International Journal of Research in Library Science (IJRLS)
ISSN: 2455-104X
DOI: 10.26761/IJRLS.11.4.2025.1977
Volume 11, Issue 4 (October-December) 2025, Page: 79-99, Paper ID: IJRLS-1977

Keyword Extraction From Manuscript Images: Scholarly Access and Sustainable Preservation of India's Classical Knowledge Systems

Shiva Kanaujia Sukula

Dr. B. R. Ambedkar Central Library, Jawaharlal Nehru University, New Delhi, India

ABSTRACT

One of the major research concerns have grown in the area of keyword extraction from manuscript images during last decade. Manuscripts, specially found in Sanskrit language have observed the keyword extraction from images and this has developed a lot curiosity among the professionals and scholars. The emergence of technological innovations as well as including them for the implementation purposes have advanced the functionality as well as research attention. These innovations are comprised of computer vision, language processing under the umbrella of artificial intelligence. The literature has shown inclination towards recognition accuracy and deep-learning; the examples include OCR systems as well as CNN-BiLSTM in the Sanskrit manuscript. The complexity of Sanskrit manuscripts and images in them reflect need of increasingly efficient techniques. The requirement of advanced preprocessing, segmentation methods, and other steps in order to improve the quality of manuscripts for further information retrieval and access are much paid attention.

To augment readability of such degraded manuscripts, few techniques have gained popularity. The literature reflected the steps, techniques and methods such as semantic binarization and Gabor filtering. There have been areas of concern such as degradation of manuscripts, word recognition and NLP-based approaches in order to enhance morphological analysis. These techniques are essential for topic modelling as well as semantic keyword identification in large-scale digitized texts as studies have highlighted. The literature has shown the glimpses of creation of benchmark datasets. These developments have been instrumental in creation of standardized evaluation frameworks, supporting multilingual accessibility through transliteration and neural translation. This study is based on review of recent literature focusing on keyword extraction from Sanskrit manuscript images, which has applied scoping review in the later section. Also, the article has identified Core Keywords which indicate the scattered subject matters across the literature. This study has also endeavoured to identify AI-powered digitization which has an active transitioned role in the preservation of cultural heritage. This study has reviewed the advancements during 2015-2025, have transformed Sanskrit manuscript studies ranging from manual transcription to intelligent digital analysis, ensuring both scholarly access and sustainable preservation of India's classical knowledge systems.

KEYWORDS: Sanskrit manuscripts, OCR, keyword extraction, images, deep learning, NLP, digitization, heritage preservation, Devanāgarī

INTRODUCTION

Extracting Keywords from Images in Sanskrit Manuscripts: narration into text

During the last decade (2015-2025), there have been various componential needs such as image processing, natural language processing, and AI-influenced preservation necessities including feature extraction and OCR sytem in order to alter the Sanskrit manuscript information access research. The observations made about the transition from character recognition towards semantic approaches specially for keyword extraction have encompassed the amalgamation of Information and communication technology and conventional knowledge system. The literature has the visuals or transformation; which is discerned from isolated OCR tasks to full-spectrum digital ecosystems. The tasks of extracting and interpreting text along with ensuring safeguards are pivoted in the direction of revitalizing the cultural heritage embedded in Sanskrit manuscripts.

1. Evolution of Optical Character Recognition (OCR) and Character Recognition Frameworks

The extraction of textual information in the recent past from Sanskrit manuscript images has been characterised by Optical Character Recognition (OCR) technologies as basic progress made. Few developments in the field, such as discussed in Kataria and Jethva (2019, 2021), initiated the application of CNN-BiLSTM and convolutional neural networks (CNNs) for recognizing Sanskrit characters. This has been instrumental in setting a strong precedent for later developed and applied deep learning applications. Also observed few years ago, Narang et al. (2019, 2020, 2021) explored various aspects such as statistical feature extraction, SIFT, and Gabor-based methods for Devanagari character recognition. As found, DeepNetDevanagari introduced as augmented feature learning for document scripts. Other recent works observed in Kulkarni et al. (2022) and Valaboju et al. (2025), have proposed semi-automatic recognition systems. Such systems are smart in identification, digitization, and translation of Sanskrit manuscripts. These technologies have been integrating OCR with linguistic models for the puspose of end-to-end text recognition.

Most recent advances, have been including Kore et al. (2025), Dhruva et al. (2023), and Jindal and Ghosh (2024) which have narrated the deep learning frameworks' contributions. It is observed that combination of CNN, LSTM, and hybrid semi-supervised models are capable of achieving utmost accuracy in recognizing degraded and handwritten Sanskrit texts. The studies and contrasting analyses by Deepthi and Seenu (2022) and Moudgil et al. (2021) have consolidated these developments. These developemtns have been able to establish a transition from traditional template-matching OCR to adaptive neural architectures. The other cross-script extensions are covering Tamil, Pali, Tulu, Grantha, and Malayalam (Subramani & Murugavalli, 2019; Jayashree et al., 2024; Maheswari et al., 2024; Samantaray et al., 2025) which have further enriched recognition pipelines and dataset diversity.

In entirely, OCR evolution has been instrumental in transforming Sanskrit manuscript interpretation. This has changed from a manual decoding process into an automated, intelligent recognition system. This capability is observed as extracting searchable text from complex historical imagery.

2. Image Preprocessing, Segmentation, and Feature Extraction

There are two components such as image preprocessing and segmentation which are responsible to form the backbone of accurate OCR and keyword extraction. There have been observations about historical Sanskrit manuscripts often reveal degradation, ink fading, or palm leaf damage, requiring advanced denoising and restoration techniques. A significant study has pioneered (Dubey, 2018) and introduced digital restoration approaches for Devanagari manuscripts, while (Sudarsan and Sankar, 2022, 2024) there have been complete denoising frameworks and ensemble neural models for Malayalam palm leaf manuscripts. Few revolutionary techniques are including tri-level semi-adaptive approaches and deep semantic binarization networks ((Shobha Rani et al., 2022; Rani, 2024). The methods observed in these studies are found effective for degraded image correction. For segmentation and feature extraction (Vijitha et al., 2024; Mehta and Doshi, 2020), there has been application of morphological and edge-based image processing to isolate handwritten Sanskrit words and text lines. The utility of Gabor filters for extracting localized text patches in Balinese palm leaf manuscripts (Kesiman and Pradnyana, 2020), as well as employing Gabor–Random Forest hybrids for degradation assessment (Kowshik et al., 2024) are few examples. Using orientation-shape context histograms to classify Grantha script, and designing ensemble classifiers to segment touching Brahmi characters (Raj et al., 2017; Mali, 2021) are other significant applications.

Hybrid optimization and feature selection strategies are observed in the recent past which are including nature-inspired algorithms, wavelet-CNN models, and histogram balancing (Khaparde et al., 2023; Shelke et al., 2025; Maheshwari et al., 2024). Together, these methods played a significant role in improved segmentation accuracy. The methods displayed across the studies have placed them as instruments bringing accuracy in segmentation as well as securing the crucial role of preprocessing and feature extraction. Such techniques are crucial for advancing manuscript images preparation for OCR, and leading towards subsequent keyword identification.

