WORLD JOURNAL OF PHARMACEUTICAL RESEARCH
Coden USA: WJPRAP Impact Factor 8.63
Volume 15, Issue 17, 1928-1937. Research Article ISSN 2277–7105

WHEN *ŚĀSTRA* MEETS ARTIFICIAL INTELLIGENCE: REIMAGINING *ĀYURVEDIC* MANUSCRIPTOLOGY

Dr. Akshita[1]*, Dr. Sapna Kaushish[2]

[1]Post Graduate Scholar, Department of Ayurveda Samhita Evum Siddhant, Shri Dhanwantry Ayurvedic College & Hospital, Chandigarh.

[2]Associate Professor, Department of Ayurveda Samhita evum Siddhant, Shri Dhanwantry Ayurvedic College & Hospital, Chandigarh.

Article Received on 05 August 2026,
Article Revised on 25 August 2026,
Article Published on 01 Sept. 2026,

https://doi.org/10.5281/zenodo.22670451

***Corresponding Author**

Dr. Akshita

Post Graduate Scholar, Department of Ayurveda Samhita Evum Siddhant, Shri Dhanwantry Ayurvedic College & Hospital, Chandigarh.

How to cite this Article: Dr. Akshita[1]*, Dr. Sapna Kaushish[2] (2026). When Śāstra Meets Artificial Intelligence: Reimagining Āyurvedic Manuscriptology. World Journal of Pharmaceutical Research, 15(17), 1928–1937

ABSTRACT

Āyurvedic manuscripts represent an extensive repository of knowledge encompassing classical principles, commentarial traditions, therapeutic practices and specialised concepts. However, access to these manuscripts remain limited due to linguistic and textual complexities. Recent advances in Optical Character Recognition (OCR), Handwritten Text Recognition (HTR), Sanskrit Natural Language Processing (NLP) and Artificial Intelligence (AI) offer new possibilities for transforming these manuscripts into searchable, analysable and globally accessible digital resources. The present study explores the potential application of contemporary technologies at various stages of manuscriptology. The study highlights their application as well as limitations including OCR errors, inadequate domain-specific datasets, contextual ambiguity, limitations of general-purpose NLP tools and hallucinations in generative AI. The integration of these technologies and *Āyurveda*-focused datasets and thorough human validation, can help disseminate traditional wisdom with modern digital science and promote its responsible sharing with a worldwide audience.

KEYWORDS: Manuscriptology, OCR, NLP, Artificial Intelligence, HTR, Digital Humanities, Generative AI.

INTRODUCTION

Āyurveda embodies a rich and ever-developing tradition of knowledge, with its principles passed down through generations via oral traditions, classical *Saṃhitās*, commentaries, treatises, and manuscripts. Although the primary texts like *Caraka Saṃhitā*, *Suśruta Saṃhitā*, and *Aṣṭāṅga Hṛdaya* form the core textual foundation of *Āyurveda*, a significant amount of additional knowledge is safeguarded in manuscripts found in libraries, repositories, private collections, and traditional learning centers. These texts contain valuable insights into pharmacology, surgery and holistic wellness. Thus, *Āyurvedic* manuscripts represent not merely historical documents but living repositories of the intellectual evolution of the *Āyurvedic* tradition.

Preservation of these manuscripts comes with several challenges. The physical state of these manuscripts is often precarious, facing degradation due to time, climate, and handling. The scriptural complexity of Sanskrit, characterized by its use of the *Devanāgarī* script and frequent compound characters, adds another layer of difficulty to its digital preservation.

Traditional manuscriptology encompasses a range of detailed procedures such as cataloguing, digitisation, transcription, transliteration, translation, collation, and critical editing. While digitisation has greatly enhanced preservation efforts, merely having a digital image of a manuscript does not automatically render its contents searchable, comparable, or easily accessible to researchers from different linguistic and geographical areas. Digitization is not merely about scanning images but involves transcribing text into formats suitable for Natural Language Processing (NLP) and knowledge graph construction.[1][2] This technological application is essential for safeguarding the Indian Knowledge System and making it accessible for modern scientific inquiry and contemporary healthcare research.

