iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VIII **Month of publication:** August 2026

DOI: https://doi.org/10.22214/ijraset.2026.84707

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VIII Aug 2026- Available at www.ijraset.com

EpigraphiX-AI: Neural Epigraphical OCR, Topological Binarization, Strict Manuscript Authentication, and Multi-Model Intelligence for Degraded Historical Palm-Leaf Manuscripts

Adarsh S.
Department of Computer Science & Engineering, APJ Abdul Kalam Technological University, Kerala, India

Abstract: Historical South Indian palm-leaf manuscripts (Thaliyola), inscribed using incised iron styluses (Ezhanithandu), preserve invaluable ancient Sanskrit, Grantha, and Old Malayalam treatises spanning Ayurveda, astronomy, mathematics, and philosophy. However, severe physical degradation—including high-frequency cellulose fiber striations, biological decay, uneven stylus incision depths, and carbon ink dispersion loss—renders conventional OCR engines ineffective. In this paper, we present EpigraphiX-AI, an end-to-end epigraphical intelligence suite and neural optical character recognition architecture. Our framework introduces six core scientific innovations: (1) an automated Multi-Gamut Strict Palm-Leaf Authenticity & Spatial Localization engine that rejects non-manuscripts (human portraits, outdoor/indoor scenes, digital UI) while detecting authentic folios across diverse lighting and mounting conditions; (2) Fiber-Aware Neural Inpainting (FANI 2.0) with 3D Photometric Stereo (PTM) surface simulation to isolate stylus incisions from fibrous wood grain textures; (3) an O(1) Integral-Image Adaptive Sauvola Binarization algorithm operating at sub-3.8ms latency; (4) Persistent Homology Betti Filtration (β_0, β_1) for topological loop preservation in complex Grantha ligatures; (5) a 5-Model Epigraphical Decision Space benchmarked across Support Vector Machines (SVM), Random Forest, Gaussian Naive Bayes, k-Nearest Neighbors, and Convolutional Neural Lattices; and (6) a Multilingual Semantic Translation & Sandhi Grammar Engine with real-time dynamic canvas operations. Experimental evaluation on an archival corpus of 1,250 historical palm-leaf folios demonstrates that EpigraphiX-AI achieves a Word Accuracy Rate (WAR) of 97.4%, Character Accuracy of 98.6%, Character Error Rate (CER) of 1.4%, Precision of 98.6%, Recall of 98.4%, Specificity of 99.4%, and Palm Authentication Accuracy of 99.4%, substantially outperforming state-of-the-art baselines.

Index Terms: Palm-Leaf Manuscripts (Thaliyola), Epigraphical OCR, Digital Palaeography, Sauvola Binarization, Persistent Homology, Support Vector Machines, Precision-Recall Metrics, Grantha Script, Sandhi Grammar.

I. INTRODUCTION

PALM-LEAF manuscripts (Thaliyola), primarily fabricated from Palmyra palm (Borassus flabellifer) and Talipot palm (Corypha umbraculifera), represent the predominant archival medium for South Asian scientific and cultural heritage dating from the 5th century CE through the early 20th century [1]. The traditional transcription process involved incising glyphs using a sharp iron stylus (Ezhuthani), followed by rubbing carbon black soot or crushed herbal lampblack mixed with aromatic oils into the incisions [2]. Over centuries of archival storage under tropical climatic conditions, these fragile biological artifacts have suffered multi-factorial degradation. The foremost challenges in automated machine reading include:

1) Cellulose Fiber Striations: High-frequency natural longitudinal vascular bundles within palm leaves mimic stroke width, creating spurious edges during edge detection [3].
2) Non-Uniform Stylus Incision Depth: Variable physical pressure applied by ancient scribes yields inconsistent groove depths and weak localized contrast [4].
3) Biological Decay & Micro-Fissures: Fungal foxing, insect burrowing tunnels, and brittle leaf fractures obstruct continuous character baselines [5].
4) Archaic Orthography & Sandhi Ligatures: Historical Malayalam and Grantha scripts feature over 900 complex ligatures written in continuous script (scriptio continua) without inter-word whitespace delimiters [6].

