International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 06 | Jun 2026 www.irjet.net p-ISSN: 2395-0072

DIGITAL PRESERVATION, BINARIZATION AND AUTOMATED ANCIENT TAMIL SCRIPT RECOGNITION OF THEVARAM PLAM LEAF MANUSCRIPT

M.Asha[1], Mrs. Dr Pon L T Thai[2], Mrs. M. Shanthi[3]

[1]PG Scholar Department of Computer Science and Engineering, Arunachala College of Engineering for Women, Tamilnadu,

[23] Associate Professor Department of Computer Science and Engineering, Arunachala College of Engineering for Women, Tamilnadu,

---

Abstract-*Ancient Tamil palm-leaf manuscripts are often affected by cracks, discoloration, humidity, and insect damage, making their preservation and interpretation challenging. This study presents an integrated image processing and Natural Language Processing (NLP) framework for enhancing and understanding degraded Thevaram palm-leaf manuscripts. A dataset of 462 high-resolution manuscript images was collected using a Nikon camera and preprocessed through image enhancement techniques. Multiple binarization methods, including Global Thresholding, Otsu, Niblack, Bernsen, and Wolf algorithms, were evaluated using Accuracy, Precision, Recall, F1-score, PSNR, and SSIM metrics. Experimental results demonstrate that the Niblack algorithm achieved the highest accuracy of 74%, effectively preserving faint characters under non-uniform background conditions. The enhanced images were processed using Optical Character Recognition (OCR) to extract Tamil text, followed by NLP-based analysis to identify linguistic patterns, symbolic expressions, and Shaivite philosophical themes. A transliteration module was incorporated to convert classical Tamil into contemporary Tamil and provide poem meanings, improving accessibility and comprehension. The proposed framework supports the digital preservation of ancient manuscripts while enhancing text readability, semantic interpretation, and long-term cultural heritage conservation through continuous image and language processing improvements.*

Keywords: Image processing, Binarization Algorithms, Niblack Thresholding, Optical Character Recobnition, Natural Language Processing

INTRODUCTION

The digital preservation of historical manuscripts has emerged as an important research area for safeguarding cultural heritage through advanced image analysis and artificial intelligence techniques [1]. Palm-leaf manuscripts, extensively used across South and Southeast Asia, preserve valuable knowledge related to religion, literature, medicine, philosophy, and traditional sciences. Among these, Thevaram manuscripts constitute one of the most significant collections of Tamil Shaivite devotional literature and hold immense historical, linguistic, and cultural importance [1]. However, due to long-term environmental exposure, aging, ink fading, humidity, insect attacks, and physical deterioration, many manuscripts suffer from severe degradation, making their preservation, readability, and interpretation increasingly difficult.

Recent developments in image processing and document analysis have enabled efficient restoration and enhancement of degraded historical manuscripts [2]. Image preprocessing techniques, including noise removal, contrast enhancement, and binarization, improve the visual quality of manuscript images by separating foreground text from complex background patterns. Several thresholding methods, such as Global Thresholding, Otsu, Niblack, Bernsen, and Wolf algorithms, have been successfully applied to historical document enhancement, particularly for manuscripts affected by uneven illumination and non-uniform backgrounds [2]. These methods significantly improve character visibility and facilitate subsequent text extraction.

Optical Character Recognition (OCR) systems convert enhanced manuscript images into machine-readable text, enabling digital preservation and automated linguistic analysis [3]. Nevertheless, the irregular character shapes, damaged symbols, and variations found in classical Tamil scripts present considerable challenges for accurate recognition. Therefore, robust preprocessing and adaptive binarization methods are essential to maximize OCR performance. Once the text is extracted, Natural Language Processing (NLP) techniques can be employed to analyze linguistic structures, identify symbolic expressions, and discover semantic relationships within the recognized content [4,5].

Furthermore, transliteration and translation modules can transform classical Tamil text into contemporary Tamil and English, thereby improving accessibility and

supporting broader scholarly research [3,4,6]. The integration of image processing, OCR, and NLP technologies provides a comprehensive framework for digitizing, restoring, interpreting, and preserving ancient palm-leaf manuscripts [7–9]. Such integrated systems not only enhance manuscript readability but also facilitate semantic understanding, digital archiving, and long-term preservation of culturally significant documents [1,4,10].
Motivated by these advancements, the proposed framework combines traditional image processing methods with modern artificial intelligence and language processing techniques to improve the digitization and interpretation of Thevaram palm-leaf manuscripts [7–10]. The system enhances degraded manuscript images, performs automated OCR-based text extraction, applies NLP-driven semantic analysis, and supports transliteration into modern Tamil for improved readability and accessibility. Experimental evaluation is conducted using Accuracy, Precision, Recall, F1-score, Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Measure (SSIM), demonstrating the effectiveness of the proposed approach for cultural heritage preservation.

