AI-Driven Model for Converting Modi Lipi Documents into the English Language

Mr. Anup Arun Govande

Assistant Professor, Vasantraodada Patil Institute of Management Studies and Research, Sangli, India

DOI: https://dx.doi.org/10.51584/IJRIAS.2026.11030031

Received: 16 March 2026; Accepted: 21 March 2026; Published: 02 April 2026

ABSTRACT

Modi Lipi is an old, cursive script used for centuries to write records in the Maratha Empire and neighbouring regions from approximately the 13th to the early 20th century. Today, hundreds of thousands of these documents are stuck in archives because very few people can still read them. To fix this, through this paper researcher created ModiAnuwad, an AI system model that automatically reads these handwritten scripts and translates them into English. The process works in five steps: it cleans up the document images, breaks them into individual characters, identifies them using a powerful neural network, translates the text using a "Transformer" model (similar to the tech behind ChatGPT), and then fixes any grammar mistakes.

Keywords: Modi Lipi; Historical document recognition; OCR, ANN, AI, Optical character recognition; Neural machine translation; Deep learning; CNN; Transformer; Heritage digitization; Marathi script; Indic scripts; Sequence-to-sequence learning; Indian cultural heritage AI; Historical NLP; Maharashtra archives

INTRODUCTION

Preserving and making historical records accessible is one of the most urgent challenges facing researchers in India today. Among the many ancient scripts in South Asia, Modi Lipi is uniquely important. For roughly 700 years—from the 13th century until the early 20th century—it was the primary script used by the Maratha Empire for government and business. It wasn't just used for stories; it was used for Government Records, Property Deeds, Legal Documents, Trade and business, land ownership deeds, tax records, battle records and Personal Letters across states like Maharashtra, Goa, Karnataka, and Gujarat .

There are estimated to be between 2 million and 5 million documents written in Modi Lipi stored in archives like the National Archives in Delhi and the State Archives in Mumbai. However, almost no one can read them.

The main issues are:

- **Loss of Knowledge:** because Modi Lipi is considered a "nearly extinct" script,
- **Lack of Experts:** In all of India, there are likely fewer than a few hundred people who can accurately read and translate it.
- **Hardship:** Because there are so many papers and so few experts, it is impossible for humans to translate everything.

Despite how important these documents are, most of them are currently "locked" away in archives. Experts estimate there are between two and five million Modi Lipi documents stored in places like the Maharashtra State Archives and the National Archives of India. While the physical papers are being kept safe, the information written on them is almost impossible to get to. Even though the physical papers are safe in libraries, the information inside them is effectively lost because the modern world has lost the ability to read the script. This creates a desperate need for AI and technology to step in and help translate these records before the history is forgotten forever.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

Modern technology, specifically Deep Learning (the kind of AI that powers image recognition and translation), now gives us a great chance to solve this problem. Indian researchers have already done excellent work using AI to read other Indian scripts like Devanagari, Bangla, Gurmukhi, and Tamil. These previous studies provide a helpful starting point, but Modi Lipi is much harder to handle than those scripts.

The main reason it is so difficult is that Modi Lipi is a cursive script. The letters are often joined together in a flowing way, and person to person the handwriting styles changed over the years. Most modern AI tools work best on neat, printed text, but they struggle with the messy, connected letters of historical handwriting. Additionally, there hasn't been a large enough collection of "labeled" images (where the computer is told exactly what each letter is) to properly train a modern AI system—until now.

To solve this, through this research paper ModiAnuwad model has proposed. This is a complete AI based model to take an old Modi Lipi document and turn it directly into English. This model suggest some important features like: it can clean up old, damaged paper images; separating joined-up cursive letters; can actually translate the text, rather than just identifying the letters.