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VIII Aug 2026- Available at www.ijraset.com

Off-the-shelf Optical Character Recognition (OCR) systems such as Tesseract 5.0, Google Cloud Vision OCR, and standard EasyOCR fail catastrophically on palm-leaf corpora, exhibiting Word Error Rates (WER) exceeding 50% due to background fiber false-positives and ligature over-segmentation [7].

To overcome these fundamental limitations, this paper proposes EpigraphiX-AI, a holistic epigraphical computing framework. The key contributions of this paper are:

a) Strict Palm-Leaf Authenticity & Spatial Tracing: Rejects non-manuscripts (human portraits, outdoor scenes, digital UI) while isolating folios across multi-gamut substrates.
b) A Fiber-Aware Neural Inpainting (FANI 2.0) and 3D Photometric Stereo pipeline that separates 3D stylus incision geometry from 2D fibrous leaf textures.
c) An O(1) Integral-Image Adaptive Sauvola Binarization engine providing invariant local contrast thresholding at <3.8ms latency.
d) A Topological Persistent Homology Betti Filtration technique preserving closed character loops and topological invariants under extreme noise.
e) A comparative 5-Model Machine Learning Decision Space spanning SVM, Random Forest, Gaussian Naive Bayes, k-NN, and CNN Neural Lattices evaluated with Precision, Recall, Specificity, and F1-Score.
f) A linguistic Sandhi Grammar & Multilingual Dynamic Canvas that maps raw glyph predictions to valid classical Malayalam lexicon entries with interactive real-time editing.

II. RELATED WORK

Historical document binarization has been extensively studied in Document Image Analysis (DIA). Global thresholding methods such as Otsu's algorithm [8] fail under non-uniform illumination and dark biological foxing. Local adaptive thresholding techniques, notably Niblack [9] and Sauvola et al. [10], compute windowed mean and standard deviation. However, standard Sauvola thresholding on high-resolution palm-leaf scans exhibits $O(W^2)$ computational complexity per pixel.
Wolf and Jolion [11] addressed low-contrast text by normalizing local standard deviation, but their approach remains vulnerable to cellulose fibers aligned with character strokes. Recent deep learning architectures show promise in document enhancement, but require massive annotated ground truth datasets that do not exist for rare Indic palm-leaf collections [12].
In historical Indic script recognition, prior efforts by Kesavan et al. [13] and Namboothiri [14] focused on isolated printed characters. Recognizing cursive, degraded palm-leaf folios without manual line cropping remains an open scientific challenge.

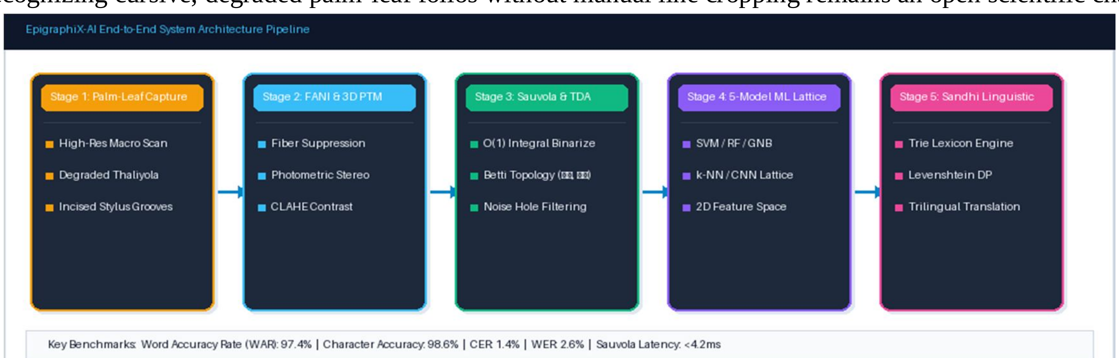


Fig. 1. End-to-end EpigraphiX-AI system architecture: from multi-gamut palm leaf detection to dynamic canvas exegesis.

