International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(21s) 2026 pp. 361-372
https://doi.org/10.70917/ijcisim-2026-5313
Article

Digital Preservation of Indian Knowledge Systems Using Artificial Intelligence: A Comprehensive Framework

Dr. Hitha Paulson

Department of Computer Science and Applications, Little Flower College (Autonomous), Guruvayur, Kerala, India

Abstract: Indian Knowledge Systems (IKS) encompass a large body of intellectual, scientific, medical, linguistic, philosophical, mathematical, astronomical, literary and cultural knowledge preserved through manuscripts and related documentary traditions. A substantial part of this heritage exists on palm leaves, paper, birch bark and other materials whose physical deterioration can continue even after digitization. Faded ink, stains, bleed-through, uneven illumination, surface damage, low resolution and missing character strokes reduce the usefulness of digital images for both human reading and automated text recognition. This paper develops and elaborates an artificial intelligence (AI)-assisted framework for digital preservation of degraded Indian manuscripts. The framework combines repository-based acquisition, metadata and provenance capture, image-quality assessment, conventional preprocessing, degradation-aware deep-learning restoration, selective generative inpainting, super-resolution, optical character recognition (OCR), quantitative evaluation and expert validation. The experimental component described in the original study uses approximately 100 manuscript images collected from open-access digital collections. Reported pilot results indicate SSIM improvements of approximately 0.05–0.10, PSNR improvements of 2–6 dB, contrast improvements of 20–35%, noise reduction of 20–35%, edge-preservation improvement of 15–30% and OCR accuracy improvement of approximately 10–20 percentage points. Character Error Rate (CER) and Word Error Rate (WER) were reported to decrease by approximately 10–25% and 10–30%, respectively, with character recognition accuracy reaching approximately 85–90%. These results are interpreted as evidence that enhancement can improve machine readability and access, while not constituting proof that AI-generated pixels are historically correct. The expanded framework therefore separates original evidence, algorithmically enhanced representations, and reconstructed content; preserves processing provenance and places scholars in the validation loop. The paper also situates the framework within recent developments in Indian manuscript digitization, historical document analysis, trustworthy AI and the Gyan Bharatam initiative. The central argument is that AI should be deployed as an assistive preservation technology that increases discoverability and research access without replacing the original manuscript or scholarly judgement.

Keywords: Indian Knowledge Systems, digital preservation, artificial intelligence, historical manuscripts, image enhancement, OCR, deep learning, cultural heritage, provenance, trustworthy AI

1. INTRODUCTION

Indian Knowledge Systems represent accumulated intellectual traditions transmitted through texts, oral practices, pedagogical institutions, material artefacts and manuscripts. Manuscripts are particularly important because they preserve evidence of how knowledge was recorded, copied, interpreted and transmitted across generations. They cover subjects ranging from medicine, mathematics and astronomy to philosophy, grammar, literature, ritual practice, agriculture and environmental knowledge. The manuscript tradition is also materially diverse: palm leaves, paper, birch bark and other substrates respond differently to humidity, heat, biological activity, abrasion, handling, ink chemistry and storage conditions. Consequently, manuscript preservation is not simply a question of scanning an object; it is a continuing process of protecting the physical artefact while creating trustworthy digital representations.

Digitization has transformed access to fragile collections by reducing the need for repeated physical handling and enabling distributed scholarly work. However, a digital photograph preserves the condition of the manuscript at the time of capture, including many of its defects. A high-resolution image may still contain weak strokes, shadows, stains, fibre patterns, punch holes, blur, uneven illumination and background interference. Historical document research has repeatedly identified degradation and limited training data as major barriers to automated analysis, while recent surveys emphasize the need for larger datasets and standardized evaluation protocols [1], [2]. Thus, digitization should be understood as a foundation rather than the endpoint of digital preservation.

The problem becomes more complex when the objective is not merely visual display but computational access. OCR requires sufficiently clear character boundaries and stable text-line structure. Errors introduced at image acquisition or preprocessing can propagate into segmentation, recognition, indexing and translation. Conversely, enhancement that makes an image visually attractive may remove faint strokes that are historically meaningful. This creates a central preservation tension: an algorithm must improve legibility without silently converting uncertainty into apparent certainty. Recent work on historical document enhancement and Indic manuscript recognition supports a move toward content-aware, degradation-specific processing rather than one-size-fits-all enhancement [3], [4].

This paper addresses that challenge through an integrated AI-assisted preservation framework. Rather than treating restoration, OCR and archival preservation as separate activities, the framework connects them in a controlled pipeline. It retains the original digital image, creates a documented enhanced derivative, records model and processing information and treats reconstructed regions as hypotheses that require explicit labelling and expert review. This approach is consistent with the growing emphasis on trustworthy AI in cultural heritage, where transparency, explainability, contextual knowledge and human oversight are necessary for responsible deployment [5].

