International Journal of Research in Engineering & Science
Available online on http://rspublication.com/IJRES/IJRE.html
DOI: 10.5281/zenodo.16277831

ISSN:(P) 2572-4274 (O) 2572-4304
volume 9 Issue 4 2025
*Original Article*

# “A Comprehensive Survey of OCR for Devanagari Script Based Languages”

**Arvind Kaur**[1] ✉ **Gurpreet Singh Lehal**[2]
[1] Department of Computer Science, Punjabi University, Patiala, Punjab, India,

✉ arvindkaur.nibber@gmail.com ·
[2] Senior Project Consultant, IIIT Hyderabad, India; Formerly Professor, Department of Computer Science, Punjabi University, Patiala, Punjab, India
✉gslehal@gmail.com

## Abstract

The Devanagari script, one of the most widely used writing systems in India, serves as the foundation for several languages, including Hindi, Marathi, Sanskrit, and Nepali. Optical Character Recognition (OCR) for Devanagari presents significant challenges due to the script's intricate structure, which features conjunct characters, vowel modifiers, and diacritical marks. This survey paper offers a comprehensive review of techniques developed for OCR systems across all Devanagari-based languages, encompassing traditional rule-based methods and modern approaches utilizing machine learning and deep learning. Uniquely, this survey goes beyond previous studies that primarily focused on Hindi, providing a broader perspective on the progress of OCR technology for diverse languages using Devanagari. The insights presented here aim to guide researchers and practitioners in advancing efficient and accurate OCR solutions, ensuring the digital inclusion of these languages in the evolving AI landscape.

**Keywords**: Devanagari, Convolutional Neural Network, Recurrent Neural Network, Transformers, Bidirectional Long Short-Term Memory (BLSTM), Optical Character Recognition

## 1 Introduction

Optical Character Recognition (OCR) is a transformative technology that converts scanned images of typed or printed documents into machine-readable text formats, such as ASCII or Unicode. This converted text can be utilized in a variety of applications, including digital storage, text search, editing, and integration with advanced AI systems. OCR also plays a pivotal role in enabling technologies like text-to-speech systems, where digital text is transformed into spoken words, and machine translation, facilitating automatic translation of text from images across different languages. By converting text images (e.g., .bmp, .jpg) into editable formats (e.g., .txt, .doc), OCR serves as an essential tool for advanced tasks in language processing, data analysis, and enhancing accessibility, thereby bridging the gap between physical and digital text.

Although OCR technologies have been successfully implemented for languages such as English, Korean, Chinese, German, and Japanese, significant challenges remain for certain

**Cite This Paper :** Arvind Kaur ✉ Gurpreet Singh Lehal (2025). "A Comprehensive Survey of OCR for Devanagari Script based languages". *INTERNATIONAL JOURNAL OF RESEARCH IN ENGINEERING & SCIENCE (IJRES) {ISSN- (PRINT) 2572-4274 (ONLINE) 2572-4304}*, vol. 9, no.4, 2025, pp. 34-47. DOI: https://dx.doi.org/10.5281/zenodo.16277831

languages [1]. The recognition of texts in languages using scripts like Devanagari is particularly difficult due to factors such as the large character set, language-specific complexities, and the unique structural feature of the Shirorekha (horizontal line), all of which complicate accurate document analysis and recognition. The field of Indic OCR has evolved from traditional methodologies to modern approaches that leverage machine learning and deep learning algorithms. These advancements have significantly enhanced the robustness and efficiency of OCR systems, allowing for the effective processing of Devanagari documents [2].

This research paper is structured into several sections, each addressing critical aspects of Devanagari OCR. Section 2 provides an in-depth analysis of the Devanagari script, highlighting its distinctive features and the challenges they present for OCR systems. Section 3 examines various OCR techniques applied to the Devanagari script, showcasing both traditional and modern approaches. Section 4 outlines the stages of OCR implementation. Section 5 offers a comprehensive literature review of prior research in Devanagari OCR, evaluating methods, datasets, and performance metrics. Section 6 presents a concise overview of publicly available OCR systems designed for Devanagari-based languages. Finally, Section 7 summarizes the findings of the survey, offering insights and guidance for future research and development in this domain.

## 2 Introduction to Devanagari Script

The Devanagari script is one of the most widely used writing systems in South Asia, with over 120 languages using it across India, Nepal, and surrounding regions. Major languages written in Devanagari include Hindi, Sanskrit, Marathi, Nepali, Bhojpuri, Konkani, Maithili, Sindhi, Dogri, Awadhi, Magahi, and several other regional languages. As an alphasyllabary, Devanagari combines both alphabetic and syllabic elements, featuring distinct characters for vowels and consonants. A defining characteristic of the script is the horizontal line that runs along the top of the letters, linking them together. Devanagari is not only used for everyday communication but also plays a vital role in sacred texts, literature, and official documents. The script is integral to the preservation and dissemination of rich cultural, philosophical, and religious traditions, particularly within Hinduism and Buddhism. The Devanagari Character Set including Consonants, vowels and Modifiers are presented in Table 1 [3, 4]. Few examples of languages based on Devanagari script are presented in Table 2.

### 2.1 Complexity of Devanagari Script

The Devanagari script is intricate due to its phonetic structure and distinctive elements. Here are several key aspects of the complexity of the Devanagari script [5, 6]:

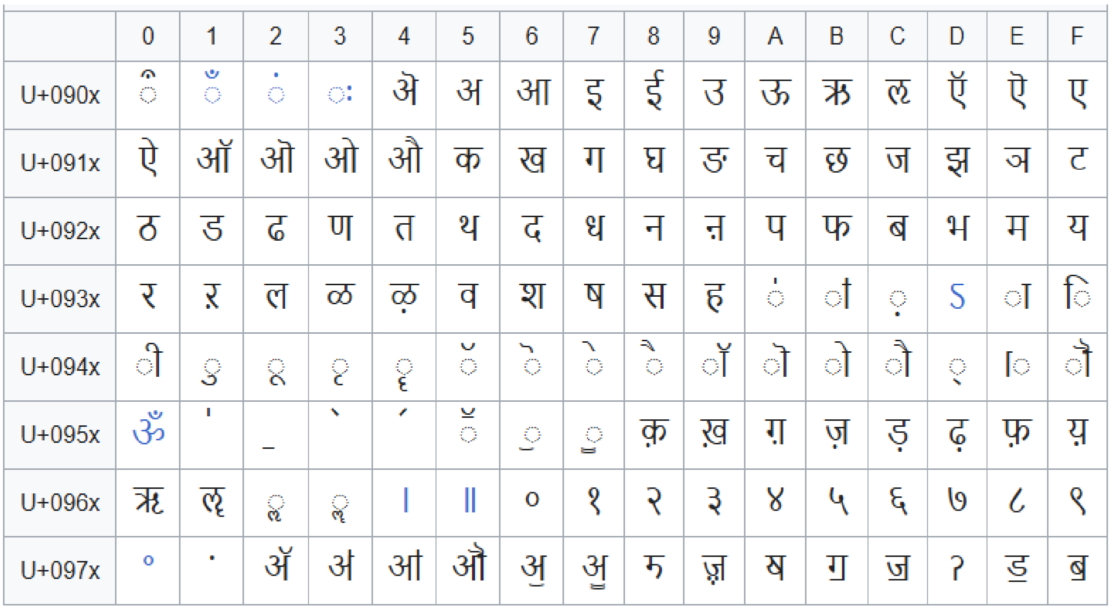
a. *Shirorekha*: In Devanagari script, multiple characters and modifiers are joined together as a single unit, connected by the common Shirorekha.

