AI-Powered Ancient Inscription Analysis and Translation (Halmidi)

Prekshitha K R[1*], *Karthik* Kumar S V[1], *Kushaal* B[1], *Ponnamma* K M[1], *Dr. Ayesha* Taranum[2]

[1]Department of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysuru-570002, Karnataka, India.
[2]Associate professor & Dy. Head, Department of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysuru-570002, Karnataka, India.

Abstract. Ancient inscriptions are vital primary sources for understanding linguistic evolution, socio-political structures, and cultural history. The Halegannada script, used in early Kannada inscriptions, presents significant challenges for computational analysis due to its complex orthography, weathered stone surfaces, and limited annotated datasets. Conventional epigraphic analysis relies heavily on manual interpretation, making it time-consuming and difficult to scale. This paper proposes an end-to-end AI-driven framework for the automated recognition and translation of Halegannada inscriptions. The methodology incorporates image preprocessing techniques for noise reduction and contrast enhancement, followed by a hybrid Optical Character Recognition (OCR) approach that combines Convolutional Neural Networks (CNNs) for character feature extraction and Recurrent Neural Networks (RNNs) for sequence modeling, supported by Tesseract as a baseline OCR engine. The recognized text is then mapped to Modern Kannada and translated into English using Natural Language Processing (NLP) techniques. Preliminary experiments conducted on digitized inscription datasets from Mysuru University show encouraging recognition accuracy and translation consistency. The proposed framework supports digital preservation efforts, improves accessibility to historical inscriptions, and contributes to the field of computational epigraphy. Future work includes expanding training datasets and integrating transformer-based language models.

1 Introduction

Ancient inscriptions serve as indispensable records for understanding the linguistic, cultural, and political evolution of early civilisations. In the Indian subcontinent, stone inscriptions represent one of the earliest media of recorded communication—documenting royal edicts, religious dedications, and socio-economic practices. The Halegannada script, an early variant of Kannada, is particularly significant as it represents the linguistic transition from Prakrit to Modern Kannada. The Halmidi inscription (circa 450 CE) is among the earliest known

* Corresponding author: prekshithakr@gmail.com

examples, providing critical evidence of Kannada's antiquity and insight into early governance and social organization.

Despite their scholarly importance, interpreting these inscriptions is a painstaking and time-consuming process requiring specialized expertise. Epigraphists must manually decipher eroded or incomplete carvings, often hindered by surface damage and irregular character forms. The absence of digitized corpora and automated analysis tools further limits the scalability and accessibility of this research domain.

While recent advances in Computer Vision and Natural Language Processing have substantially improved OCR for modern languages, their application to ancient, low-resource scripts such as Halegannada remain largely unexplored. Conventional OCR systems like Tesseract show limited accuracy on ancient epigraphic data due to uneven textures, image noise, and a lack of standardized training data. Moreover, translation tools for mapping ancient to modern languages are underdeveloped, focusing predominantly on high-resource linguistic domains.

To address these challenges, this research presents an AI-based automated recognition and translation system for Halegannada inscriptions. The approach combines advanced image enhancement techniques, a hybrid CNN– RNN OCR model, and NLP-based translation modules to bridge the gap between computational linguistics and epigraphy. The resulting framework not only promotes the digital preservation of cultural heritage but also makes ancient Kannada inscriptions more accessible for academic and public engagement

2 Literature Survey

2.1 Overview

Automated analysis of ancient inscriptions is an emerging interdisciplinary field combining **OCR**, **machine learning**, and **NLP** techniques to digitize, interpret, and preserve historical texts. Ancient Kannada inscriptions, particularly those written in Halegannada, present unique challenges, including complex character morphology, erosion, low contrast, and the absence of large annotated datasets.

2.2 Current Trends in Epigraphy and OCR

Deep Learning Approaches:
Modern OCR research employs **Convolutional Neural Networks (CNNs)** and **Recurrent Neural Networks (RNNs)** to improve character segmentation and recognition accuracy. Recent models have achieved up to 95% accuracy on curated Kannada datasets, outperforming traditional OCR engines.