Recent developments in the areas of Artificial Intelligence (AI), Natural Language Processing (NLP) and Optical Character Recognition (OCR) have opened up new possibilities in this regard. Sanskrit OCR can potentially convert digitised manuscript images into editable and searchable text, while AI-assisted technologies can support transcription, transliteration, translation, textual comparison, terminology identification and semantic retrieval. These technologies may enable researchers to examine large collections of manuscripts more efficiently and identify relationships within texts that would otherwise require extensive manual searching. A manuscript could therefore progress from being a preserved image to becoming a searchable and structured knowledge resource.

The emerging technology should not be viewed as a replacement of traditional manuscript scholarship by artificial intelligence. Rather, it should be understood as a collaborative model in which computational tools enhance the capabilities of the *Āyurvedic* scholar. AI can offer speed, large-scale text processing, pattern recognition and information retrieval, whereas the scholar contributes textual judgement, contextual understanding, knowledge of *śāstric* principles and critical interpretation. Such approach may allow technological advancement while retaining the authenticity and integrity of *Āyurvedic* knowledge.

Against this background, the present article explores the emerging role of Sanskrit OCR and Artificial Intelligence in *Āyurvedic* manuscriptology. It examines their potential applications in manuscript digitisation, transcription, transliteration, translation, textual comparison and *Āyurvedic* knowledge extraction, while also considering their limitations and scholarly risks. Further, it proposes a *Śāstra* - AI assisted Scholar framework through which artificial intelligence can assist in transforming manuscript-based knowledge into reliable, searchable and structured digital resources.

METHODS

The present article adopts a narrative review and conceptual framework approach to examine the emerging role of Artificial Intelligence (AI), Sanskrit Optical Character Recognition (OCR) and related digital technologies in *Āyurvedic* manuscriptology.

Academic databases and scholarly digital resources were searched with combinations of keywords related to *Āyurveda*, manuscriptology, Sanskrit computational linguistics and artificial intelligence to identify relevant literature.

Relevant publications on Sanskrit language technologies, digital humanities, preservation of manuscripts, OCR, NLP, machine translation and AI-assisted textual analysis were considered. The practical applicability of these technologies was also explored by studying digital repositories and publicly available documentation on digitization of manuscripts and computational processing of Sanskrit texts.

The analysis also took into account potential risks of use of AI in manuscriptology such as textual errors due to OCR, inaccurate translation, loss of contextual meaning, misinterpretation of technical *Āyurvedic* terms, propagation of errors to digital databases and over-reliance on automated output.

In this sense the methodology aimed at proposing technological options as well as studying the possibility of incorporating AI in *Āyurvedic* manuscriptology without endangering the textual authenticity, contextual interpretation and traditional scholarly principles of *Śāstra*-based knowledge transmission.

Current OCR Technologies

The conversion of a manuscript from a visual document into machine-readable text is one of the most significant technological steps in digital manuscriptology. Optical Character Recognition (OCR) is a technology that allows computers to recognize characters in an image and convert them into editable and searchable text [3]. With advances in machine learning and artificial intelligence, OCR has developed from simple character recognition to systems that can recognize complex document structures and handwritten material. This development opens up great opportunities for the digitization and study of the *Āyurvedic* manuscripts.

These OCR technologies can be open-source as well as with premium subscriptions and special AI- based systems for historical and handwritten documents.

Open-source options include Tesseract OCR, OCRmyPDF, Calamari OCR, EasyOCR, and Kraken OCR, which offer varying levels of customisation and can be particularly useful for research and specialised text-recognition workflows. Proprietary or premium solutions include ABBYY FineReader, Adobe Acrobat Pro DC, Readiris, Nuance OmniPage, Google Cloud Vision API, and Microsoft Azure Computer Vision, which generally provide user-friendly interfaces and advanced document-processing capabilities.[4]

Among the widely and commonly utilized OCR technologies, Tesseract OCR stands out as a significant open-source choice. Its models trained for specific languages enable researchers to conduct OCR for various scripts and languages, such as Sanskrit. This technology requires elaborate training and pre-processed data.[5] It can be especially beneficial for relatively clear *Devanāgarī* content and provides the extra benefit of customization and model training. This flexibility could be advantageous for creating OCR systems designed for specialized Ayurvedic texts.

