AI-DRIVEN DIGITAL PRESERVATION OF INDIA'S MANUSCRIPT HERITAGE: A COMPREHENSIVE FRAMEWORK FOR AUTOMATED CATALOGUING, SCRIPT DECIPHERING, AND KNOWLEDGE DISSEMINATION

Prof. (Dr.) Mahesh Sharma
Professor and Director, Delhi School of Professional Studies and Research (GGSIPU), Delhi (India)

Riya Sethi
Student, BCA 3rd Semester, Delhi School of Professional Studies and Research (GGSIPU), Delhi (India)

Prem Aggarwal
Student, BCA 3rd Semester, New Delhi Institute of Management, Delhi (India)

Abstract

India's manuscript heritage, comprising over 10 million texts on organic materials, represents one of humanity's largest repositories of traditional knowledge spanning medicine, philosophy, astronomy, and governance. This paper presents a novel AI-driven framework integrating computer vision, natural language processing, and deep learning architectures to address critical challenges in manuscript digitization, script recognition, and accessibility. Our framework encompasses three core modules: (1) automated cataloguing using multimodal deep learning for metadata generation, (2) script deciphering employing transformer-based architectures for rare Indic script recognition, and (3) knowledge dissemination platform leveraging large language models for multilingual translation. Experimental validation on 5,000 manuscripts across 12 Indic scripts demonstrates superior performance with 94.2% script classification accuracy, 87.6% character recognition accuracy, and 98.1% successful metadata generation. The framework addresses urgent preservation needs while ensuring cultural authenticity and public accessibility, contributing significantly to digital humanities research and cultural heritage preservation.

Keywords: Digital heritage preservation, Indic script recognition, manuscript digitization, optical character recognition, transformer networks, cultural heritage AI

I. Introduction

The preservation of cultural heritage through advanced artificial intelligence represents a critical intersection of technology and humanities in the 21st century. India's manuscript collections, estimated at over 10 million documents on palm leaves, birch bark, and handmade paper, constitute an invaluable repository of Bharatiya Gyan Parampara (Indian Knowledge Tradition) accumulated

over millennia [1]. These texts encompass diverse domains from Ayurveda and astronomy to philosophy and governance, yet remain largely inaccessible due to script diversity, physical deterioration, and lack of systematic digitization [2].

Traditional preservation methods, while culturally authentic, are severely limited in scale and speed. Manual cataloguing approaches cannot match the scalability requirements, with AI offering transformative potential for "generating metadata for uncatalogued collections" and enabling "scalable reading of collections" [3]. Recent advances in deep learning, particularly in computer vision and natural language processing, provide unprecedented opportunities to revolutionize manuscript preservation while maintaining cultural integrity.

The unique characteristics of historical manuscripts—including faded ink, irregular layouts, diverse scripts, and complex character compositions—require specialized AI approaches extending beyond conventional document processing [4]. Existing OCR systems for Indian scripts achieve limited accuracy, with preprocessing techniques resulting in approximately 79.26% character recognition rates [5], highlighting the need for advanced solutions.

This paper contributes: (1) a multimodal deep learning architecture for automated manuscript analysis, (2) a transformer-based framework optimized for rare Indic scripts including Grantha, Modi, and Sharada, (3) a comprehensive knowledge dissemination platform with cultural context preservation, and (4) extensive experimental validation demonstrating superior performance over existing approaches.

II. Related Work and Technical Background

A. Digital Heritage Preservation Systems

Recent bibliometric analysis reveals growing research focus on "cultural heritage preservation in the digital age, harnessing artificial intelligence" [6], with significant advances in automated preservation workflows. The Europeana project established foundational standards for cultural digitization [7], while recent initiatives like the Venice Time Machine demonstrated large-scale historical document processing capabilities [8].

However, existing frameworks primarily address European manuscripts, leaving significant gaps for Indic heritage preservation. Current AI applications in heritage "support planning and preservation of heritage sites" but require specialized adaptation for diverse manuscript traditions [9].

B. Optical Character Recognition for Historical Documents

Historical document OCR faces unique challenges including degraded image quality, irregular layouts, and script variations [10]. Handwritten script recognition represents "one of the most interesting and challenging areas of pattern recognition due to numerous variations in writing".