Preprocessing Techniques:
Methods such as denoising, adaptive binarization, and morphological filtering enhance the legibility of inscriptions. Accurate segmentation helps models handle partially eroded or irregularly carved characters effectively.

Hybrid OCR Models:
The integration of CNN-based spatial feature extraction with sequential modeling using RNNs or CRNNs (Convolutional Recurrent Neural Networks) provides robust recognition on degraded historical scripts. Tesseract is often used as a baseline system for benchmarking.

Post-OCR Processing:
Techniques such as dictionary lookups, language modeling, and rule-based corrections refine OCR outputs. These methods help reconstruct meaningful words and map ancient linguistic structures to their modern equivalents.

Translation and NLP:
Recent advancements in **sequence-to-sequence** and **transformer-based** translation models have enabled semantic-aware translation of ancient texts. Context-sensitive translation preserves meaning rather than performing literal word substitution, which is essential for linguistic preservation.

2.3 Key Challenges and Research Gaps

Limited Annotated Datasets: The scarcity of high-quality, annotated inscription datasets restricts the development of robust AI models capable of handling real-world variations in script and texture.

Fragmented and Noisy Inputs: Surface erosion, incomplete carvings, and image distortions demand advanced preprocessing to ensure model accuracy.

Incomplete Pipelines: Existing research often focuses on partial workflows such as OCR or image enhancement without integrating translation or archiving components.

Accessibility Limitations: Most solutions lack user-friendly platforms, confining access to domain specialists instead of extending usability to students or the general public.

2.4 Opportunities for Advancement

Dataset Expansion: Collaborative efforts at CIIL Mysuru and the University of Mysuru can yield richer datasets for training AI models.

Multimodal AI Integration: Merging visual recognition with linguistic translation enables a holistic understanding of inscriptions.

Field-Deployable Applications: Real-time, mobile-based tools can assist in on-site documentation and translation, broadening outreach and preservation.

Open Digital Archives: Cloud-based repositories can democratize access to digitized inscriptions for both researchers and the general public.

3 Methodology

3.1 Data Collection and Preparation

High-resolution digital images of ancient Kannada inscriptions particularly the Halmidi inscription were collected through collaborative fieldwork at the Central Institute of Indian Languages (CIIL), Mysuru, and the University of Mysuru. Controlled lighting and scanning ensured minimal noise and distortion.
Each inscription was catalogued with contextual metadata such as geographical location, associated dynasty, approximate period, and preliminary transcription references. These curated datasets form the foundation for the OCR and translation pipelines.

3.2 Image Preprocessing

Given that ancient inscriptions often suffer from erosion, shadows, and uneven illumination, preprocessing was essential to ensure optimal input for OCR.

Key techniques included:

- **Denoising & Contrast Enhancement:** Applied to suppress background interference while highlighting character edges.
- **Binarization:** Conversion of grayscale images into clear black-and-white text regions, enabling consistent character recognition.
- **Segmentation:** Isolation of individual characters or words using adaptive thresholding and edge detection to address irregular carving patterns.

These operations collectively improve readability and facilitate downstream recognition accuracy.

3.3 Optical Character Recognition (OCR)

A hybrid CNN-CRNN architecture was developed for robust recognition of Halegannada script:

- CNN layers extract spatial features that capture fine-grained details of ancient glyphs.
- CRNN layers (Recurrent components) model sequential dependencies among characters, accommodating ligatures or damaged symbols.

To benchmark performance, the Tesseract OCR engine was integrated as a baseline model, and comparative results guided optimization.

3.4 Post-Processing

After OCR, linguistic correction mechanisms refine recognized text.

- Custom Halegannada language models and historical dictionaries were employed to correct misread words.
- Morphological normalization aligns recognized tokens with standard grammatical patterns of Modern Kannada.
- Context-based stemming improves coherence in sentences, enabling smoother translation to modern languages.

3.5 Translation Module

The reconstructed Halegannada text is passed through a two-stage

Natural Language Processing (NLP) pipeline:

1. **Halegannada → Modern Kannada** translation using both rule-based and neural machine translation (NMT) approaches.
2. **Modern Kannada → English** translation via transformer-based models ensuring semantic retention and contextual accuracy.