Cloud-based AI platforms like Google Cloud Vision OCR offer automated text recognition and presently feature Sanskrit/*Devanāgarī* within their experimental language options.[6] It also requires programming (eg-python script) to interact with Google Cloud Vision API.

Although it possesses utility for large scale projects but it has high subscription costs.[7]

Likewise, user-focused tools like Google Lens can retrieve and translate text from images or scanned documents. These tools are beneficial for initial exploration and quick text retrieval, though their results should not be seen as comparable to a scholarly transcription, especially in relation to historical manuscripts.

A major advancement beyond traditional OCR is Handwritten Text Recognition (HTR). In contrast to conventional OCR, which mainly focuses on printed text, HTR employs machine-learning models that have been trained using handwriting samples.[8] Platforms like Transkribus offer AI-driven recognition for historical and handwritten documents, enabling researchers to utilize or create models tailored to specific collections. The Devanagari Lithography NKP M4 model accessible via Transkribus[9] is especially significant for Sanskrit manuscript studies. It was developed using *Devanāgarī* lithographic texts from the Naval Kishore Press collection, dating from the late 19th to early 20th century, and incorporates Sanskrit as one of its related languages. The reported character error rate for validation is 5.58%, but the developers specifically warn that performance on unfamiliar documents may vary based on handwriting, document quality, and resemblance to the training data.

Sanskrit OCR projects focused on research are also surfacing. Models specifically for Sanskrit have been created to overcome the constraints of general-purpose OCR by training recognition systems using images of Sanskrit documents. These methods are crucial since Sanskrit features intricate conjunct characters and writing patterns that can complicate standard *Devanāgarī* OCR.

Softwares such as Indic NLP Library can be used. It is an open-source Python-based toolkit developed for Natural Language Processing of Indian languages. It provides foundational text-processing capabilities including text normalization, sentence splitting, word tokenization and detokenization, word segmentation, syllabification, script conversion, transliteration, romanization and indicization.[10] For *Āyurvedic* manuscriptology, such functions can become useful after OCR has converted a manuscript image into machine-readable text. The Indic NLP Library is not designed specifically for *Āyurveda*, and its general functions should not be regarded as offering accurate *śāstric* interpretation.

Specialised resources may be needed for *Āyurvedic* technical terms and the analysis of Sanskrit grammar.

Another software is Sanskritshala, it is a neural Sanskrit Natural Language Processing (NLP) toolkit developed by researchers associated with IIT Kanpur and IIT Kharagpur.[11] It provides a web-based interface for computational analysis of Sanskrit and includes modules for word segmentation, morphological tagging, dependency parsing, and compound-type identification. It also provides interactive annotation features through which users can correct the system's predictions, making it useful for creating and refining annotated Sanskrit datasets.

The Sanskrit Heritage platform, linked to the INRIA Sanskrit computational linguistics initiative, offers multiple useful services, such as lexical resources, morphological analysis, segmentation and sandhi division. Its Sanskrit Reader can conduct machine-aided sentence analysis that includes segmentation, sandhi separation and morphological tagging.[12] In *Āyurvedic* manuscriptology, such tools may aid in breaking down a continuous Sanskrit sentence into analytically beneficial elements.

The University of Hyderabad's Sanskrit Computational Linguistics resources similarly provide computational tools for analysing Sanskrit according to the Indian grammatical tradition[13], including *pada pāṭha*, *pada paricaya*, *samasta-pada paricaya* and *karaka anvaya*. Such technologies could be particularly valuable for *Āyurveda* because the meaning of an *Āyurvedic* statement often depends upon grammatical relationships rather than isolated words.

Apart from these, Generative AI platforms such as ChatGPT, Google Gemini, Claude, and DeepSeek can offer additional support in manuscriptology through tasks such as preliminary translation, summarisation, text comparison, terminology exploration, research ideation, and organisation of large textual datasets; however, their outputs require source verification and expert scholarly validation, particularly when dealing with technically specialised *Āyurvedic* and Sanskrit texts.

