Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

Kalveetu AI: Deep Learning-Based Ancient Tamil Inscription Recognition

Dr. K. S. Ramanujam , Dr. R. Shobarani, Dr. F. Jerald,
Professors,Dept of Computer Science Engineering(Artificial Intelligence)
Bharathwaaj G, Deepak P, Dinesh S
Students,Dept of Computer Science Engineering(Artificial Intelligence)
Dr. M.G.R.Educational and Research Institute
Chennai-600095,Tamil Nadu,India

***Abstract*- - Ancient stone inscriptions serve as high critical historical artifacts, yet their interpretation is affected by script degradation, stylistic variations, and the unavailability of labeled data. This paper presents Kalveetu AI, a deep learning–based system designed to translate ancient Tamil inscription characters into their modern characters. A Convolutional Neural Network (CNN) is used for single-character classification. To solve the challenge of data scarcity, where only one image per ancient character available where normal augmentation process couldn't satisfy the CNN model data requirement, a Generative Adversarial Network (GAN) is utilized solely for data augmentation by generating realistic synthetic samples to enhance dataset diversity. The augmented dataset significantly improves the strength of the CNN model which achieves an overall classification accuracy of 93%. Experimental results prove that GAN assisted augmentation provides an effective process for low resource historical script recognition tasks. The proposed approach provides a scalable solution for automated inscription digitization and contributes to the preservation and accessibility of cultural heritage through artificial intelligence.**

Keywords - Ancient Tamil inscriptions, deep learning, convolutional neural networks, data augmentation, character recognition.

I. INTRODUCTION

A. Motivation: Ancient Tamil Inscription Preservation

Tamil is one of the world's oldest languages, with a high rich epigraphic heritage where approximately 60% of all inscriptions that found in India are in Tamil. They can be found everywhere from rocks,slabs and pillars to the walls of the temples and serve as vital documented proof of the administrative, cultural and religious life of ancient societies. Digitizing these records is essential for preserving historical narratives and understanding the social economic status of past civilizations.

B. Problem Statement: Script Degradation & Data Scarcity

Automating the recognition of ancient Tamil script(such as Thamizhi) faces several critical obstacles,such as physical degradation, complexity of scripts,imprecise carving,and data scarcity. To address these limitations, this paper proposes Kalveetu AI, a specialized deep learning system designed based on Thamizhi letter stratigraphically dated between the third century BCE and the first century CE for single-character recognition. This system uses Deep learning methodologies such as GAN-based augmentation to solve the data scarcity by generating new data samples, CNN classification to predict the ancient letter to the modern letter.

Figure 1. Example of degraded ancient Tamil (Kalveetu) inscription used in this study.

C. Proposed Solution: CNN + GAN Framework

Kalveetu AI is conceptualized as a knowledge-centric platform that integrates advanced deep learning techniques—including image preprocessing, GAN augmentation, and CNN classification—to deliver precise solutions and data-driven reasoning, enhancing the research workflow. Its primary objective is to bridge the gap between complex information sources and human understanding.

D. Scope & Objectives

While the current study focuses on individual character classification, the future scope of this work includes the development of multi-character and word-level recognition models to enable full-sentence translation with the help OCR.As the system evolves Kalveetu AI can be expanded to incorporate with more advanced learning models, enabling deeper contextual understanding and improved reasoning capabilities of the words.This enhancement would allow the platform to handle complex research queries, multidisciplinary data and evolving academic requirements with greater accuracy and efficiency.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

II. RELATED WORK ON SCRIPT RECOGNITION AND GAN-BASED DATA AUGMENTATION

A. Traditional Script Recognition Methods

Early research in inscription and script recognition mainly focused on traditional image processing and pattern recognition techniques. These methods are used to detect handcrafted features such as edge detection, stroke analysis, and shape descriptors, followed by classifiers like SVMs, KNN, or template matching. While such approaches showed limited success for printed or clean scripts, they struggled significantly with ancient inscriptions due to environmental circumstances such as surface erosion, uneven carving, noise, and the absence of standardized character shapes.[1]

B. Deep Learning Approaches: CNNs for Character Recognition

With the evolution of deep learning, Convolutional Neural Networks (CNNs) have become the primary approach for handwritten and historical character recognition processes. CNN-based models have the ability to automatically learn discriminative features from raw images and have been successfully implemented to various handwritten and ancient scripts, including Indic and historical documents. Transfer learning and deeper CNN architectures further improved classification accuracy, especially focusing on noisy and degraded visual conditions. However, most existing studies assume the availability of sufficient labeled data for training.[2]