3. Keyword Extraction, Word Recognition, and Natural Language Processing Integration

In addition to character-level recognition, semantic extraction of keywords from Sanskrit manuscripts relies on several aspects. These aspects are generally identified as word-level recognition, natural language integration, and contextual understanding. Few studies have explored word spotting and retrieval techniques for Devanagari manuscripts, which has paved the way for content-based image retrieval (Varghese and Govilkar, 2015, 2023). As observed (Kesima, 201; Mohammed et al., 2021) this has been advanced by developing pattern detection and word recognition systems that made manuscript images searchable even without full textual transcription.

Various recent NLP-integrated models provide a crucial leap forward. Recently (Krishnan et al., 2025) normalized datasets for Sanskrit morphological parsing and segmentation are observed which are significant for enabling machines to process syntactic structures essential for accurate keyword extraction. There has been demonstration of the use of NLP in extracting medical and philosophical concepts from Sanskrit classics such as Charaka Samhita and Sushruta Samhita , application of topic modeling and naming entity recognition for semantic analysis of Ramayana, and abstractive summarization for Sanskrit prose (Bagchi et al., (n.d.); Jain et al., 202; Thottempudi, 2021; Sinha and Jha, 2020).

Cross-linguistic NLP efforts on neural machine translation, universal networking grammar, and morphological excellence (Kumar et al., 2024; Sitender and Bawa, 2022; Tapaswi , 2025) expanded Sanskrit keyword extraction to multilingual and semantic networks. Marking a paradigm shift from techniques such as surface-level recognition to

methods as meaning-based retrieval, have been able in transforming Sanskrit manuscripts into machine-understandable linguistic resources.

4. Digitization, Preservation, and Cultural Informatics of Sanskrit Heritage

The culmination of OCR, segmentation, and NLP advances have been emerging from the digitization and preservation of Sanskrit heritage manuscripts. Major dataset contributions, such as standardized benchmark corpora for Indic manuscripts, introduction of HMPLMD for Malayalam texts continued to invite attention (Kesiman et al., 2018; Dhruva et al., 2023; Nair and Rani, 2023). Heritage computing and technological initiatives as in the studies (Borthakur, 2021; Acri, 2024) emphasized upon AI-assisted manuscript conservation, preventive measures and contextual cataloguing. As significant for image-based conservation technologies, which are including automatic damage identification using 'SegFormer' (Wang et al., 2024) and spectralization with VGG-16 (Nair et al., 2023), have provided basis for improved digital archive quality.

After these many applications and experimentations, few cross-disciplinary applications have emerged including machine learning and NLP for Pali manuscript conservation, integration of deep learning with cultural emotion analysis; and ancient astrological manuscripts with predictive ML models (Gudadhe et al., 2024; Geethanjali and Valarmathi, 2025; Ray, 2025). For enhanced multilingual accessibility, there have ben roles of transliteration and script conversion systems (Mubarakkaa et al., 2024; Pradeep et al., 2024). A convergence of computer vision, linguistics, and cultural preservation has paved way for transformation of fragile Sanskrit manuscripts into user friendly interactive, searchable, and educational resources. The aspects such as digitization initiatives and highly acclaimed advancements are dancing in the field as well as inviting attention towards the ethical preservation of Sanskrit's intellectual legacy for future generations.

Core Keywords (Grouped by Theme)

For the present study, the data corpus (as retrieved references) is ranging from AI-based extraction, OCR, and NLP approaches in the context of Sanskrit and related Indic manuscripts (2015–2025). The focus has been maintained on the theme "extracting keywords from images in Sanskrit manuscripts" and this has been used for discerning the subject matter.

Core Keywords (Grouped by Theme)

1. Optical Character Recognition (OCR) and Deep Learning

- OCR
- Handwritten Character Recognition
- Devanagari OCR
- Sanskrit Manuscript OCR
- Convolutional Neural Networks (CNN)
- Recurrent Neural Networks (RNN)
- Bidirectional LSTM (BiLSTM)
- Capsule Networks (CapsNet)
- Ensemble Neural Network
- Hybrid Deep Learning

- Semi-Supervised Learning
- DeepNetDevanagari
- VGG-16 Model
- CNN-LSTM Architecture

Key References:

Kataria & Jethva (2019, 2021), Kore et al. (2025), Narang et al. (2021, 2022), Jindal & Ghosh (2024), Moudgil et al. (2023), Shelke et al. (2025), Valaboju et al. (2025), Kulkarni et al. (2022).

2. Image Preprocessing and Segmentation

- Image Processing
- Binarization
- Image Restoration
- Noise Removal
- Image Denoising
- Thresholding
- Gamma Variation
- Histogram Balancing
- False Color Spectralization
- Gabor Filters
- Discrete Wavelet Transform (DWT)
- Damage Detection
- Palm Leaf Degradation Assessment
- Feature Extraction
- Dimensionality Reduction

Key References:

Sudarsan & Sankar (2022, 2024), Dubey (2018), Nair et al. (2023), Unnikrishnan et al. (2025), Kowshik et al. (2024), Maheshwari et al. (n.d.), Tomar et al. (2015), Mehta & Doshi (2020).

3. Word and Character Segmentation

- Word Segmentation
- Character Segmentation
- Line Segmentation
- Touching Character Separation
- Nandinagari Recognition
- Indic Script Segmentation

Key References:

Vijitha et al. (2024), Krishnan et al. (2025), Guruprasad & Rao (2021), Mali (2021), Mehta & Doshi (2020), Lomte & Doye (2022).

4. Dataset Development and Benchmarking

- Dataset Generation
- Normalized Dataset
- Benchmark Databases
- Sanskrit Character Dataset
- Palm Leaf Manuscript Dataset
- Multilingual Indic Script Dataset

Key References:

Dhruva et al. (2023), Kore et al. (2025), Kesiman et al. (2018, 2020), Nair & Rani (2023), Valy et al. (2017), Singh et al. (2018), Narang et al. (2022).

5. Keyword Extraction, Feature Learning, and Information Retrieval

- Keyword Extraction
- Feature Extraction
- Information Retrieval
- Word Spotting
- Pattern Detection
- Semantic Analysis
- Morphological Parsing
- Named Entity Recognition (NER)
- NLP-Based Knowledge Extraction
- Topic Modeling
- Cultural Emotion Analysis

Key References:

Singh & Ahuja (2019), Varghese & Govilkar (2015, n.d.), Mohammed et al. (2021), Thottempudi (2021), Bagchi et al. (n.d.), Jain et al. (2025), Tapaswi (2025), Geethanjali & Valarmathi (2025).

6. Script Diversity and Multilingual Processing

- Devanagari
- Sanskrit
- Tamil
- Malayalam
- Brahmi
- Grantha
- Balinese
- Tulu
- Pali

- Nandinagari
- Multilingual Indic OCR
- Script Identification
- Transliteration

Key References:

Subramani & Murugavalli (2019), Maheswari et al. (2024), Mubarakkaa et al. (2024), Poddar & Gupta (2023), Agarwal et al. (2023), Raj et al. (2017), Vijayalakshmi & Gnanasekar (2022), Jayashree et al. (2024), Gudadhe et al. (2024).