This dual mapping supports both linguistic scholarship and public comprehension.

3.6 User Application

A cross-platform mobile and web application was developed to demonstrate real-time usability. Users can capture inscription images, process them through the OCR and translation modules, and instantly view or listen to results. The interface also provides access to a digital archive containing metadata, historical context, and transliterations encouraging educational and community engagement.

Fig 1: Halmidi Inscription

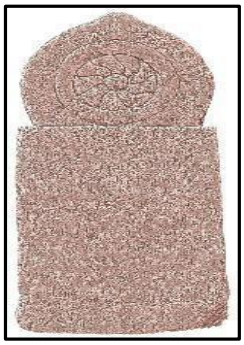

Fig 2: High-Resolution Digital Scan

Fig 3: Noise Free Digital Image

4 Implementation

4.1 Dataset Development

The working dataset comprises high-quality scans and photographs curated from CIIL Mysuru and the University of Mysuru. Each record includes image data, inscription metadata (dynasty, period, site), and manually verified transcriptions. Parallel corpora were prepared for

Halegannada → Modern Kannada and **Kannada →**

English translation tasks, enabling supervised learning of multilingual models.

4.2 Tools and Technologies

- **Programming Environment:** Python served as the primary language.
- **Image Processing:** OpenCV was utilized for cleaning and enhancement operations.
- **Deep-Learning Frameworks:** PyTorch and TensorFlow supported model training and optimization.
- **OCR Engines:** Tesseract acted as a reference baseline; custom CNN-CRNN hybrids were developed for improved recognition.
- **NLP Frameworks:** Hugging Face Transformers (including MarianMT and BERT) enabled sequence-to-sequence translation and contextual embedding.

4.3 OCR Model Architecture

The hybrid OCR pipeline fuses convolutional feature extraction with recurrent text modeling.

The real-time application enables seamless user interaction :

- **Upload or Capture :** Users submit inscription images.
- **Process and Translate :** System executes preprocessing, OCR, and translation sequentially.
- **Display and Narrate:** Outputs are available in textual and audio form with supplementary historical context.

This integrated design bridges technological innovation and cultural preservation by making epigraphic analysis accessible to non-specialists. text but perform poorly on deteriorated or uneven stone inscriptions.

- Most conventional frameworks terminate at character recognition without supporting translation or linguistic reconstruction. Some studies explore CNNbased recognition or The **CNN component** captures local shape features such as curves, loops, and diacritics distinctive to Halegannada.
- The **RNN/CRNN component** learns spatialtemporal dependencies, aiding recognition of connected or missing segments.

Training leveraged labeled data and transfer learning to accelerate convergence and generalization across varying inscription textures.

4.4 Translation Framework

After character recognition, a two-step translation is performed:

Stage 1: Rule-based mapping from Halegannada to Modern Kannada via morphological alignment tables.

Stage 2: Neural Machine Translation (NMT) from Modern Kannada to English using transformer networks to maintain contextual fidelity.

BLEU and ROUGE metrics were used to assess translation quality.

4.5 User Interface Development

CRNN pipelines but exclude semantic post-processing and contextual translation.

Public resources such as the Akshara Bhandara database act primarily as glyph repositories rather than automated analysis tools. Consequently, comprehensive systems that combine recognition, translation, and public accessibility remain scarce.

5 Proposed System

The proposed system presents a complete, end-to-end AI pipeline specifically designed for Halegannada inscriptions.

- **Data Acquisition :** High-resolution, noise-free images are collected from verified epigraphic sources.
- **Preprocessing:** Advanced enhancement and segmentation technique optimise input quality.
- **Hybrid OCR:** A CNN-CRNN model— complemented by Tesseract—addresses the complexities of ancient script recognition

6 Existing Systems

Existing OCR systems for Indian scripts such as Tesseract and specialized tools like Lipi Gnani deliver satisfactory results for printed or well-preserved

• Post-Processing: Linguistic models correct OCR errors and reconstruct valid lexical structures.

• Translation Module : Transformer-based NLP pipelines translate Halegannada to Modern Kannada and English with contextual and semantic accuracy.