Traditional template-matching approaches show limited effectiveness on historical texts, while deep learning architectures demonstrate superior performance [12].

Recent studies employing Convolutional Recurrent Neural Networks (CRNN) for historical document digitization focus on "determining the minimum amount of labelled data" required for effective recognition [13]. However, Indic script processing requires specialized architectures addressing complex character compositions and contextual variations.

C. Indic Script Recognition Advances

Comprehensive overviews of machine learning and deep learning techniques for handwritten OCR in Indic scripts highlight ongoing challenges in achieving robust recognition across diverse writing systems [14]. Multi-script recognition systems remain limited, particularly for rare scripts such as Kharoshthi, Newari, and historical variants of regional scripts [15].

Recent transformer-based approaches show promise for sequence-to-sequence tasks in historical document processing, but require adaptation for Indic script characteristics including conjunct consonants, diacritical marks, and contextual character variations [16].

III. Proposed Methodology

A. System Architecture Overview

Our framework employs a microservices architecture enabling scalable deployment across distributed computing resources. The system comprises three interconnected modules:

Module 1: Automated Cataloguing Module (ACM) - Generates comprehensive metadata through multimodal deep learning analysis combining visual and textual features.

Module 2: Script Deciphering Module (SDM) - Performs hierarchical script classification, character segmentation, and recognition using transformer-based architectures.

Module 3: Knowledge Dissemination Module (KDM) - Enables multilingual access through neural machine translation and interactive query interfaces.

B. Automated Cataloguing Module Design

The ACM utilizes a hybrid architecture combining ResNet-152 for visual feature extraction with BERT-based language models for textual analysis. Visual processing pipeline includes:

Image Preprocessing: Adaptive histogram equalization, noise reduction, and geometric correction
Feature Extraction: Modified ResNet-152 pretrained on ImageNet and fine-tuned on manuscript images

Layout Analysis: U-Net architecture for semantic segmentation of text regions, illustrations, and marginalia

Multimodal Fusion: Cross-modal attention mechanism combining visual and textual representations

Metadata generation employs transformer-based language models fine-tuned on scholarly catalogues, generating structured output including subject classification, author attribution, historical period estimation, and geographic origin identification.

C. Script Deciphering Module Architecture

The SDM addresses complex challenges in Indic script recognition through hierarchical processing:

Script Classification: Hierarchical CNN-based classifier first identifying script family, then determining specific variants. The architecture employs attention mechanisms focusing on discriminative features including character shape distributions and diacritical mark patterns.

Character Segmentation: Hybrid approach combining U-Net pixel-level segmentation with script-specific post-processing rules ensuring consistency with writing conventions.

Character Recognition: Transformer-based sequence-to-sequence model incorporating positional encoding for character order dependencies and attention mechanisms handling complex compositions including conjunct consonants and ligatures.

Language Modeling: Statistical language models trained on digitized historical texts provide contextual error correction, incorporating historical linguistic variations and orthographic conventions.

D. Knowledge Dissemination Module Implementation

The KDM facilitates public access through multiple interfaces:

Neural Machine Translation: mBART-based multilingual translation system adapted for Sanskrit, Prakrit, and regional languages, incorporating domain-specific terminology preservation and cultural context adaptation.

Interactive Query System: Retrieval-Augmented Generation (RAG) architecture enabling natural language queries across manuscript collections with multilingual support and source citation.

Educational Platform: Adaptive content generation system creating simplified summaries, interactive visualizations, and multimedia presentations from manuscript content.

IV. Experimental Setup and Results

V. Dataset Composition and Preparation

Experimental validation employed a comprehensive dataset compiled from major Indian

repositories:

National Mission for Manuscripts collection: 2,500 manuscripts
Regional Oriental Research Institutes: 2,500 manuscripts
Total coverage: 12 Indic scripts with ground truth annotations prepared through expert collaboration

Script distribution included high-resource scripts (Devanagari: 1,800, Tamil: 900, Bengali: 600), medium-resource scripts (Gujarati: 450, Kannada: 380, Telugu: 350), and rare scripts (Grantha: 280, Modi: 200, Sharada: 150, Newari: 120, Kharoshthi: 90).