b. *Compound Characters:* Compound characters are formed by joining two or more distinct elements, such as consonants, vowels, or both, to create a new symbol. These combinations can include consonants and vowels (such as का (kā) or ◌◌े (e) attached to a consonant) or entirely new forms, often through ligatures.

c. *Conjunct Characters:* Conjunct characters are created by merging two or more consonants into a single glyph to represent a sound cluster. These characters are formed when consonants combine, typically without any vowel sounds in between. For example, in Devanagari, the combination of क (ka) and ष (ṣa) results in the conjunct character क्ष (kṣa).

d. *Use of Matras (Vowel Modifiers):* Matras, or vowel signs, modify consonants to change their pronunciation. These modifiers can appear in various positions before, after, above, or below the consonant adding complexity to character recognition in OCR systems.

**Table 1:** Devanagari Character Dataset



| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | A | B | C | D | E | F |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| U+090x | ◌ऀ | ◌ँ | ◌ं | ◌ः | ऄ | अ | आ | इ | ई | उ | ऊ | ऋ | ऌ | ऍ | ऎ | ए |
| U+091x | ऐ | ऑ | ऒ | ओ | औ | क | ख | ग | घ | ङ | च | छ | ज | झ | ञ | ट |
| U+092x | ठ | ड | ढ | ण | त | थ | द | ध | न | ऩ | प | फ | ब | भ | म | य |
| U+093x | र | ऱ | ल | ळ | ऴ | व | श | ष | स | ह | ◌ऺ | ◌ऻ | ◌़ | ऽ | ◌ा | ◌ि |
| U+094x | ◌ी | ◌ु | ◌ू | ◌ृ | ◌ॄ | ◌ॅ | ◌ॆ | ◌े | ◌ै | ◌ॉ | ◌ॊ | ◌ो | ◌ौ | ◌् | ◌ॎ | ◌ॏ |
| U+095x | ॐ | ◌॑ | ◌॒ | ◌॓ | ◌॔ | ◌ॕ | ◌ॖ | ◌ॗ | क़ | ख़ | ग़ | ज़ | ड़ | ढ़ | फ़ | य़ |
| U+096x | ॠ | ॡ | ◌ॢ | ◌ॣ | । | ॥ | ० | १ | २ | ३ | ४ | ५ | ६ | ७ | ८ | ९ |
| U+097x | ॰ | ॱ | ॲ | ॳ | ॴ | ॵ | ॶ | ॷ | ॸ | ॹ | ॺ | ॻ | ॼ | ॽ | ॾ | ॿ |

**Table 2:** Example sentences in different languages using Devanagari script

| Language | Example Sentence |
|---|---|
| Hindi | भारत एक महान देश है |
| Maithili | भारत एक महान देश अछि |
| Dogri | भारत इक महान देश ऐ |
| Konkani | भारत एक म्हान देश आसा |
| Sanskrit | भारतम् एकं महन् देश् अस्ति |
| Nepali | भारत एक महान देश हो |
| Sindhi | भारत एकु महानु देश आहे |
| Bodo | भारतआ मोनसे महान देश |
| Marathi | भारत एक महान देश आ |

## 3 Overview of Devanagari OCR Techniques

### 3.1 Traditional Methods for Devanagari Script

Traditional Optical Character Recognition (OCR) techniques have served as the foundation for automated text recognition systems for Indic scripts. These methods typically involve a series of sequential stages, including preprocessing to enhance image quality, feature extraction to identify distinctive textual patterns, and classification to assign text to predefined categories. Despite their foundational importance, traditional methods often face limitations in handling complex scripts like Devanagari, which features conjunct characters, vowel modifiers, and diacritics.

### 3.2 Modern OCR Techniques and Their Application in Devanagari Script

Advancements in machine learning and deep learning have significantly enhanced the accuracy and adaptability of OCR systems for Devanagari script. Modern OCR approaches leverage sophisticated models, including:

a. **Convolutional Neural Networks (CNNs) -** CNN-based architectures have proven highly successful in OCR tasks, including those involving Indic scripts. These models automatically extract relevant features from raw pixel representations of characters through hierarchical learning. This capability results in accurate and robust recognition, making CNNs particularly effective for feature extraction from images. However, their high computational cost remains a consideration [7].

b. **Recurrent Neural Networks (RNNs) -** RNNs, especially Long Short-Term Memory (LSTM) networks, are well-suited for sequential data, making them ideal for recognizing Devanagari script's structural dependencies and contextual variations. By capturing temporal and contextual relationships, RNNs enhance precision in OCR tasks. Nevertheless, challenges such as vanishing gradients can arise when processing very long sequences [8, 9].

c. **Transformer-Based Models -** Transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers), have revolutionized Natural Language Processing (NLP) tasks. Their ability to manage long-range dependencies and capture contextual information across characters makes them exceptionally suited for Indic script recognition, including Devanagari [10].

These modern approaches have redefined the capabilities of OCR systems, enabling efficient processing of complex scripts like Devanagari while addressing the script's unique challenges.