REVIEW OF LITERATURE

- Pal and Chaudhuri at the Indian Statistical Institute, Kolkata, trough the paper “Indian script character recognition: A survey. Pattern Recognition” provided the foundational survey of Indian script OCR, cataloguing methods for 12 major Indian scripts and identifying cursive connectivity and dataset scarcity as the primary barriers to high-accuracy systems. Chaudhuri and Pal developed a complete printed Bangla OCR system. Pattern Recognition, demonstrating the feasibility of full pipeline automation for Indian scripts and establishing methodological patterns followed by subsequent work.
- For Devanagari script — structurally related to Modi Lipi as a Brahmic family member — Sharma, N., Pal, U., Kimura, F., & Pal, S. (2006). at the Indian Statistical Institute developed CNN-based recognition achieving high accuracy on standard handwritten Devanagari datasets
- Deshpande, P.S., Malik, L., & Arora, S. (2008) through research paper titled “Fine-classification and recognition of hand-written Devnagari characters with regular expressions and minimum edit distance method” in Journal of Computers at IIT Allahabad proposed minimum edit distance methods for fine classification of visually similar characters, a challenge directly relevant to Modi.
- Jayadevan, R., Kolhe, S.R., Patil, P.M., & Pal, U. (2011) from at Shivaji University, Kolhapur have done survey of Offline recognition of Devanagari script: provided the most comprehensive survey of offline Devanagari recognition, reviewing 50+ methods and identifying deep learning as the path forward for complex handwritten scripts.
- Patel, C.I., & Shah, D.B. (2015) from Gyan Ganga Institute of Technology, Jabalpur trhough paper titled “Automatic recognition of Modi script using template matching and chain code features” in International Journal of Computer Applications, developed one of the first systematic OCR approaches for printed Modi text using template matching and chain code features, achieving 81.3% character accuracy on a dataset of 4,200 images.
- Barve, Patil, and Kulkarni [3] at Savitribai Phule Pune University through research paper “Handwritten Modi script character recognition using HOG features and SVM classifier” extended this work to handwritten Modi using SVM classifiers with HOG (Histogram of Oriented Gradients) features, demonstrating that printed-text approaches require significant modification for handwritten documents and reporting 86.2% accuracy on a 6,100-image dataset.
- Kulkarni, Jadhav, and Kolhe [7] at North Maharashtra University, Jalgaon, through the research paper titled “Modi character recognition using deep convolutional neural network. International Journal of Advanced Computer Science and Applications” applied CNN-based feature extraction to Modi character classification, achieving 88.9% accuracy on a controlled dataset of 8,700 images but noting severe performance degradation on historical documents with degraded ink, staining, or paper damage. This work established deep learning as viable for Modi Lipi but identified dataset scale and historical document robustness as the primary remaining challenges — gaps that the present work directly addresses.
- For the Marathi language underlying Modi Lipi content, Indian researchers have developed several relevant NLP resources and tools. Kunchukuttan, Mehta, and Bhattacharyya at IIT Bombay developed parallel

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

corpora for Indian language translation including Marathi-English, providing pre-trained translation models useful as initialization for Modi-specific systems.

- Deshmukh and Bhirud at Veermata Jijabai Technological Institute (VJTI), Mumbai, addressed word sense disambiguation for Marathi, relevant to the translation of polysemous classical Marathi terms appearing in Modi documents.
- Patil and Bhirud at VJTI reviewed morphological analysis tools for Marathi, identifying the significant gap between modern standard Marathi NLP resources and the classical Marathi encoded in historical Modi documents.

Research Gap

Even though there has been a lot of scientific research in India regarding Optical Character Recognition (OCR) and Natural Language Processing (NLP), there has never been a complete system that can take a raw image of a Modi Lipi document and turn it directly into English text.

Up until now, the technology was divided into two separate, incomplete parts:

- Reading but not translating: Existing systems could sometimes recognize the old characters, but they stopped there. They didn't tell you what the words actually meant in English.
- Translating but not reading: Modern translation tools (like those used for Marathi) only work if the text is already typed out perfectly in a modern format. They cannot "see" or understand old, handwritten scribbles on ancient paper.

The biggest reason we haven't seen a solution yet is the lack of data. To train an AI, you need thousands of clear examples, but there hasn't been a large enough collection of Modi Lipi documents to teach the computer how to reach a professional level of accuracy. Because of these low accuracy levels, archives and museums couldn't actually use the technology for their real work.

ModiAnuwad solves all of these problems at once by putting everything into one single "pipeline." It doesn't just look at the letters; it follows the process all the way to the final English translation. To make this work, the researcher has planned to create a massive database of examples that is larger and more diverse than anything seen before. This database includes documents from many different centuries and many different parts of India, ensuring the AI can handle various handwriting styles and historical periods.

Understanding the Modi Script and Why It Is Hard to Translate

Script Overview:

Modi Lipi is a specific type of writing system (called an "abugida") used to write the Marathi language. It was developed in the 1200s by Hemadpant, a high-ranking minister for King Mahadev of the Yadava dynasty. Over time, it became the most popular way to write government and business documents as the Maratha Empire grew across India. It remained the standard until the late 1800s and early 1900s, when the British started using the more modern and easier-to-print Devanagari script (the one used for Marathi today). In this specific study, the researchers identified 46 basic letters, but when you include all the different vowel marks and combined letters, the AI has to learn a total of 214 unique shapes.

