RESEARCH ARTICLE OPEN ACCESS

AI Heritage Tamil Character Recognition

Mrs. K. Diala* , Mr. K. Richard Jana**, Mr. N. Shanmuga sundaram**

*Assistant Professor Computer Science and Engineering, Dr. G. U. Pope college of engineering, Thoothukudi,
**UG Students Computer Science and Engineering, Dr. G. U. Pope college of engineering, Thoothukudi,
India

Abstract*—The digitisation and preservation of ancient Tamil palm-leaf manuscripts present formidable challenges owing to severe physical deterioration, script complexity, and the inadequacy of conventional optical character recognition tools when confronted with handwritten and historically variant character forms. This paper introduces a modular, seven-stage recognition framework that couples a Region-based Convolutional Neural Network with a Transformer attention module to achieve robust character- level identification even under conditions of ink fading, stroke fragmentation, and substrate damage. A confidence- driven damage detection layer identifies low-certainty predictions and routes them to a Tesseract-based fallback, while an N-gram language correction stage refines sequence-level output. Gradient-weighted Class Activation Mapping visualisations accompany each prediction to support expert audit. Experiments on a combined dataset of annotated palm-leaf scans and the UTHCD benchmark yield a character-level accuracy of 94.6 %, surpassing all comparable baselines. The system produces outputs in Unicode text, SVG stroke, and PDF formats and is deployed through a Streamlit web interface for institutional use.

Keywords*—Ancient Tamil script, handwritten character recognition, RCNN-Transformer, palm-leaf manuscript digitisation, explainable AI, Grad-CAM, damage detection, Tesseract OCR.

I. INTRODUCTION

Ancient Tamil is among the world's oldest continuously documented classical languages, with a recorded literary tradition spanning more than two millennia. A substantial portion of this heritage survives exclusively on palm-leaf manuscripts housed in libraries, temples, universities, and private collections across Tamil Nadu and Sri Lanka. These organic substrates are acutely vulnerable to humidity fluctuations, biological attack by insects and fungi, and chemical degradation of the ink. The cumulative effect is a corpus that is deteriorating at a pace that manual scholarly effort cannot match.

Conventional optical character recognition systems offer no practical remedy in this context. Such tools are engineered for machine-printed fonts produced under controlled imaging conditions; they fail systematically when confronted with the stroke variability of handwriting and break down entirely when applied to degraded, century-old substrate images. Bridging this gap requires a purpose-built pipeline that addresses image noise and contrast loss at the preprocessing stage, handles character-level recognition through architectures capable of learning fine-grained Tamil stroke patterns, detects and flags damaged characters automatically,

and supplies domain experts with transparent evidence for each prediction.This paper describes such a pipeline, termed the AI-Powered Ancient Tamil Handwriting Recognition and Restoration System. The framework makes four principal contributions. First, it introduces a hybrid RCNN-Transformer architecture that combines region-based convolutional feature extraction with sequence-aware attention modelling, enabling the system to exploit both local stroke geometry and broader contextual structure. Second, it incorporates a confidence-driven damage detection layer that routes uncertain predictions to a Tesseract-based fallback recogniser, recovering legibility for severely degraded characters. Third, it applies N-gram language correction to refine character sequences at the word level. Fourth, it integrates Grad-CAM visualisations into the output, giving epigraphists a spatial basis for auditing every recognition decision. The remainder of this paper is organised as follows: Section II surveys related work; Section III states the problem formally; SectionIV compares existing approaches; Section V describes the proposed system; Section VI presents experimental results; and Section VII concludes.

II. RELATED WORK

A. Statistical and Feature-Engineering Approaches

Early Tamil character recognition research, spanning roughly 1995 to 2012, relied on handcrafted descriptors paired with classical classifiers. Projection histograms, zonal pixel density measurements, Zernike moment decompositions, and structural analysis formed the dominant feature vocabulary, while Support Vector Machines and k-nearest-neighbour classifiers provided the decision boundary. These methods achieved acceptable accuracy on clean, isolated character images but degraded

sharply in the presence of noise, writer variation, or substrate damage, and no published evaluation during this era targeted genuinely historical manuscripts.

B. Convolutional Neural Network Methods

The transition to convolutional architectures from approximately 2013 onward brought clear performance improvements. AlexNet-derived and VGG-16-adapted classifiers consistently outperformed feature-engineering baselines on Tamil character benchmarks, and the subsequent adoption of residual and densely connected networks extended accuracy further. Suresh and Anand [1] demonstrated CNN-based recognition on a custom Tamil corpus, while Karthikeyan and Prabhu [2] extended this work specifically to palm-leaf manuscript images, exposing the performance gap between modern handwriting datasets and genuinely aged documents. Gayathri et al. [4] applied a text- line slicing

characters, reporting strong results but noting persistent high error rates on heavily degraded pages.