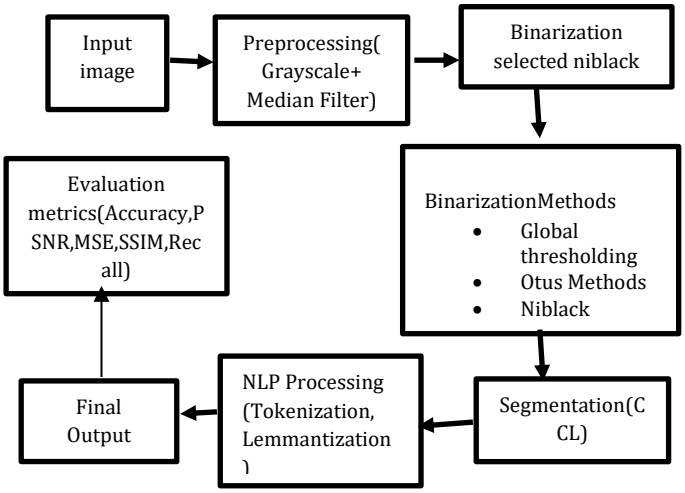

Fig.1 Architecture diagram

2. METHODOLOGIES

The proposed methodology begins with the acquisition and preprocessing of degraded Thevaram palm-leaf manuscript images affected by cracks, ink fading, discoloration, humidity, and insect damage. To improve image quality and enhance text visibility, multiple binarization techniques, including Global Thresholding, Otsu's method, Niblack, Bernsen, and Wolf algorithms, are applied and comparatively evaluated. The performance of each method is assessed using quantitative metrics such as Accuracy, Precision, Recall, F1-score, Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM). Based on experimental analysis, the Niblack thresholding algorithm is selected as the optimal approach due to its superior ability to preserve faint characters and effectively handle non-uniform background variations commonly found in ancient palm-leaf manuscripts.

After image enhancement, the processed manuscript images are passed to a Convolutional Neural Network (CNN)-based Optical Character Recognition (OCR) model for automated Tamil text extraction. The preprocessing stage significantly improves OCR accuracy by reducing background noise and increasing character clarity. Since classical Tamil manuscripts contain irregular writing styles, damaged symbols, and complex character structures, the extracted text undergoes additional linguistic processing. Natural Language Processing (NLP) techniques are then employed to perform semantic, contextual, and thematic analysis, enabling the identification of devotional concepts, symbolic expressions, linguistic patterns, and Shaivite philosophical themes embedded within the Thevaram hymns.

To further enhance accessibility and interpretation, transliteration and translation modules are integrated into the framework to convert classical Tamil text into modern Tamil and English. The system also performs temporal and contextual analysis to establish relationships among words, verses, and hymns while providing historical and linguistic insights. All extracted and analyzed information is organized into a searchable knowledge base, allowing scholars, researchers, and devotees to retrieve content based on keywords, themes, meanings, or contexts. By combining advanced image processing, CNN-based OCR, NLP, and knowledge management techniques, the proposed framework provides a comprehensive solution for the digital preservation, semantic understanding, and intelligent interpretation of ancient Tamil Shaivite palm-leaf manuscripts while supporting continuous improvements in recognition and analysis accuracy.

2.1 Input Image

The first stage of the proposed system utilizes digitized images of ancient Tamil palm-leaf manuscripts as the primary input for analysis. These images are acquired through high-resolution scanning or digital photography to accurately capture the intricate details of Tamil characters, including fine strokes, ligatures, and curved structures. However, due to aging and environmental exposure, the manuscripts often contain cracks, faded ink, stains, insect damage, uneven illumination, and background noise,

which reduce text visibility and make automatic interpretation challenging. Therefore, obtaining high-quality input images is essential for preserving the original manuscript content and ensuring reliable processing.