## 4 Phases while implementing OCR

Developing an Optical Character Recognition (OCR) system entails several essential stages:

a. **Dataset Creation -** The initial step involves compiling an extensive dataset of printed text in the Devanagari script. This dataset should encompass diverse samples, including various fonts, styles, and text variations.

b. **Data Preprocessing** - To prepare the dataset for training, it undergoes preprocessing to improve quality.

c. **Feature Extraction and Classification** - Features are extracted from the text image and characters and words are classified.

d. **Model Training -** A model is trained using the preprocessed data.

e. **Testing and Validation -** The model is evaluated with a separate dataset reserved for testing and validation, which the model has not encountered before. This phase measures performance metrics, such as Character Error Rate (CER) and Word Error Rate (WER), by comparing predicted text against ground truth labels.

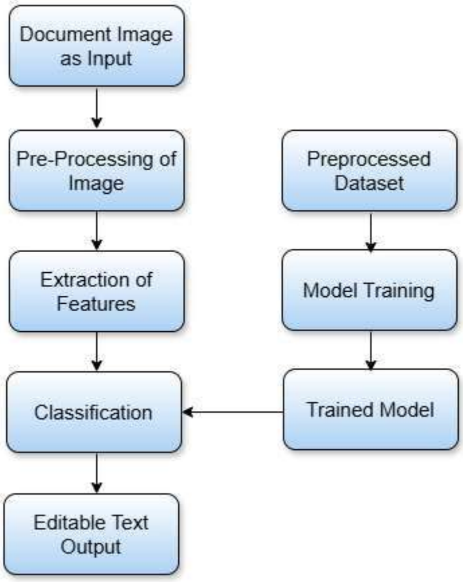
f. **Model Evaluation and Refinement -** Based on testing results, the model's performance is assessed. If it meets accuracy requirements, it can proceed to deployment. Otherwise, adjustments like parameter tuning, data augmentation, or alternative architectures may be applied to enhance performance.

The various stages of the Optical Character Recognition are presented in Figure 2. The model is trained using a preprocessed dataset. Once the document image is provided, it undergoes preprocessing, segmentation, feature extraction, and classification using the trained model. Finally, editable text is extracted from the document image.

## 5 Evolution of Devanagari OCR Techniques: From Traditional Methods to Advanced Deep Learning Models

This section provides a comprehensive overview of the evolution of techniques and methods employed by researchers in the field of Devanagari Optical Character Recognition (OCR). It highlights the progression from traditional rule-based systems to modern machine learning and deep learning approaches, emphasizing their application across various languages that use the Devanagari script. While the Devanagari script encompasses a vast array of consonants, vowels, and modifiers, the specific character sets vary by language.

**Figure 2:** Various Stages of OCR



For instance, the Sindhi language utilizes unique characters such as ॾ, ॿ, ॻ, and ॼ; Sanskrit includes distinctive symbols like ◌ॢ, ◌ॣ, ॠ, ऌ, and ॡ; and Kashmiri incorporates characters such as ◌ॖ, ◌ॗ, ॶ, ॷ, and ॵ [11]. Given these linguistic variations, researchers have tailored their OCR efforts to address the needs of specific Devanagari-based languages, including Hindi, Marathi, Nepali, Sanskrit, and Konkani.

To provide a structured and in-depth analysis, this section is organized into subsections, each focusing on a specific language that uses the Devanagari script. The discussion begins with Hindi, the most widely spoken and researched language among those using Devanagari, and expands to cover other significant languages, illustrating the unique challenges and advancements in OCR technologies for each.

## 5.1 Hindi

Hindi, the official language of India, occupies a prominent position both nationally and globally due to its extensive use and cultural significance. It is the most spoken language in India and ranks as the fourth most spoken language worldwide. According to the 2011 Census of India, Hindi is spoken by 528 million people, constituting 43.63% of the country's total population.

Hindi's vast reach and its rich literary, cultural, and historical heritage emphasize its importance in India's sociolinguistic fabric. This prominence also underscores the need for ongoing efforts to preserve and promote the language through technological advancements. One such critical innovation is the development of Optical Character Recognition (OCR) systems tailored specifically for Hindi. These systems play a vital role in digitizing and processing Hindi text, ensuring its seamless integration into modern digital platforms.

The character set of Hindi, as relevant to OCR applications, is detailed in Table 3. This set forms the foundation for developing robust OCR systems capable of accurately recognizing and processing Hindi text across diverse use cases.

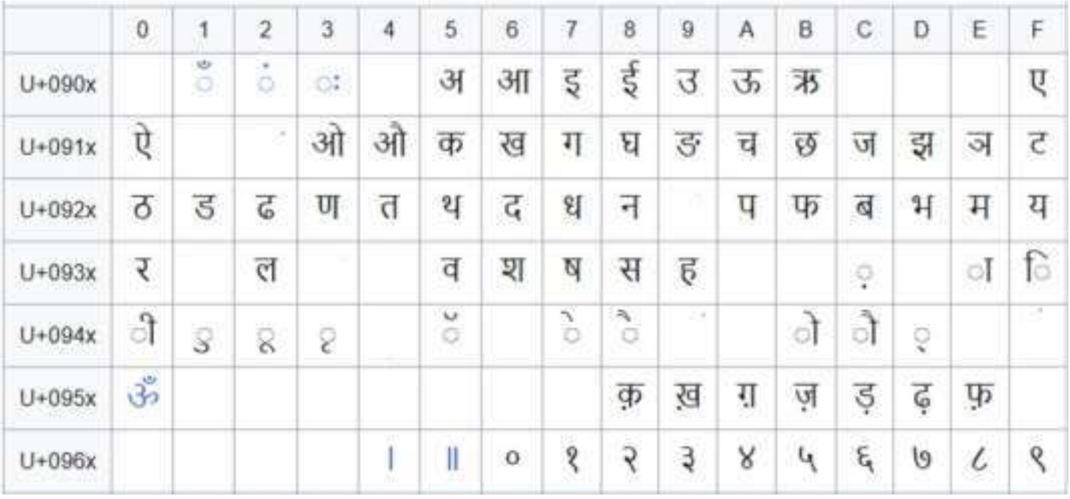
Hindi Optical Character Recognition (OCR), often incorrectly referred to as Devanagari OCR, has been a major focus of research due to the prominence of Hindi as mentioned above. This interchangeable usage arises from the fact that Hindi is written in Devanagari, leading to the misconception that research in Hindi OCR inherently addresses the entirety of Devanagari OCR. However, this notion is flawed, as the Devanagari script is used by several Indian languages, including Marathi, Sanskrit, Nepali, and others as discussed earlier, and contains over 35 additional characters not utilized in standard Hindi (Table 1 and Table 3). These unique characters are essential for other languages, underscoring the need for distinct OCR solutions tailored to each language. Despite this linguistic diversity, OCR research has predominantly focused on Hindi, with significant advancements in addressing its specific challenges, though only recently has attention expanded to other Devanagari-based languages.