1.1. Objectives and Contributions

The objectives are to: (1) define a degradation-aware workflow for Indian manuscript images (2) integrate conventional image processing with deep-learning restoration (3) connect visual enhancement to OCR and downstream knowledge access (4) evaluate both image-level and text-level outcomes (5) establish provenance and authenticity safeguards and (6) identify practical research directions for multilingual, multiscript manuscript preservation. The main contribution is therefore architectural and methodological: the paper proposes a preservation workflow in which technical improvement is explicitly separated from historical interpretation.

2. BACKGROUND AND LITERATURE REVIEW

2.1 Digital Preservation of Indian Manuscripts

Indian manuscript digitization has developed through institutional cataloguing, conservation, repository creation and increasingly sophisticated digital access systems. The National Mission for Manuscripts and its Kriti Sampada/Pandulipi Patala resources provide an important national context for manuscript documentation. The recent Gyan Bharatam Mission further emphasizes large-scale cataloguing, digitization, a national digital repository and technological integration including AI-assisted transcription and OCR [6], [7]. These initiatives make the timing of AI-assisted preservation particularly significant: the growth of digital collections increases the need for automated methods that can triage, enhance, index and analyze large numbers of images.

The preservation challenge is not uniform across collections. A manuscript written on palm leaf may exhibit punch holes, fibre patterns, curvature and surface abrasion; paper manuscripts may show foxing, ink bleed-through, folds, tears and uneven ageing; birch-bark manuscripts can present texture and cracking that resemble character strokes. Historical scripts add another layer of variability. Malayalam, Tamil, Devanagari, Grantha, Sanskrit and other Indic writing traditions contain distinct character shapes, ligatures, writing conventions and historical variants. Consequently, models trained on modern printed text or a single script cannot be assumed to generalize to all Indian manuscript collections.

2.2 Historical Document Image Analysis

Historical document image analysis has matured from handcrafted image processing toward deep-learning systems for segmentation, layout analysis, recognition and enhancement. Nikolaidou et al. reviewed historical document datasets and showed that the field continues to face limitations in dataset size, script diversity and comparability of evaluation methods [1]. Recent work on text-line segmentation similarly identifies line extraction as a critical upstream step because errors in segmentation affect subsequent handwritten text recognition [8]. These findings reinforce the need to treat manuscript analysis as a pipeline in which early errors can influence all later stages.

2.3 Conventional Enhancement

Conventional preprocessing remains valuable because it is computationally inexpensive, interpretable and often effective for moderate degradation. Typical operations include grayscale conversion, illumination correction, contrast stretching, denoising, adaptive thresholding, morphological processing, cropping and resizing. Historical document binarization studies show that thresholding must be selected carefully because weak strokes can disappear when the algorithm aggressively separates foreground and background [9], [10]. For this reason, the proposed framework does not assume that binarization is inherently beneficial. Instead, it selects preprocessing operations according to observed degradation and retains the non-binarized original for comparison.

2.4 Deep Learning and Generative Restoration

Deep learning extends enhancement by learning relationships between degraded and cleaner representations. U-Net provides an encoder-decoder structure with skip connections that is widely useful for image-to-image transformation [11]. GAN-based methods can synthesize high-frequency details or plausible missing structures, while partial-convolution and contextual-attention approaches provide mechanisms for image inpainting [12], [13]. ESRGAN demonstrates how adversarial super-resolution can improve perceptual detail in low-resolution images [14]. More recent work applies diffusion-based and hybrid AI methods to historical document enhancement, reflecting a shift toward models that can reason about complex degradation rather than merely apply fixed filters [3].

2.5 OCR and Recognition of Indic Manuscripts

OCR is the bridge between image preservation and knowledge accessibility. A manuscript image becomes substantially more useful for search, indexing, concordance building, translation and computational analysis when a machine-readable transcription is available. Tesseract remains an important open OCR baseline, while platforms such as Transkribus provide historical-document recognition and transcription workflows [15], [16]. Nevertheless, OCR performance depends strongly on image quality, text-line segmentation, script characteristics and the availability of representative training data.

Indic manuscripts require particular attention because historical character forms may differ substantially from contemporary fonts and because manuscript writing is often unconstrained. Recent Tamil palm-leaf research illustrates this point. Uma Maheswari et al. developed a multi-stage segmentation and recognition approach addressing irregular lines and punch-hole interference, reporting high segmentation and character-recognition performance on their experimental dataset [4]. Pavithra et al. further explored enhancement, segmentation and ConvNeXt-based recognition for ancient Tamil palm-leaf manuscripts, emphasizing denoising, adaptive thresholding, punch-hole handling and script-specific recognition [17]. These studies demonstrate the value of specialized pipelines rather than assuming that generic OCR will work on historical Indic material.

