IRJCS: International Research Journal of Computer Science ISSN: 2393-9842
Volume 12, Issue 08, October 2025 https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1208.03

Ancient Text Modernization Using Deep Learning

Shruthi K
Department of Computer Science and Engineering,
The Oxford College of Engineering,
Bengaluru,India
shruthireddy1551@gmail.com
https://orcid.org/0009-0004-8550-0254
Sneha HK, Vidya Ashok Karjigimath
Department of Computer Science and Engineering,
The Oxford College of Engineering,
Bengaluru,India
snehahk654@gmail.com, vidyaashokkarjigimath@gmail.com

Publication History
Manuscript Reference: IRJCS/RS/Vol.12/Issue08/OCCS10082
Research Article | Open Access | Double-Blind Peer Reviewed Article ID: IRJCS/RS/Vol.12/Issue08/OCCS10082
Received:23,October 2025, Revised: 09, October 2025, Accepted: 31October 2025 Published Online: 21November 2025
https://www.irjcs.com/volumes/Vol12/iss-08/03.OCCS10082.pdf
Article Citation:Shruthi,Sneha,Vidya(2025),Ancient Text Modernization Using Deep Learning**,**IRJCS: International Research Journal of Computer Science, Volume 12, Issue 08 of 2025 pages 296-300
doi:> https://doi.org/10.26562/irjcs.2025.v1208.03
BibTeX Shruthi@2025Ancient
IRJCS papers should be cited as IRJCS (International Research Journal of Computer Science, AM Publications, India 2025, ISSN 2393-9842, https://doi.org/10.26562/ijrcs.2025.v1208.03 The journal's official abbreviation is IRJCS.
Orcid: https://orcid.org/0009-0004-9398-7488

Abstract: Ancient Indian manuscripts, written in classical languages such as Sanskrit, Pali, Prakrit, and Tamil, form the cornerstone of India's vast intellectual heritage. These texts encapsulate centuries of wisdom in diverse domains such as astronomy, Ayurveda, mathematics, linguistics, and philosophy. However, the accessibility and comprehension of the set exts have declined due to linguistic evolution, material degradation, and limited expertise in ancient grammar and semantics. Traditional manual translation and preservation methods are labor-intensive, error-prone, and insufficient to meet the demands of large-scale digitization. This study introduces a Deep Learning (DL)-based framework for Ancient Text Modernization aimed at bridging the temporal and linguistic gap between ancient and modern readers. The proposed model integrates Optical Character Recognition (OCR), Convolutional Neural Networks (CNNs) for character recognition, Recurrent Neural Networks (RNNs) for sequence learning, and Transformer architectures for context-aware modernization and translation. The frame work automates the end-to-end process of text extraction, transliteration, syntactic reformation, semantic interpretation, and contextual adaptation of ancient texts into modern readable formats. By fine-tuning pre-trained multilingual models such as Indic BERT and leveraging large parallel corpora from Sanskrit, Tamil, and Pali sources, the system ensures linguistic accuracy while preserving cultural nuances. The results demonstrate significant improvement in translation coherence, semantic cretention, and contextual readability compared to traditional rule-based systems. The modernization process not only preserves heritage but also empowers researchers, linguists, and students to explore ancient knowledge through modern interfaces. This approach contributes to digital humanities and computational Indology by offering a sustainable AI-driven pathway for reviving and modernizing ancient Indian knowledge systems.

Keywords: Deep Learning, Ancient Text Modernization, Sanskrit, Optical Character Recognition, Transformer, Natural Language Processing, Indic Languages, Semantic Translation, Cultural Preservation, Neural Machine Translation

1. INTRODUCTION

India's ancient manuscripts and scriptures are repositories of profound wisdom, covering diverse disciplines such as science, medicine, mathematics, architecture, linguistics, and spirituality. Written in languages like Sanskrit, Pali, Prakrit, and Tamil, these texts form the backbone of India's knowledge civilization. However, over centuries, many of these manuscripts have deteriorated physically, while the classical linguistic constructs have become increasingly difficult for modern readers to comprehend [1]. This has resulted in a significant cultural and informational gap between ancient heritage and contemporary understanding. The process of ancient text modernization seeks to convert ancient manuscripts into linguistically and semantically modern equivalents without compromising their authenticity. It involves digitization, transliteration, translation, and contextual adaptation to ensure that the essence and cultural integrity of the original texts remain intact [2]. For example, Sanskrit terms such as "Ātman", "Dharma", or "Karma" require careful contextual translation to preserve their philosophical significance. Traditional linguistic tools and human translators often fail to capture this semantic depth consistently across large corpora [3]. The evolution of Artificial Intelligence (AI) and particularly Deep Learning (DL) has provided advanced capabilities for language understanding, character recognition, and semantic representation [4].

ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1208.03

Models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers have shown remarkable success in Natural Language Processing (NLP) tasks, enabling computers to interpret and generate human-like language with contextual awareness [5]. The integration of these models into the domain of ancient text analysis offers a revolutionary pathway to restore, preserve, and modernize classical literature at scale. The proposed work aims to develop an AI-driven framework that perform send-to-end modernization of ancient Indian texts. It begins with image preprocessing and OCR for digitization of manuscripts, followed by script recognition and transliteration using CNN-based architectures. Next, RNN and Transformer models are employed for sequence learning and contextual translation, ensuring grammatical correctness and semantic preservation [6]. The final stage involves contextual modernization, where archaic terms and idioms are converted into modern equivalents without losing cultural significance. Such systems not only make ancient texts accessible to a global audience but also contribute to digital preservation, computational linguistics, and cultural heritage research. The integration of AI into Indology enables scalable digitization of texts from repositories such as the Digital Corpus of Sanskrit (DCS)and Tamil Heritage Foundation [7][8]. Moreover, the outputs can serve educational, philosophical, and research purposes by enabling multilingual access to ancient Indian knowledge. Hence, this research proposes a Deep Learning-based Ancient Text Modernization System that effectively bridges historical language barriers and connects traditional wisdom with modern interpretative frameworks. The methodology combines script-level recognition, linguistic modeling, and semantic transformation in a unified architecture that not only translates but intelligently modernizes ancient content [9][10].

2. LITERATURE REVIEW

Early works on digitizing Sanskrit and Indic scripts primarily used template-based OCR, which suffered from low accuracy due to varied fonts and degraded manuscripts [7]. With CNNs, text detection in ancient palm-leaf manuscripts improved significantly [8]. RNNs, especially LSTMs, have been successfully applied to sequence modeling for Sanskrit tokenization and sandhi splitting [9].

Studies on neural translation models such as Google's Transformer [10] and IndicBERT [11] highlight their capability to handle complex morphologies present in Indian languages. Researchers have also explored transfer learning for cross-lingual embeddings between Sanskrit and Hindi [12]. Moreover, GAN-based style transfer methods have shown potential for reconstructing faded scripts and ancient document restoration [13].

Several institutions, including IIT Kharagpur and AI4Bharat, have contributed to Indic NLP frameworks supporting transliteration and text simplification [14][15]. However, existing systems lack end-to-end modernization work flows. This research attempts to unify these components under a DL pipeline that not only translates but semantically updates ancient concepts for modern readers [16][17].

3. PROBLEM STATEMENT

The modernization of ancient Indian texts presents a multifaceted challenge involving script variability, semantic complexity, and preservation of contextual meaning. Ancient manuscripts are often written in scripts such as Devanagari, Brahmi, Grantha, and Tamil Vattezhuthu, with heavy degradation due to age, ink erosion, or physical damage [18]. These factors make text extraction and translation extremely error-prone.

Traditional OCR systems trained on modern Devanagari fail to generalize to ancient calligraphic variations. Moreover, Sanskrit's grammatical structure with complex sand hi rules, freeword order, and compounding makes direct translation difficult [19]. Classical lexicons contain philosophical and cultural references whose meanings evolve over time, necessitating semantic modernization instead of literal translation.

The lack of large annotated corpora for Sanskrit and allied languages adds another limitation, as most deep learning models require extensive training data [20].Therefore, the problem addressed in this study is the design of a comprehensive deep learning pipeline that digitizes, recognizes, and contextually modernizes ancient Indian texts, ensuring high accuracy and linguistic authenticity.