Early efforts in Hindi OCR, initiated around the mid-1970s, were primarily theoretical and did not result in practical systems. Sinha [12] developed a syntactic pattern analysis system for recognizing Hindi text, while Sinha and Mahabala [13] expanded on this by embedding a picture language into their system for recognizing handwritten and machine-printed Devanagari characters. Sinha [14] further highlighted the importance of spatial relationships among symbols in interpreting Hindi words written in Devanagari.

The development of practical Hindi OCR systems gained momentum in the mid-1990s. Palit and Chaudhuri [15] as well as Pal and Chaudhuri [16], introduced techniques such as headline deletion, text line zoning, and hybrid recognition methods, achieving approximately 96% accuracy in their systems. This marked a pivotal milestone in Hindi OCR research. Bansal and Sinha [17] addressed error correction by proposing a partitioned word dictionary and penalty-based mismatch matrix, significantly enhancing recognition accuracy. Jawahar et al. [18] developed a bilingual OCR system for Hindi and Telugu using Principal Component Analysis and support vector classification, achieving 96.7% accuracy on a dataset of approximately 2 lakh characters.

Segmentation-based approaches dominated early Hindi OCR research. Dhurandhar et al. [19] addressed the cursive nature of Devanagari by using contour extraction and noise removal methods, coupled with a prioritization scheme for handling similar contours. Kompalli et al. [20] employed neural networks for recognizing machine-printed, multi-font Hindi text by segmenting characters into ascenders, core components, and descenders. They later introduced a graph-based recognition-driven segmentation methodology Kompalli et al. [21], incorporating stochastic finite state automata (SFSA) for word recognition and improved language models for post-processing.

**Table 3:** Hindi Character Dataset



| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | A | B | C | D | E | F |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| U+090x | | ◌ँ | ◌ं | ◌ः | | अ | आ | इ | ई | उ | ऊ | ऋ | | | | ए |
| U+091x | ऐ | | | ओ | औ | क | ख | ग | घ | ङ | च | छ | ज | झ | ञ | ट |
| U+092x | ठ | ड | ढ | ण | त | थ | द | ध | न | | प | फ | ब | भ | म | य |
| U+093x | र | | ल | | | व | श | ष | स | ह | | | ◌़ | | ◌ा | ◌ि |
| U+094x | ◌ी | ◌ु | ◌ू | ◌ृ | | ◌ॅ | | ◌े | ◌ै | | | ◌ो | ◌ौ | ◌् | | |
| U+095x | ॐ | | | | | | | | क़ | ख़ | ग़ | ज़ | ड़ | ढ़ | फ़ | |
| U+096x | | | | | । | ॥ | ० | १ | २ | ३ | ४ | ५ | ६ | ७ | ८ | ९ |

Holambe et al. [22] compared SVM and KNN classifiers for printed and handwritten Devanagari script, while Yadav et al. [23] addressed segmentation errors due to touching characters in scanned Hindi documents using a back-propagation neural network, achieving a 90% recognition rate. Choksi et al. [24] proposed a fuzzy KNN classifier with geometric and wavelet feature extraction techniques to tackle touching character issues, enhancing segmentation and recognition accuracy. Puri and Singh [25] developed a classification model using SVM for mono-lingual Hindi, Sanskrit, and Marathi documents, reporting average accuracies of 99.54% for printed and 98.35% for handwritten text. Habib et al. [26] explored neural networks for recognizing Urdu and Devanagari characters, showcasing cross-linguistic applicability.

## 5.1.1 The Transformative Role of Deep Learning

The advent of deep learning in 2012 fundamentally transformed Optical Character Recognition (OCR) by significantly improving accuracy, scalability, and adaptability across diverse languages, scripts, and applications. Unlike traditional OCR methods that relied heavily on manual feature engineering, deep learning models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers have revolutionized the field by learning features directly from raw data. These models enable end-to-end text recognition without requiring explicit segmentation, overcoming one of the most common challenges in OCR for complex scripts like Devanagari.

Sankaran and Jawahar [27] were pioneers in applying deep learning to Devanagari OCR. They introduced the use of Bidirectional Long Short-Term Memory (BLSTM) networks for segmentation-free text recognition, addressing the high word error rates caused by complexities in character segmentation. Their approach reduced word error rates by more than 20% and character error rates by over 9% compared to the best available systems. Similarly, Karayil et al. [28] highlighted the effectiveness of LSTMs for segmentation-free OCR of Devanagari script, achieving error rates between 1.2% and 9.0%, depending on data complexity. They also demonstrated the superior performance of LSTMs compared to the Tesseract OCR system.

Mathew et al. [29] addressed the need for a multilingual OCR system in India to accommodate the prevalence of bilingualism and trilingualism. They proposed an end-to-end RNN-based architecture capable of script identification at the word level and segmentation-free text recognition. Supporting 12 Indian languages and English, their system demonstrated strong performance, with Hindi OCR results surpassing those of other popular systems when tested on a large corpus.

Chavan et al. [30] evaluated BLSTM and Multi-Dimensional LSTM (MDLSTM) architectures for OCR tasks on printed documents in seven Indian languages, including Hindi. Their study demonstrated that MDLSTM consistently outperformed BLSTM and Tesseract, showcasing its robustness in handling complex scripts. Similarly, Kundaikar and Pawar [31] proposed a multi-font Devanagari OCR (MFD_OCR) model based on LSTM networks. Their detailed error analysis revealed consistent insertion and deletion errors across different font styles, indicating room for further refinement.

Gupta et al. [32] introduced CMViT, a cutting-edge model combining ConvMixer and modified Vision Transformer attention (SVTR), for OCR tasks in 11 Indian languages, including Hindi. By leveraging advancements in AI, transitioning from RNNs and LSTMs to sophisticated encoder-decoder architectures, their model achieved Character Error Rates (CER) between 0.15% and 1.95%, highlighting its effectiveness for multilingual and complex script recognition.

These advancements underscore the transformative impact of deep learning on Hindi OCR, enabling significant improvements in accuracy, robustness, and multilingual adaptability, while laying the foundation for future innovations in Devanagari and other Indic script OCR systems.