7. Natural Language Processing (NLP) and Digital Humanities

- Sanskrit Computational Linguistics
- Sanskrit Word Segmentation
- Translation (English ↔ Sanskrit, Brahmi → Tamil)
- Text Summarization
- Semantic Parsing
- Morphological Analysis
- NMT (Neural Machine Translation)
- Digital Preservation
- Cultural Heritage Computing
- Sanskrit Spellcheck
- Sanskrit Digital Infrastructure

Key References:

Sinha & Jha (2020), Kumar et al. (2024), Pradeep & Mamidi (2025), Sitender & Bawa (2022), Ray (2025), Chand et al. (2023), PRADEEP et al. (2024).

8. Preservation, Accessibility, and Heritage Studies

- Manuscript Digitization
- Palm Leaf Conservation
- Heritage Informatics
- Preventive Conservation
- Digital Archives
- Cultural Tourism
- Historical Text Restoration
- Damage Identification

Key References:

Borthakur (2021), Kesiman et al. (2018), Nair et al. (2023), Wang et al. (2024), Gudadhe et al. (2024), Acri (2024).

Consolidated Keyword List (Alphabetically)

AI, Ancient Manuscripts, Binarization, CNN, Character Recognition, Cultural Heritage, Denoising, Devanagari Script, Digital Preservation, Feature Extraction, Handwritten Text Recognition, Heritage Informatics, Image Processing, Indic Scripts, Information Retrieval, Keyword Extraction, LSTM, Machine Learning, Manuscript Digitization, Morphological Parsing, NLP, OCR, Palm Leaf Manuscripts, Sanskrit, Segmentation, Semantic Analysis, Topic Modeling, Transliteration, Word Recognition.

DATA PRESENTATION AND RELATION

Sl. No.	Section	Theme	Focus Area Covered by References	Representative Reference Numbers
1.	Manuscript Heritage and Conservation	Historical and cultural studies of manuscripts	Region, script diversity, traditional heritage practices	1, 5, 11, 33
2.	Manuscript Heritage and Conservation	Physical conservation and degradation analysis	Preservation, damage assessment, degradation levels	4, 28, 35, 45, 56, 68, 73, 80
3.	Manuscript Heritage and Conservation	Digital preservation initiatives	Making manuscripts digitally accessible	6, 13, 52, 62
4.	Pre-processing and Restoration for OCR	Noise removal, binarization, enhancement	Improving readability of degraded manuscripts	10, 35, 45, 56, 61, 68, 73
5.	Pre-processing and Restoration for OCR	Segmentation of characters and words	Separating touching, noisy, ancient characters	37, 39, 79
6.	Pre-processing and Restoration for OCR	Image patching and zoning	Specific feature extraction areas	12, 24
7.	OCR and Character Recognition in Indic Scripts	Deep learning models for OCR	CNN, hybrid CNN, LSTM, GAN models	7, 18, 19, 21, 22, 26, 36, 42, 47, 60, 69
8.	OCR and Character Recognition in Indic Scripts	Script-specific character recognition studies	Sanskrit, Tamil, Malayalam, Gujarati, Brahmi, Balinese, Nandinagari	4, 14, 17, 23, 30, 34, 37, 38, 43, 46, 55, 67
9.	OCR and Character Recognition in Indic Scripts	Word-level recognition	Word prediction, segmentation solutions	9, 23, 32
10.	Datasets and Benchmarks	Benchmark datasets for Indic scripts	Public datasets for palm leaf and historic manuscripts	8, 25, 44, 63, 75

11.	Datasets and Benchmarks	Dataset generation frameworks	Systematic frameworks for dataset building	27, 48
12.	Feature Extraction and ML Techniques	Feature engineering and hybrid methods	SIFT, Gabor filters, wavelet transforms	24, 37, 46, 54, 60
13.1	Feature Extraction and ML Techniques	Script identification and retrieval	Auto identification of multilingual Indic scripts	50, 76, 77
14.	NLP and Computational Linguistics on Sanskrit and Ancient Texts	NLP for Sanskrit and Ayurveda texts	Extraction, spellcheck, segmentation	3, 16, 29, 53, 65, 70
15.	NLP and Computational Linguistics on Sanskrit and Ancient Texts	Translation and transliteration	Sanskrit and Indian languages to modern scripts	30, 31, 43, 66
16.	NLP and Computational Linguistics on Sanskrit and Ancient Texts	Text summarization and retrieval	Knowledge mining, topic modeling	59, 71
17.	NLP and Computational Linguistics on Sanskrit and Ancient Texts	Cultural tourism and multimedia use cases	Enhancing public access via audio, multilingual support	6
18.	Surveys and State-of-the-Art Reviews	OCR surveys on manuscripts and Indic languages	Focused reviews on recognition challenges	7, 20, 41, 58, 64, 72, 78
19.	Surveys and State-of-the-Art Reviews	Digital humanities and Sanskrit infrastructure surveys	Infrastructure, tools, challenges	51
20.	Cross-disciplinary and Emerging Applications	Heritage emotion analysis, astrology ML models	Cultural computing with ML and social science integration	11, 57

1. **Table 1:** Various Sections and Themes Scattered throughout the Retrieved Literature

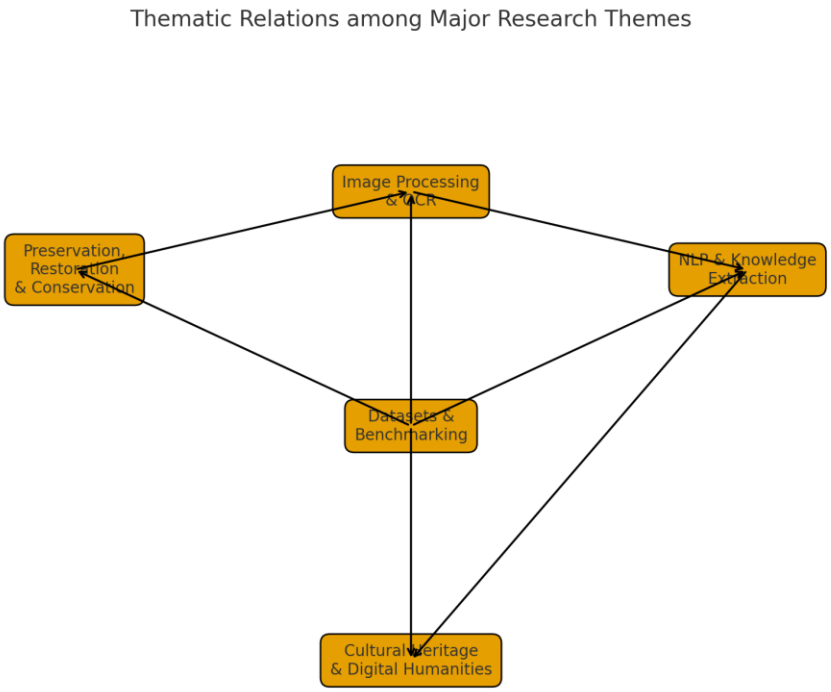

Figure: 1 Thematic relational framework illustrating how preservation, digitization, language processing, and cultural applications integrate to support digital heritage research.

Figure 1 conveys and reflects the relationships between the various components where the reservation enables image processing and OCR; where the role of OCR kicks in, it feeds the NLP and information extraction workflows; while NLP drives the Cultural Heritage and digital humanities advantages, and datasets support and strengthen each component at various stages.