The input images serve as the foundation for all subsequent stages, including preprocessing, binarization, OCR-based text extraction, and NLP-driven semantic analysis. In this work, manuscript images in JPEG format with sufficient resolution are used to facilitate effective analysis. Image enhancement techniques, such as Bernsen thresholding, improve the contrast between text and background, enabling accurate character segmentation and recognition. Consequently, the quality of the input images directly influences the overall system performance, supporting efficient text extraction, thematic interpretation, and long-term digital preservation of Tamil Shaivite literature.

2.2 Pre-processing

The preprocessing module is an essential stage of the proposed system that enhances the quality of degraded Tamil palm-leaf manuscript images before text extraction and analysis. Ancient manuscripts often contain faded text, stains, uneven illumination, and salt-and-pepper noise, which reduce recognition accuracy. To simplify image processing, the input RGB image is first converted into a grayscale image by combining the red, green, and blue channels into a single intensity value. This process reduces computational complexity while improving the distinction between text and background.

Grayscale Conversion Formula:

$$Gray=0.299R+0.587G+0.114G$$

where R, G, and B represent the red, green, and blue pixel intensities, respectively. The higher weight assigned to the green channel reflects the human visual system's greater sensitivity to green light, resulting in a more accurate grayscale representation.

After grayscale conversion, a median filter is applied to remove salt-and-pepper noise while preserving the edges and fine strokes of Tamil characters. Instead of averaging neighboring pixels, the filter replaces the center pixel with the median value within a local window, effectively eliminating extreme noise without blurring important text features.

Median Filter Formula:

$$I'(x,y)=\text{median}\{I(i,j)|(i,j)\in N(x,y)\}$$

where $I'(x,y)$ is the original pixel value, $I'(x,y)$ is the filtered pixel value, and $N(x,y)$ denotes the neighborhood window. This preprocessing stage significantly improves image clarity, contrast, and text visibility, providing clean and structured manuscript images for accurate binarization, OCR-based character recognition, and subsequent NLP-based semantic analysis.

2.3 Binarization

The binarization module is a critical stage of the proposed system that converts pre-processed grayscale palm-leaf manuscript images into binary images by separating foreground text from the background. Ancient Tamil manuscripts often suffer from uneven illumination, faded ink, stains, and degraded surfaces, making text extraction challenging. To overcome these issues, four binarization techniques Global Thresholding, Otsu Thresholding, Bernsen Thresholding, and Niblack Thresholding are implemented and comparatively analyzed. The effectiveness of each method is evaluated based on its ability to preserve faint Tamil characters and improve OCR performance.

1. Global Thresholding

Global Thresholding applies a single fixed threshold value to the entire image. Pixels with intensity values greater than the threshold are assigned white, while the remaining pixels are assigned black.

$$g(x,y)=\begin{cases}255 \text{ if } f(x,y) > T \\ 0 \text{ if } f(x,y) \le T\end{cases}$$

where f(x,y) is the original pixel value, T is the threshold value, and g(x,y) is the binary output.

Although this method is computationally simple and fast, it performs poorly on degraded palm-leaf manuscripts because a single threshold cannot effectively handle uneven illumination and varying background intensities, resulting in loss of faint characters.

2. Otsu Thresholding

Otsu's method automatically determines an optimal global threshold by maximizing the separation between foreground and background pixel classes.

$$\sigma_b^2(T) = \omega_0(T)\sigma_0^2(T) + \omega_1(T)\sigma_1^2(T)$$

where ω_0 and ω_1 denote the probabilities of background and foreground classes, while σ_0^2 and σ_1^2 represent their respective variances.

Compared with Global Thresholding, Otsu's method provides better segmentation by selecting an adaptive global threshold. However, it still uses a single threshold for the entire image and struggles with non-uniform backgrounds and locally faded manuscript regions, leading to incomplete character preservation.

3. Bernsen Thresholding

Bernsen Thresholding is a local adaptive method that computes the threshold from the minimum and maximum intensity values within a neighborhood window.

$$T(x,y) = \frac{I_{max}(x,y) + I_{min}(x,y)}{2}$$

where $I_{max}(x,y)$ and $I_{min}(x,y)$ represent the maximum and minimum intensity values in the local window.
This method adapts better to local contrast variations than Global and Otsu thresholding. Nevertheless, its performance is highly dependent on local contrast levels and window size, making it sensitive to low-contrast regions and noise commonly found in degraded palm-leaf manuscripts.