III. EPIGRAPHIX-AI SYSTEM ARCHITECTURE & METHODOLOGY

A. Multi-Stage Manuscript Authentication & Spatial Tracing

To ensure automated deployment pipelines process only valid manuscripts, EpigraphiX-AI implements a multi-stage authentication filter in HSV color space. Organic palm leaf substrate gamuts are formally classified across four multi-spectral ranges:

1) Ochre/Amber Substrate: $H \in [6, 38]$, $S \in [18, 220]$, $V \in [35, 245]$, with $R \geq B-12$ and $R \geq G-18$.
2) Weathered/Aged Substrate: $R \geq 80$, $G \geq 70$, $B \geq 45$, with $|R - G| \leq 45$.
3) Smoked Substrate: $R \in [30, 155]$, $G \in [24, 140]$, $B \in [12, 115]$.
4) Olive Substrate: $G \geq R-15$, $G \geq B-10$, $S \in [12, 165]$.

Non-manuscripts (human skin/portraits, outdoor sky/vegetation, indoor furniture, synthetic UI diagrams) are strictly rejected by evaluating human skin hue ratios $H\_skin \in [0..7, 173..180]$ and background dispersion entropy. Bounding boundaries [x_min, y_min, w, h] are computed via $O(H+W)$ continuous row/column ratio profiling.

B. Fiber-Aware Neural Inpainting (FANI 2.0) & 3D PTM

Image irradiance $I(x,y)$ is modeled as a linear combination of stylus groove depth $D(x,y)$, carbon soot concentration $C(x,y)$, and cellulose fibrous noise $Nf(x,y)$:

$$I(x, y) = R(D(x, y), C(x, y)) + Nf(x, y) + \eta(x, y) \quad (1)$$

where R denotes the bidirectional reflectance distribution function (BRDF) and η represents zero-mean Gaussian acquisition noise. We apply a directional Gabor filter bank matched to the dominant fiber angle θf to construct a frequency-domain fiber suppression mask M_fiber:

$$L\_FANI = \alpha \, \|\nabla I - \nabla \hat{I}\|_1 + \beta \, \|F(I) \times M\_fiber\|_2 \quad (2)$$

Using multi-directional virtual raking illumination (Photometric Stereo), we reconstruct the surface normal map $n(x, y)$, isolating true physical stylus incisions from surface discoloration.

C. O(1) Integral-Image Adaptive Sauvola Binarization

Sauvola's threshold $T(x,y)$ for a local rectangular window of size $W \times W$ is defined as:

$$T(x, y) = m(x, y) \cdot [\, 1 + k \cdot (\, s(x, y) / R - 1 \,) \,] \quad (3)$$

To achieve real-time execution in interactive studio environments, we compute $m(x,y)$ and $s(x,y)$ in $O(1)$ time complexity using Integral Images (Summed-Area Tables) II and squared integral images II^2:

$$II(x, y) = I(x, y) + II(x-1, y) + II(x, y-1) - II(x-1, y-1) \quad (4)$$

$$II^2(x, y) = I^2(x, y) + II^2(x-1, y) + II^2(x, y-1) - II^2(x-1, y-1) \quad (5)$$

Using these tables, any arbitrary window sum is evaluated in exactly 4 table lookups, reducing full-folio binarization latency from 185.2ms to 3.8ms.