## 5.2 Marathi

Marathi is a classical Indo-Aryan language spoken primarily by the Marathi people in Maharashtra, India, and in other states such as Goa, Karnataka, Tamil Nadu, and Gujarat. It is the official language of Maharashtra and an additional official language in Goa. With 83 million speakers as of 2011, Marathi ranks 13th in the world by the number of native speakers and has the third-largest number of native speakers in India, after Hindi and Bengali [33].

The Marathi language, like Hindi, utilizes the Devanagari script and shares a similar character set. However, one key distinction is the inclusion of the character ळ (ḷ) (Unicode: U+0933), which is not found in Hindi. This unique character results in Marathi Optical Character Recognition (OCR) systems often being based on Hindi OCR models, with adjustments made to account for the specific Marathi characters. As a result, research focused exclusively on Marathi OCR remains limited, and in many cases, Marathi OCR is integrated within multilingual OCR systems, as seen in works by Chavan et al. [30].

In the existing literature, there are only a few studies dedicated to printed OCR specifically for Marathi. Amarjot et al. [34] proposed a threshold based pre-processing methodology with enhancement support for word spotting in Marathi printed bimodal images, using image segmentation techniques. Vibhute and Deshpande [35] developed a statistical method based on template matching and modified template matching to improve the recognition of Marathi text. Additionally, Sonawane et al. [36] introduced an advanced preprocessing technique for degraded Marathi characters, demonstrating significant improvements in recognition accuracy and performance. Their approach outperformed previous methods, with evaluation metrics such as Mean Square Error, Mutual Information, and Peak Signal to Noise Ratio emphasizing the effectiveness of their solution.

## 5.3 Nepali

Nepali is an Indo-Aryan language native to the Himalayas and is the official language of Nepal, where it is also the most widely spoken and serves as a lingua franca. It has official status in Sikkim and the Gorkhaland Territorial Administration of West Bengal in India, and is spoken by significant populations in Bhutan and several Indian states, as well as by Nepali diaspora communities worldwide, with approximately 19 million native speakers and 14 million second language speakers [37].

Nepali, like Hindi, uses the Devanagari script and shares the same core character set. However, the unique phonetic and grammatical structures of Nepali introduce specific character combinations and usages not found in Hindi. These distinctions necessitate adaptations to existing Hindi OCR models or the development of dedicated OCR systems tailored to Nepali. Such differences present challenges in text recognition, requiring fine-tuning of models to ensure accuracy.

Several research efforts have been directed at developing OCR solutions specifically for Nepali. Early works explored traditional machine learning approaches. Pant et al. [38] applied Histogram of Oriented Gradients (HOG) features with an SVM classifier on 400 images of license plates, achieving 75% accuracy. Pant [39] extended this approach with 519 Nepali words and 417 characters, using HOG descriptors and a Random Forest classifier, resulting in accuracies ranging from 93% to 98%. Similarly, Pant and Bal [40] employed a Nepali word dataset and HOG descriptors, achieving 94.80% accuracy with a Random Forest classifier. Pandey et al. [41] used a Multi-layer Feedforward Back Propagation Artificial Neural Network (ANN) to process handwritten and printed Nepali characters, achieving accuracies of 90% for simple words, 60% for complex words, and 50% for handwritten text.

Other studies examined existing OCR frameworks for Nepali. Hengaju and Bal [42] evaluated Tesseract's built-in neural network methods on Nepali text images, with accuracy varying based on image quality: 90.69%-94.84% for high-quality images, 54.34%-71.15% for medium quality, and 38.45%-51.21% for low-quality images.

Recent advancements have leveraged deep learning to improve Nepali OCR. Sharma and Bhattarai [43] used the Devanagari Handwritten Character Dataset and artificial data synthesis to train a CNN-based model with a histogram-based approach, achieving an accuracy of 99.31%. Prajapati et al. [44] extracted pixel-based feature vectors from a dataset of 2,484 segmented character samples, employing an ANN classifier with 81% accuracy. Expanding on this, Prajapati et al. [45] applied template matching using polygonal approximations on the same dataset, achieving an accuracy of 69%.

Nakarmi et al. [46] developed a Convolutional Recurrent Neural Network (CRNN) for both handwritten and printed Nepali characters, achieving a Character Error Rate (CER) of 6.65% and a Word Error Rate (WER) of 13.11%.

## 5.4 Sanskrit

Sanskrit is a classical language of the Indo-Aryan branch of the Indo-European languages, originating in South Asia. It is the sacred language of Hinduism, classical Hindu philosophy, and key historical texts of Buddhism and Jainism. As a link language in ancient and medieval South Asia, Sanskrit spread to Southeast Asia, East Asia, and Central Asia, influencing the religious, cultural, and political elites. Its impact is still evident in the formal and learned vocabularies of languages across South Asia, Southeast Asia, and East Asia.

The development of Optical Character Recognition (OCR) for Sanskrit is a complex and challenging research problem due to several inherent difficulties. One of the primary challenges is the image degradation often encountered in historical and ancient manuscripts, where text may be faded, distorted, or damaged, making it difficult for OCR systems to accurately recognize characters. Additionally, there is a lack of sufficiently large, high-quality datasets specifically tailored for Sanskrit OCR, which significantly hinders the development and training of robust recognition models. Another significant challenge is the presence of long-length words and complex compound characters in Sanskrit text. These compound characters are often formed by the combination of half-letter and full-letter consonants, creating unique characters that do not exist in other languages. These characters are not only more complex to recognize, but they also occur less frequently or are entirely absent in other languages like Hindi, which shares the Devanagari script with Sanskrit.

Given that Hindi OCR systems are typically trained on the simpler structure of Hindi text, they are not equipped to handle these complex compound characters inherent in Sanskrit. As a result, Hindi-based OCR models fail to accurately segment and classify these characters, leading to suboptimal performance when applied to Sanskrit texts. This issue significantly impacts the overall accuracy and reliability of OCR for Sanskrit, necessitating the development of specialized systems designed to address the unique challenges of Sanskrit script recognition.

Despite these challenges, significant strides have been made in developing OCR systems tailored to Sanskrit. Avadesh and Goyal [47] trained a ConvNet-based model using a dataset of 10,106 letter images, achieving a recognition accuracy of 93.32%. Kataria and Jethva [48] utilized a CNN-BLSTM model for both handwritten and printed Sanskrit characters, achieving a character error rate (CER) of 16.67%. Building upon these efforts, Dwivedi et al. [49] used a dataset of 24,000 Sanskrit lines and developed a CNN-BLSTM-based OCR system, reporting a CER of 3.71% and a word error rate (WER) of 15.97%. Shah et al. [50] focused on degraded Sanskrit manuscripts, applying attention-based encoder-decoder models combined with LSTM, achieving an impressive accuracy of 99.44%. Most recently, Madake et al. [51] applied Discrete Cosine Transform (DCT) features and a neural network classifier to Sanskrit characters, achieving an accuracy of 98.7%.

