Thamizhi- Ancient Tamil Script To Modern Tamil Script In Stone Inscription

S.GAYATHRI THANUS S[1], ROHAN Y[2], SUBISH RAJ S[3]
[1, 2, 3] Dept of Computer Science and Bussiness Systems
[1, 2, 3] Jerusalem College of Engineering, Chennai, Tamilnadu, India

Abstract- *The identification of characters from ancient Tamil stone inscriptions is addressed through the integration of image processing and machine learning, specifically utilizing the XGBoost algorithm. The primary objective is to facilitate the preservation and digitization of these inscriptions, ensuring accessibility for historical studies and cultural conservation. A structured workflow is employed, beginning with image preprocessing techniques such as brightness and contrast optimization, noise removal, and contour-based segmentation to enhance clarity. Following preprocessing, individual characters are isolated and analyzed using an XGBoost classifier, trained for the accurate recognition of Tamil and English characters.*

Once recognition is complete, the extracted text is converted into modern Tamil and English formats and compiled into digital PDF documents. This approach ensures ease of storage, retrieval, and further analysis. By transforming inscriptions into structured digital formats, historical knowledge and cultural heritage are safeguarded, providing valuable resources for researchers and historians. The fusion of machine learning and image processing presents a robust solution for the conservation and study of ancient civilizations.

I. INTRODUCTION

Automating the recognition of characters in ancient Tamil stone inscriptions is facilitated by leveraging modern machine learning techniques. The initial stage involves preprocessing images by adjusting brightness and contrast to enhance visibility and removing unwanted noise that may interfere with accurate recognition. Character segmentation is performed using contour detection, ensuring precise isolation of individual characters from the stone inscriptions. This step is crucial for the accurate identification of complex and often eroded scripts.

The XGBoost algorithm is employed for classification, leveraging extracted features to ensure robust identification of both Tamil and English inscriptions. The output is structured into a searchable PDF format, enhancing accessibility for historical documentation and academic research. The integration of machine learning techniques with image processing significantly improves the preservation and analysis of ancient texts, providing a valuable tool for historians and researchers.

II. OVERVIEW

An intelligent character recognition system is developed for ancient Tamil stone inscriptions using advanced image processing techniques and the XGBoost algorithm. The core objective is to digitize and preserve these inscriptions, facilitating historical and linguistic research.

The process begins with preprocessing techniques, including contrast enhancement, noise reduction, and contour-based segmentation to enhance character clarity. Segmented characters are classified using an XGBoost model trained to recognize both Tamil and English scripts. The recognized text is formatted into modern Tamil and English versions and compiled into a structured digital PDF document.

By automating the recognition and conversion of inscriptions into digital text, long-term preservation of historical knowledge is ensured. A structured, searchable format benefits historians and researchers, offering a systematic approach to understanding ancient Tamil scripts through cutting-edge technological solutions.

III. EXISTINGSYSTEM

The preservation and deciphering of ancient stone inscriptions have historically depended on traditional manual techniques. Scholars and archaeologists have relied on methods such as visual examination, ink rubbings, tracings, and photography. Visual examination involves experts analyzing inscriptions through direct observation, often sketching or documenting scripts manually. However, this method is highly subjective and prone to human error. Ink rubbings and tracings require pressing paper or fabric against the stone surface and applying ink to capture the inscriptions. While useful for clearer engravings, it becomes ineffective when dealing with eroded or fragmented inscriptions. Photography and digital enhancement techniques involve capturing high-resolution images and applying contrast or brightness adjustments to improve visibility. Despite their

usefulness in improving clarity, these methods do not provide an automated mechanism for character recognition.

Traditional approaches present several challenges. They are time-consuming and labor-intensive, requiring significant effort to decipher and document inscriptions, which slows down historical research. Additionally, manual interpretation can be inconsistent, leading to discrepancies in transcription and translation. Moreover, when inscriptions are degraded due to erosion, missing portions, or unclear markings, manual methods become unreliable and difficult to apply effectively.