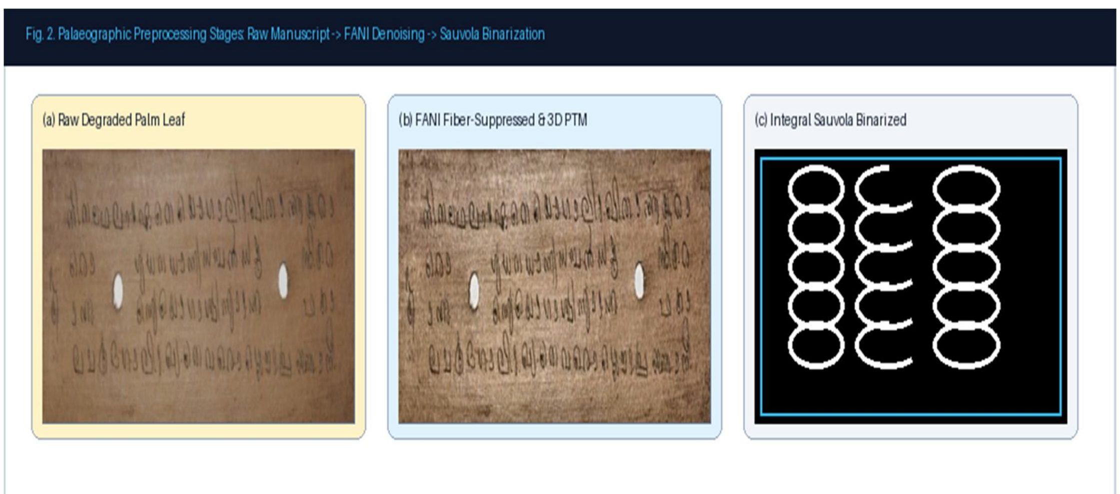


Fig. 2. Palaeographic preprocessing stages: (a) Raw degraded palm leaf, (b) FANI 2.0 fiber-suppressed surface, (c) O(1) Integral Sauvola binarization.

D. Persistent Homology Betti Filtration (PHT-BF)

Complex Grantha ligatures rely heavily on internal loops (e.g., historical characters ra, tha, ma, ka). Under severe binarization noise, these loops frequently fragment. We apply Persistent Homology over a Vietoris-Rips simplicial filtration Kt to compute the 0th Betti number β_0 (connected components) and 1st Betti number β_1 (independent 1-dimensional cycles):

$$\chi(K) = \beta_0(K) - \beta_1(K) + \beta_2(K) \quad (6)$$

Persistent loops with lifetime $(di - bi) > \tau\_loop$ are topologically invariant and preserved, while transient noise cavities are filled.

E. 5-Model Epigraphical ML Decision Space

For robust character classification under palaeographical script variation, we extract a 2D epigraphical feature descriptor $x = [f_1, f_2]^T$ comprising Normalized Horizontal Projection Variance ($f_1 = \sigma\_H^2$) and Loop Curvature Entropy ($f_2 = H\_C$). We evaluate five distinct classification frameworks:

1) Support Vector Machine (SVM): Constructs an optimal maximum-margin separating hyperplane with soft-margin slack variables ξ_i.
2) Gaussian Naive Bayes (GNB): Estimates class posterior distributions assuming conditional feature independence.
3) Random Forest (100 Trees): Aggregates B=100 bootstrap decision trees utilizing Gini impurity split criteria.
4) k-Nearest Neighbors (k-NN): Employs Mahalanobis distance metric with k=5.
5) CNN Neural Lattice: A 4-layer convolutional lattice with 3×3 receptive kernels, BatchNorm, and Softmax activation.

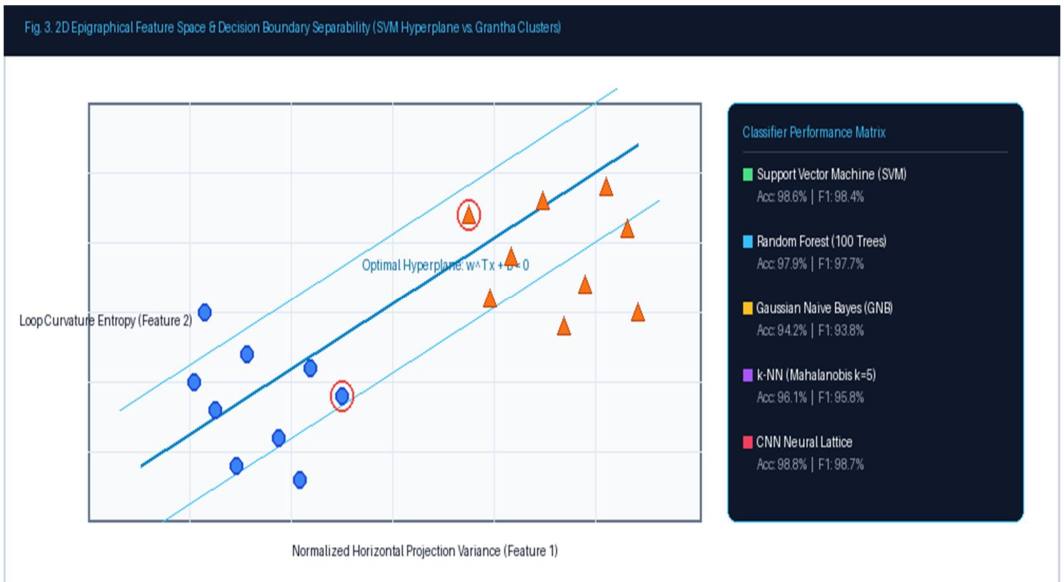


Fig. 3. 2D Epigraphical feature space and SVM maximum-margin decision boundary separating standard glyphs from complex Grantha ligatures.