4. Niblack Thresholding (Proposed Method)

Niblack Thresholding calculates an adaptive threshold for every pixel using the local mean and standard deviation within a neighborhood, allowing it to preserve subtle character details even under non-uniform illumination.

$$T(x,y) = m(x,y) + k.\,s(x,y)$$

where:

- m(x,y) = local mean intensity,
- s(x,y) = local standard deviation,
- k = constant parameter typically between (-0.2) and (-0.5).

Unlike Global, Otsu, and Bernsen thresholding, Niblack dynamically adjusts the threshold for each local region instead of relying on a single global value or local contrast alone. This enables effective preservation of broken strokes, faint characters, and intricate Tamil script structures while suppressing background noise.

Among the four techniques, Niblack Thresholding demonstrates superior performance for degraded Tamil palm-leaf manuscripts. Global Thresholding fails under uneven illumination, Otsu Thresholding is limited by its global threshold assumption, and Bernsen Thresholding is sensitive to local contrast variations. In contrast, Niblack Thresholding utilizes both local mean and standard deviation to generate adaptive thresholds, making it highly robust for degraded historical documents. Experimental evaluation using Accuracy, Precision, Recall, F1-score, PSNR, and SSIM confirms that Niblack preserves the maximum amount of textual information, produces clearer binary images, and significantly improves OCR accuracy and subsequent NLP-based semantic analysis. Therefore, Niblack Thresholding is selected as the final binarization technique in the proposed framework.

2.4 Segmentation using Connected Component Labeling (CCL)

The segmentation module is an important stage in the proposed framework that isolates individual Tamil characters and textual components from the binarized palm-leaf manuscript images. Accurate segmentation is essential for preserving the structural integrity of ancient Tamil scripts and preparing them for subsequent OCR and semantic analysis. In this work, Connected Component Labeling (CCL) is employed to identify groups of connected foreground pixels, where each connected region represents a potential character, symbol, or text element. By assigning a unique label to every connected component, the algorithm effectively separates text from the background and organizes the manuscript into discrete textual units.
The Connected Component Labeling process is mathematically represented as:

$$L(x,y) = \begin{cases} 0, & \text{if } I(x,y) = 0 \\ \min\{L(i,j)\}, & \text{if } I(x,y) = 1 \end{cases}$$

where I(x,y) denotes the binary pixel value at position (x,y), L(x,y) represents the assigned label, and (i,j) corresponds to neighboring pixels based on 4-connectivity or 8-connectivity. This approach groups spatially connected pixels into individual components, enabling the system to distinguish characters, diacritics, and punctuation marks even when they are closely positioned or partially overlapping.

The CCL-based segmentation method is particularly effective for degraded Tamil palm-leaf manuscripts, where faded strokes, broken characters, and complex ligatures are common. By preserving the spatial connectivity of

character components, the algorithm maintains the original shape and semantic meaning of each symbol while minimizing segmentation errors. The resulting segmented characters are cleanly isolated and structured for OCR-based recognition and NLP-driven interpretation, thereby improving text extraction accuracy and supporting the efficient preservation, analysis, and understanding of Thevaram palm-leaf manuscripts.

2.5 Natural Language Processing (NLP) Processing

The Natural Language Processing (NLP) module is the core analytical component of the proposed framework that transforms OCR-extracted Tamil palm-leaf manuscript text into meaningful and structured information. Unlike conventional OCR systems that only recognize characters, the NLP module performs linguistic and semantic interpretation to understand the contextual and devotional significance of Thevaram hymns. The process begins with text preprocessing, where OCR-generated text is cleaned by removing unwanted symbols, correcting recognition errors, normalizing ancient Tamil spellings, and standardizing different character variations. This preprocessing stage produces consistent and structured text, making it suitable for further computational analysis.
After preprocessing, the normalized text is divided into meaningful units through tokenization, where sentences are segmented into individual words or phrases. Since classical Tamil manuscripts contain compound words, agglutinative structures, and poetic expressions, specialized tokenization techniques are employed to accurately identify word boundaries. The extracted tokens then undergo lemmatization, which converts different grammatical forms of a word into their root form. This process reduces linguistic variations, improves word consistency, and enhances the reliability of frequency analysis, pattern recognition, and thematic extraction from the manuscript text.