## 5.5 Sindhi

Sindhi, an Indo-Aryan language, is spoken by approximately 30 million people in Pakistan's Sindh province, where it holds official status, and by an additional 1.7 million people in India, where it is recognized as a scheduled language without state-level official status. The primary writing system for Sindhi is the Perso-Arabic script, which dominates literature and is exclusively used in Pakistan. In India, Sindhi is written in both the Perso-Arabic and Devanagari scripts. The Devanagari representation includes four additional characters (ॾ, ॿ, ॻ, and ॼ) that are not present in Hindi, making Hindi OCR unsuitable for recognizing Sindhi text.

Efforts to develop Sindhi OCR for the Devanagari script have only recently begun, with research initiatives emerging from institutions such as Punjabi University, Patiala, and IIIT Hyderabad. One significant contribution is the work by Kaur and Lehal [52], focused on enhancing Sindhi (Devanagari) OCR accuracy through a post-processing model based on Masked Language Modeling with BERT (MLM-BERT). Their research evaluated the model’s performance on two distinct datasets: one from the same domain and another from a different domain. The trained MLM-BERT model improved OCR accuracy by 4.01% on the same-domain dataset and by 1.90% on the different-domain dataset, showcasing its ability to enhance OCR accuracy across varying contexts.

## 5.6 Other Devanagari-based Constituent Indian Languages

For the remaining Indian languages that use the Devanagari script - Bodo, Maithili, Konkani, Dogri, and Kashmiri, relatively little research has been reported in the domain of Optical Character Recognition (OCR). Each of these languages presents unique characteristics and challenges in text recognition, depending on their specific linguistic and orthographic features.

a. **Kashmiri** - While Kashmiri is predominantly written in the Perso-Arabic script, the Devanagari script is also employed, especially in India. Kashmiri in Devanagari incorporates 15 additional characters (Table 4) that are not part of the standard Hindi character set. These unique characters present significant challenges for OCR systems designed exclusively for Hindi, requiring specialized adaptations to accommodate the expanded character set and phonetic requirements.

b. **Maithili** - Similar to Kashmiri, Maithili written in Devanagari includes several additional characters from Table 4 that are absent in Hindi. These unique characters reflect Maithili's distinct phonetic structure and grammatical features, necessitating modifications to OCR models to achieve accurate text recognition.

c. **Dogri** - Although Dogri largely shares its character set with Hindi, it includes a few additional characters specific to the language. These subtle differences require slight modifications to existing Hindi OCR systems to ensure optimal performance on Dogri text.

**d.** **Bodo and Konkani** - Both Bodo and Konkani use the same Devanagari character set as Hindi. Consequently, these languages can be effectively processed by existing Hindi OCR systems without the need for significant changes, making their integration relatively straightforward.

Despite the availability of Hindi OCR systems that can effectively process Bodo and Konkani text, the lack of dedicated research and development for Maithili, Dogri, and Kashmiri highlights a critical gap in the field of Optical Character Recognition (OCR). These languages, with their unique linguistic and orthographic features and additional character sets, require specialized OCR solutions that go beyond the capabilities of Hindi-centric models.

**Table 4**: Kashmiri-Specific Devanagari Characters

| Character | Unicode Code Point | Character | Unicode Code Point |
|---|---|---|---|
| ऄ | U+0904 | ◌ॖ | U+0956 |
| ऎ | U+090E | ◌ॗ | U+0957 |
| ऒ | U+0912 | ॳ | U+0973 |
| ◌ऺ | U+093A | ॴ | U+0974 |
| ◌ऻ | U+093B | ॵ | U+0975 |
| ◌ॆ | U+0946 | ॶ | U+0976 |
| ◌ॊ | U+094A | ॷ | U+0977 |
| ◌ॏ | U+094F | | |

Recent research, such as the work by Sarkar et al. [11] at IIIT Hyderabad, represents a promising step in this direction. They have fine-tuned a pre-existing OCR model to accommodate the additional Devanagari characters specific to these languages. However, a significant challenge in advancing OCR for these languages lies in the development of robust datasets. The lack of expert linguists and the considerable time and resources required to compile such datasets further compounds the difficulty.

To address this, Sarkar et al. have introduced two innovative datasets: Mozhi-LR(S), a synthetic dataset, and Mozhi-LR(R), a real-world dataset. Both datasets provide word-level images paired with textual transcriptions, offering valuable resources for training, and evaluating OCR systems tailored to these languages. These initiatives underscore the importance of dedicated efforts in this domain and pave the way for improved OCR solutions that can preserve and digitize the linguistic diversity of these languages.

## 6 Publicly Available OCR Tools for Devanagari Script

Thanks to advancements in deep learning technology, a number of highly accurate Optical Character Recognition (OCR) systems are now available for recognizing Hindi text. Prominent examples include **Google Cloud Vision**, **Microsoft Azure Computer Vision**, **Tesseract**, **ABBYY FineReader**, and **i2OCR**. These systems leverage state-of-the-art algorithms and deep learning models to deliver exceptional accuracy and reliability in processing Hindi text. These systems have also developed separate OCR solutions for **Marathi** and **Sanskrit**. Additionally, these OCRs can be used to recognize languages such as **Nepali**, **Bodo**, and **Konkani**, as these languages share the same Devanagari character set as Hindi.

However, for other Devanagari-based languages, such as **Sindhi (Devanagari)**, **Kashmiri (Devanagari)**, **Maithili**, and **Dogri**, no separate OCR systems are currently available. The existing OCR systems cannot be directly used to recognize these languages, as they contain additional characters not found in Hindi, as already discussed in previous sections.

Recognizing this gap, efforts are underway under the **Government of India's Bhashini Project** to develop OCR systems for these underrepresented languages. Leading these initiatives, **IIIT Hyderabad** is actively working on creating OCR solutions that can handle the extra characters and unique requirements of these languages. These developments aim to bridge the technological divide and ensure the inclusion of all Indian languages in the digital era, advancing the scope of OCR technologies for Devanagari-based scripts.