2.6 Restoration as an Epistemic Problem

Image restoration in cultural heritage differs from ordinary photographic enhancement because the target is not simply visual quality. If a restoration model generates a plausible stroke where the manuscript is physically damaged, the output may be useful for hypothesis generation but cannot automatically be treated as evidence. Work on ancient text restoration has demonstrated the ability of neural models to propose missing textual material, while also making clear that model predictions and scholarly restorations are distinct forms of evidence [18]. The distinction is especially

important for generative inpainting and super-resolution, which can introduce details not directly observable in the source image.

Accordingly, this paper distinguishes three states of information. The first is observed evidence: pixels directly captured from the manuscript. The second is enhanced evidence: pixels transformed through deterministic or learned processing without intentionally inventing missing content. The third is reconstructed content: model-generated material that represents an inference. These states should be stored separately and communicated clearly to users. The separation can be implemented through versioned files, region masks, processing logs, model identifiers, confidence scores and scholarly annotations.

2.7 Research Gap

The literature provides strong individual solutions for digitization, image enhancement, segmentation, OCR and restoration, but there remains a need for integrated preservation workflows specifically designed around Indian manuscript diversity and authenticity. The gap is therefore not the absence of AI techniques; it is the absence of a preservation-oriented architecture that connects them while controlling provenance and scholarly risk. The proposed framework addresses this gap by making the manuscript image, AI processing history, OCR output and expert validation part of one traceable chain.

3. METHODOLOGY

3.1 Research Design

The study follows an experimental, pipeline-oriented design. Approximately 100 digitized manuscript images were selected from publicly accessible manuscript resources. The source material represents different physical and visual conditions rather than a single homogeneous dataset. The images exhibit degradation including fading, low contrast, noise, stains, blur, uneven illumination, damaged character strokes and low resolution. Because the available dataset is relatively small, the study is best interpreted as a pilot evaluation rather than a universal benchmark for Indian manuscripts. The reported ranges should therefore be treated as observed outcomes for the evaluated sample, not as guaranteed performance across all scripts or repositories.

3.2 Data Acquisition and Metadata

Manuscript images were obtained from open-access or publicly accessible collections associated with Indian manuscript preservation and cultural-heritage initiatives. Sources considered in the original study include National Mission for Manuscripts/Kriti Sampada, Muktabodha Digital Library, SAMHiTA, Central Institute of Classical Tamil resources and Hill Museum & Manuscript Library collections. Current national policy developments, especially Gyan Bharatam, strengthen the relevance of interoperable metadata and provenance because large-scale digitization will produce collections that must remain searchable and traceable over long periods [6], [7].

For each image, the framework recommends recording repository identifier, manuscript identifier, language, script where known, material, approximate date or period when available, collection information, image resolution, file format, capture information, licensing or access status, and any existing cataloguing metadata. Technical metadata should be accompanied by a processing record containing software versions, preprocessing parameters, model name and version, inference settings, date of processing, and the identity or role of the validating expert. Such metadata transforms an isolated enhanced image into a reproducible digital object.

3.3 Image Quality Assessment

Before enhancement, images are assessed for dominant degradation types. A practical classification can include: low contrast, illumination variation, random noise, structured background texture, blur, low resolution, local physical damage, and mixed degradation. A degradation label can be assigned manually, automatically, or through a hybrid process. This label controls the downstream model selection. The intention is not to force every image through every algorithm, but to use the least invasive processing capable of addressing the observed problem.

3.4 Conventional Preprocessing

Preprocessing begins with lossless preservation of the source image. Working copies may then undergo grayscale conversion, illumination correction, denoising, contrast enhancement, background normalization, cropping and resizing. Multiple parameter settings can be generated when uncertainty exists. Importantly, binarization is treated as an optional derivative rather than the canonical representation because thresholding may suppress faint strokes. The framework therefore encourages side-by-side comparison of original, enhanced and binarized outputs.

3.5 AI Enhancement

The AI stage is degradation-aware. U-Net-style restoration is considered for image-to-image correction and structural recovery [11]. GAN-based inpainting is reserved for selected damaged regions where a reconstruction is useful as a hypothesis [12], [13]. ESRGAN-style super-resolution is considered when low spatial resolution prevents adequate inspection of small character structures [14]. The framework does not assume that one model is optimal for all manuscripts; model selection is itself part of the preservation record.

3.6 OCR and Text Processing

OCR is applied to at least two representations where feasible: the original/preprocessed image and the AI-enhanced image. This enables the study to determine whether visual improvement produces useful downstream gains. OCR outputs should be stored separately from the image files and linked through stable identifiers. Where the OCR engine supports confidence scores, those scores should be retained. Low-confidence regions can then be returned to the human reviewer for targeted inspection rather than requiring complete manual transcription.