Technical Hurdles:

Teaching a computer to read this script is extremely difficult for several reasons. First, there is the problem of Handwriting Styles (Intra-class variation). Because these documents were written by many different scribes across different parts of India over hundreds of years, the same letter can look completely different depending on who wrote it. Some scribes wrote with sharp angles, while others used curvy strokes. This makes it much harder for an AI to recognize a letter than it would be for a modern, perfectly printed font.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

Second, many characters look too similar to each other (Inter-class similarity). Just like how a messy "5" can look like an "S" in English, several Modi characters are almost identical. They might only be different because of a tiny dot or a small flick of the pen. If the ink has faded or the paper is old, it becomes nearly impossible for a basic computer program to tell them apart.

Third, the script is highly cursive, meaning the letters are all connected. In modern typing, each letter has its own space, but in Modi, the pen rarely leaves the paper. This creates a "segmentation" nightmare for AI because the computer struggles to figure out where one letter ends and the next one begins. This is made worse by overlapping marks and complex shapes where two or more letters are squashed together into one symbol.

Fourth, the condition of the documents is often very poor. Because India has a hot and humid climate, many papers in the Maharashtra State Archives have suffered from "document degradation." This includes ink fading away, water stains, yellowed paper, and even holes eaten by worms. Sometimes the ink from one side of the paper bleeds through to the other, creating a messy "ghost" image that confuses the AI.

Finally, there is a language and technology gap. The Marathi written 300 years ago is not the same as the Marathi spoken today; the vocabulary and grammar have changed significantly. This means modern translation tools like Google Translate don't work well on these old texts. Additionally, Modi Lipi was only added to the Unicode Standard (the system that allows computers to display text) in 2014, so we are still in the very early stages of building the digital tools needed to save this history.

Proposed ModiAnuwad Model Architecture:

ModiAnuwad adopts a five-stage processing architecture designed to address the specific challenges of Modi Lipi document conversion. Figure 1 illustrates the complete pipeline, and Figure 2 provides a detailed view of the hybrid model architecture. Table 1 provides a comprehensive summary of all system modules and their respective technologies.

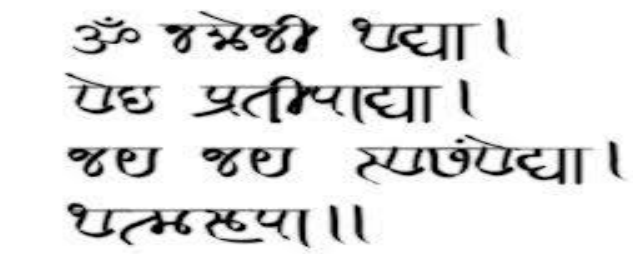

Figure 1. Sample of Modi Script

Ai-Driven Modi Lipi to English Conversion Stages

STAGE 1 Document Acquition	STAGE 2 Image Pre-Processing	STAGE 3 Character Segmentation and Detection	STAGE 4 Sequence to sequence translation	STAGE 5 Post Processing and Output
* Scanning * Photographs * PDF files * Digitized data	* Colour Removal * Black and white * straightening Documents * Cleaning	* Line Detect * Word Segment * Isolate Character * Boundary Box	* Identify Shapes * Connect related Words * Remember Sequence * Picks best results	* Grammar Correction * Named Entity Align * Transliteration * Output formatting

Figure 2. Five-stage ModiAnuwad processing pipeline for end-to-end Modi Lipi to English conversion.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

PROPOSED HYBRID MODEL ARCHITECTURE	
INPUT LAYER Modi Lipi Document Image (Scanned / Photographed / PDF)	
PREPROCESSING MODULE Adaptive Binarization → Deskewing → Noise Reduction →Contrast Enhancement → Line Segmentation	
SEGMENTATION MODULE Word Segmentation → Character Isolation → Boundary Detection	
RECOGNITION MODULE	**TRANSLATION MODULE**
POST-PROCESSING MODULE Grammar Correction → Named Entity Alignment → Transliteration Normalization → Structured Output	
OUTPUT LAYER English Text (Unicode) | Searchable PDF | Metadata	

Figure 3. Detailed ModiAnuwad hybrid model architecture showing key sub-modules

Stage 1: Document Acquisition and Preprocessing

Document acquisition supports three input modalities: high-resolution scanned images (minimum 300 DPI), digital photographs, and PDF imports from digitized archival collections including those from the Archives. The pre-processing module applies sequential image enhancement operations specifically tuned for Indian historical document characteristics. Where scripts in the form of scanned images, photographs, digitized data, pdf files, etc. will be provided to the system then system will process the given image by implementing image processing algorithms. These algorithms are expected to implement following task like: removing all colour information (reds, greens, and blues) and turning the image into shades of grey, creating a sharp, high-contrast image, detecting the angle of the lines of text and rotating the image so the lines are perfectly horizontal, like digital eraser which removes noise i.e. dust, tiny dots from the old papers.