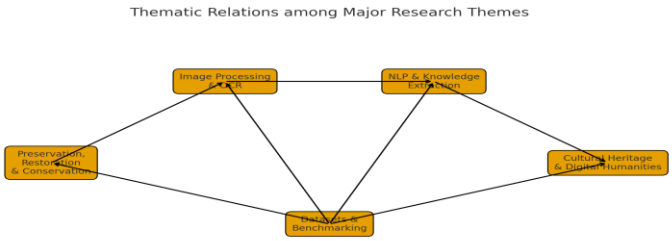

Figure 2: Process Flow

Figure 2 illustrates the stages of the process flow: from preservation to OCR, then to NLP, to the next stage, and to Cultural Heritage impact. How is this possible? Datasets and benchmarks are crucial, as these two components

support all processes and the status achieved at different stages, and by being accountable for the tasks, they improve results throughout the entire process.

OBSERVATIONS

Though there have been various developments yet few major advances occurred during 2015–2025 concerned with extracting keywords from Sanskrit manuscript images are discussed as following:

1. OCR improvements: Deep learning-based OCR models (such as CNN-BiLSTM and hybrid LSTM networks) have been observed to improve the recognition accuracy of Sanskrit and Indic scripts, even for degraded manuscripts.
2. Image preprocessing: Improved denoising, three-level thresholding, and semantic binarization techniques have proven useful in restoring the intelligibility and clarity of palm leaf and paper manuscripts and improving text readability.
3. Segmentation and feature extraction: Gabor filters, wavelet transforms, and joint classifiers have been developed, enabling accurate segmentation of complex characters and ligatures specific to Sanskrit scripts.
4. Word Recognition: The techniques such as word spotting and content-based retrieval systems have been applied for direct keyword identification from manuscript images without full transcription.
5. NLP Integration: In order to achieve morphological analysis, topic modeling, and named entity recognition, combining OCR with natural language processing has been useful. Such applications have helped in enriching semantic keyword extraction.
6. Dataset Development: Standardized datasets (HMPLMD and DeepNetDevanagari) have contributed in augmenting model training and benchmarking for manuscript digitization.
7. Multilingual Access: For the expansion of accessible Sanskrit, Hindi, Tamil, and Pali manuscripts, transliteration and neural translation systems have significantly supported.
8. Digital Preservation: During the recent past, studies have shown the value of AI-powered digitization, damage detection, and cultural informatics platforms. These components in the ongoing revolution have ensured the long-term conservation. Such developments have led towards the interactive accessibility of Sanskrit texts.

Scoping Review: Extraction of Keywords from Images in Sanskrit Manuscripts

The present study has further deliberated on comprehensive and narrowly focused corpus for a scoping review on "Extraction of Keywords from Images in Sanskrit Manuscripts" is based on 80 works related with the specific purpose.

1. Introduction

The significant technological transformation regarding the extraction of meaningful textual and semantic content from Sanskrit manuscripts is observed between 2015 and 2025. Since the ancient times, Indian manuscripts, were often written on palm leaves or birch bark. These were written in scripts such as Devanāgarī, Grantha, Nandināgarī, or Tamil, and embodying vast cultural, philosophical, and scientific knowledge. By the passing of time, however, there has been degradation. Also factors such as non-standardized writing, and intricate ligatureshave been responsible for grave challenges for keyword extraction.

New mechanisms for image preprocessing, character recognition, and text segmentation has been possible through the recent advances in Artificial Intelligence (AI), Optical Character Recognition (OCR), Natural Language Processing (NLP). In these areas, Deep Learning (DL) has been instrumental further in the directions of transliteration, and semantic indexing.

Using a diverse and rich set of articles, various studies in this scoing review has finalised and included 80 review articles which were published during 2015 and 2025. Help of Google scholar platform was taken to collect data and browse literature during October 2025. The core areas among these articles, studies and entire set of browsed literature were identified as applicable techniques, datasets and stereotyping for keyword extraction in the context of textual, image form and linguistic information from digitised Sanskrit manuscripts and few others.

2. Methodological Overview

It is observed that most of the studies employ a three-stage workflow:

1. Image pre-processing and enhancement: There are various aspects such as managing manuscript distortions as well as degradation. There are needs of readability restorations as well as maintaining legibility in the ancient texts.
2. Text recognition and segmentation: Text recognition and the segmentation associate with OCR models as well as other methods such as BiLSTM, in the context of Digital library architectures. To recognise characters, texts in image forms, such techniques are needed.
3. Keyword extraction and NLP-based semantic analysis: For the purpose of morphological analysis, parsing and text summarization This phase involves linguistic labeling to extract key concepts.

There are studies where datasets such as DeepNetDevanagari, HMPLMD, and Sleukrith Set (Narang et al., 2021; Nair and Rani, 2023; Valy et al., 2017) are discussed as contributing developments as benchmarks for Indic writing systems. Also, it is observed that Sanskrit-specific corpora in published literature (Krishnan et al., 2025; Dhruva et al., 2023) ahs a significant role in downstream linguistic analysis.

3. Theme I: Image Preprocessing, Restoration, and Segmentation

3.1 Image Restoration and Noise Reduction

There are observations that palm leaf manuscripts often suffer from ink fading, fungal spots, and cracks. There are methods like gamma variation and histogram balancing, semi-adaptive thresholding, and deep semantic binarization (Maheshwari et al., 2023; Shobha Rani et al., 2022; Rani, 2024) which have improved text clarity.

The works shared such as Sudarsan & Sankar (2022, 2024) and Unnikrishnan et al. (2025) have proposed denoising pipelines for Malayalam and Tamil manuscripts, and they have demonstrated up to 95% restoration efficiency. It is found that machine learning–based degradation assessment (Kowshik et al., 2024; Wang et al., 2024) further automates damage detection using techniques such as CNN and transformer models like SegFormer.

3.2 Text Segmentation and Layout Analysis

For accessibility purpose, precise segmentation is significantly vital for isolating Sanskrit words and ligatures. In few studies (Vijitha et al., 2024; Mehta & Doshi, 2020), there have been discussions on morphological segmentation and projection profile techniques. These steps are useful for word boundary detection.

Hybrid algorithms are there which are combining connected component analysis, zone detection, and clustering. These algorithms have been reflected to outperform traditional heuristics in ancient manuscripts/texts layouts (Tomar et al., 2015; Kulkarni et al., 2022).

4. Theme II: Optical Character Recognition (OCR) and Deep Learning Approaches

4.1 Neural OCR and Convolutional Models

During the long phase of years 2019 and 2025, most of the Sanskrit manuscript recognition studies have adopted Convolutional Neural Networks (CNNs). These are often paired with Bidirectional Long Short-Term Memory (BiLSTM) layers (Kataria & Jethva, 2021; Kore et al., 2025). It is found that these architectures achieved over 90% accuracy for segmented character images.

For improved feature extraction from degraded characters, variants such as CapsNet (Moudgil et al., 2023) and VGG-16 fine-tuning (Nair et al., 2023) have been instrumental. OCR tools as discussed in few works (Guruprasad & Rao, 2021; Shelke et al., 2025) have introduced specialized recognition systems for Nandināgarī and Sanskrit scripts.

4.2 Feature Engineering and Traditional Classifiers

Early models relied on handcrafted features like SIFT, Gabor filters, and zoning (Narang et al., 2019; Puri & Singh, 2019). Later on, deep hybrid systems (Narang et al., 2021; Narang et al., 2022) have optimized feature sets for ancient Devanāgarī, resulting in the DeepNetDevanagari dataset.