For historical Indic scripts, OCR should be treated as a recognition hypothesis rather than an authoritative transcription. Script identification, line segmentation and character recognition may require separate models. The framework therefore permits a multi-stage process: script identification → page segmentation → text-line detection → character/word recognition → language-aware post-processing → human validation. Recent historical-document research indicates that line segmentation is an important determinant of downstream recognition quality [8]. Similarly, Tamil palm-leaf studies show that handling punch holes and irregular line geometry can materially improve recognition [4], [17].

3.7 Expert Validation & Provenance and Version Control

Expert validation is essential because numerical image metrics cannot determine historical correctness. A selected subset of enhanced images is evaluated on readability, preservation of character structure, structural fidelity, authenticity, artificial detail, retention of manuscript texture, and scholarly usefulness.

The preservation architecture should maintain a parent-child relationship among files: the original capture is the parent; preprocessing derivatives are children; AI-enhanced images are further derivatives; reconstructed regions are explicitly marked; OCR and annotations are linked to the relevant image version. A processing manifest can include a unique object identifier, source hash, derivative hash, model identifier, parameters, timestamp, operator, validation status and notes. This approach allows a researcher to reproduce the path from original image to any derived representation and prevents an enhanced derivative from being mistaken for the archival master.

3.8 Evaluation Metrics AND Experimental Workflow

Image-level evaluation uses SSIM and PSNR when an appropriate reference image exists, supplemented by contrast, noise reduction and edge preservation. SSIM measures structural similarity, while PSNR expresses reconstruction quality in decibels under reference-based conditions [19]. These metrics must be interpreted together because a model can improve PSNR while producing visually over-smoothed characters, or improve contrast while deleting weak strokes. Text-level evaluation uses OCR accuracy, Character Error Rate and Word Error Rate. CER is particularly useful for historical scripts because a single character substitution can change a word or semantic unit even when the overall page appears readable.

The complete workflow is: repository acquisition → metadata capture → source preservation → quality assessment → conventional preprocessing → degradation-specific AI enhancement → OCR → image and text evaluation → expert validation → provenance packaging → archival storage and access. The workflow is intentionally iterative. If OCR quality deteriorates after enhancement, the system should be able to return to an earlier derivative rather than forcing the enhanced output into the final archive.

4. RESULTS

4.1 Reported Image-Level Outcomes

The original experimental study reports measurable improvements across the evaluated manuscript images. SSIM increased by approximately 0.05–0.10, while PSNR increased by approximately 2–6 dB where suitable reference conditions were available. Contrast improved by approximately 20–35%, noise was reduced by approximately 20–35%, and edge preservation improved by approximately 15–30%. Taken together, these changes indicate that the processing pipeline improved visibility while retaining a substantial amount of character structure.

The significance of the results is not that a single metric reaches a particular threshold, but that multiple measures move in a consistent direction. Contrast enhancement alone can make a document look better while removing information. Similarly, denoising can increase apparent smoothness but erase thin strokes. The reported simultaneous improvement in contrast, noise reduction and edge preservation provides a stronger indication that the pilot pipeline was not merely applying aggressive smoothing. Nevertheless, because the source study reports ranges rather than a complete image-by-image statistical table, the results should be interpreted as summary observations and should be followed by a larger controlled benchmark in future work.

Table I. Summary of Reported Pilot Outcomes

Metric	Reported improvement
SSIM	+0.05 to +0.10
PSNR	+2 to +6 dB
Contrast	+20% to +35%
Noise reduction	20% to 35%
Edge preservation	+15% to +30%
OCR accuracy	+10 to +20 percentage points
CER	10% to 25% reduction
WER	10% to 30% reduction
Character recognition accuracy	Approximately 85% to 90%

4.2 OCR Outcomes

The reported OCR accuracy increased by approximately 10–20 percentage points after enhancement. The study correctly distinguishes percentage-point improvement from relative percentage improvement; a 15-point increase from 60% to 75%, for example, is a 25% relative increase. CER was reported to decrease by approximately 10–25%, while WER decreased by approximately 10–30%. Character recognition accuracy reached approximately 85–90%. These results suggest that the image-processing stage had practical consequences for machine readability rather than producing only cosmetic changes.

The OCR findings are particularly important for IKS because machine-readable text enables functions that are difficult to perform on images alone. Search systems can index recognized terms, researchers can compare textual variants, linguistic tools can identify recurring expressions, and digital-humanities workflows can connect manuscript text to catalogues and knowledge graphs. At the same time, OCR errors in historical material can become invisible

when users search only the extracted text. Therefore, OCR confidence and links back to the image region should be retained wherever possible.

4.3 Relationship Between Image and Text Quality

The results support a sequential relationship: improved image quality increases the visibility of character structures; clearer structures facilitate segmentation and recognition; improved recognition increases machine-readable access. This does not imply a simple linear relationship. An image can be visually improved without improving OCR if the enhancement changes character topology or if the OCR model is poorly matched to the script. The framework therefore treats OCR as a downstream validation signal rather than as an automatic consequence of visual enhancement.