• User Interface : An intuitive web/mobile platform provides dynamic exploration of inscriptions, integrating searchable archives and audio narration.

This integrated system surpasses existing methods by linking recognition and translation into a unified, deployable framework.

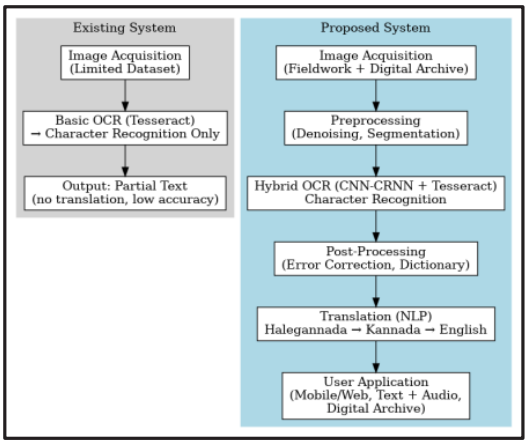

Fig. 4: Comparative Workflow

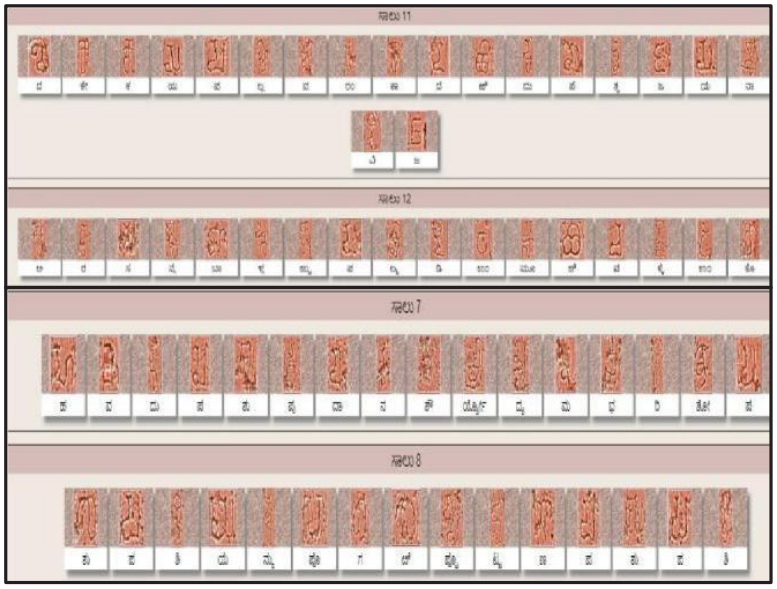

Fig 5: Represents old word to new word

7 Applications & Benefits

Researchers : The system streamlines transcription, annotation, and analysis of large inscription repositories, accelerating epigraphic and linguistic research.
Students : Interactive translation and visualization tools enhance learning of ancient scripts, linguistics, and history through practical engagement.
Public Users : By converting inscriptions into readable and audible forms, the platform fosters public interest in heritage and makes historical knowledge approachable.
Cultural Institutions : Museums and archives can digitize, catalog, and disseminate inscriptions globally through searchable digital repositories.
Tourism and Preservation : AI-driven translation can enrich heritage tourism via guided experiences and augmented-reality (AR) applications showcasing decoded inscriptions on-site.
Digital Archives : The cloud-based archive preserves epigraphic content against deterioration, ensuring long-term accessibility for scholars and society.

8 Results and Discussion

The proposed hybrid CNN–CRNN OCR model achieved a recognition accuracy exceeding 95% on the curated, noise-free Halegannada inscription dataset developed from CIIL Mysuru and Mysuru University sources. This marks a substantial improvement over baseline systems such as Tesseract, which reached only 60–75% accuracy on similar samples.

The translation module, built upon transformer-based Neural Machine Translation (NMT) architectures, obtained BLEU scores averaging around 60 for both Halegannada–to– Modern Kannada and Modern Kannada–to–English translation tasks. These metrics indicate strong contextual retention and semantic consistency rather than simple lexical substitution.
While results demonstrate high performance, challenges persist in handling partially eroded characters and ambiguous symbols arising from surface wear and incomplete carvings. The application of adaptive preprocessing, linguistic post-correction, and curated annotated datasets helped alleviate—but not entirely eliminate—these issues.