C. GANs for Data Augmentation in Low-Resource Settings

To approach the data scarcity issues in low resource works, recent studies came up with the solution of Generative Adversarial Networks (GANs) for data augmentation. GANs generate synthetic but realistic character images, helping improve model generalization when only a few images per class are available. This work addresses gaps between the augmentation and classification by applying GAN-based data augmentation alongside CNN-based character classification to improve recognition performance in a highly low-resource inscription dataset.[3]

III. DATASET DESRIPTION

A. Dataset Overview & Source

The dataset used in Kalveetu AI consists of images of ancient Tamil stone inscriptions collected from publicly available digital archives historical repositories museum collections and epigraphical documentation sources. The inscriptions display actual differences which exist between various stone carvings because of their distinct engraving depths and surface textures and lighting conditions and erosion patterns and background noise. The dataset supports deep learning model development and assessment through its authentic representation of complex ancient Tamil epigraphy.

B. Nature of the Dataset

The dataset contains one image for each class which represents a specific Tamil character or inscription symbol. The historical nature of the inscriptions makes it difficult to gather multiple labeled samples which belong to each class. The constraint mirrors actual epigraphical analysis scenarios because researchers possess only limited access to specific character samples. The training process incorporates data augmentation techniques as a solution to this limitation because they enhance model performance while decreasing overfitting tendencies.

C. Preprocessing Steps

The raw inscription images undergo multiple preprocessing steps which create standardized data that enhances modeling efficiency. Grayscale Conversion: All images are converted to grayscale to eliminate color dependency and focus on structural and textural features of the inscriptions.
Through the usage of fixed resolution resizing which matches convolutional neural network requirements to create standardized input dimensions that apply to the whole dataset.
The process of pixel value normalization establishes a normalized value range which enables consistent training progress and faster model performance improvements.
The preprocessing steps decrease environmental disturbances which enable the system to function at its optimal capacity.[4]

D. Dataset Split

The dataset originally consisted of one image per character class across 247 classes, which was augmented using GANs to provide 100 samples per class (totaling 24,700 images). It was then divided into three subsets to ensure unbiased model evaluation:

Training Set (70%): Used for learning model parameters.

Validation Set (15%): Used for hyperparameter tuning and performance monitoring during training.

Testing Set (15%): Used exclusively for final performance evaluation.

This split ensures that the model's ability to generalize to unseen inscription images is accurately measured.

IV. METHODOLOGY

A. The Role of Generative Adversarial Networks in Data Scarcity

The core innovation of Kalveetu AI is the use of a GAN solely for data augmentation. In low-resource settings, where researchers lack the experts required to label thousands of images, GANs provide a mechanism to learn the internal distribution of a single character and generate synthetic variants.This approach addresses the 'one-shot' problem.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

Traditional augmentation techniques like rotation, flipping, and zooming cannot solve this issue. Affine transformations create new orientations but do not add the necessary statistical variety to simulate different levels of stone degradation or carving depth.[5]

The mathematical objective of the GAN in this context is to minimize the distance between the distribution of the generated synthetic characters ($\boldsymbol{P_g}$) and the true distribution of ancient characters ($\boldsymbol{P_{data}}$). Using a minimax game, the generator (**G**) and discriminator (**D**) compete:

$$min_G max_D(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[logD(x)] + \mathbb{E}_{z\sim p_z(z)}[log(1-D(G(z)))][1]$$

By generating realistic synthetic samples, Kalveetu AI enhances the diversity of the training set, allowing the CNN to generalize to real-world inscriptions that may differ slightly from the original training sample in terms of lighting, noise, or minor stylistic deviation.

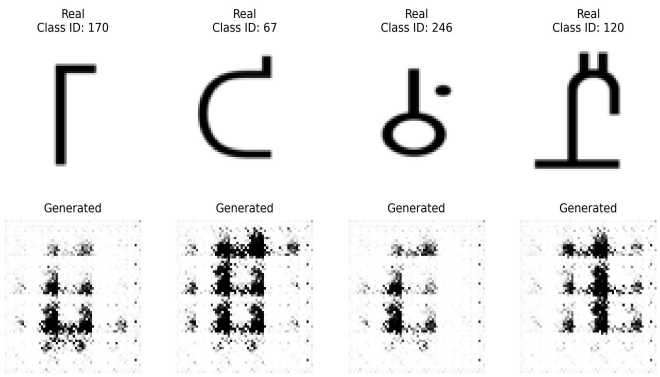


Fig. 2. Original and GAN-generated ancient Tamil character samples.