## 7 Conclusion and Future Work

This paper presents a comprehensive review of the progress in Optical Character Recognition (OCR) technology for languages using the Devanagari script. Research conducted over the past 50 years, primarily focusing on Hindi, has been thoroughly examined. For Hindi, the primary language, as well as other languages with a similar character set, such as Marathi, Nepali, Bodo, and Konkani, high quality OCR systems have been developed and are readily available to the public.

For Sanskrit, despite its larger character set and long, complex word structures that make OCR development significantly more challenging than Hindi, several companies have successfully created reliable OCR systems.

However, for languages like Sindhi, Dogri, Maithili, and Kashmiri (Devanagari), limited research has been conducted, and no dedicated OCR systems for these languages are currently publicly accessible. Recognizing this gap, efforts are now underway as part of the Government of India's **Bhashini Project** to develop OCR systems tailored to these underrepresented languages, ensuring their inclusion in the evolving digital and AI landscape.

## Declarations

*Ethical Approval:* This declaration is not applicable.

*Funding:* No funding was received to assist with the preparation of this manuscript.

*Financial Interests:* The authors declare that there are no financial interests for this research work.

*Conflicts of Interest:* The authors declare that they have no conflicts of Interest to report regarding the present study.

*Data Availability:* There are no datasets available for sharing.

## References

[1] Sonkusare, M., Gupta, R., Moghe, A., A Review on Handwritten Devanagari Character Recognition. EasyChair Preprints (2019)

[2] Singh, T.P., Gupta, S., Garg, M., Machine learning: A Review on Supervised Classification Algorithms and Their Applications to Optical Character Recognition in Indic Scripts. ECS Transactions 107(1), 6233 (2022)

[3] Wikipedia contributors, Sindhi Language. https://en.wikipedia.org/wiki/Sindhi_language Accessed 2024-11-14

[4] Wikipedia contributors, Devanagari (Unicode block). https://en.wikipedia.org/wiki/Devanagari_(Unicode_block) Accessed: 2024-12-05.

[5] Girdher, H., Sharma, H., Gupta, A., Comprehensive Survey on Devanagari OCR. In Proceedings of the International Conference on Innovative Computing & Communication (ICICC) (2022)

[6] Acharya, S., Pant, A.K., Gyawali, P.K., Deep Learning Based Large Scale Hand-written Devanagari Character Recognition. In: 2015 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), pp. 1–6 (2015). IEEE

[7] Shirkande, A.S., Sawant, S.S., Shinde, N.V., Rao, S.S., Study on the OCR of the Devanagari Script using CNN. International Journal for Research in Applied Science & Engineering Technology. 10(IX), pp. 1502–1506 (2022)

[8] Voigtlaender, P., Doetsch, P., Ney, H., Handwriting Recognition with Large Multidimensional Long Short-Term Memory Recurrent Neural Networks. In: 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 228–233 (2016). IEEE

[9] Sansowa, R., Abraham, V., Patel, M.I., Gajjar, R., OCR for Devanagari Script using a Deep Hybrid CNN-RNN Network. In Emerging Technology Trends in Electronics, Communication and Networking: Select Proceedings of the Fourth International Conference, ET2ECN 2021, pp. 263–274 (2022). Springer

[10] Ahmed, S.M., Abdul, W., Advances and Challenges in Multilingual OCR for Indic Scripts: A Comprehensive Literature Review. Journal of Current Research in Engineering and Science. 6(2), pp. 1–9 (2023)

[11] Sarkar, A., Mondal, A., Lehal, G.S. and Jawahar, C.V., Printed OCR for Extremely Low-resource Indic Languages." In Proceedings of the International Conference on Language Processing (ICLP), vol. 10, pp. 89-101 (2024)

[12] Sinha, R.M.K., A Syntactic Pattern Analysis System and its Application to Devnagari Script Recognition, Ph.D. Thesis, Electrical Engineering Department, Indian Institute of Technology, India, (1973)

[13] Sinha, R.M.K. and Mahabala, H.N., Machine Recognition of Devanagari Script, IEEE Trans on Systems, Man and Cybernetics, Vol. 9, pp. 435-449 (1979).

[14] Sinha, R.M.K., Rule Based Contextual Post-processing for Devnagari Text Recognition, Pattern Recognition, 20, pp. 475-485 (1987)

[15] Palit, S., Chaudhuri, B.B., Das, P.P. and Chatterjee, B.N., A Feature-Based Scheme for the Machine Recognition of Printed Devanagari Script, Pattern Recognition, Image Processing and Computer Vision, Narosa Publishing House: New Delhi, India, pp. 163–168 (1995)

[16] Pal, U. and Chaudhuri, B.B., Printed Devnagari Script OCR System. Vivek, Vol. 10, pp. 12-24 (1997).

[17] Bansal, V. and Sinha, R.M.K., Partitioning and Searching Dictionary for Correction of Optically read Devnagari Characters Strings, Proceedings of the Fifth International Conference on Document Analysis and Recognition, pp. 653–656 (1999).

[18] Jawahar, C., Kumar, M.P., Kiran, S.R., A Bilingual OCR for Hindi-Telugu Documents and its Applications. In Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings., pp. 408–412 (2003). IEEE

[19] Dhurandhar, A., Shankarnarayanan, K. and Jawale, R. Robust Pattern Recognition Scheme for Devanagari Script, In Proceedings of the International Conference on Computational Intelligence and Security, 1021–1026 (2005)

[20] Kompalli, S., Nayak, S., Setlur, S. and Govindaraju, V. Challenges in OCR of Devanagari Documents, In Proceedings of the International Conference on Document Analysis and Recognition, pp. 327–331, (2005). IEEE

[21] Kompalli, S., Setlur, S. and Govindaraju, V. Devanagari OCR Using a Recognition Driven Segmentation Framework and Stochastic Language Models. International Journal on Document Analysis and Recognition (IJDAR). 12, pp. 123–138 (2009)

[22] Holambe, A.N. and Thool, R.C., Printed and Handwritten Character & Number Recognition of Devanagari Script using SVM and KNN. International Journal of Recent Trends in Engineering and Technology 3(2), pp. 163–166 (2010)

[23] Yadav, D., Sánchez-Cuadrado, S. and Morato, J., Optical Character Recognition for Hindi Language using a Neural-Network Approach. Journal of Information Processing Systems 9(1), 117–140 (2013)

[24] Choksi, A., Kumari, K., Kanojiya, S., Sahu, P., Rindani, N., Hindi Optical Character Recognition for Printed Documents using Fuzzy k-Nearest Neighbor Algorithm: A Problem Approach in Character Segmentation. IOSR Journal of VLSI and Signal Processing (IOSR-JVSP), 8(1), pp. 25–24 (2019)