As observed (Raj et al., 2017; Mali, 2021), the segmentation of touching characters and complex ligatures is possible by using ensemble classifiers and histogram orientation techniques.

5. Theme III: Dataset Development and Benchmarking

While going through literature, it is discerned that the advancement in the past decade has been in the form of dataset creation for Indic scripts.

- DeepNetDevanagari standardized Devanāgarī OCR benchmarks (Narang et al., 2021).
- HMPLMD provided handwritten Malayalam manuscript data (Nair & Rani, 2023).
- Building a comprehensive Sanskrit character dataset (Dhruva et al., 2023).
- Valy et al. (2017) and Kesiman et al. (2018, 2019) focused on Southeast Asian palm leaf manuscripts.

Benchmarking initiatives also compared models across mixed scripts like Bangla, Roman, and Devanāgarī. It is observed that such corpus expansion facilitates fair comparison. This further includes transfer learning across Indic OCR systems (Singh et al., 2018; Kesiman et al., 2018).

6. Theme IV: Natural Language Processing and Keyword Extraction

6.1 Morphological and Lexical Analysis

There are few examples of advanced studies which have shown glimpses into integrated OCR outputs with NLP for morphological parsing, word segmentation, leading towards named entity recognition. Krishnan et al. (2025) created a "normalized Sanskrit dataset" for morphological tagging, while few other recent works (Tapaswi, 2025; Pradeep & Mamidi, 2025) have surveyed computational linguistics tools supporting Sanskrit lemmatization and dependency parsing. Studies (Sinha & Jha, 2020; Thottempudi, 2021) have shown the application of extractive and abstractive text summarization in the context of Sanskrit prose and epics, which are very useful for keyword generation.

6.2 Semantic and Thematic Keyword Extraction

It is found that keyword extraction from digitized manuscripts are now capable to leverage the topic modeling (LDA), and semantic embeddings. These steps are also associated with context-free grammar–based parsing (Sitender & Bawa, 2022; Bagchi et al., 2024). NLP-driven retrieval from Ayurvedic texts reflect how computational

tools aid domain-specific indexing. Such examples are observed like Charaka Samhita and Sushruta Samhita (Jain et al., 2025). To support multilingual access and transliteration and linking Sanskrit with modern Indic and English corpora has been possible by such objectives of semantic and specific keyword extraction (Mubarakkaa et al., 2024; Kumar et al., 2024).

7. Theme V: Heritage Preservation, Accessibility, and Cross-Script Digitization

7.1 Multilingual and Transliteration Efforts

Recent work have showcased the stages and forms of transliteration across scripts, Brahmi to Tamil, English to Sanskrit NMT, and multi-script identification, facilitating accessibility of digitized archives (Mubarakkaa et al., 2024; Kumar et al., 2024; Poddar & Gupta, 2023). It is also observed that AI models are enabling cross-lingual search and translation of Sanskrit content (Chand et al., 2023).

7.2 Digital Humanities and Cultural Informatics

during last two years, few examples have exemplified digital heritage approaches combining computational linguistics with cultural analytics (Acri, 2024; Ray, 2025). In the recent past, semi-automatic systems merged machine learning and NLP for manuscript conservation, metadata and structural tagging for Sanskrit textual transmission also took place (Preisendanz, 2018; Valaboju et al., 2025, Gudadhe et al., 2024).

8. Emerging Trends and Research Gaps

There have been several steps and stages for the techniques and methods in this context reflecting the following trends which emerged during last decade (2015-2025):

1. Shifting from classical image processing and retrieval to deep neural architectures, especially CNN-BiLSTM hybrids.
2. Augmented availability of datasets for Sanskrit as well as related Indic scripts.
3. Integrating OCR outputs with NLP pipelines for the task of keyword and theme extraction.
4. Expanding into various meaningful aims such as transliteration, and semantic retrieval; the focus has shifted towards multilingual alignment using neural translation.
5. Adopting transformer-based architectures (e.g., SegFormer, ViT) for degradation and damage analysis.
6. The enhanced interdisciplinarity linking computer vision with subjects such as philology, archaeology, and digital humanities.

Observing the trends over the time duration, there are still few gaps such as:

- Limited end-to-end Sanskrit OCR-to-keyword pipelines.
- Disarrayed and inconsistent ground truth annotation for rare ligatures.
- Inquisitiveness for context-aware keyword extraction which is possible by using large language models, specially fine-tuned on Sanskrit corpora.
- Lack of availability of open-source tools in the context of enhancing multimodal retrieval across text and image.

9. Future Directions

The next phase of research should address interoperability between OCR, natural language processing (NLP), and semantic knowledge graphs to enable automatic keyword indexing of Sanskrit manuscripts.

Possible directions include:

1. Exploiting transformation-based OCR models for complex Indic scripts.
2. Developing large-scale Sanskrit LLM programs that also include philological metadata.
3. Applying AI tools for repair purpose related to missing segments.
4. Reconstructing the incomplete texts.
5. Creating the unified databases for access at larger scale.
6. Integrating various tools such as augmented reality (AR) and AI visualization tools for learning access as well as dissemination.

With the context of global endeavours, such directions and actions are contributing in digital preservation and computational philology. The purposes are ensuring that Sanskrit manuscripts remain accessible and semantically searchable for future generations.

10. Single Unique Finding

With the application of scoping review, the clarity in the ongoing researches is visible form the 80 publications from 2015–2025. This review reveals a decisive transformation in Sanskrit manuscript research. The shifting from manual transcription toward AI-driven digital interpretation has been a message for the future.

ROLE OF ARTIFICIAL INTELLIGENCE IN THE RESEARCH TRENDS

AI is not only shaping but also driving the research trends. This comprehensive technological advancement has been playing pivotal role in many areas yet in this study. It is also discussed in the literature as the same has emerged as the enabler in preservation of Sanskrit manuscripts. The roles of AI are ranging from revolutionizing the traditional workflows of digitization, recognition, and semantic analysis. There are few elaborated aspects given as following:

1. Intelligent Image Enhancement and Restoration

AI-based image processing models such as deep convolutional autoencoders, GANs, and transformer vision networks, have been instrumental in replacing manual restoration techniques. These models trained on degraded manuscript datasets automatically identify cracks, noise, and ink fading, restoring lost clarity (Sudarsan & Sankar, 2022; Unnikrishnan et al., 2025). Such methods are useful in permitting the accurate OCR input preparation and reduction in the dependency on human conservators.

2. Deep Learning in Character Recognition

There is a clear transition from handcrafted feature extraction to the deep neural architectures (for example, CNN–BiLSTM, CapsNet, VGG-16). This transition characterizes Sanskrit OCR research as observed in few studies (Kataria & Jethva, 2021; Kore et al., 2025). These models are instrumental in enabling the recognition of complex ligatures, and compound consonants. It is also possible to recognise irregular handwriting thus historically unsolved challenges in Sanskrit and other Indic scripts.

3. AI-Driven Keyword and Concept Extraction

AI-powered Natural Language Processing (NLP) methods include semantic embeddings, topic modeling, and contextual word clustering. These are capable to identify key philosophical or textual terms from OCR outputs. AI-based keyword extraction and Latent Dirichlet Allocation (LDA) (Bagchi et al., 2024; Sitender & Bawa, 2022) are now able to detect thematic structures in Sanskrit texts. Such advancements are now improving digital retrieval accuracy.