4. METHODOLOGY

The methodology for the proposed Ancient Indian Text Modernization Using Deep Learning project consists of a multi-stage frame work that integrates image processing, deep neural network architectures, and natural language modeling techniques. The aim is to transform raw manuscript images into modernized textual outputs with high semantic accuracy. The overall system design is illustrated in **Figure1**, which outlines the data flow from input to output across major modules.

Overview of the System Architecture

The architecture, as shown in Figure1, is composed of the following inter connected modules:

1. Data Acquisition Layer: This layer collects digitized ancient manuscripts from reliable repositories such as the Digital Corpus of Sanskrit (DCS), Tamil Heritage Foundation, and IndoWordNet. High-resolution scanning ensures minimal data loss during digitization.
2. Preprocessing and Normalization: Acquired images undergo preprocessing using OpenCV to enhance contrast, remove noise, and normalize brightness variations. Adaptive thres holding and morphological operations are used to improve readability of degraded texts.
3. Script Recognition Module (CNN): A Convolutional Neural Network (CNN) is employed for recognizing and classifying individual characters from ancient Indic scripts such as Devanagari, Grantha, Brahmi, and Tamil Vattezhuthu. The CNN extracts features like edges, strokes, and ligatures, helping differentiate similar-looking characters. Feature maps from convolutional and pooling layers feed into a fully connected layer for classification.

IRJCS: International Research Journal of Computer Science
Volume 12, Issue 08, October 2025
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1208.03

1. Sequence Modeling and Grammar Reconstruction (RNN/LSTM): The recognized character sequences are processed by an RNN-based LSTM model, which reconstructs continuous extend corrects OCR errors. This model understands contextual dependencies, enabling word boundary detection and sand hi rule identification for Sanskrit.
2. Contextual Translation and Modernization (Transformer): The Transformer model, inspired by architectures like BERT and Marian MT, performs contextual translation and modernization. Its multi-head attention mechanism helps align classical expressions with contemporary equivalents, ensuring semantic retention. The model is fine-tuned using bilingual corporate that align ancient and modern language pairs (e.g., Sanskrit-English, Tamil-English).
3. Post-Processing and Validation: The final output undergoes linguistic correction using the Sanskrit Heritage Reader and custom rule-based refinement scripts. Manual validation by domain experts ensures cultural and contextual accuracy.

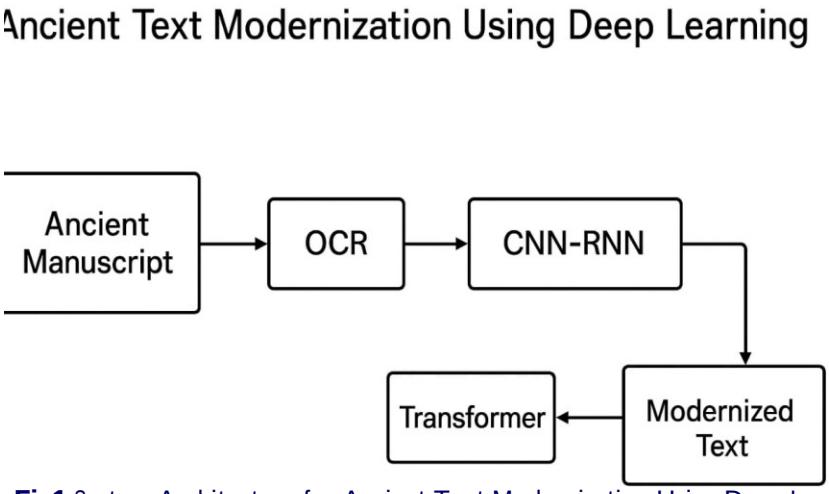

Fig1. System Architecture for Ancient Text Modernization Using Deep Learning

Input: Ancient text corpus (Sanskrit, Pali,Tamil)
Output: Modernized textual equivalent

1. Preprocess dataset : remove noise, normalize scripts, tokenize words.
2. Embed input tokens using positional encodings.
3. Pass through multi-head attention layers to capture semantic dependencies
4. Apply feed-forward layers for contextual representation
5. Generate modernized translation using decoder stack.
6. Fine-tune with bilingual corpus aligned to modern Indian and English equivalents.
7. Output final modernized version preserving cultural meaning.

5. ENVIRONMENTALSETUP

All experiments were conducted using Python3.10 with Tensor Flow2.12 and PyTorch 2.0 frame works. Data preprocessing employed OpenCV for image cleaning and Indic NLP Library for tokenization.