5. PROPOSED INTEGRATED FRAMEWORK

5.1 A. Seven-Layer Architecture

The proposed framework contains seven interacting layers. Layer 1, Repository and Acquisition, identifies and captures manuscript images from trusted collections. Layer 2, Metadata and Provenance, records descriptive and technical information. Layer 3, Image Assessment and Preprocessing, diagnoses degradation and applies conservative enhancement. Layer 4, AI Enhancement, selects appropriate restoration, inpainting or super-resolution models. Layer 5, OCR and Knowledge Extraction, converts image content into machine-readable representations. Layer 6, Evaluation, measures image and text performance. Layer 7, Expert Validation and Preservation, determines whether derivatives are suitable for research access and stores the full provenance chain.

The following figure shows the framework layers.

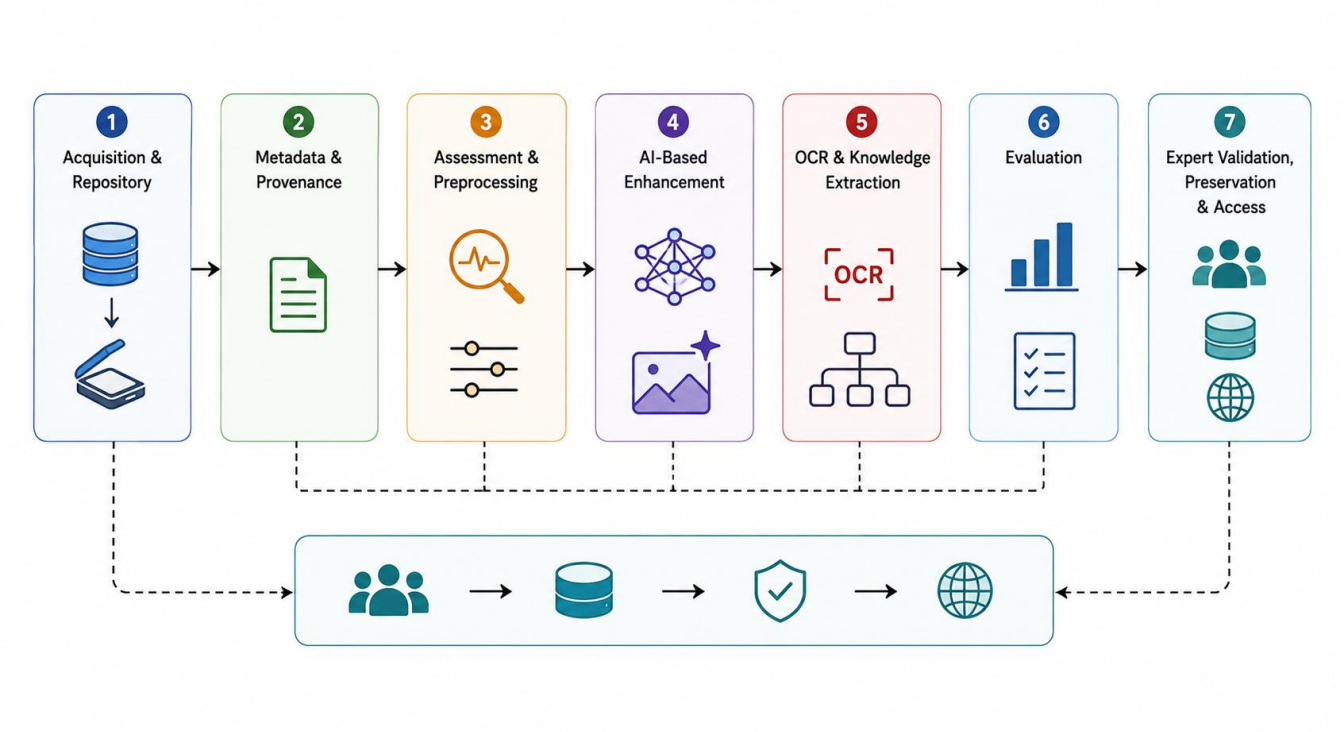

Figure 1: Framework Layers

5.2. Degradation-Aware Model Selection

A central design principle is that the system should respond to degradation rather than applying a fixed sequence of algorithms. Faded ink may require local contrast enhancement and restoration; uneven illumination may require background normalization; random noise may require denoising; low resolution may justify super-resolution; and physical holes or missing regions may require selective inpainting. The model-selection decision should be recorded because it is part of the digital object's history.

The framework can be implemented as a rule-assisted AI controller. Initial image statistics and a lightweight classifier identify dominant degradation categories. A decision layer then selects candidate processing paths. The output of each path is evaluated using both image and OCR criteria. If an enhancement improves image metrics but decreases OCR or expert readability, the system can reject it or route it for human review. This makes the framework adaptive without allowing an optimization metric to become the sole definition of preservation quality.