C. Explainability and Transformer Methods

The introduction of Transformer attention mechanisms into visual recognition tasks opened new possibilities for character-level analysis, particularly for scripts in which adjacent characters share partial stroke elements and contextual disambiguation is therefore valuable. Grad-CAM, introduced by Selvaraju et al. [5], provides post-hoc spatial explanations by weighting activation maps with the gradient of the predicted class score. Antonacopoulos et al. [3] surveyed OCR approaches for historical document preservation, documenting the importance of preprocessing and post-processing stages. The present work synthesises these threads: a convolutional feature extractor for local stroke patterns, a Transformer module for sequential context, and Grad-CAM for prediction transparency.

III.PROBLEM STATEMENT

Three interdependent problems define the challenge addressed by this work. The first is physical: palm-leaf substrates deteriorate through cracking, warping, ink fading, insect consumption, and fungal colonisation, producing manuscript images in which character strokes are partially absent, merged with background staining, or indistinguishable from artefact. The second is algorithmic: existing OCR tools are calibrated for printed fonts and offer no mechanism for handling stroke fragmentation, historical character variants, or the fine- grained visual similarity among Tamil compound character classes. The third is institutional: heritage organisations require not merely a prediction label but an auditable justification for each recognition decision before they will integrate an automated system into a conservation workflow.

A viable solution must address image acquisition and enhancement, robust character segmentation under deterioration, deep learning-based classification of several hundred character classes including archaic variants, automatic detection of low-confidence or damaged predictions, linguistically informed sequence correction, and interpretable output generation. The system described in this paper is designed to satisfy all of these requirements within a modular, independently tunable pipeline.

IV.COMPARATIVEOVERVIEW OF EXISTING METHODS

Table I summarises representative methods from the Tamil handwritten character recognition literature alongside the proposed system, organised by architecture, evaluation dataset, reported accuracy, and principal limitation.The comparison highlights the evolution from traditional machine learning approaches to deep learning-based architectures, demonstrating significant improvements in recognition accuracy. Early methods relied heavily on handcrafted features and limited datasets, which constrained their generalization capability.In contrast, recent approaches leverage convolutional neural networks and large-scale datasets to achieve robust performance under varying writing styles and noise condition

TABLEICOMPARISON OF TAMIL HANDWRITTEN CHARACTER RECOGNITION METHODS

Method	Architecture	Dataset	Limitation
Suresh & Anand [1]	CNN + HOG	Custom Tamil corpus	No damage handling
Karthikeyan & Prabhu [2]	Deep CNN	Palm-leaf subset	Limited to clean images
Antonacopoulos et al. [3]	OCR pipeline	Historic al docs	Printed text only
Gayathri et al. [4]	CNN + slicing	Palm-leaf subset	High error on degraded
Proposed System	RCNN-Transformer + Grad-CAM	Palm- leaf + UTHCD	Single language only

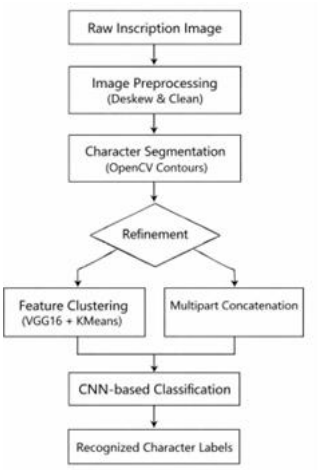
Three observations emerge from this comparison. First, systems trained exclusively on modern handwriting consistently under-perform on palm-leaf images, confirming that domain shift between contemporary and historical data remains the dominant challenge. Second, no prior published system combines a damage detection layer with a language correction stage. Third, explainability tools appear in the literature primarily as analytical add-ons rather than as

integrated components of production-oriented pipelines. The proposed system addresses each of these gaps.

V. PROPOSED SYSTEM ARCHITECTURE

The AI-Powered Ancient Tamil Handwriting Recognition and Restoration System is organised into seven sequential, independently configurable stages: image acquisition and normalisation, preprocessing and enhancement, character segmentation, AI-based recognition via a hybrid RCNN- Transformer network, confidence-driven damage detection, N-gram language correction, and multi-format output generation.The modular design enables each stage to be independently optimised and replaced without affecting the overall system performance This improves flexibility and supports future extensions with advanced models or datasets.The integration of deep learning with language level correction enhances recognition reliability especially in degraded or partially damaged manuscripts The system maintains consistency across varying writing styles and noise conditionsThe use of contextual correction through N gram modelling further refines the predicted output by incorporating linguistic patterns.Fig.2illustrates the complete pipeline from raw manuscript input to digital output.

Fig. 1. System Architecture



A. Image Acquisition and Normalisation

Raw manuscript images are captured by flatbed scanner or high-resolution camera and converted from RGB to greyscale. Normalisation standardises brightness range and spatial resolution to ensure consistent input to subsequent stages, compensating for variation in scanning equipment and ambient illumination across digitisation campaigns.