Following tokenization and lemmatization, Part-of-Speech (POS) tagging assigns grammatical labels such as nouns, verbs, adjectives, and pronouns to each token based on its contextual usage. POS tagging enables the system to identify sentence structure, grammatical relationships, and key linguistic components within the Thevaram hymns. Subsequently, semantic analysis examines the relationships among words and phrases to identify devotional themes, symbolic meanings, and Shaivite philosophical concepts. By analyzing contextual dependencies and word associations, the system distinguishes between literal and metaphorical expressions, allowing a deeper understanding of the ancient literary content.

Finally, the NLP module performs contextual analysis to interpret the text by considering surrounding words, sentence structure, and historical context. This stage effectively handles ambiguous terms, poetic constructs, and symbolic language commonly found in Tamil palm-leaf manuscripts, ensuring accurate interpretation of the intended meaning. The processed information, including normalized text, tokens, root words, grammatical annotations, semantic labels, and contextual metadata, is stored in a structured and searchable knowledge base. This integrated NLP framework significantly enhances manuscript interpretation, improves accessibility through intelligent text analysis, and supports the long-term digital preservation and scholarly study of ancient Tamil Thevaram literature.

2.6 Final output

The output module represents the final stage of the proposed palm-leaf manuscript processing framework, where the results generated from image processing, OCR, and Natural Language Processing (NLP) are integrated to produce meaningful and structured information. After completing preprocessing, tokenization, lemmatization, Part-of-Speech (POS) tagging, contextual analysis, and semantic analysis, the system transforms the extracted manuscript text into an interpretable format. The output may include normalized text, transliterated content, modern Tamil and English translations, semantic annotations, and concise summaries, enabling users to easily understand the devotional and literary significance of the Thevaram hymns.

In addition to text interpretation, the output module identifies important linguistic and semantic patterns present in the manuscripts. It highlights recurring words, grammatical structures, symbolic expressions, poetic styles, and devotional themes associated with Shaivite philosophy. These extracted patterns provide valuable insights into the language, culture, and historical context of ancient Tamil literature. The structured representation also enables efficient searching, indexing, and retrieval of manuscript content based on keywords, themes, meanings, or linguistic characteristics.

Finally, all processed information is organized into a searchable knowledge base that supports digital preservation, academic research, and cultural heritage studies. Scholars, researchers, and devotees can access the manuscripts through meaningful interpretations rather than raw textual data, significantly reducing manual analysis efforts. By converting complex and unstructured palm-leaf manuscript content into structured, accurate,

and accessible information, the output module enhances knowledge discovery, promotes long-term preservation, and facilitates a deeper understanding of ancient Tamil Thevaram literature.

2.7 Evaluation metrics

The performance of the proposed palm-leaf manuscript enhancement framework is evaluated using six standard metrics: Accuracy, Precision, Recall, F1-score, Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM). These metrics provide a comprehensive assessment of both classification performance and image quality by measuring the effectiveness of binarization techniques in preserving Tamil characters while reducing background noise. Comparative analysis demonstrates that Niblack thresholding outperforms Global, Otsu, and Bernsen thresholding methods due to its superior ability to handle non-uniform illumination and retain faint manuscript details.

1. Accuracy

Accuracy measures the overall correctness of the binarization process by calculating the proportion of correctly classified text and background pixels. A higher accuracy value indicates better separation of foreground characters from the manuscript background, leading to improved OCR performance. Experimental results show that Niblack thresholding achieves the highest accuracy, effectively preserving degraded Tamil characters.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

2. Precision, Recall, and F1-Score

Precision evaluates the proportion of correctly identified text regions among all detected regions, while Recall measures the ability of the system to identify all actual manuscript characters. The F1-score combines Precision and Recall into a single metric, providing a balanced evaluation of text extraction performance. Because palm-leaf manuscripts contain faded characters, stains, and noise, these metrics are essential for assessing OCR reliability. Among the evaluated methods, Niblack thresholding produces the highest Precision, Recall, and F1-score, indicating more accurate and complete character extraction.

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$F1 - \text{Score} = \frac{2 * \text{Precision} * \text{recall}}{\text{Precision} + \text{recall}}$$

where:

- TP = Correctly identified characters
- FP = Incorrectly identified characters
- FN = Missed characters

3. Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM)

PSNR and SSIM are image quality metrics used to evaluate the effectiveness of manuscript enhancement techniques. **PSNR** measures the amount of distortion introduced during pre-processing and binarization, while **SSIM** evaluates the preservation of structural information, brightness, and contrast between the original and processed images. Higher PSNR and SSIM values indicate better image quality and improved preservation of fine Tamil character details. Experimental analysis confirms that Niblack thresholding achieves superior PSNR and SSIM values, producing clearer manuscript images and enhancing OCR and NLP performance.