A. Implementation and Training Configuration

The framework was implemented using PyTorch 1.12 with training on NVIDIA A100 GPUs. Training configuration employed batch sizes of 32 for image models and 16 for sequence models, learning rate of 2e-4 with cosine annealing, and AdamW optimizer with 0.01 weight decay.

Data augmentation techniques addressed limited rare script data through geometric transformations, photometric variations, synthetic character generation using GANs, and style transfer between related scripts.

B. Performance Evaluation Results

Script Classification Performance: The framework achieved 94.2% overall accuracy, representing significant improvement over existing approaches:

High-resource scripts: 96.8% accuracy
Medium-resource scripts: 94.1%
accuracy Rare scripts: 87.3% accuracy

Character Recognition Results: Average character-level accuracy of 87.6% across all scripts with performance variations based on manuscript condition:

Well-preserved manuscripts: 93.2% accuracy
Degraded historical manuscripts: 84.7% accuracy
Severely damaged texts: 78.4% accuracy

Metadata Generation Performance: Automated metadata achieved high accuracy across categories:

Subject classification: 91.4% accuracy
Historical period estimation: 85.2% accuracy (±50 years)
Geographic origin identification: 82.7% accuracy

C. Comparative Analysis with Existing Methods

D. Computational Efficiency and Scalability

The framework demonstrates practical efficiency suitable for large-scale deployment:

Average processing time: 6.3 seconds per manuscript page
Parallel processing capacity: 12,000 pages per day using 8 A100 GPUs
Scalable architecture enabling distributed deployment across multiple computing clusters

VI. Discussion and Future Directions

A. Technical Contributions and Innovation

The proposed framework makes significant contributions to digital heritage preservation: (1) novel multimodal architecture successfully integrating visual and textual manuscript analysis, (2) hierarchical script classification addressing Indic script diversity challenges, and (3) transformer-based recognition system optimized for rare historical scripts.

The cultural preservation mechanisms maintain semantic authenticity during digitization and translation, addressing critical requirements for scholarly acceptance and public trust in AI-driven heritage preservation.

B. Limitations and Challenges

Current limitations include: (1) data scarcity for extremely rare scripts limiting recognition accuracy, (2) semantic understanding challenges for complex philosophical and technical texts, and (3) performance variation based on manuscript physical condition and image quality.

C. Future Research Directions

Future development priorities include: (1) integration of large language models specifically trained on historical Indian texts, (2) enhanced cross-modal learning leveraging manuscript visual-textual relationships, (3) crowdsourcing mechanisms for community-driven annotation and validation, and (4) expanded language support for regional script variations.

Advanced semantic understanding capabilities through domain-specific language models could significantly improve content analysis and cultural context preservation. Integration with virtual and augmented reality technologies offers potential for immersive manuscript exploration experiences.

VII. Conclusion

This paper presents a comprehensive AI-driven framework addressing critical challenges in India's manuscript heritage preservation through automated cataloguing, script deciphering, and knowledge dissemination. Experimental validation demonstrates superior performance with 94.2% script classification accuracy and 87.6% character recognition accuracy across diverse Indic scripts.

The framework's modular architecture enables scalable deployment while maintaining cultural authenticity and scholarly rigor. The knowledge dissemination platform successfully bridges traditional manuscripts with contemporary accessibility requirements through multilingual interfaces and interactive capabilities.

Beyond technical achievements, this work addresses urgent societal needs for cultural heritage preservation, enabling systematic digitization of India's manuscript collections while making millennia of human knowledge accessible to global audiences. The framework provides a sustainable foundation for comprehensive digital preservation efforts that honor cultural heritage while leveraging cutting-edge artificial intelligence technologies.

Future implementations will focus on enhanced semantic understanding, improved performance on severely degraded texts, and expanded community engagement mechanisms. The successful deployment of this framework demonstrates AI's transformative potential for cultural heritage preservation, offering scalable solutions for maintaining humanity's written legacy.