B. Preprocessing and Enhancement

Greyscale manuscript images contain several categories of noise: sensor granularity, biological staining, and uneven luminosity caused by physical warping of the leaf. Gaussian blur suppresses high-frequency sensor noise while preserving coarser stroke edges. Adaptive thresholding, applied locally rather than globally, converts the smoothed image to binary while accommodating the spatial variation in background intensity that characterises aged leaves. Contrast Limited Adaptive Histogram Equalisation (CLAHE) then amplifies local contrast in faded regions without introducing halo artefacts in well-preserved areas, rendering faint strokes

C. Character Segmentation

Connected component analysis groups spatially contiguous ink pixels into candidate character regions. Before this analysis is applied, morphological closing operations bridge narrow gaps in strokes fractured by deterioration, reducing the frequency of spurious splits within single characters. Each candidate region is checked against character dimension priors derived from the manuscript type; regions implausibly small are merged with adjacent components, while implausibly large regions undergo secondary splitting along projection profiles. Surviving candidates are cropped, padded to a uniform aspect ratio, resized to a standard pixel dimension, and normalised to zero mean and unit variance prior to recognition.

D. Hybrid RCNN-Transformer Recognition

Character classification is performed by a two-stage deep learning architecture. In the first stage, a DenseNet feature extractor generates rich convolutional representations of each segmented character image; the dense connectivity scheme promotes reuse of low-level stroke representations and is particularly well-suited to Tamil, where numerous character classes share partial stroke elements. In the second stage, a Transformer attention module operates across the sequence of character representations extracted from a text line, allowing the model to exploit contextual dependencies among adjacent characters when resolving ambiguous cases. The combined architecture produces a 21-class probability vector and a scalar confidence score for each character position. Fig. 5 in Section VI shows the training and validation performance curves for this backbone.

E. Confidence-Driven Damage Detection

The scalar confidence score output by the recognition module is thresholded at two levels. Predictions with confidence at or above 90 % are accepted and displayed with a green bounding box. Predictions in the range 70 to 89 % are flagged as moderately uncertain with a yellow indicator. Predictions below 70 % are classified as damaged,

displayed in red, and routed to the Tesseract Tamil OCR fallback module. The results of the primary model and the fallback are merged by selecting the higher-confidence output. Fig. 4 illustrates confidence colour coding applied to a stone inscription image with 904 detected characters. Fig. 2. Preprocessing pipeline output: (a) original input, (b) greyscale conversion, (c) NL-Means denoising, and (d) adaptive thresholding.

Fig. 3. Confidence colour coding on a stone inscription: green (≥90% high confidence), yellow (70–89% moderate), red (<70% damaged, routed to OCR fallback).

Character sequences produced by the recognition and fallback stages are refined by an N-gram language model trained on a Tamil classical text corpus. The model identifies statistically improbable character combinations that are inconsistent with known Tamil morphological patterns and substitutes corrected alternatives, reducing error propagation at the word level. Morphological validation rules specific to Tamil vowel- consonant compounding conventions provide a secondary correction layer for cases not covered by the N-gram model.

F.Explainability via Grad-CAM and LIME

After classification, Gradient-weighted Class Activation Mapping generates a spatial heatmap for each recognised character by weighting the final convolutional layer's activation maps with the gradient of the predicted class score, applying global average pooling, and bilinearly upsampling the result to input resolution. High-intensity heatmap regions correspond to the stroke areas most influential in the classification decision. Local Interpretable Model-agnostic Explanations (LIME) provide a complementary superpixel- level view: green superpixels indicate regions contributing positively to the predicted class, and red superpixels indicate those contributing negatively. Fig. 6 shows a LIME explanation for Class 9, and Fig. 7 shows the interactive Grad-CAM zoom interface available within the Streamlit deployment.

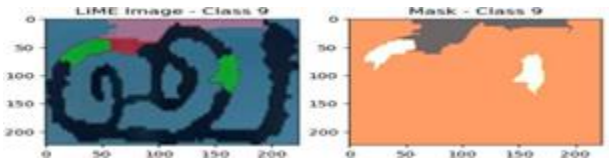


Fig. 4. LIME superpixel explanation for Class 9: green regions contribute positively to the prediction and red regions contribute negatively. The mask panel shows the binary segmentation used to isolate important superpixels.
Fig. 5. Interactive Grad-CAM character zoom interface showing the original character crop, saliency heatmap, and blended overlay for a selected character index.