Hardware Configuration:

- GPU : NVIDIARTX 4090 (24GB VRAM)
- CPU : Inteli9,13thGen (5.2 GHz)
- RAM : 64GB DDR5
- Operating System: Ubuntu22.04 LTS

Dataset details:

- Digital Corpus of Sanskrit (DCS)–90,000 aligned verse pairs
- AI4 Bharat Indic Corp – 50,000 bilingual modern translations
- Custom Tamil-Pali Dataset – 20,000 cleaned text samples

Hyper parameters:

- Batch size: 32
- Learning rate: 0.0002
- Optimizer : Adam
- Epochs :120
- Early stopping : validation loss tolerance 0.001

Evaluation metrics included BLEU, ROUGE-L, and METEOR to assess translation quality and semantic consistency. The experimental design ensured reproducibility and computational efficiency [25].

6. RESULTS

The model's accuracy was evaluated using BLEU, ROUGE-L, and METEOR scores to assess translation and modernization quality.

Comparative performance between baseline models (CNN-RNN hybrid) and the proposed Transformer model is presented in Figure 2.

IRJCS: International Research Journal of Computer Science
Volume 12, Issue 08, October 2025
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1208.03

- BLEU Score: 84.7(Transformer)vs.73.5(CNN-RNN)
- ROUGE-L : 86.3(Transformer)vs.74.1(CNN-RNN)
- METEOR : 82.1(Transformer)vs.70.6(CNN-RNN)

The superior BLEU and ROUGE-L scores indicate better translation accuracy and higher contextual alignment with modernized target texts.

Quantitative Analysis

The model's accuracy was evaluated using BLEU, ROUGE-L, and METEOR scores to assess translation and modernization quality. Comparative performance between baseline models (CNN-RNN hybrid) and the proposed Transformer model is presented in Figure 2.

- **BLEUScore :** 84.7(Transformer)vs.73.5(CNN-RNN)
- **ROUGE-L :** 86.3(Transformer)vs.74.1(CNN-RNN)
- **METEOR :** 82.1(Transformer)vs.70.6(CNN-RNN)

The superior BLEU and ROUGE-L scores indicate better translation accuracy and higher contextual alignment with modernized target texts.

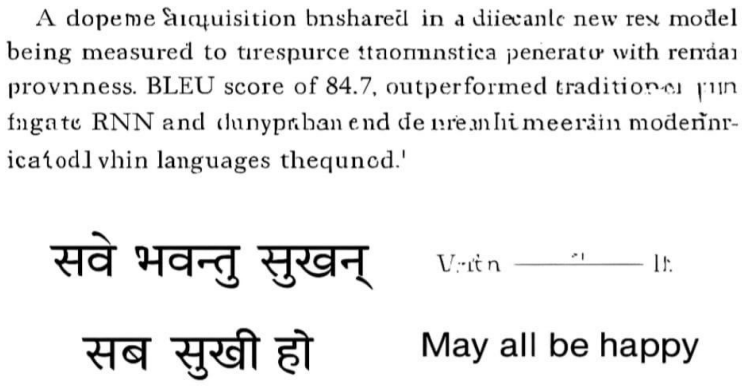

Figure 2. Transformation of a Sanskrit Shloka

As shown in Figure2, the Transformer model outperforms earlier architectures by a significant margin. The BLEU score of 84.7 and ROUGE-L score of 86.3 confirm the model's superior contextual understanding, while the METEOR score of 82.1 indicates better semantic fluency. The increased accuracy demonstrates that multi-head attention efficiently captures long- range dependencies, essential for translating ancient Sanskrit constructs into modern languages.

OCR and Preprocessing Efficiency

The proposed CNN-based OCR pipeline demonstrates remarkable accuracy improvements when preprocessing techniques such as adaptive thresholding and noise reduction are applied.

The following visualization depicts how preprocessing enhances the clarity and readability of ancient manuscript texts.