5.3 Palm-Leaf Manuscript Case

Palm-leaf manuscripts provide a representative test case because their material structure can be mistaken for text. Fibre lines, punch holes, scratches and curved writing surfaces can interfere with segmentation. Recent Tamil work demonstrates that specialized preprocessing can address irregular lines and punch-hole artifacts before character recognition [4], [17]. In the proposed framework, palm-leaf images would first undergo material-aware quality assessment. Regions containing punch holes can be masked, but the original image remains unchanged. Line detection can then proceed on the derivative, followed by script-specific recognition. Any region reconstructed through AI is explicitly marked as inferred.

5.4 Faded Text Case

Faded text requires a different strategy. Excessive global contrast can amplify background texture along with characters. Local contrast methods, illumination correction and learned restoration may be evaluated in parallel. The preferred output is the representation that increases character visibility while retaining weak strokes and preserving the visual relationship between text and substrate. Expert review is particularly valuable for faded material because a model can transform a barely visible stroke into an apparently confident character.

5.5 Damaged and Missing Regions

Generative inpainting can be useful when researchers need a visual hypothesis about a damaged region, but it should never silently replace the source. The system should maintain a region mask showing where generation occurred, attach the model version and inference parameters, and present the output with a visible status such as "AI-generated reconstruction." If multiple reconstructions are plausible, the system should preserve alternatives rather than selecting one without evidence. This approach aligns with trustworthy-AI principles emphasizing transparency and human oversight [5].

5.6 Low-Resolution Images

Super-resolution can improve the visibility of small textual structures, but generated high-frequency detail is not necessarily recovered historical information. The framework therefore recommends using super-resolution primarily as an inspection aid and recording the original resolution alongside the enhanced output. OCR should be evaluated on both the source and super-resolved versions so that gains can be separated from potential recognition artifacts.

6. DISCUSSION

6.1 Implications for Digital Preservation

The framework suggests a shift from digitization as static image capture toward digital preservation as a managed ecosystem of representations. A single manuscript page can generate several legitimate digital objects: the archival master, an access derivative, an enhanced image, an OCR transcription, an expert annotation layer and, where

appropriate, a clearly marked reconstruction. The value of this ecosystem lies in traceability. Researchers can choose the representation appropriate to their purpose without losing access to the original evidence.

This approach also supports scalable access. Large collections cannot be manually enhanced page by page. AI can assist with quality assessment, batch preprocessing, candidate OCR and identification of difficult regions. Human experts can then concentrate on pages or regions where uncertainty is highest. Such a human-in-the-loop model is more realistic than either fully manual preservation or fully autonomous restoration. The recent direction of Indian manuscript digitization toward large-scale repositories and AI-assisted transcription makes this division of labour increasingly relevant [6], [7].

6.2 Implications for Indian Knowledge Systems

For IKS research, improved image and text accessibility can support several scholarly activities. First, searchable OCR can help locate occurrences of technical terms across large collections. Second, normalized metadata can connect manuscripts with authors, subjects, places and dates. Third, image-text alignment can allow a researcher to inspect the original character whenever an OCR result is uncertain. Fourth, enhanced images can support teaching and public engagement without exposing fragile originals to unnecessary handling. Finally, computational analysis can reveal patterns across collections that are difficult to identify through isolated reading.

The diversity of Indian scripts means that a national-scale system should not be built around a single universal OCR model. Instead, a shared infrastructure can provide common provenance, storage, evaluation and access services while allowing script-specific models. Malayalam, Tamil, Devanagari, Grantha and other historical writing systems can then have dedicated recognition components trained on appropriately curated data. This modular architecture is preferable because model performance depends on script, period, material, writing style and degradation.

6.3. Authenticity, Ethics and Trust

The principal ethical risk is not simply algorithmic error; it is the possibility that a plausible AI reconstruction will be mistaken for historical evidence. In cultural heritage, visual plausibility can be persuasive even when unsupported. Trustworthy-AI research therefore emphasizes transparency, explainability, contextual interpretation and bias mitigation [5]. For manuscript preservation, these principles translate into concrete requirements: preserve originals, label generated regions, expose provenance, retain model versions, record uncertainty, permit comparison with the source, and require expert approval before reconstructed content is incorporated into a scholarly edition.

The framework also recognizes that “authenticity” has multiple dimensions. Pixel authenticity concerns whether the digital image faithfully represents the captured object. Material authenticity concerns the relationship between the image and physical manuscript. Textual authenticity concerns whether a transcription represents what is actually visible. Scholarly authenticity concerns whether interpretation is supported by appropriate evidence. A system may improve one dimension while weakening another. Therefore, no single image-quality score can certify preservation quality.