[25] Puri, S., Singh, S.P., An Efficient Devanagari Character Classification in Printed and Handwritten Documents using SVM. Procedia Computer Science 152, 111–121 (2019)

[26] Habib, S., Shukla, M.K., Kapoor, R., OCR Recognition System for Degraded Urdu and Devnagari Script. In: 2019 International Conference on Contemporary Computing and Informatics (IC3I), pp. 245–251 (2019). IEEE

[27] Sankaran, N., Jawahar, C., Recognition of Printed Devanagari Text using BLSTM Neural Network. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), pp. 322–325 (2012). IEEE

[28] Karayil, T., Ul-Hasan, A., Breuel, T.M., A Segmentation-free Approach for Printed Devanagari Script Recognition. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 946–950 (2015). IEEE

[29] Mathew, M., Singh, A.K., Jawahar, C., Multilingual OCR for Indic scripts. In: 2016 12th IAPR Workshop on Document Analysis Systems (DAS), pp. 186–191 (2016). IEEE

[30] Chavan, V., Malage, A., Mehrotra, K., Gupta, M.K., Printed Text Recognition using BLSTM and MDLSTM for Indian Languages. In: 2017 Fourth International Conference on Image Information Processing (ICIIP), pp. 1–6 (2017). IEEE

[31] Kundaikar, T., Pawar, J.D., Multi-font Devanagari Text Recognition using LSTM Neural Networks. In: First International Conference on Sustainable Technologies for Computational Intelligence: Proceedings of ICTSCI 2019, pp. 495–506 (2020). Springer

[32] Gupta, M.K., Dhawan, S., Kumar, A., CMViT OCR: Printed Indian Language Recognition using CMViT. In: 2024 11th International Conference on Signal Processing and Integrated Networks (SPIN), pp. 28–33 (2024). IEEE

[33] Wikipedia contributors, Marathi Language. https://en.wikipedia.org/wiki/Marathi_language. Accessed 2024-11-14

[34] Singh, A., Bacchuwar, K. and Choubey, A., "A Novel GA Based OCR Enhancement and Segmentation Methodology for Marathi Language in Bimodal Framework." In Information Systems for Indian Languages: International Conference, ICISIL 2011, Patiala, India, March 9-11, 2011. Proceedings, pp. 271-277. Springer Berlin Heidelberg (2011).

[35] Vibhute, P. M., & Deshpande, M. S., Optical Character Recognition (OCR) of Marathi Printed Documents Using Statistical Approach. In Advances in Computing and Data Sciences: Second International Conference, ICACDS 2018, Dehradun, India, April 20-21, 2018, Revised Selected Papers, Part I 2, pp. 489-498 (2018) Springer Singapore.

[36] Sonawane, M. S., Dhawale, C. A., & Patil, C. H., Enhanced Preprocessing Technique for Degraded Printed Marathi Characters. In Intelligent Systems and Applications: Select Proceedings of ICISA 2022 (pp. 319-329). (2023) Singapore: Springer Nature Singapore.

[37] Wikipedia contributors, Nepali Language. https://en.wikipedia.org/wiki/Nepali_language Accessed 2024-11-14

[38] Pant, A.K., Gyawali, P.K., Acharya, S., Automatic Nepali Number Plate Recognition with Support Vector Machines. In Proceedings of the 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), pp. 92–99 (2015)

[39] Pant, N., Nepali OCR using Hybrid Approach of Recognition. PhD thesis. May 2016. DOI: 10.13140/RG. 2.2. 33676.72327 (2016)

[40] Pant, N., Bal, B.K., Improving Nepali OCR Performance by using Hybrid Recognition Approaches. In: 2016 7th International Conference on Information, Intelligence, Systems & Applications (IISA), pp. 1–6 (2016). IEEE

[41] Pandey, R.C., Dawadi, B.R., Sharma, S., Basnet, A., Dictionary Based Nepali Word Recognition using Neural Network. Int. J. Sci. Eng. Res, 473–479 (2017)

[42] Hengaju, U., Bal, B.K., Improving the Recognition Accuracy of Tesseract-OCR engine on Nepali Text Images via Preprocessing. Advancement in Image Processing and Pattern Recognition, 3, 40–52 (2023)

[43] Sharma, M.K., Bhattarai, B., Optical Character Recognition System for Nepali Language using Convnet. In Proceedings of the 9th International Conference on Machine Learning and Computing, pp. 184–189 (2017)

[44] Prajapati, S., Joshi, S.R., Maharjan, A., Balami, B., Evaluating Performance of Nepali Script OCR Using Tesseract and Artificial Neural Network. In: 2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS), pp. 104–107 (2018). IEEE

[45] Prajapati, S., Maharjan, A., Joshi, S.R., Balami, B., Assessing and Analyzing Tesseract Based Nepali Script OCR. Deerwalk Journal of Computer Science Information Technology, 1, 37–46 (2019)

[46] Nakarmi, S., Sthapit, S., Shakya, A., Chulyadyo, R., Bal, B.K., Nepal Script Text Recognition using CRNN CTC Architecture. In Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages@ LREC- COLING 2024, pp. 244–251 (2024)

[47] Avadesh, M., Goyal, N., Optical Character Recognition for Sanskrit using Convolution Neural Networks. In: 2018 13th IAPR International Workshop on Document Analysis Systems (DAS), pp. 447–452 (2018). IEEE

[48] Kataria, B., Jethva, H.B., 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 (2019)

[49] Dwivedi, A., Saluja, R., Sarvadevabhatla, R.K., An OCR for Classical Indic Documents Containing Arbitrarily Long Words. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 560–561 (2020)

[50] Shah, R., Gupta, M.K., Kumar, A., Ancient Sanskrit Line-Level OCR using Open-NMT Architecture. In: 2021 Sixth International Conference on Image Information Processing (ICIIP), vol. 6, pp. 347–352 (2021). IEEE

[51] Madake, J., Yedle, Y., Shahabade, V., Bhatlawande, S., Sanskrit OCR System. In International Conference on Advanced Communication and Intelligent Systems, pp. 188–200 (2023). Springer

[52] Kaur, A. and Lehal, G.S., Enhancing Sindhi (Devanagari) OCR Performance Through MLM-BERT-Based Error Correction Model, NanoTechnology Perceptions, 20(6), pp. 4441-4459 (2024)