4. Cross-Script and Multilingual Learning

AI's neural machine translation (NMT) and transliteration systems are smartly representing as a bridge Sanskrit. Models (Mubarakkaa et al., 2024; Kumar et al., 2024) apply attention-based transformers for cross-lingual mapping. These models and techniques are making Sanskrit manuscripts accessible.

5. Heritage Preservation and Predictive Restoration

In a significant development areas, AI-driven damage prediction models (Wang et al., 2024) evaluate the manuscript fragility and degradation using visual data. This further escalate towards the knowledge graphs and metadata tagging (Valaboju et al., 2025) in the direction of automated cataloging. Transformation of cultural heritage management from reactive to preventive preservation has been possible due to AI-driven predictive approaches.

6. Integrative AI Ecosystems and Semantic Interoperability

The entire visualisation of emerging frameworks link OCR, NLP, and Knowledge Graphs into unified pipelines. This single picture permits for further allowing machines not just to read but to understand Sanskrit manuscripts. There are continuous transitions into "Transformer-based models and large language frameworks" for being trained on Sanskrit corpora in order to achieve future semantic indexing.

CONCLUSION

AI is the core research driver redefining how Sanskrit manuscripts are restored, read, and interpreted, thus believing AI no longer an auxiliary tool, only. With the applications of deep learning, and semantic reasoning it is leading way towards the multilingual intelligence. The entire spectrum is regarding the AI enabled end-to-end workflows from image to insight. The whole exercise is towards bridging the ancient and the digital worlds (Sukula, 2025). The examples of chapter naming and the transmission of embedded texts as well as Devanagari character classification in printed and handwritten documents using SVM, the Image Patch Extraction and handwritten Vedic Sanskrit text recognition using deep learning as research moves toward integrated deep-learning ecosystems and semantic linking. Here, the insights are conveying and it becomes clear that the field holds promise not only for efficient information access but also for the revival of ancient Indian knowledge through digital intelligence. Through innovations in OCR, image restoration, NLP, and dataset standardization, the extraction of Sanskrit keywords from manuscript images has become a cornerstone of computational heritage studies.

REFERENCES

[1] Acri, A. (2024). Sanskrit Inscriptions in Northeastern Indian Scripts in Premodern Java and the Maritime Asian Networks of Mahāyāna Buddhist Tantra. *Acta Via Serica*, *9*(1), 91-138.

[2] Agarwal, M., Indu, S., & Jayanthi, N. (2023, April). An Approach to the Classification of Ancient Indian Scripts Using the CNN Model. In *International Conference on Women Researchers in Electronics and Computing* (pp. 367-377). Singapore: Springer Nature Singapore.

[3] Bagchi, P., Jain, V., & Kharat, A. NLP-Based Knowledge Extraction from Charak Samhita for Navigating Ancient Wisdom: A Django Framework Approach.

[4] Bipin Nair, B. J., Shobha Rani, N., & Khan, M. (2023). Deteriorated image classification model for malayalam palm leaf manuscripts. *Journal of Intelligent & Fuzzy Systems*, *45*(3), 4031-4049.

[5] Borthakur, P. (2021). A study of sanchipat manuscripts found in Assam: Techniques adopted for preventive conservation of manuscripts by different institutes of this region. *Library Philosophy and Practice*, 1-8.

[6] Chand, A., Agarwal, P., & Sharma, S. (2023, January). Real-Time Retrieving Vedic Sanskrit Text into Multi-Lingual Text and Audio for Cultural Tourism Motivation. In *2023 International Conference for Advancement in Technology (ICONAT)* (pp. 1-6). IEEE.

[7] Deepthi, C. V. S., & Seenu, A. (2022, December). A Systematic Review on OCRs for Indic Documents & Manuscripts. In *2022 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)* (Vol. 1, pp. 1-4). IEEE.

[8] Dhruva, G., Kore, V., Vijitha, M., Rao, S., & Preethi, P. (2023, December). Comprehensive dataset building of isolated handwritten Sanskrit characters. In *International Conference on Applied Soft Computing and Communication Networks* (pp. 489-503). Singapore: Springer Nature Singapore.

[9] Dinesh, P. M., Sruthi, A. L., Praveen, S., & Manjunathan, A. (2023, March). Word prediction using CNN for ancient manuscripts. In *AIP Conference Proceedings* (Vol. 2690, No. 1, p. 020048). AIP Publishing LLC.

[10] Dubey, N. (2018). Digital image restoration of historical devanagari manuscripts. In *Computational Intelligence: Theories, Applications and Future Directions-Volume II: ICCI-2017* (pp. 571-583). Singapore: Springer Singapore.

[11] Geethanjali, R., & Valarmathi, A. (2025). Innovative deep learning-based CEA-MMSA framework for cultural emotion analysis of Tamil and Sanskrit Siddha palm leaf manuscripts. *Journal of Computational Social Science*, *8*(3), 65.

[12] Giridharan, R., Vellingiriraj, E. K., & Balasubramanie, P. (2016, April). Identification of Tamil ancient characters and information retrieval from temple epigraphy using image zoning. In *2016 International conference on recent trends in information technology (ICRTIT)* (pp. 1-7). IEEE.

[13] Gudadhe, S. R., Bardekar, A. A., & Ranit, A. B. (2024, October). A novel approach using machine learning and NLP for revolutionizing Pali manuscript conservation. In *2024 4th International Conference on Sustainable Expert Systems (ICSES)* (pp. 844-848). IEEE.

[14] Guruprasad, P., & KS Rao, G. (2021). Recognition of Handwritten Nandinagari Palm Leaf Manuscript Text. In *Computational Intelligence Methods for Super-Resolution in Image Processing Applications* (pp. 177-190). Cham: Springer International Publishing.

[15] Jain, R., Dwivedi, S., & Gopalan, K. (2024, December). Converting Gujarati Text in Custom-Embedded Subsetted Non-unicode Fonts to Searchable Formats: A Case Study Using Jain Religious Texts. In *International Conference on Computer Vision and Image Processing* (pp. 366-380). Cham: Springer Nature Switzerland.

[16] Jain, V., Bagchi, P., Kharat, A., & Shivani, V. (2025). Extracting Invaluable Insights from Sushruta Samhita Using Natural Language Processing. *International Journal of Public Mental Health and Neurosciences*, *12*(2), 10-14.

[17] Jayashree, K. R., Manisha, K., Sudarshan, K., & Kannadaguli, P. (2024, May). Tulu Manuscript OCR: Preserving Ancient Wisdom through Character Recognition. In *2024 Second International Conference on Data Science and Information System (ICDSIS)* (pp. 1-7). IEEE.

[18] Jindal, A., & Ghosh, R. (2024). A hybrid deep learning model to recognize handwritten characters in ancient documents in Devanagari and Maithili scripts. *Multimedia Tools and Applications*, *83*(3), 8389-8412.

[19] Jindal, A., & Ghosh, R. (2024). A semi-self-supervised learning model to recognize handwritten characters in ancient documents in Indian scripts. *Neural Computing and Applications*, *36*(20), 11791-11808.

[20] Kataria, B., & Jethva, H. B. (2019). CNN-bidirectional LSTM based optical character recognition of Sanskrit manuscripts: a comprehensive systematic literature review. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.(IJSRCSEIT)*, *5*(2), 2456-3307.