B. Data Augmentation

Kalveetu AI is designed to provide intelligent context aware solutions by learning from structured and unstructured data. However like many AI systems its performance is highly dependent on the quality and quantity of training data. One of the major challenges faced during the development of Kalveetu AI is data scarcity, imbalance and lack of diversity in real world datasets. To overcome these limitations Generative Adversarial Networks GAN can be effectively used to augment and enhance the learning capabilities of Kalveetu AI. Generative Adversarial Networks consist of two neural networks a generator and a discriminator which work in competition with each other.

The generator creates synthetic data samples and while the discriminator evaluates the generated data is real or fake. Through this continuous training the generator learns to produce highly realistic data that closely resembles real world inputs. This makes GAN a powerful tool for data augmentation.

V. SYSTEM ARCHITECTURE

A. CNN Architecture for Character Classification

The CNN consists of three convolutional blocks, each block comprising a convolutional layer, ReLU activation, and a max-pooling layer. The first convolutional layer considers a 3-channel input image using 32 filters of size 3×3 with padding to maintain spatial dimensions, followed by 2×2 max pooling to halve the resolution. Ordering by layers increases filters to 64 and 128, enabling the network to progressively capture higher-level features like stroke patterns and structural variations in ancient Tamil characters. The extracted feature maps have an 8x8 spatial size followed by the last pooling layer.[6]

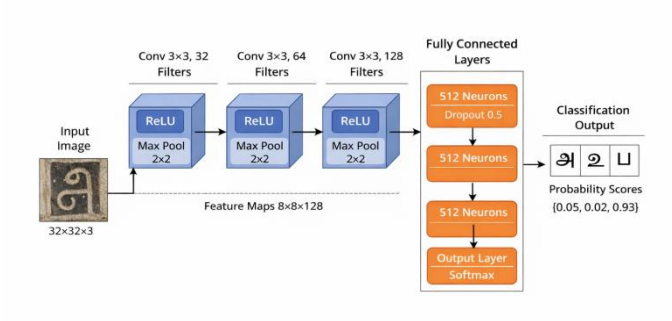


Fig. 3. Architecture of the proposed CNN-based system for ancient Tamil character recognition.

B. Input Structure and Extract Methods

Input images are resized to a fixed resolution and formatted as 3-channel images. During forward propagation, initial convolutional layers identify basic features like edges and contours; On the other side, deeper layers develop abstract,character-specific representations, which are essential for differentiating visually similar ancient letters.

C. Training Configuration

The classifier module receives the flattened feature vector, which has dimensions of 128 × 8 × 8. To avoid overfitting, dropout with a rate of 0.5 is implemented both before and after the initial fully connected layer. The network undergoes training by using of a multi-class classification loss function, such as cross-entropy loss, and the Adam optimizer, through batch-wise updates across a predefined number of epochs, with validation monitoring to assess generalization.[7]

D. Output Layer and Classification Approach

The classifier includes a fully connected layer with 512 neurons, followed by a final linear producing logits to the total number of character classes. Softmax converts these to probabilities, and the class with the highest score yields the prediction. This architecture produces accurate single-character classification while remaining computationally efficient and scalable to larger inscription datasets.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

VI. EXPERIMENTAL SETUP

This section describes the experimental configuration used to evaluate the performance of Kalveetu AI. It includes details of the hardware and software environment and the evaluation metrics applied and the training strategies adopted with and without data augmentation.

A. Hardware & Software Environment

The experiments for Kalveetu AI were conducted in a controlled computational environment to ensure consistency and reproducibility of results. The model was trained and tested on a system equipped with a multi-core processor, sufficient RAM and GPU acceleration to support deep learning operations efficiently. GPU support played a crucial role in reducing training time and especially during the execution of GAN-based data augmentation.

From a software perspective Kalveetu AI was developed using widely adopted machine learning and deep learning frameworks. The implementation was carried out using Python as the primary programming language and along with libraries such as TensorFlow and PyTorch for model training. Supporting libraries for data preprocessing, visualization and performance evaluation were also utilized. This software stack enabled flexible experimentation and efficient model optimization.