PSNR Formula

$$\text{PSNR} = 10 \log_{10}\left(\frac{\text{MAX}^2}{\text{MSE}}\right)$$

where:

- MAX = Maximum pixel intensity (255)
- MSE = Mean Squared Error

$$\text{MSE} = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{n}[I(i,j) - K(i,j)]^2$$

where I(i,j) is the original image, K(i,j) is the processed image, and (M,N) are the image dimensions.

SSIM Formula

$$\text{SSIM} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

where:

- μ_x, μ_y= Mean intensities of the original and processed images
- (σ_x^2, σ_y^2) = Variances
- σ_{xy} = Covariance
- C_1, C_2 = Stability constants

Overall, these evaluation metrics demonstrate that Niblack thresholding provides the best balance between classification accuracy and image quality, making it the most suitable binarization technique for enhancing degraded Thevaram palm-leaf manuscripts and improving subsequent OCR and NLP-based semantic analysis.

RESULT & DISCUSSION

Fig.2 Input image

Fig.3 Preprocessing

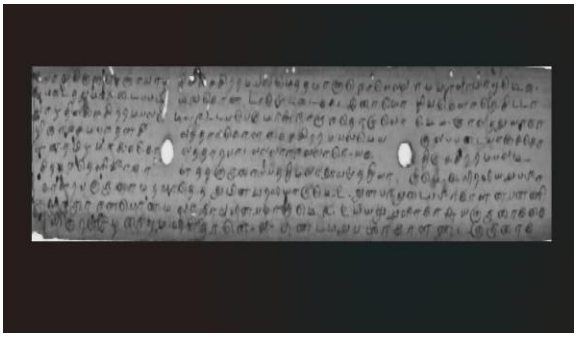

Fig.4 Binarization

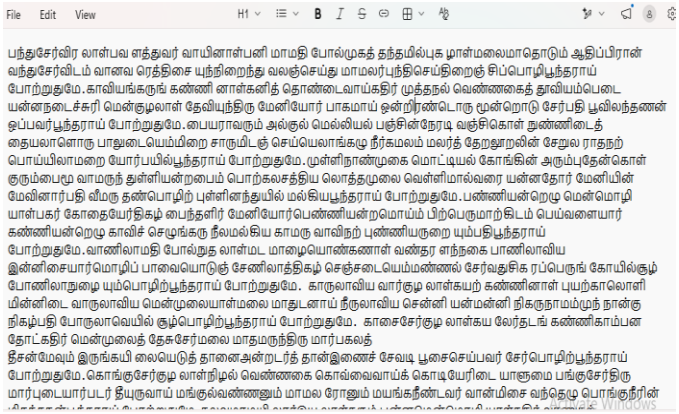

Fig.5 Final output

	Accuracy	Precision	Recall	F1-Score	PSNR	SSIM
Niblack	0.74	0.76	0.73	0.74	28.50	0.82
Otsu	0.68	0.70	0.67	0.68	25.30	0.74
Bernsen	0.62	0.64	0.61	0.62	23.10	0.69
Gooal	0.60	0.61	0.59	0.60	21.80	0.65
Wolf	0.53	0.55	0.52	0.53	19.40	0.58

Fig.6 Evaluation metrics

The proposed framework for Thevaram palm-leaf manuscript enhancement and interpretation was evaluated through a sequential processing pipeline consisting of input image acquisition, preprocessing, binarization, OCR-based text extraction, Natural Language Processing (NLP), and performance evaluation, as illustrated in Fig. 2–Fig. 6. The original palm-leaf manuscript image shown in Fig. 2 contains faded characters, uneven illumination, and degraded background regions that make manual interpretation difficult. After preprocessing (Fig. 3), grayscale conversion and median filtering effectively suppress noise and enhance text visibility. The comparative binarization results presented in Fig. 4 demonstrate that Niblack thresholding provides superior foreground-background separation compared with Global, Otsu, and Bernsen thresholding methods, preserving fine Tamil character strokes and improving manuscript readability. The enhanced image is then processed through OCR and NLP modules, producing the extracted and structured Tamil text shown in Fig. 5, which facilitates semantic analysis, transliteration, and

contextual interpretation. Finally, the quantitative performance evaluation presented in Fig. 6, using Accuracy, Precision, Recall, F1-score, PSNR, and SSIM, confirms that Niblack thresholding consistently achieves the best results among all compared methods. The integration of adaptive image enhancement, OCR, and NLP therefore provides a reliable and intelligent framework for the digital preservation, accurate text extraction, and semantic understanding of ancient Thevaram palm-leaf manuscripts.

