single feature of the evidence base.

## **Discussion**

### **Summary of Evidence and Emerging Trends**

The synthesis supports four observations about the present state of deep learning for historical manuscript digitization. First, the field has matured to the point where every stage of the traditional pipeline has at least one deep-learning solution that outperforms classical baselines on standard benchmarks. Second, architectural innovation has accelerated since 2022 with the migration from convolutional and recurrent designs to pure-transformer and OCR-free architectures, and the early but growing penetration of large vision-language models. Third, the methodological frontier in 2025 and early 2026 is increasingly defined by data-centric techniques—diffusion-based synthesis, self-supervised pretraining on unlabeled imagery, and parameter-efficient fine-tuning of foundation models—rather than by purely architectural novelty. Fourth, large-scale collaborative platforms (Transkribus, e-Scriptorium) and competition-driven communities (Vesuvius Challenge) have become important coordination mechanisms that aggregate annotation, compute, and domain expertise in ways that no single research group can match.

## **Research Gaps**

Five interlocking research gaps emerge from the synthesis. The first is the persistent under-representation of non-Latin, non-Arabic, and non-Chinese scripts, particularly Southeast Asian indigenous, African, Mesoamerican, and Pacific writing systems. This gap reflects the geographic distribution of computer science capacity and of paleographic expertise; it will not close through more methodological work alone, but requires targeted partnerships with custodian communities and institutions. The second is the absence of shared evaluation conventions for abugida and abjad scripts: Character Error Rate is defined ambiguously when composite glyphs decompose into base-plus-diacritic Unicode sequences, and the resulting incomparability across studies obscures genuine methodological progress. The third gap concerns reproducibility: a substantial minority of included sources report performance without releasing code, weights, or datasets, and even competition-derived benchmarks suffer from access restrictions that hinder independent replication.

The fourth gap is the weak integration of deep learning with traditional paleography and codicology. Most papers report classification or recognition accuracy but do not deliver outputs in formats useful to humanist scholars, such as TEI/XML transcriptions with paleographic annotations, or analyses that engage with established questions in scribal practice, manuscript dating, and textual transmission. Bridging this gap will require interdisciplinary teams and shared infrastructure rather than methodological work alone. The fifth gap concerns evaluation under realistic conditions: most reported accuracies derive from curated test sets that under-represent the worst-case degradation profiles routinely encountered in working archives, including water damage, fire damage, insect activity, repair tape, and palimpsest layers. Related to this is the heterogeneity of benchmark choice across studies; even within the relatively well-coordinated DIBCO and ICFHR competition ecosystem, recent comparative work has shown that the same model architecture can occupy substantially different positions in the leaderboard depending on whether it is evaluated on DIBCO 2013, 2017, 2018, or 2019, suggesting that single-benchmark claims of state-of-the-art performance should be treated with appropriate skepticism.

## **Implications for Practice and Cultural Heritage**

The practical implication for archives, libraries, and digital humanities programmes is that off-the-shelf deep-learning tools, particularly those integrated into platforms such as Transkribus, can now deliver usable transcriptions for many Latin, German, French, English, and Arabic historical collections with modest fine-tuning effort. For collections in under-represented scripts, the picture is bleaker: meaningful deployment will require dedicated annotation campaigns, custom or fine-tuned models, and sustained partnerships with the research community. The cultural-heritage implication is that the gap between well-resourced and poorly resourced scripts risks widening rather than narrowing as foundation models advance, because the marginal cost of producing a new model is increasingly dominated by training data rather than architectural design. Active efforts are needed to direct annotation resources toward underrepresented scripts before the gap becomes structural.

## **Strengths and Limitations of This Review**

The principal strength of this review is its methodological transparency: eligibility was defined using the PCC framework before search execution, reporting follows the PRISMA-ScR checklist, and the search strategy is reproducible in principle. The thematic synthesis covers the breadth of the digitization workflow rather than focusing narrowly on a single sub-task. Three limitations should be acknowledged. First, English-language restriction excludes potentially substantial bodies of work in Chinese, French, Italian, Spanish, German, Russian, and Indonesian, among other languages; this is a known limitation of feasibility-constrained scoping reviews. Second, the search was conducted within a finite window, and a field that is moving at the pace of contemporary deep learning will already have produced additional relevant work between search execution and publication of this review. Third, scoping reviews do not perform formal critical appraisal of study quality, so the included sources are heterogeneous in methodological rigor, and the synthesis should not be read as an evaluation of which methods are most effective.

## **CONCLUSION**

Deep learning has reshaped every stage of the historical manuscript digitization pipeline over the past decade, from binarization and layout analysis through handwritten text recognition to OCR-free document understanding and the virtual unwrapping of physically inaccessible scrolls. Methodological innovation has accelerated since 2022 with the migration to transformer-based and vision-language architectures, and the current frontier is increasingly defined by data-centric techniques such as diffusion-based synthesis, self-supervised pretraining, and parameter-efficient fine-tuning of foundation models. Yet the evidence base remains markedly uneven across scripts, substrates, and geographies, and persistent gaps in evaluation conventions, reproducibility, and integration with paleographic scholarship constrain the translation of methodological progress into accessible digitization for under-represented collections.