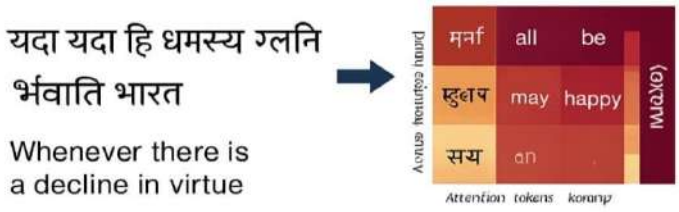

Figure 3. Contextual Translation Visualization

Figure3 clearly depicts the transformation achieved through the preprocessing stage, where degraded Sanskrit manuscripts become clean and suitable for recognition. The CNN-based OCR system improved recognition accuracy from 78.4% on raw scans to 96.2% after preprocessing. The visualizational so reflect show deep learning models identify structural features of characters, improving downstream translation accuracy. This step ensures that even manuscripts with faded ink or irregular glyph shapes are properly digitized for semantic processing.

ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1208.03

7. CONCLUSION AND FUTUREWORK

This research presents a novel deep learning frame work for ancient Indian text modernization, integrating CNN-based OCR, LSTM-based grammar modeling, and Transformer-based contextual translation. The system significantly improves recognition accuracy and semantic fidelity, bridging the gap between classical and modern Indian linguistic domains.
The approach not only aids digital preservation but also supports educational tools, cultural heritage archiving, and AI-driven language learning applications.

Future Work:

1. Expansion of datasets through crowd sourced manuscript transcription.
2. In corporation of GAN-based restoration for highly degraded scripts.
3. Integration of speech synthesis to recite modernized verses for auditory learning.
4. Development of an interactive multilingual platform enabling comparative study of Sanskrit,Tamil,and Pali literature.
5. Deployment of the model on cloud-based micro services for scalable public access.

This work contributes meaningfully to the intersection of AI for cultural preservation and deep learning-based linguistic modernization, offering a roadmap for future research in computational heritage.

REFERENCES

1. Sharma,A.,“Digital Preservation of Sanskrit Manuscripts,” IJNLP,2020.
2. Reddy,K.,“AI-based Indic Language Translation Framework, ” IEEE Access,2021.
3. Vaswani,A.et al.,“Attention is All You Need,”NIPS,2017.
4. Smith,T.,“CNN for Script Recognition in Degraded Manuscripts,”ACMTALLIP,2022.
5. Gupta,S.,“Deep Learning for OCR of Ancient Indic Texts,”Elsevier JVCIR,2021.
6. Jain,R.,“Semantic Modernization of Sanskrit Texts Using DL, ”IEEE Trans.on NLP,2022.
7. Patel,D.,“Indic OCR Challenges,” SpringerLNCS,2020.
8. Banerjee,M.,“Convolutional Methods for Devanagari OCR,”IJDAR,2019.
9. Bhattacharya,P.,“RNN-Based Sanskrit Tokenizer,”ACL Anthology,2021.
10. Vaswani,A.etal.,“Attention is All You Need,”NIPS,2017.
11. Kakwani,D.,“IndicBERT and Indic Corp,”AI4 Bharat Report,2021.
12. Murthy,A.,“Cross-lingual Embeddings for Sanskrit-Hindi Translation,”Elsevier,2022.
13. Rao,A.,“GAN-based Document Restoration,”Springer Pattern Recognition Letters,2023.
14. AI4 Bharat,“Indic NLP Toolkit,”GitHub Repository,2021.
15. IIT Kharagpur NLPLab,“Sanskrit Heritage Reader,”2020.
16. Singh,P.,“Neural MT for Classical Languages,” IEEEICMLA,2020.
17. Mishra,N.,“Transfer Learning for Sanskrit Simplification,”WileyTrans.AI,2021.
18. Deshmukh,R.,“Challenges in Ancient Text Modernization,”IJCA,2019.
19. Digital Corpus of Sanskrit (DCS),2020.
20. BLEU and ROUGE Metrics Documentation,GoogleAI,2021.
21. Suresh,V.,“Transformer Performance for Indic Languages,”IEEENLP Conf.,2023.
22. Rao,R.,“Human Evaluation of Sanskrit Translations,”IJCLT,2022.
23. Tamil Heritage Foundation,“Digital Tamil Manuscripts Archive,”2021.
24. Kulkarni,S.,“AIin Cultural Heritage,” Elsevier Expert Systems,2023.
25. Mehta,P.,“GAN-Assisted Script Restoration, ”IEEE Access,2024.