Recent digital approaches have attempted to enhance inscription legibility using basic image processing techniques such as contrast adjustment and noise reduction. These methods help in improving visibility but lack the ability to accurately identify and classify characters. Optical Character Recognition (OCR) technologies, commonly used for modern printed and handwritten text, struggle to process ancient scripts. Existing OCR systems are designed for contemporary languages and cannot handle the unique curvature and stylistic variations of ancient Tamil inscriptions. Furthermore, they fail to recognize partially eroded characters, as standard OCR models lack the ability to reconstruct missing portions of text. Additionally, OCR solutions have limited feature extraction capabilities and do not incorporate robust techniques to distinguish complex character shapes and textures found in ancient inscriptions.

Given the shortcomings of existing methods, there is a need for a more sophisticated approach that incorporates machine learning and advanced image processing techniques. A system that integrates deep feature extraction, noise-resistant segmentation, and machine learning-based classification can significantly improve the accuracy and efficiency of ancient inscription recognition. By leveraging these technologies, it becomes possible to automate the interpretation of historical scripts while preserving linguistic heritage in digital formats.

IV. PROPOSEDSYSTEM

Character recognition in ancient Tamil stone inscriptions is enhanced through the integration of advanced machine learning techniques. The workflow begins with an image preprocessing phase that optimizes inscription quality through brightness and contrast adjustments and noise reduction. Character segmentation is performed using contour detection techniques, accurately isolating individual characters for analysis.

The XGBoost algorithm is implemented due to its efficiency and high accuracy in processing large datasets. This model extracts key features from each character, ensuring reliable classification of Tamil and English scripts in inscriptions containing multiple languages. Once recognized, the text is converted into a structured PDF format, facilitating easy accessibility and long-term archival for researchers and historians.

V. ARCHITECTUR DIAGRAM

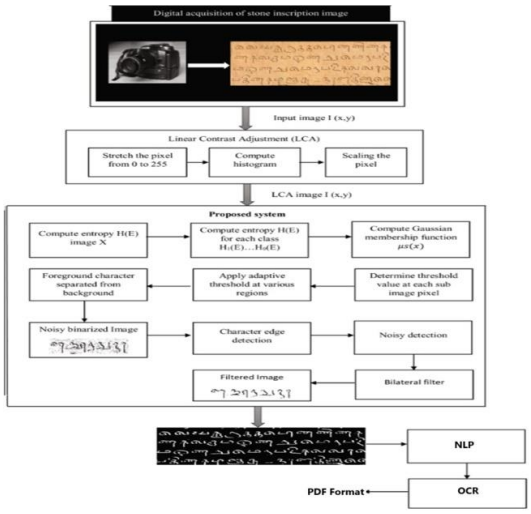

Fig.1.ArchitectureDiagram

we introduce the significance of designing a robust system for Tamil character recognition in ancient stone inscriptions. The goal is to create an efficient and accurate tool that leverages machine learning techniques to automate the recognition process, ensuring the preservation of cultural heritage

Theproposedworkingmodelhasthefollowingphases:

Datacollection and Pre processing: To build a robust recognition system, gather a diverse dataset of high-resolution images of ancient Tamil stone inscriptions. This dataset should capture a wide range of inscriptions with varying degrees of wear and erosion, simulating real-world challenges in inscription clarity. Include images taken under different lighting conditions and with a variety of stone textures and backgrounds to ensure the system's adaptability and accuracy. This diversity will help improve the model's resilience to environmental variations, enabling more accurate and consistent character recognition across different types of inscriptions. The preprocessing stage enhances character clarity by first adjusting brightness and contrast to ensure consistent illumination across the dataset. Noise removal techniques, like Gaussian or median filters, reduce artifacts

while preserving character edges. Finally, contour-based segmentation isolates individual characters from the background, enabling accurate analysis in subsequent stages.

Feature Extraction

Feature extraction is a crucial step in recognizing ancient Tamil inscriptions, ensuring that key characteristics of segmented characters are identified for accurate classification. Given the complexity of ancient scripts, multiple techniques are used to capture essential attributes such as shape, edges, texture, and structural properties.