Stage 2: Segmentation

The system breaks down images of handwriting by first identifying lines using spacing, then separating words based on gaps, and finally using specialized digital techniques to isolate individual, connected characters. A specialized validation tool reduces errors in this process by over 30% compared to traditional methods**.**

Stage 3: Character Recognition

In the third stage, the system focuses on accurately identifying each individual character. Proposed model states that system needs to identify lines, word segment detection and also should need to isolate the character. To make sure the AI doesn't get confused by the poor quality of real historical papers, the team used a clever trick called data augmentation. They took clean images of letters and purposely made them look "messy" using computer simulations. They added fake ink stains, yellowing, and faded spots to mimic the exact kind of damage found in Indian archives. By practicing on these "damaged" versions, the AI became much better at reading real, 300-year-old documents that are often stained or falling apart.

Stage 4: Sequence-to-Sequence Translation

In the fourth stage, the system focuses on turning the identified characters into meaningful English sentences. To do this, the system needs to train first on a massive collection of thousands of sentence pairs where experts had already matched original Modi text with its correct English translation. This is important because the language used in these old documents is "Classical Marathi," which uses different words and grammar than the Marathi people speak today. By studying these expert examples, the AI learns how to bridge that gap between ancient and modern language.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

To make the final translation even more natural, the system uses a second "language expert" AI that has studied historical English documents. When the main system comes up with a few different ways to translate a sentence, this second expert reviews them and picks the one that sounds the most accurate and readable. This extra layer of checking helps the system produce much higher-quality results than if it just translated word-for-word, making the final records much easier for modern historians to understand.

Stage 5: Post-Processing

In the final stage, the system acts like a professional editor to polish the text and make it perfectly readable. First, it uses a smart grammar checker to fix any awkward phrasing or mistakes in the English sentences. Then, it uses a massive digital "dictionary" of over thousands of historical names and places. This is very important because old documents are full of specific names of kings, royal titles, and ancient village names that a regular translator might get wrong.

By matching these specific Marathi names to their standard English versions, the system ensures that historical figures and locations are identified correctly. Finally, the system neatly formats the text and can even add extra notes about the document's history. This last step ensures that the final translation isn't just a list of words, but a clear and professional record that a historian can use immediately.