[21] Kataria, B., & Jethva, H. B. (2021). Optical character recognition of indian language manuscripts using convolutional neural networks. *Design Engineering*, *3*.

[22] Kataria, D. B., & Jethva, H. B. (2021). Optical Character Recognition of Sanskrit Manuscripts Using Convolution Neural Networks. *Webology (ISSN: 1735-188X) Volume*, *18*.

[23] Kesiman, M. W. A. (2019, August). Word recognition for the balinese palm leaf manuscripts. In *2019 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)* (pp. 72-76). IEEE.

[24] Kesiman, M. W. A., & Pradnyana, G. A. (2020, January). Image Patch Extraction in Text Area of Balinese Palm Leaf Manuscripts with Gabor Filters. In *3rd International Conference on Innovative Research Across Disciplines (ICIRAD 2019)* (pp. 19-23). Atlantis Press.

[25] Kesiman, M. W. A., Valy, D., Burie, J. C., Paulus, E., Suryani, M., Hadi, S., ... & Ogier, J. M. (2018). Benchmarking of document image analysis tasks for palm leaf manuscripts from southeast asia. *Journal of Imaging*, *4*(2), 43.

[26] Khaparde, A., Deshmukh, V., & Kowdiki, M. (2023). Enhanced Nature-Inspired Algorithm-based Hybrid Deep Learning for Character Recognition in Sanskrit Language. *Sensing and Imaging*, *24*(1), 23.

[27] Kore, V., Dhruva, G., Rao, S., Vijitha, M., & Preethi, P. (2025). A systematic framework for Sanskrit character recognition using deep learning. *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, *24*(1), 81-103.

[28] Kowshik, A. S. S., Sindhur, A. S., Reddy, A. G., Ganesh, M., Sivan, R., & Pati, P. B. (2024, December). Assessing Degradation Levels of Palm Leaf Manuscripts with Random Forest Using Gabor. In *Computer, Communication, and Signal Processing. Smart Solutions Towards SDG: 8th IFIP TC 12 International Conference, ICCCSP 2024, Chennai, India, March 20-22, 2024, Revised Selected Papers* (Vol. 723, p. 239). Springer Nature.

[29] Krishnan, S., Kulkarni, A., & Huet, G. (2025). Normalized dataset for Sanskrit word segmentation and morphological parsing. *Language Resources and Evaluation*, *59*(2), 1279-1330.

[30] Kulkarni, I., Tikkal, S., Chaware, S., Kharate, P., & Pandit, A. (2022, February). Proposed Design to Recognize Ancient Sanskrit Manuscripts with Translation Using Machine Learning. In *Proceedings of the International Conference on Innovative Computing & Communication (ICICC)*.

[31] Kumar, R., Tewari, P., Thakur, R. K., & Kumar, R. (2024). ENGLISH TO SANSKRIT TRANSLATION USING NMT. *Available at SSRN 4938136*.

[32] Kundu, S., Paul, S., Singh, P. K., Sarkar, R., & Nasipuri, M. (2020). Understanding NFC-Net: a deep learning approach to word-level handwritten Indic script recognition. *Neural Computing and Applications*, *32*(12), 7879-7895.

[33] Laskar, I., & Ansari, S. (2021). Illustrated manuscripts at auniati satra of majuli island, assam. *Heritage: Journal of Multidisciplinary Studies in Archaeology*.

[34] Lomte, M. V. M., & Doye, D. D. (2022). Handwritten Vedic Sanskrit text recognition using deep learning. *Journal of Algebraic Statistics*, *13*(3), 2190-2198.

[35] Maheshwari, N., Maloo, A., & Parihar, P. S. A Novel Approach of Data Extraction from Indian Degraded Historical Documents using Gamma Variation and Histogram Balancing Method.

[36] Maheswari, S. U., Maheswari, P. U., & Aakaash, G. S. (2024). An intelligent character segmentation system coupled with deep learning based recognition for the digitization of ancient Tamil palm leaf manuscripts. *Heritage Science*, *12*(1), 342.

[37] Mali, S. (2021). Identification and Segmentation of Touching Brahmi Characters from Degraded Digital Estampage Images Using Ensemble Classifier. *Indian Journal of Computer Science and Engineering*.

[38] Manigandan, T. V. V. D. V. N. B., Vidhya, V., Dhanalakshmi, V., & Nirmala, B. (2017, August). Tamil character recognition from ancient epigraphical inscription using OCR and NLP. In *2017 international conference on energy, communication, data analytics and soft computing (ICECDS)* (pp. 1008-1011). IEEE.

[39] Mehta, N., & Doshi, J. (2020, August). Text line segmentation for medieval Devnagari manuscript. In *Proceedings of International Conference on Communication and Computational Technologies: ICCCT-2019* (pp. 405-412). Singapore: Springer Singapore.

[40] Mohammed, H., Märgner, V., & Ciotti, G. (2021). Learning-free pattern detection for manuscript research: An efficient approach toward making manuscript images searchable. *International Journal on Document Analysis and Recognition (IJDAR)*, *24*(3), 167-179.

[41] Moudgil, A., Singh, S., & Gautam, V. (2021). An overview of recent trends in OCR systems for manuscripts. *Cyber Intelligence and Information Retrieval: Proceedings of CIIR 2021*, 525-533.

[42] Moudgil, A., Singh, S., Gautam, V., Rani, S., & Shah, S. H. (2023). Handwritten devanagari manuscript characters recognition using capsnet. *International Journal of Cognitive Computing in Engineering*, *4*, 47-54.

[43] Mubarakkaa, M. F., Nandhini, M., Keerthika, M., & Ganapathy, H. (2024, October). OCR based transliteration of Brahmi to Tamil using CNN. In *2024 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS)* (pp. 1-5). IEEE.

[44] Nair, B. B., & Rani, N. S. (2023). HMPLMD: Handwritten Malayalam palm leaf manuscript dataset. *Data in Brief*, *47*, 108960.

[45] Nair, B. B., Raj, K. A., Kedar, M., Vaishak, S. P., & Sreejil, E. V. (2023). Ancient Epic manuscript binarization and classification using false color spectralization and VGG-16 model. *Procedia Computer Science*, *218*, 631-643.

[46] Narang, S. R., Jindal, M. K., Ahuja, S., & Kumar, M. (2020). On the recognition of Devanagari ancient handwritten characters using SIFT and Gabor features. *Soft Computing*, *24*(22), 17279-17289.

[47] Narang, S. R., Kumar, M., & Jindal, M. K. (2021). DeepNetDevanagari: a deep learning model for Devanagari ancient character recognition. *Multimedia Tools and Applications*, *80*(13), 20671-20686.

[48] Narang, S. R., Kumar, M., & Jindal, M. K. (2022, December). Optimization of Character Classes in Devanagari Ancient Manuscripts and Dataset Generation. In *International Conference on Frontiers in Computing and Systems* (pp. 59-69). Singapore: Springer Nature Singapore.

[49] Narang, S., Jindal, M. K., & Kumar, M. (2019). Devanagari ancient documents recognition using statistical feature extraction techniques. *Sādhanā*, *44*(6), 141.

[50] Poddar, S., & Gupta, R. (2023). Optical Script Identification for multi-lingual Indic-script. *arXiv preprint arXiv:2308.05780*.