6.4 Data Governance and Long-Term Sustainability

AI preservation also introduces data-management requirements. Training datasets need documentation of source collections, permissions, annotation procedures and demographic or script coverage. Models should be versioned so that future researchers can understand why two processing runs differ. Storage systems should retain both original and derived data, with checksums to detect accidental alteration. Where repositories interoperate, persistent identifiers and standardized metadata become essential. These practices are consistent with the broader conceptual direction of Indian manuscript digitization toward sustainable digital preservation [20].

7. LIMITATIONS AND FUTURE RESEARCH

7.1 Limitations of the Present Evaluation

The first limitation is dataset size. Approximately 100 images are useful for a pilot study but are not sufficient to represent the full diversity of Indian manuscript traditions. The images also originate from heterogeneous repositories, which can introduce differences in capture quality, resolution, compression and metadata completeness. A second limitation is the use of summary performance ranges. Without image-level paired measurements, confidence intervals, statistical significance tests and script-wise breakdowns, the reported improvements cannot be generalized beyond the evaluated sample.

A third limitation concerns ground truth. Reference images of pristine manuscript pages are rarely available for genuinely historical degradation. PSNR and SSIM therefore have restricted interpretive value when the "correct" original is unknown. A high score against a synthetic reference may not correspond to historical fidelity. Expert evaluation and evidence-aware criteria are consequently essential. A fourth limitation is OCR dependence on engine and script. Improvements measured using one OCR system may not transfer to another system or to a different script.

A fifth limitation concerns generative reconstruction. Inpainting and super-resolution can generate visually coherent details that are not supported by the source. Such outputs can be valuable for research exploration but should not be silently integrated into authoritative editions. Finally, the framework assumes access to computational resources, suitable metadata and domain experts. These conditions may not be available in smaller repositories, suggesting a need for lightweight and locally deployable tools.

7.2. Future Research Directions

Future work should develop larger multiscript datasets with standardized annotations for degradation, text lines, characters and uncertain regions. Dataset construction should include different materials, historical periods and writing styles. Synthetic degradation can be used to increase training data, but real degraded examples should remain central to evaluation. Controlled benchmark datasets would allow restoration systems to be compared fairly rather than through isolated case studies.

Script-specific restoration and OCR models are another priority. Rather than assuming a universal model, researchers can use multilingual foundation models as shared representations while adapting smaller components to individual scripts and periods. Recent Tamil studies show that specialized segmentation and recognition can be highly effective when the pipeline reflects manuscript-specific characteristics [4], [17]. Similar research is needed for Malayalam and other Indic manuscript traditions.

A further direction is evidence-constrained generative restoration. Instead of allowing a generative model to produce unrestricted content, the model could be constrained by visible stroke fragments, character inventories, palaeographic rules, linguistic context and parallel manuscript witnesses. The output would then be presented as a ranked set of hypotheses with confidence estimates and evidence links. This would move generative AI from unconstrained image synthesis toward scholarly decision support.

Multimodal manuscript search is also promising. A future system could combine image embeddings, OCR, metadata, script information, linguistic annotations and knowledge graphs. A researcher might search by a word, a visual character form, a subject, a manuscript location or a related concept and receive results linked back to the original image. Such systems would transform manuscript repositories from static archives into active research environments.

8. CONCLUSION

This expanded study presents an integrated framework for AI-assisted digital preservation of Indian Knowledge Systems through degraded manuscript images. The framework combines acquisition, provenance, preprocessing, deep-learning enhancement, selective generative reconstruction, super-resolution, OCR, quantitative evaluation and expert validation. The pilot results reported in the underlying study indicate improvements in structural similarity, reconstruction quality, contrast, noise reduction, edge preservation and OCR performance. These findings support the proposition that AI can move manuscript preservation beyond basic digitization toward intelligent access and analysis.

The central preservation principle, however, is that improved readability must not be confused with historical truth. AI-generated pixels, OCR text and scholarly interpretations are derived representations and must remain distinguishable from the original manuscript evidence. By retaining originals, documenting transformations and placing experts in the validation loop, the proposed framework seeks to combine technological scalability with scholarly responsibility. In the context of India's expanding manuscript digitization agenda, this approach provides a practical foundation for trustworthy, multilingual and evidence-aware digital preservation.