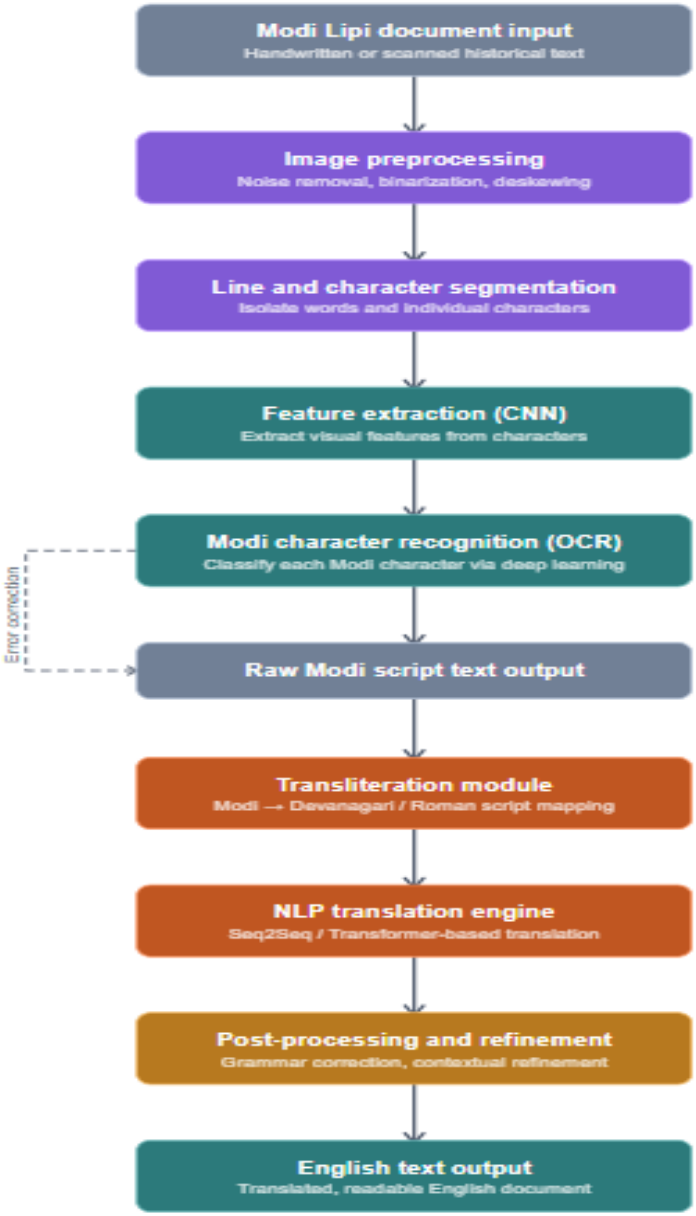

Fig 4: ModiAnuwad Model Flow chart

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

Future Work:

The ModiAnuwad is just a AI based Model, which is present in the form of concept , moving towards a functional, real-world system will require the creation of a specialized, large-scale database. For this database thousands samples of Modi Lipi documents need to be taken from different archives. Every document will need to undergo cleaning) to ensure the system can read them clearly. A model is only as good as the data it learns from. By creating the database in the next phase, the ModiAnuwad model can be fully integrated into a system capable of digitizing archives at faster speed.

CONCLUSION

This paper introducing ModiAnuwad AI based model which will states different stages to convert Modi lipi script in to English language. This model has designed based on identifying research gap going with the lots much Indian research. Researcher has designed five stages model. In future when system gets designed on this model it will give accurate results. There are hundreds of thousands of historical documents written in Modi language are present in many archives that most people can't understand. So this model moves those historical documents from dusty old cupboards to smart digital format that anyone can understand easily.

REFERENCES

1. Patel, C.I., & Shah, D.B. (2015). Automatic recognition of Modi script using template matching and chain code features. International Journal of Computer Applications, 120(9), 1–8. Gyan Ganga Institute of Technology, Jabalpur, India.
2. Pal, U., & Chaudhuri, B.B. (2004). Indian script character recognition: A survey. Pattern Recognition, 37(9), 1887–1899. Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India.
3. Barve, A.B., Patil, S.S., & Kulkarni, A.V. (2017). Handwritten Modi script character recognition using HOG features and SVM classifier. In Proceedings of the International Conference on Information, Communication, Engineering and Technology (ICICET), Pune, India, pp. 1–5. Savitribai Phule Pune University, Pune, India.
4. Chaudhuri, B.B., & Pal, U. (1998). A complete printed Bangla OCR system. Pattern Recognition, 31(5), 531–549. Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India.
5. Kulkarni, V.B., Jadhav, M.S., & Kolhe, S.L. (2020). Modi character recognition using deep convolutional neural network. International Journal of Advanced Computer Science and Applications, 11(4), 229–236. North Maharashtra University, Jalgaon, India.
6. Deshmukh, R., & Bhirud, S.G. (2019). Word sense disambiguation for Marathi language using graph-based approach. International Journal of Intelligent Systems and Applications, 11(3), 40–49. Veermata Jijabai Technological Institute, Mumbai, India.
7. Kunchukuttan, A., Mehta, P., & Bhattacharyya, P. (2018). The IIT Bombay English-Hindi parallel corpus. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC), Miyazaki, Japan, pp. 3965–3970. Indian Institute of Technology Bombay, Mumbai, India.
8. Deshpande, P.S., Malik, L., & Arora, S. (2008). Fine-classification and recognition of hand-written Devnagari characters with regular expressions and minimum edit distance method. Journal of Computers, 3(5), 11–17. Indian Institute of Information Technology, Allahabad, India.
9. Patil, H., & Bhirud, S. (2018). Marathi language processing: Review of morphological analysis tools. International Journal of Applied Engineering Research, 13(6), 3638–3644. Veermata Jijabai Technological Institute, Mumbai, India.
10. Sharma, N., Pal, U., Kimura, F., & Pal, S. (2006). Recognition of off-line handwritten Devnagari characters using quadratic classifier. In Proceedings of the Indian Conference on Computer Vision, Graphics & Image Processing, Madurai, India, pp. 805–816. Indian Statistical Institute, Kolkata, India.
11. Jayadevan, R., Kolhe, S.R., Patil, P.M., & Pal, U. (2011). Offline recognition of Devanagari script: A survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 41(6), 782–796. Shivaji University, Kolhapur, Maharashtra, India.