RESULTS

The review identified a range of contemporary digital technologies that can potentially support different stages of *Āyurvedic* manuscriptology. These technologies demonstrate a

progressive movement from image recognition to linguistic processing and finally to AI-assisted knowledge extraction. Digitisation is the first step in creating a computationally accessible manuscript. High-resolution imaging preserves the physical content of the manuscript in digital form; however, a scanned image remains inaccessible to computational text analysis until its contents are converted into machine-readable text. However, the above discussed software bridges this gap in various aspects. After a manuscript is transformed into machine-readable text, Natural Language Processing (NLP) can be employed to conduct linguistic analysis.[14] This marks a significant shift from recognizing characters to processing language.

Artificial Intelligence can further assist in converting machine-readable Sanskrit into formats accessible to a wider scholarly community. AI-based systems can potentially support transcription correction, transliteration and preliminary translation. Transliteration can convert *Devanāgarī* or other scripts into standard Roman transliteration systems such as IAST, facilitating searching and cross-linguistic research. Translation systems and large language models can then generate preliminary translations into English and other languages.

Recent developments in large language models (LLMs) have expanded the potential applications of AI beyond conventional OCR and NLP.

However, the use of generative AI introduces the possibility of hallucination, in which an apparently convincing response may not be supported by the source text. This is especially concerning in manuscriptology because an AI system may generate a plausible Sanskrit reading, translation or explanation that is absent from the manuscript.

A proposed *Sāstra*-AI assisted Scholar framework that can be used for future developments can be.

Manuscript

↓

Manuscript metadata and cataloguing

↓

Digital preservation and image enhancement

↓

OCR / Handwritten Text Recognition

↓

Human-verified transcription

↓
Transliteration
↓
Indic/Sanskrit NLP
↓
AI-assisted translation and terminology extraction
↓
Textual comparison and variant identification
↓
Expert scholarly validation
↓
Āyurvedic knowledge mapping / knowledge graph
↓
Final Expert Validation and Provenance recording
↓
Multilingual digital dissemination

This framework demonstrates that AI can intervene at multiple stages of manuscriptology, but human scholarly validation must remain embedded throughout the process rather than being treated merely as a final checking step.

DISCUSSION

The results of this review suggest that Artificial Intelligence and similar computational technologies could change *Āyurvedic* manuscriptology from mainly a preservation-focused discipline into a more accessible, searchable, and interconnected domain of digital scholarship. The advancement from digitization and Optical Character Recognition (OCR) to Sanskrit Natural Language Processing (NLP), AI-enhanced translation, and knowledge extraction presents a chance to render a vast and varied collection of *Āyurvedic* textual heritage available to scholars outside the geographical and linguistic limits traditionally associated with its study.

Nonetheless, this change should not be perceived as a substitute for traditional manuscriptology. Instead, AI ought to be seen as a supportive element within the existing academic process. The manuscript continues to be the main source, while computational technologies aid in its processing, and the knowledgeable scholar is accountable for determining the text's reliability and significance.

The most appropriate model for integrating AI into manuscriptology is not "AI versus scholar", but "AI with scholar." AI can identify patterns, but the scholar determines their *śāstric* significance. The scholar contributes capabilities that cannot be reduced to pattern recognition. These include knowledge of Sanskrit grammar, manuscript traditions, textual criticism, *Āyurvedic* principles, commentarial literature and contextual interpretation. This distinction becomes especially important when dealing with technical *Āyurvedic* concepts whose meanings depend upon context and established textual traditions. The present technological environment highlights a significant gap: the majority of existing OCR, NLP, and generative AI tools are general-use or generally focused on Sanskrit/Indic rather than being specifically tailored for *Āyurvedic* texts. *Āyurvedic* literature includes specialized vocabulary, intricate compounds, synonyms, formulations, technical classifications, and meanings that depend on context. A general Sanskrit language model might identify a word but not grasp its particular *Āyurvedic* meaning.