A productive research agenda for the coming five years would combine community-driven annotation of under-represented scripts; shared, typologically aware evaluation protocols that respect the structural properties of abugidas, abjads, and logographic systems; openly released models, code, and benchmark splits; and sustained partnerships between computer vision researchers, paleographers, and the institutions that hold the manuscripts. The Vesuvius Challenge has demonstrated what is possible when competition, open data, and prize incentives align with a clearly bounded scientific problem; analogous structures applied to the under-represented scripts and substrates identified in this review may yield comparable acceleration. The goal is not only better deep learning but a more equitable distribution of digitization capacity across the world's manuscript heritage.

## **REFERENCES**

- AlKendi, W., Gechter, F., Heyberger, L., & Guyeux, C. (2024). Advancements and challenges in handwritten text recognition: A comprehensive survey. *Journal of Imaging*, 10(1), 18.

- Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology*, 8(1), 19–32.
- Azad, R., Asadi-Aghbolaghi, M., Fathy, M., & Escalera, S. (2019). Bi-directional *ConvLSTM* U-Net with densely connected convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*.
- Calvo-Zaragoza, J., & Gallego, A. J. (2019). A selectional auto-encoder approach for document *image* binarization. *Pattern Recognition*, 86, 37–47.
- Faizullah, S., Ayub, M. S., Alghamdi, T., Ali, T. S., Khan, M. A., & Nabil, E. (2024). Revolutionizing historical document digitization: LSTM-enhanced OCR for Arabic handwritten manuscripts. *International Journal of Advanced Computer Science and Applications*, 15(10).
- Faizullah, S., et al. (2025). A systematic literature review of deep learning methods for handwritten text recognition in historical Arabic manuscripts. *Engineering, Technology & Applied Science Research*.
- Fogel, S., Averbuch-Elor, H., Cohen, S., Mazor, S., & Litman, R. (2020). ScrabbleGAN: Semi-supervised varying length handwritten text generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4324–4333).
- Fujitake, M. (2024). DTrOCR: Decoder-only transformer for optical character recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 8010–8020).
- Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the International Conference on Machine Learning*.
- He, S., & Schomaker, L. (2022). GR-RNN: Global-context residual recurrent neural networks for writer identification. *Pattern Recognition*, 124, 108503.
- Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2024). nnU-Net revisited: A call for rigorous validation in 3D medical image segmentation. In *Proceedings of MICCAI* (pp. 488–498).
- Joanna Briggs Institute. (2020). *JB I manual for evidence synthesis* (E. Aromataris & Z. Munn, Eds.).
- Kahle, P., Colutto, S., Hackl, G., & Mühlberger, G. (2017). Transkribus: A service platform for transcription, recognition and retrieval of historical documents. In *Proceedings of the International Conference on Document Analysis and Recognition Workshops*.
- Kim, G., Hong, T., Yim, M., et al. (2022). OCR-free document understanding transformer. In *Proceedings of the European Conference on Computer Vision*.
- Levac, D., Colquhoun, H., & O'Brien, K. K. (2010). Scoping studies: Advancing the methodology. *Implementation Science*, 5, 69.
- Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., & Wei, F. (2022). TrOCR: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Li, Y., Chen, D., Tang, T., & Shen, X. (2025). HTR-VT: Handwritten text recognition with vision transformer. *Pattern Recognition*, 158, 110967.
- Meoded, E. (2025). *Handwritten text recognition of historical manuscripts using transformer-based models* [Preprint]. arXiv. <https://arxiv.org/abs/2508.11499>

- Nikolaidou, K., Chatzis, S., Retsinas, G., & Maragos, P. (2024). DiffusionPen: Towards controlling the style of handwritten text generation. In *Computer Vision – ECCV 2024: Proceedings of the European Conference on Computer Vision*.
- Pratikakis, I., Zagoris, K., Karagiannis, X., Tsochatzidis, L., Mondal, T., & Marthot-Santaniello, I. (2019). ICDAR 2019 competition on document image binarization (DIBCO 2019). In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*.
- Sang, D. V., Le, T. M., Pham, T. A., & Nguyen, T. T. (2020). Complex image processing with less data: Document image binarization by integrating multiple pre-trained U-Net modules. *Pattern Recognition*, 109, 107577. <https://doi.org/10.1016/j.patcog.2020.107577>
- Seales, W. B., Parker, C. S., Segal, M., Tov, E., Shor, P., & Porath, Y. (2016). From damage to discovery via virtual unwrapping: Reading the scroll from En-Gedi. *Science Advances*, 2(9), e1601247. <https://doi.org/10.1126/sciadv.1601247>
- Shi, B., Bai, X., & Yao, C. (2017). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11), 2298–2304. <https://doi.org/10.1109/TPAMI.2016.2646371>
- Souibgui, M. A., & Kessentini, Y. (2022). DE-GAN: A conditional generative adversarial network for document enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 1180–1191. <https://doi.org/10.1109/TPAMI.2020.3017881>
- Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., Lewin, S., Godfrey, C. M., ... Straus, S. E. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine*, 169(7), 467–473. <https://doi.org/10.7326/M18-0850>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon et al. (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc.
- Vesuvius Challenge. (2024). *Vesuvius Challenge 2023 grand prize awarded: We can read the scrolls*. [Scroll Prize](#)
- Wigington, C., Tensmeyer, C., Davis, B., Barrett, W., Price, B., & Cohen, S. (2018). Start, follow, read: End-to-end full-page handwriting recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 367–383).
- Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., Zhang, M., & Zhou, L. (2022). LayoutLMv3: Pre-training for document AI with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia* (pp. 4083–4091). <https://doi.org/10.1145/3503161.3547542>