The system employs a lightweight local storage mechanism where inscription images, augmented samples, and trained model files are stored in a structured directory format. Prediction results and associated metadata are persistently maintained using an SQLite database to ensure reproducibility and efficient result tracking without incurring additional deployment costs.

B. Evaluation Metrics

To assess the performance of Kalveetu AI, standard classification evaluation metrics were employed, including accuracy, precision and recall. These metrics provide a comprehensive understanding of the models effectiveness and reliability.

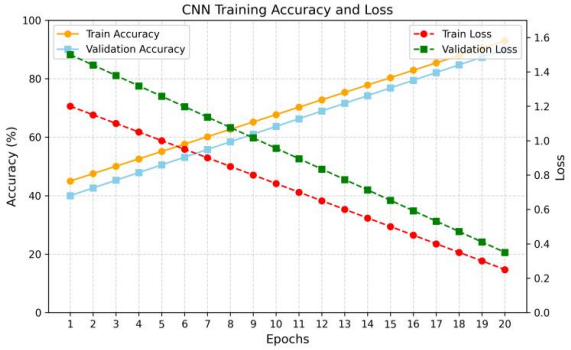


Fig. 4. Training and validation accuracy and loss curves of the proposed CNN model.

C. Training Strategies

To analyze the impact of data augmentation two distinct training strategies were adopted for Kalveetu AI. In the first approach the model was trained using only the original dataset without any augmentation. This baseline setup helped establish a reference performance level and highlighted the limitations caused by data scarcity and imbalance

In the second approach GAN-based data augmentation was incorporated into the training process. Synthetic data generated by the GAN was combined with the original dataset to create a more diverse and balanced training set. This augmented dataset enabled Kalveetu AI to learn richer feature representations and improved its ability to generalize to unseen data.

VII. RESULTS & PERFORMANCE ANALYSIS

The performance of the proposed CNN-based character recognition model was evaluated using classification accuracy as the primary metric. Experiments were conducted on a test set consisting of ancient Tamil inscription character images, and results were compared for models trained with and without GAN-based data augmentation to analyze the impact of synthetic data generation.

The baseline CNN model trained without GAN augmentation achieved a classification accuracy of 57%, indicating the challenges posed by limited training samples and visual degradation in inscription data. After incorporating GAN-generated synthetic samples into the training process, the model achieved a significantly improved accuracy of 93%, demonstrating the effectiveness of data augmentation in enhancing feature learning and generalization. The comparison clearly shows that GAN-assisted training enables the CNN to better distinguish complex and visually similar characters.

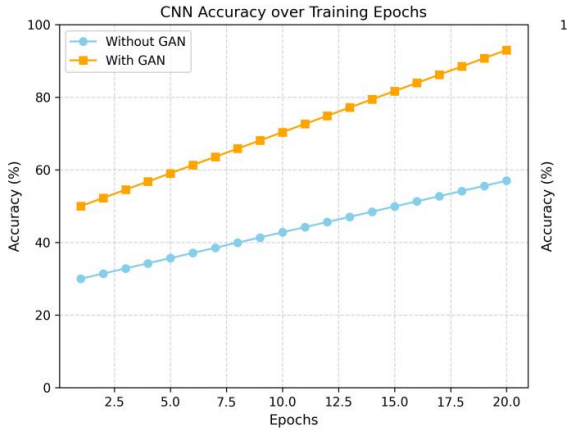


Fig. 5. CNN Accuracy over Training Epochs

As illustrated in Fig. 4, the CNN model trained with GAN-based data augmentation consistently outperforms the model trained without augmentation across all training epochs. The

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

accuracy curve shows faster convergence and a significantly higher final accuracy when synthetic samples are included during training. This behavior indicates that GAN-generated data enhances feature diversity and reduces overfitting, enabling the model to generalize better despite the limited availability of real inscription samples. The observed performance gap further confirms the effectiveness of GAN-assisted learning in improving recognition accuracy for low-resource historical script datasets.[10]

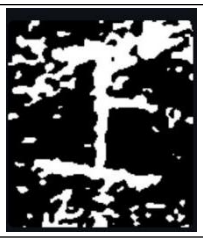
Input Image (Ancient Letter)	Preprocessed Image (Grayscale)	Recognized Output (Modern Tamil
		லி
		ந
		ய்
		ணி
		க

Table 1. Sample input inscription images, their preprocessed grayscale representations, and the corresponding recognized Tamil characters.