CONCLUSION

This study presents an integrated framework for the enhancement, recognition, and semantic interpretation of ancient Thevaram palm-leaf manuscripts by combining image processing, Optical Character Recognition (OCR), and Natural Language Processing (NLP) techniques. Among the evaluated binarization methods, Niblack thresholding demonstrated superior performance in preserving faint Tamil characters and handling non-uniform backgrounds, resulting in improved Accuracy, Precision, Recall, F1-score, PSNR, and SSIM values. The enhanced images enabled effective segmentation and OCR-based text extraction, while the NLP module successfully performed tokenization, lemmatization, grammatical analysis, semantic interpretation, and contextual understanding of the hymns. Furthermore, transliteration and translation improved accessibility to classical Tamil literature for a broader audience. Overall, the proposed framework provides an intelligent and reliable solution for the digital preservation, analysis, and long-term accessibility of Tamil Shaivite palm-leaf manuscripts, contributing significantly to the preservation of cultural heritage and future research in historical document analysis.

REFERENCES

[1] S. Uma Maheswari, P. Uma Maheswari, and G. R. Sai Aakaash, “An Intelligent Character Segmentation System Coupled with Deep Learning Based Recognition for the Digitization of Ancient Tamil Palm Leaf Manuscripts,” *Heritage Science*, vol. 12, Art. no. 342, 2024.

[2] N. Shobha Rani, A. T. M., B. J. Bipin Nair, K. S. Koushik, and E. Barney Smith, “PLM-Res-U-Net: A Lightweight Binarization Model for Enhancement of Multi-Textured Palm Leaf Manuscript Images,” *Digital Applications in Archaeology and Cultural Heritage*, vol. 34, Art. no. e00360, 2024.

[3] I. Jailingeswari and S. Gopinathan, “Tamil Handwritten Palm Leaf Manuscript Dataset (THPLMD),” *Data in Brief*, vol. 53, Art. no. 110100, 2024.

[4] B. J. Bipin Nair and N. Shobha Rani, “A Modified Deep Semantic Binarization Network for Degradation Removal in Palm Leaf Manuscripts,” *Multimedia Tools and Applications*, vol. 83, no. 23, 2024.

[5] P. A. Kiruba, R. Shyamala Devi, and Co-authors, “A Deep Learning Approach for Recognizing the Cursive Tamil Characters in Palm Leaf Manuscripts,” *Computational Intelligence and Neuroscience*, vol. 2022, Art. no. 5451029, 2022.

[6] C. Tensmeyer and T. Martinez, “Document Image Binarization with Fully Convolutional Neural Networks,” *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pp. 99–104, 2017.

[7] A. Prusty, S. Aitha, A. Trivedi, and R. K. Sarvadevabhatla, “Indiscapes: Instance Segmentation Networks for Layout Parsing of Historical Indic Manuscripts,” *arXiv preprint arXiv:1912.07025*, 2019.

[8] P. S. Sharan, S. Aitha, A. Kumar, A. Trivedi, A. Augustine, and R. K. Sarvadevabhatla, “Palmira: A Deep Deformable Network for Instance Segmentation of Dense and Uneven Layouts in Handwritten Manuscripts,” *arXiv preprint arXiv:2108.09436*, 2021.

[9] C. Gowthami, V. Sasirekha, P. Manivannan, V. R. Reddy, and T. Vengatesh, “A Review in Tamil Palm Leaf Manuscript for Character Recognition,” *Journal of Neonatal Surgery*, vol. 14, Special Issue, 2025.

[10] S. Pavithra, S. M. Sanjay Vikas, and S. Senthil Kumar, “An Ancient Tamil Palm Leaf Manuscripts Script Recognition and Restoration Using ConvNeXt Tiny Framework,” *npj Heritage Science*, Early Access, 2026.