Shape analysis identifies the structural outline of each character, using descriptors like aspect ratio, stroke width, and curvature to differentiate subtle variations. Edge detection techniques, including Sobel and Canny filters, enhance character contours, improving visibility even for eroded inscriptions.

Texture analysis methods, such as Local Binary Patterns (LBP) and Gray Level Co-occurrence Matrix (GLCM), detect repetitive patterns and surface variations, helping distinguish similar-looking characters. Morphological operations like dilation and erosion refine extracted features by reducing noise and preserving character integrity.

Histogram-based techniques analyze pixel intensity distributions, while the median filter enhances fine details, making faint inscriptions clearer. These combined techniques improve feature extraction, enabling accurate classification and aiding in the preservation of ancient Tamil scripts for historical research.

XG Boost Based character recongination:

The XGBoost algorithm is employed to classify extracted features into specific Tamil characters. It is a gradient boosting method that builds an ensemble of decision trees, where each new tree aims to correct the errors made by the previous trees. The primary advantage of XGBoost lies in its ability to handle large datasets, high-dimensional feature spaces, and to perform feature selection, making it an ideal choice for character recognition tasks where the data is complex and noisy.

Feature Extraction: The first step in this process is extracting meaningful features from the segmented Tamil characters. These features may include geometric properties such as aspect ratio, contour information, and pixel intensity distributions. Additionally, more sophisticated techniques like **Histogram of Oriented Gradients (HOG)** and **Zernike Moments** may be applied to capture the structural and shape-based information of the characters, helping the model distinguish between similar-looking characters. The resulting feature vectors serve as the input for the XGBoost model.

Model Training: Once the features are extracted, they are used to train the XGBoost classifier. The training process involves feeding the labeled dataset of Tamil characters (e.g., images of the character "அ" labeled as "அ") into the model. The XGBoost algorithm builds an ensemble of decision trees, where each tree focuses on learning the distinguishing features that separate one character from another. This process allows the model to learn complex patterns and relationships in the data, which are critical for accurate character recognition.

Cross-Validation: To ensure the model's robustness and prevent overfitting, **cross-validation** techniques are applied. This involves splitting the dataset into multiple folds (e.g., K-fold cross-validation) and training the model on different subsets of the data while testing it on the remaining data. This process helps to validate the model's accuracy and ensures that it generalizes well to new, unseen data. Cross-validation also aids in identifying any biases in the model and helps ensure consistent performance across different subsets of the dataset.

HyperparameterTuning: XGBoost's performance can be significantly improved by fine-tuning various hyperparameters, such as the learning rate, max depth of trees, and regularization terms. Hyperparameter tuning is typically done using grid search or random search techniques, where different combinations of hyperparameters are tested to find the optimal settings that yield the best model performance. Proper tuning ensures that the model balances the trade-off between bias and variance, leading to more accurate predictions.

Model Evaluation: After training and tuning, the model is evaluated using metrics like **accuracy**, **precision**, **recall**, and **F1-score**. The accuracy measures how often the model correctly identifies the Tamil characters, while precision and recall provide insights into the model's ability to correctly classify characters (precision) and its ability to identify all instances of a character (recall). The F1-score combines precision and recall into a single metric, providing a balanced measure of the model's performance. Additionally, a **confusion matrix** is used to visualize the model's classification performance, identifying which characters are being misclassified and helping to fine-tune the model further.

Post Processing: The recognized Tamil characters are converted into their modern Tamil equivalents, with an option to translate into English characters as well. The final text is formatted and saved as a digital PDF, ensuring accessibility and easy archival. This PDF serves as a complete, readable document of the ancient inscriptions, supporting both preservation and broader accessibility for research and educational use.

VI. RESULT

The proposed system demonstrates high accuracy in recognizing ancient Tamil characters, surpassing the effectiveness of traditional OCR methods. Through comparative evaluations, the approach has been validated as a reliable solution for handling degraded inscriptions, ensuring efficient transcription and digital preservation.