CONCLUSION

Artificial Intelligence, Sanskrit OCR, Handwritten Text Recognition and Natural Language Processing offer new possibilities for the preservation, processing and global dissemination of *Āyurvedic* manuscripts. They can transform manuscript images into searchable text and support transcription, translation, textual comparison and structured knowledge extraction. Nonetheless, technological results cannot replace *śāstric* knowledge. Errors in OCR, constraints in Sanskrit NLP, contextual uncertainties, and generative AI inaccuracies require thorough human validation. Consequently, advancements in the future necessitate Āyurveda-specific datasets, tailored NLP tools, and scholar-guided AI systems.

The way forward is a *Śāstra*-AI- Scholar model, where AI provides scale, speed and accessibility, while the *Āyurvedic* scholar ensures textual authenticity, contextual understanding and interpretive integrity. Thus, AI should not replace engagement with *Śāstra*, but extend its reach into the digital age while preserving its intellectual foundations.

REFERENCES

1. Ashi Maheshwari, Reema Ajmera, Dinesh Kumar Dharamdasani. International Journal on recent and innovation trends in computing and communication. ijritccorg [Internet], 2023 Oct 30 [cited 2026 Aug 19]; 11(11). Available from: http://www.ijritcc.org
2. Hrishikesh Terdalkar, Bhattacharya A. Framework for question-answering in sanskrit through automated construction of knowledge graphs. ACL Anthology [Internet], 2019

Oct [cited 2026 Aug 19];97–116. Available from: https://aclanthology.org/W19-7508/

3. Digital Tools for South Asian Studies [Internet]. Prakrit.info. 2023 [cited 2026 Aug 19]. Available from: http://prakrit.info/dh/ocr.html
4. Jayaraj HT, Sourav K, Resmi B. Artificial intelligence powered OCR models for digitizing ayurveda manuscripts. Journal of Ayurveda and Holistic Medicine (JAHM), 2024 Oct 16;12(9)
5. Tesseract-ocr. Tesseract-OCR/Tesseract [Internet]. GitHub. 2019 Oct [cited 2026 Aug 19]. Available from: https://github.com/tesseract-ocr/tesseract
6. Index. Cloud vision API. google cloud documentation [Internet]. Google Cloud Documentation. 2025 [cited 2026 Aug 19]. Available from: https://docs.cloud.google.com/vision/docs/reference/rest?apix=true
7. Digital Tools for South Asian Studies [Internet]. Prakrit.info. 2023 [cited 2026 Aug 19]. Available from: http://prakrit.info/dh/ocr.html
8. What is handwritten text recognition (HTR)? | Complete guide [Internet]. Llamaindex.ai. 2026 [cited 2026 Aug 19]. Available from: https://www.llamaindex.ai/glossary/what-is-handwritten-text-recognition
9. merkel@ub.uni-heidelberg.de. Devanagari lithography NKP M4 [Internet]. Transkribus. 2026 [cited 2026 Aug 19]. Available from: https://www.transkribus.org/models/devanagari-lithography-nkp-m4
10. Anoopkunchukuttan. GitHub - anoopkunchukuttan/indic_nlp_library: Resources and tools for Indian language natural language processing [Internet]. GitHub. 2021 May [cited 2026 Aug 19]. Available from: https://github.com/anoopkunchukuttan/indic_nlp_library
11. Jivnesh. GitHub - jivnesh/SanskritShala: SanskritShala: A neural sanskrit NLP toolkit with web-based interface [Internet]. GitHub, 2023 [cited 2026 Aug 19]. Available from: https://github.com/Jivnesh/SanskritShala
12. Inria.fr. 2026 [cited 2026 Aug 19]. Available from: https://sanskrit.inria.fr/index.en.html
13. संसाधनी [Internet]. Uohyd.ac.in, 2011 [cited 2026 Aug 19]. Available from: https://sanskrit.uohyd.ac.in/scl/
14. Katru P, Sharma A. Natural language processing in Ayurveda: Unlocking the wisdom of ancient Ayurveda texts with artificial intelligence. Journal of Research in Ayurvedic Sciences, 2026 Apr; 10(Suppl 1): S136–40.