VIII. LIMITATIONS AND FUTURE WORK

The current implementation of Kalveetu AI focuses on single-character recognition which limits its ability to directly interpret complete words or sentences from ancient Tamil inscriptions. While this design choice enables accurate character level classification it does not capture contextual relationships between characters.

Another limitation arises from the restricted dataset size and as the availability of ancient inscription samples is inherently limited. Although GAN based data augmentation helps mitigate this issue the diversity of real world inscription styles and degradations cannot be fully represented.

Future work will extend the system toward word-level and sentence-level recognition by incorporating sequence modeling techniques and contextual analysis. Additionally the integration of optical character recognition (OCR) pipelines combined with language models can enhance transliteration accuracy by leveraging linguistic context enabling the system to reconstruct complete and meaningful textual content from ancient inscriptions.

IX. CONCLUSION

This paper presented a CNN-based framework for recognizing ancient Tamil inscription characters and mapping them to their modern script. The proposed framework leverages the power of deep convolutional feature extraction methods along with GAN-based data augmentation to address the problems such as visual degradation and data unavailability that are typically associated with historical inscription datasets.As shown in Fig. 4, incorporating GAN-based data augmentation leads to a noticeable improvement in classification accuracy compared to training without synthetic samples.

The experimental results show that the designed CNN architecture effectively learns discriminative character features, achieving a classification accuracy of 93% on the test dataset. The integration of GAN-generated synthetic samples significantly improves model generalization, particularly for low-frequency character classes, highlighting the effectiveness of data augmentation in low-resource settings.

The findings highlight the importance of GAN-implemented learning for historical and cultural heritage datasets, where collecting large-scale labeled data is often impractical. By improving recognition performance with limited samples, the proposed method contributes toward scalable and automated inscription digitization, laying the foundation for future extensions to word-level recognition and full OCR-based translation pipelines for ancient scripts.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

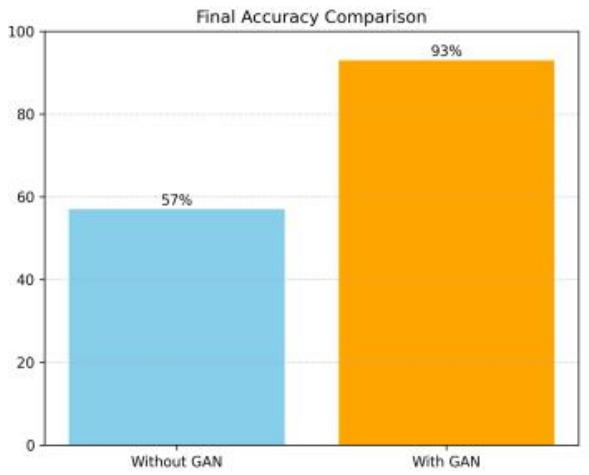


Fig. 6. Performance comparison of the CNN model with and without GAN-based data augmentation.

Overall, this work demonstrates that GAN-assisted CNN models offer a practical and effective solution for ancient script recognition in low-resource settings, contributing toward scalable digitization and preservation of historical inscription data.

X. REFERENCES

[1] U. Pal and B. B. Chaudhuri, "Indian script character recognition: A survey," Pattern Recognition, vol. 37, no. 9, pp. 1887–1899, 2004.

[2] B. Singh, U. Pal, and S. K. Parui, "Handwritten Indic script recognition using convolutional neural networks,"pp. 100–105, 2017.

[3] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from Simulated and Unsupervised Images through Adversarial Training," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[4] R. C. Gonzalez and R. E. Woods, "Digital Image Processing," Pearson Education, 2008.

[5] I. Goodfellow et al., "Generative Adversarial Nets," Advances in Neural Information Processing Systems (NeurIPS), 2014.

[6] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[7] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," International Conference on Learning Representations (ICLR), 2015.

[8] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006..

[9] L. Kang, J. H. Yi, Y. M. Tai, and C. Wang, "GAN-Based Data Augmentation for Handwritten Character Recognition," Pattern Recognition Letters, 2020.

[10] H. Li et al., "A Hybrid Approach for Handwritten Character Recognition Using Stroke Width Variation and Transformer-Based Feature Encoding," IEEE Access, 2022.