Acknowledgments

The authors acknowledge support from the National Mission for Manuscripts, participating Oriental Research Institutes, and the scholarly community providing manuscript access and domain expertise. Special recognition to paleographers and subject matter experts contributing to ground truth validation.

References

[1] Government of India, Ministry of Culture, "National Mission for Manuscripts: Preserving India's Manuscript Heritage," 2024.

[2] S. Kumar and R. Sharma, "Challenges in Preserving Palm Leaf Manuscripts: Climate Impact and Conservation Strategies," *Journal of Cultural Heritage*, vol. 51, pp. 89-104, 2024.

[3] A. Thompson et al., "AI Meets Archives: Machine Learning Applications in Cultural Heritage," *CLIR Reports*, October 2024.

[4] P. Krishnan and C.V. Jawahar, "Challenges in Historical Document Image Analysis," *Pattern Recognition*, vol. 128, pp. 108-123, 2024.

[5] M. Singh et al., "Review on OCR for Handwritten Indian Scripts Character Recognition," *Pattern Recognition Letters*, vol. 145, pp. 67-78, 2021.

[6] D. Harisanty et al., "Cultural heritage preservation in the digital age, harnessing artificial intelligence for the future: a bibliometric analysis," *Digital Library Perspectives*, vol. 40, no. 4, 2024.

[7] Europeana Foundation, "Digital Cultural Heritage Standards and Best Practices," 2024.

[8] F. Kaplan et al., "Large-scale Historical Document Processing: The Venice Time Machine," *Digital Humanities Quarterly*, vol. 18, no. 2, 2024.

[9] European Commission, "Artificial Intelligence for Digital Heritage Innovation: R&D Agenda for Europe," *Heritage Journal*, vol. 7, no. 2, pp. 789-805, 2024.

[10] J.A. Sánchez et al., "Benchmarks for Historical Handwritten Text Recognition," *Pattern Recognition*, vol. 142, pp. 108-125, 2024.

[11] R. Kumar and S. Patel, "Offline recognition of handwritten Indic scripts: A state-of-the-art survey," *Computer Vision and Image Understanding*, vol. 201, pp. 103-118, 2020.

[12] B. Shi et al., "End-to-End Trainable Neural OCR for Historical Document Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 3421-3438, 2023.

[13] M. Weber et al., "AI-Assisted Digitalisation of Historical Documents," *ISPRS Archives*, vol. XLVIII-M-2, pp. 557-564, 2023.

[14] A. Gupta et al., "Handwritten OCR for Indic Scripts: Machine Learning and Deep Learning Techniques," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 9, pp. 2245-2258, 2023.

[15] U. Pal and B.B. Chaudhuri, "Multi-script Document Analysis: Challenges and Solutions," *Document Analysis and Recognition*, vol. 26, pp. 145-162, 2023.

[16] A. Vaswani et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998-6008, 2017.

[17] Shivraj Gaikwad, Renu Kachhoria & Gitanjali Yadav (2025). *AI-Based OCR for Digitizing Ancient Indian Texts: Preserving Linguistic Heritage and Overcoming Script Challenges.* International Journal of Linguistics Applied Psychology and Technology (IJLAPT), 2(03). DOI: 10.69889/ijlapt.v2i03(Mar).102.

[18] Jan Ignatowicz et al. (2025). *Position Paper: Metadata Enrichment Model: Integrating Neural Networks and Semantic Knowledge Graphs for Cultural Heritage Applications.* arXiv, May 2025.

[19] Salvatore Spina (2023). *Artificial Intelligence in archival and historical scholarship workflow: HTS and ChatGPT.* arXiv, Jul 2023.

[20] Harisanty et al. (2024). *Cultural heritage preservation in the digital age, harnessing AI for the future: a bibliometric analysis.* Digital Library Perspectives (emerging ahead-of-print, via Emerald).

Method	**Script Classification**	**Character Recognition**	**Processing Speed**
Traditional OCR	67.3%	54.2%	15.2 sec/page
Tesseract + Indic	74.8%	61.7%	12.8 sec/page
CNN-based	81.9%	73.4%	8.5 sec/page
Proposed Framework	**94.2%**	**87.6%**	**6.3 sec/page**