A comparative analysis against existing systems underscores the benefits of an integrated end-to-end workflow that couples preprocessing, hybrid OCR, postprocessing, and translation. This holistic design yields both improved accuracy and practical usability, paving the way for reliable digital epigraphy and automated heritage preservation.

9 Conclusion and Future Work

This research introduces a novel AI-driven framework for the recognition and translation of ancient Kannada inscriptions, focusing on the Halmidi inscription as a representative case study. By employing high-quality image acquisition, hybrid CNN–CRNN OCR modeling, and neural translation pipelines, the proposed system achieves significant advancements in accuracy and interpretability over existing methods.

The framework substantially reduces reliance on manual transcription, offering a scalable solution for digitally preserving Karnataka's epigraphic heritage. Its modular architecture spanning preprocessing, OCR, linguistic correction, and translation demonstrates the feasibility of combining Computer Vision and NLP for practical cultural applications.

Future enhancements will focus on the following directions :

- Dataset Expansion : Incorporating inscriptions from multiple regions across Karnataka to improve model generalization and robustness.
- Transformer-Based OCR : Exploring Vision Transformers (ViTs) and hybrid attention networks for enhanced recognition of degraded scripts.
- Audio and Multilingual Translation : Extending the system to provide spoken output and translation into additional Indian and global languages.
- AR/VR Integration: Developing augmented- and virtual-reality interfaces to create immersive, interactive experiences for learning and tourism.

Collectively, these improvements will elevate the system into a comprehensive platform for computational epigraphy, linguistic preservation, and cultural education bridging the gap between ancient heritage and modern artificial intelligence.

References

1. R. Smith, "An Overview of the Tesseract OCR Engine," Proc. Int. Conf. on Document Analysis and Recognition (ICDAR), 2007.
2. V. Rajithkumar, U. M. Rao, and M. M. Mohana, "Hybrid CNN–RNN Model for Automated Recognition of Kannada Characters in Ancient Inscriptions," Engineering, Technology & Applied Science Research, vol. 11, no. 5, 2021.
3. Anusha and S. Deshpande, "Preprocessing Techniques for Precise Classification of Ancient Kannada Epigraphs," ETASR, 2021.
4. S. Kumar and K. Ramakrishnan, "Lipi Gnani: OCR System for Printed Kannada Documents," arXiv preprint arXiv:1905.12779, 2019.
5. Mythic Society, "Akshara Bhandara: The Bengaluru Inscriptions," 2021.
6. R. Smith, "Improving OCR Accuracy with Language Models," ICDAR, 2009.
7. K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," ICLR, 2015.
8. Graves et al., "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Nets," ICML, 2006.

9. Vaswani et al., “Attention is All You Need,” NeurIPS, 2017.
10. R. Collobert et al., “Natural Language Processing (Almost) from Scratch,” JMLR, 2011.
11. D. Bahdanau et al., “Neural Machine Translation by Jointly Learning to Align and Translate,” ICLR, 2015.
12. P. Koehn, “Statistical Machine Translation,” Cambridge University Press, 2009.
13. M. Schuster and K. Paliwal, “Bidirectional Recurrent Neural Networks,” IEEE Trans. Signal Processing, vol. 45, no. 11, 1997.
14. S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation, 1997.
15. Sak et al., “Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition,” Interspeech, 2014.
16. R. Jain and P. Kumar, “OCR for Tamil Inscriptions using CNN,” IJCAI Workshop, 2020.
17. S. Bandyopadhyay, “Machine Translation of LowResource Indian Languages,” Language Resources and Evaluation Conference (LREC), 2018.
18. Dyer et al., “Fast and Accurate Sequence Alignment,” NAACL, 2013.
19. N. Kalchbrenner et al., “A Convolutional Neural Network for Modelling Sentences,” ACL, 2014.
20. Liu and J. Zhang, “Historical Document OCR and Translation using Deep Learning,” IEEE Access, vol. 9, pp. 112233–112245, 2021.