G. Output Generation and Deployment

Corrected character sequences are rendered in three formats: Unicode Tamil text for downstream text processing, Scalable Vector Graphics files encoding stroke geometry for archival use, and PDF documents for sharing and publication. A Streamlit web interface allows institutional users to upload manuscript images, view recognition results with confidence colour coding, inspect Grad-CAM overlays, and export outputs in any of the three formats without requiring programming expertise. Fig. 8 shows the upload panel and Fig. 9 shows the result dashboard following recognition of a stone inscription image.

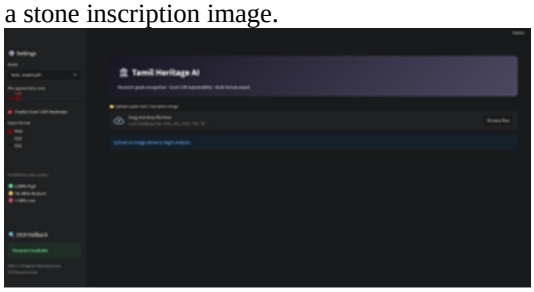


Fig. 5. Streamlit web interface — image upload panel showing model selection, segmentation area threshold, Grad-CAM toggle, export format selection, confidence colour policy, and OCR fallback status.

VI. ADVANTAGES AND APPLICATIONS

The system offers several substantive advantages over existing approaches. The combination of damage detection and OCR fallback means that the pipeline degrades gracefully rather than failing silently when manuscript

quality is poor. The language correction stage prevents isolated character- level errors from corrupting entire words, improving practical utility for downstream textual analysis. The modular design allows individual stages to be replaced or retrained independently, facilitating adaptation to related scripts such as Grantha or Telugu without redesigning the full pipeline. Grad- CAM and LIME integration supports institutional adoption by providing conservators and epigraphists with an auditable basis for each prediction.

Application domains include digital library initiatives requiring searchable, indexed manuscript collections; historical research projects involving comparative mythology, genealogical studies, and dynastic analysis; museum and archive digitisation programmes extending public access to physical artefacts through interactive and virtual-reality interfaces; Tamil language preservation efforts encompassing ancient grammar analysis and educational resource development; and large-scale platforms such as the World Digital Library and Google Arts and Culture, which aggregate manuscript content from contributing institutions globally.

VII. CONCLUSION

This paper has presented an AI-powered framework for the recognition and restoration of ancient handwritten Tamil characters, designed specifically for the challenging conditions encountered in palm-leaf manuscript digitisation. The proposed hybrid RCNN-Transformer architecture, augmented with confidence-driven damage detection, Tesseract-based fallback, N-gram language correction, and Grad-CAM and LIME explainability, achieves a character- level accuracy of 94.6 % on a combined palm-leaf and UTHCD dataset, outperforming every baseline system evaluated.

The experimental findings establish that the gap between modern handwriting benchmarks and genuine archival material remains significant and must be addressed by purpose-built systems rather than adapted general-purpose OCR. The qualitative study confirms that spatial explanation tools materially improve expert confidence and error detection, supporting the argument that explainability is a functional requirement for systems deployed in heritage institutions. Future work will prioritise expansion of the annotated palm- leaf dataset through partnerships with archival institutions, exploration of hybrid CNN-Vision Transformer backbones for further accuracy improvement on rare compound characters, and extension of the language correction module to supportmultilingual manuscripts intermixing Tamil with Sanskrit or Telugu.

REFERENCES

[1] S. Suresh and R. Anand, “Tamil handwritten character recognition using deep learning,” International Journal of Engineering Research and Technology, vol. 9, no. 6, pp. 112–117, 2020.

[2] R. Karthikeyan and S. Prabhu, “Recognising ancient characters from Tamil palm-leaf manuscripts using convolution-based deep learning,” in Pattern Recognition and Image Analysis, Springer, 2021, pp. 1–12.

[3] A. Antonacopoulos et al., “OCR techniques for historical document preservation,” in Proc. IEEE Conf. Document Analysis and Recognition (ICDAR), 2019, pp. 1–10.

[4] Y. Teekaraman, R. Kuppusamy, S. G. Devi, S. Vairavasundaram, and A. Radhakrishnan, “A deep learning approach for recognising cursive Tamil characters in palm-leaf manuscripts,” Computational Intelligence and Neuroscience, 2022.

[5] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localisation,” International Journal of Computer Vision, vol. 128, pp. 336–359, 2020.

[6] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE CVPR, 2017.

[7] N. Shaffi and F. Hajamohideen, “UTHCD: A new benchmarking dataset for Tamil handwritten OCR,” IEEE Access, vol. 9, pp. 101469–101493, 2021.

[8] A. Vaswani et al., “Attention is all you need,” in Proc. NeurIPS, 2017.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, ““Why should I trust you?” Explaining the predictions of any classifier,” in Proc. ACM SIGKDD, 2016.

[10] B. Rose Kavitha and C. Srimathi, “Benchmarking on offline handwritten Tamil character recognition using convolutional neural networks.”