[51] Pradeep, A., & Mamidi, R. (2025). Sandarśana: A Survey on Sanskrit Computational Linguistics and Digital Infrastructure for Sanskrit. *ACM Computing Surveys*, *57*(10), 1-38.

[52] Pradeep, N., Subramanian, D., & Ganapathy, M. K. (2024). Digitizing India's Ancient Texts: AI for Tamil Palm Leaf Manuscript Preservation and Accessibility.

[53] Preisendanz, K. (2018). Text segmentation, chapter naming and the transmission of embedded texts in South Asia, with special reference to the medical and philosophical traditions as exemplified by the Carakasaṃhitā and the Nyāyasūtra. In *Pieces and parts in scientific texts* (pp. 159-220). Cham: Springer International Publishing.

[54] Puri, S., & Singh, S. P. (2019). An efficient Devanagari character classification in printed and handwritten documents using SVM. *Procedia Computer Science*, *152*, 111-121.

[55] Raj, V. A., Jyothi, R. L., & Anilkumar, A. (2017, July). Grantha script recognition from ancient palm leaves using histogram of orientation shape context. In *2017 International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 790-794). IEEE.

[56] Rani, N. S. (2024). A modified deep semantic binarization network for degradation removal in palm leaf manuscripts. *Multimedia Tools and Applications*, *83*(23), 62937-62969.

[57] Ray, P. P. (2025). NādīML: Aligning Ancient Nādī Astrology with Machine Learning Techniques. *Authorea Preprints*.

[58] Samantaray, S., Mohapatra, S. K., & Mohapatra, S. (2025, April). A Systematic Literature Review of Recognizing Handwritten Vedic Text on Palm Leaves Using Machine Learning. In *International Conference on Green Artificial Intelligence and Industrial Applications* (pp. 146-160). Cham: Springer Nature Switzerland.

[59] Shankar, B., Mishra, P., Sagnika, S., & Pattanaik, A. Engaging with an Indian Epic: A Digital Approach. *International Journal of Computer Applications*, *975*, 8887.

[60] Shelke, S. V., Chandwadkar, D. M., Ugale, S. P., & Chothe, R. V. (2025). Discrete wavelet transform and convolutional neural network based handwritten Sanskrit character recognition. *Indonesian Journal of Electrical Engineering and Computer Science*, *38*(2), 1367-1375.

[61] Shobha Rani, N., Bipin Nair, B. J., Chandrajith, M., Hemantha Kumar, G., & Fortuny, J. (2022). Restoration of deteriorated text sections in ancient document images using atri-level semi-adaptive thresholding technique. *Automatika: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije*, *63*(2), 378-398.

[62] Singh, B., & Ahuja, N. J. (2019). Mining the treasure of palm leaf manuscripts through information retrieval techniques. *Digital Library Perspectives*, *35*(3-4), 146-156.

[63] Singh, P. K., Sarkar, R., Das, N., Basu, S., Kundu, M., & Nasipuri, M. (2018). Benchmark databases of handwritten Bangla-Roman and Devanagari-Roman mixed-script document images. *Multimedia Tools and Applications*, *77*(7), 8441-8473.

[64] Singh, S., Garg, N. K., & Kumar, M. (2023). Feature extraction and classification techniques for handwritten Devanagari text recognition: a survey. *Multimedia Tools and Applications*, *82*(1), 747-775.

[65] Sinha, S., & Jha, G. N. (2020, May). Abstractive text summarization for Sanskrit prose: a study of methods and approaches. In *Proceedings of the WILDRE5–5th Workshop on Indian Language Data: Resources and Evaluation* (pp. 60-65).

[66] Sitender, & Bawa, S. (2022). Sanskrit to universal networking language EnConverter system based on deep learning and context-free grammar. *Multimedia Systems*, *28*(6), 2105-2121.

[67] Subramani, K., & Murugavalli, S. (2019). Recognizing ancient characters from tamil palm leaf manuscripts using convolution based deep learning. *International Journal of Recent Technology and Engineering*, *8*(3), 6873-6880.

[68] Sudarsan, D., & Sankar, D. (2022). A novel complete denoising solution for old Malayalam palm leaf manuscripts. *Pattern Recognition and Image Analysis*, *32*(1), 187-204.

[69] Sudarsan, D., & Sankar, D. (2024). An ensemble neural network model for Malayalam character recognition from palm leaf manuscripts. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

[70] Tapaswi, N. (2025). Shabda sculptor: carving morphological excellence in Sanskrit spellcheck. *International Journal of Information Technology*, *17*(1), 591-597.

[71] Thottempudi, S. G. (2021). A visual narrative of ramayana using extractive summarization topic modeling and named entity recognition. In *CEUR Workshop Proc.* (Vol. 2823, pp. 3-10).

[72] Tomar, A., Choudhary, M., & Yerpude, A. (2015). Ancient Indian scripts image pre-processing and dimensionality reduction for feature extraction and classification: a survey. *International Journal of Computer Trends and Technology (IJCTT)*, *21*(2), 101-124.

[73] Unnikrishnan, D., Sudarsan, D., & Vignesh, R. (2025). A novel method of absolute noise removal from the degraded palm leaf manuscripts. *Indian Journal of Traditional Knowledge (IJTK)*, *24*(6), 595-605.

[74] Valaboju, B., Dwivedi, S., Chincholikar, K., Gopalan, K., & Vidwans, V. (2025, May). A Semi-Automatic Text Recognition Tool for Pre-Colonial Handwritten Manuscripts in Devanāgari Script. In *International Conference on Human-Computer Interaction* (pp. 152-160). Cham: Springer Nature Switzerland.

[75] Valy, D., Verleysen, M., Chhun, S., & Burie, J. C. (2017, November). A new khmer palm leaf manuscript dataset for document analysis and recognition: Sleukrith set. In *Proceedings of the 4th International Workshop on Historical Document Imaging and Processing* (pp. 1-6).

[76] Varghese, B., & Govilkar, S. (2015). A survey on various word spotting techniques for content based document image retrieval. *International Journal of Computer Science and Information Technologies*, *6*, 2682-2686.

[77] Varghese, B., & Govilkar, S. A. Novel Approach For Word Retrieval From Devanagari Document Images.

[78] Vijayalakshmi, R., & Gnanasekar, J. M. (2022, April). A review on character recognition and information retrieval from ancient inscriptions. In *2022 8th International Conference on Smart Structures and Systems (ICSSS)* (pp. 1-7). IEEE.

[79] Vijitha, M., Vrinda, K., Dhruva, G., Sahana, R., & Preethi, P. (2024, March). Segmentation of handwritten Sanskrit words using Image-Processing techniques. In *International Conference on Innovations in Cybersecurity and Data Science Proceedings of ICICDS* (pp. 13-27). Singapore: Springer Nature Singapore.

[80] Wang, Y., Wen, M., Zhou, X., Gao, F., Tian, S., Jue, D., ... & Zhang, Z. (2024). Automatic damage identification of Sanskrit palm leaf manuscripts with SegFormer. *Heritage Science*, *12*(1).

[81] Sukula, Shiva Kanaujia (2025). Librarians' Endeavors in Manuscript Preservation and Future Ahead: The Context of Indian Experiences and Artificial Intelligence. *SLA-ASIA NEWSLETTER*, 21(1), pp. 35-36.