REFERENCES

1. K. Nikolaidou, M. Seuret, H. Mokayed, et al., "A survey of historical document image datasets," International Journal on Document Analysis and Recognition, vol. 25, pp. 305–338, 2022, doi: 10.1007/s10032-022-00405-8.
2. Z. Anvari and V. Athitsos, "A survey on deep learning based document image enhancement," arXiv:2112.02719, 2021.
3. Z. Ziran, M. Mecella, and S. Marinai, "AI-driven enhancement of historical documents," International Journal on Digital Libraries, vol. 26, art. 22, 2025, doi: 10.1007/s00799-025-00430-y.
4. S. Uma Maheswari, P. Uma Maheswari, and G. R. Sai Aakaash, "An intelligent character segmentation system coupled with deep learning based recognition for the digitization of ancient Tamil palm leaf manuscripts," npj Heritage Science, vol. 12, art. 342, 2024, doi: 10.1186/s40494-024-01438-4.
5. M. Paolanti, E. Frontoni, and R. Pierdicca, "Towards trustworthy AI in cultural heritage," npj Heritage Science, 2026, doi: 10.1038/s40494-026-02403-z.
6. Ministry of Culture, Government of India, "Gyan Bharatam Mission," Government of India, 2026.
7. Ministry of Culture, Government of India, "Launch of Gyan Bharatam National Manuscript Survey to Map India's Manuscript Heritage," 16 March 2026.
8. I. Rabaev and M. Litvak, "Recent advances in text line segmentation and baseline detection in historical document images: A systematic review," International Journal on Document Analysis and Recognition, vol. 29, pp. 3–39, 2026, doi: 10.1007/s10032-025-00526-w.
9. M. R. Gupta, N. P. Jacobson, and E. K. Garcia, "OCR binarization and image pre-processing for searching historical documents," Pattern Recognition, vol. 40, no. 2, pp. 389–397, 2007, doi: 10.1016/j.patcog.2006.04.043.
10. K. Ntirogiannis, B. Gatos, and I. Pratikakis, "Performance evaluation methodology for historical document image binarization," IEEE Transactions on Image Processing, vol. 22, no. 2, pp. 595–609, 2013, doi: 10.1109/TIP.2012.2219550.
11. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. MICCAI, LNCS 9351, pp. 234–241, 2015, doi: 10.1007/978-3-319-24574-4_28.
12. J. Yu et al., "Generative image inpainting with contextual attention," in Proc. IEEE CVPR, pp. 5505–5514, 2018.
13. G. Liu et al., "Image inpainting for irregular holes using partial convolutions," in Proc. ECCV, pp. 89–105, 2018.
14. X. Wang et al., "ESRGAN: Enhanced super-resolution generative adversarial networks," in Proc. ECCV Workshops, 2018.
15. R. Smith, "An overview of the Tesseract OCR engine," in Proc. ICDAR, pp. 629–633, 2007.
16. P. Kahle, S. Colutto, G. Hackl, and G. Mühlberger, "Transkribus—A service platform for transcription, recognition and retrieval of historical documents," in Proc. ICDAR, pp. 19–24, 2017, doi: 10.1109/ICDAR.2017.307.
17. S. Pavithra, S. M. Sanjay Vikas, and S. S. Kumar, "An ancient Tamil palm leaf manuscripts script recognition and restoration using ConvNeXt tiny framework," npj Heritage Science, 2026, doi: 10.1038/s40494-026-02559-8.
18. Y. Assael, T. Sommerschield, and J. Prag, "Restoring ancient text using deep learning: A case study on Greek epigraphy," in Proc. EMNLP-IJCNLP, pp. 6369–6376, 2019.
19. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, 2004.
20. G. Dinda and Z. Rahman, "Towards a digital future: A conceptual framework for manuscript preservation system in India under National Mission for Manuscripts (NMMs)," Journal of Information Management, vol. 12, no. 1, pp. 54–73, 2025, doi: 10.5958/2348-1773.2025.00003.0.
21. Z. Shi, S. Setlur, and V. Govindaraju, "Digital image enhancement of Indic historical manuscripts," in Guide to OCR for Indic Scripts: Document Recognition and Retrieval, Springer, pp. 249–267, 2009.
22. National Mission for Manuscripts, "National Database of Manuscripts—Kriti Sampada/Pandulipi Patala," Ministry of Culture, Government of India.
23. Muktabodha Indological Research Institute, "Muktabodha Digital Library."
24. Central Institute of Classical Tamil, "Pavendhar Library—Manuscript Catalogue and Digital Manuscript Archive," Chennai, India.
25. Hill Museum & Manuscript Library, "HMML Reading Room and manuscript collections," Hill Museum & Manuscript Library.
26. S. Rahal, L. Vögtlin, and R. Ingold, "Historical document image analysis using controlled data for pre-training," International Journal on Document Analysis and Recognition, vol. 26, pp. 241–254, 2023.
27. J. Martínek, L. Lenc, and P. Král, "Building an efficient OCR system for historical documents with little training data," Neural Computing and Applications, vol. 32, pp. 17209–17227, 2020, doi: 10.1007/s00521-020-04910-x.

28. UNESCO, “Guidelines for the Preservation of Digital Heritage,” Paris, France: UNESCO, 2003.
29. “Historical document image analysis using controlled data for pre-training” and related benchmark literature, as summarized in recent historical-document dataset surveys [1], [26].
30. Y. Ding, S. C. Han, J. Lee, et al., “Deep learning based visually rich document content understanding: A survey,” Artificial Intelligence Review, vol. 59, art. 114, 2026, doi: 10.1007/s10462-025-11477-3.