VII. CONCLUSION

A comprehensive framework for recognizing Tamil characters from ancient stone inscriptions is established. Through data collection, preprocessing, and feature extraction, image quality is refined, individual characters are isolated, and accurate recognition is facilitated. XGBoost-based character recognition ensures precise classification of Tamil and English characters. Post-processing organizes the recognized characters into a structured PDF output, ensuring accessibility and ease of reference. The methodology bridges ancient scripts with modern digital formats, supporting cultural preservation and enabling further historical research.

VIII. FUTURE SCOPE

Potential enhancements include:

Expanding the dataset: Increasing the variety of inscriptions in the dataset will improve the system's ability to recognize different styles of ancient Tamil script, enhancing overall accuracy and robustness.

Adding text-to-speech functionality: Converting recognized Tamil inscriptions into audio output will make historical content more accessible to visually impaired users and enable interactive learning for researchers and students.

Enhancing multilingual support: Expanding recognition capabilities to include Sanskrit, Telugu, and other regional scripts found in inscriptions will make the system more versatile and useful for broader historical research.

Developing a mobile application: Creating a user-friendly mobile app will allow researchers and history enthusiasts to capture inscriptions on-site and receive real-time translation and analysis.

Integrating cloud-based storage: Enabling cloud support will facilitate seamless access to recognized inscriptions, allowing researchers to collaborate and contribute to a shared digital archive of historical texts

REFERENCES

[1] R. D. R *et al*., "Self-Adaptive Hybridized Lion Optimization Algorithm With Transfer Learning for Ancient Tamil Character Recognition in Stone Inscriptions," in *IEEE Access*, vol. 11, pp. 39621-39634, 2023, doi: 10.1109/ACCESS.2023.3268545.

[2] M. Tang, S. Xie and X. Liu, "Ancient Character Recognition: A Novel Image Dataset of Shui Manuscript Characters and Classification Model," in *Chinese Journal of Electronics*, vol. 32, no. 1, pp. 64-75, January 2023, doi: 10.23919/cje.2022.00.077.

[3] R. Krithiga, S. Varsini, R. Gabriel Joshua and C. U. Om Kumar, "Ancient Character Recognition: A Comprehensive Review," in *IEEE Access*, doi: 10.1109/ACCESS.2023.3341352.

[4] A. Yavariabdi, H. Kusetogullari, T. Celik, S. Thummanapally, S. Rijwan and J. Hall, "CArDIS: A Swedish Historical Handwritten Character and Word Dataset," in IEEE Access, vol. 10, pp. 55338-55349, 2022, doi: 10.1109/ACCESS.2022.3175197.

[5] A. Rasheed, N. Ali, B. Zafar, A. Shabbir, M. Sajid and M. T. Mahmood, "Handwritten Urdu Characters and Digits Recognition Using Transfer Learning and Augmentation With AlexNet," in IEEE Access, vol. 10, pp. 102629-102645, 2022, doi: 10.1109/ACCESS.2022.3208959.

[6] G. Zhao, W. Wang, X. Wang, X. Bao, H. Li and M. Liu, "Incremental Recognition of Multi-Style Tibetan Character Based on Transfer Learning," in IEEE Access, vol. 12, pp. 44190-44206, 2024, doi: 10.1109/ACCESS.2024.3381039.

[7] F. Yan and H. Zhang, "SMFNet: One-Shot Recognition of Chinese Character Font Based on Siamese Metric Model," in IEEE Access, vol. 12, pp. 38473-38489, 2024, doi: 10.1109/ACCESS.2024.3370574.

[8] R. Buoy, M. Iwamura, S. Srun and K. Kise, "Toward a Low-Resource Non-Latin-Complete Baseline: An Exploration of Khmer Optical Character Recognition," in IEEE Access, vol. 11, pp. 128044-128060, 2023, doi: 10.1109/ACCESS.2023.3332361.

[9] P. Wojcicki and T. Zientarski, "Polish Word Recognition Based on n-Gram Methods," in IEEE Access, vol. 12, pp.

49817-49825, 2024, doi: 10.1109/ACCESS.2024.3385113.

[10] R. Malhotra and M. T. Addis, "Handwritten Amharic Word Recognition With Additive Attention Mechanism," in IEEE Access, vol. 12, pp. 114645-114657, 2024, doi: 10.1109/ACCESS.2024.3444897.