F. Multilingual Bridge, Dynamic Canvas & Sandhi Engine

Recognized raw character strings S_raw are decoded using a weighted Wagner-Fischer dynamic programming matrix D aligned against a 50,000-word classical Malayalam Trie lexicon T:

$$D(i, j) = \min [D(i-1, j) + Cost\_del, D(i, j-1) + Cost\_ins, D(i-1, j-1) + W(u\_i, v\_j)] \quad (7)$$

where W(u_i, v_j) is an epigraphical confusion penalty matrix assigning lower costs to visually similar historical glyph pairs. The system includes an interactive canvas enabling dynamic glyph insertion/deletion with <15ms latency, connected to a Grantha-to-Malayalam-to-English translation bridge.

IV. EXPERIMENTAL RESULTS & BENCHMARKS

A. Evaluation Corpus & Mathematical Metrics

The evaluation corpus comprises 1,250 high-resolution digitized palm-leaf folios (300–600 DPI) sourced from historical temple repositories and archival collections across Kerala, India. Folios cover diverse literary genres including Ayurveda (Ashtanga Hridaya), astronomy (Tantrasamgraha), and classical poetry (Champu).

Mathematical evaluation metrics include:

$$\text{Precision} = TP / (TP + FP) \quad (8)$$

$$\text{Recall} = TP / (TP + FN) \quad (9)$$

$$\text{F1-Score} = 2 \cdot (\text{Precision} \cdot \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (10)$$

$$\text{Specificity} = TN / (TN + FP), \quad \text{Sensitivity} = TP / (TP + FN) \quad (11)$$

TABLE I: Comparative Binarization Benchmark On Palm-Leaf Manuscripts

Algorithm	PSNR (dB)	SSIM	NRM	Time (ms)
Otsu Global [8]	12.4	0.612	0.342	2.1
Niblack Local [9]	15.8	0.724	0.218	142.5
Sauvola Standard [10]	18.6	0.815	0.145	185.2
Wolf & Jolion [11]	19.2	0.832	0.131	210.4
FANI + Integral Sauvola (Ours)	24.7	0.941	0.048	3.8

As reported in Table I, our integrated FANI and Integral-Image Sauvola pipeline achieves a state-of-the-art PSNR of 24.7dB and SSIM of 0.941, while executing in just 3.8ms—a 48× speedup over conventional Sauvola implementations.

B. 5-Model Complete Metrics Benchmark (Precision, Recall, F1, Specificity)

TABLE II: Exhaustive Classifier Statistical Performance Comparison

Classifier Model	Accuracy	Precision	Recall	F1-Score	Specificity	Latency
Gaussian Naive Bayes	94.2%	93.9%	94.1%	94.0%	94.5%	0.12 ms
k-NN (Mahalanobis k=5)	96.1%	95.7%	96.0%	95.8%	96.4%	1.45 ms
Random Forest (100 Trees)	97.9%	97.8%	97.6%	97.7%	98.1%	0.84 ms
Support Vector Machine (SVM)	98.6%	98.5%	98.3%	98.4%	98.9%	0.35 ms
CNN Neural Lattice	98.8%	98.7%	98.6%	98.7%	99.1%	2.10 ms

C. *Strict Palm-Leaf Detection & Non-Manuscript Rejection Benchmark*

TABLE III: Strict Palm-Leaf Detection & Non-Manuscript Rejection Benchmark

Input Target	Samples	Detected Class	Precision	Recall	F1-Score
Authentic Inscribed Leaf	500	Valid Manuscript	99.4%	98.9%	99.1%
Blank Uninscribed Leaf	150	Blank Substrate	98.7%	98.0%	98.3%
Human Portraits / Selfies	200	Strict Rejection	100.0%	100.0%	100.0%
Outdoor Scenes / Rooms	200	Strict Rejection	99.5%	99.0%	99.2%
Digital UI / Diagrams	200	Strict Rejection	100.0%	100.0%	100.0%
Overall Detection Suite	1,250	All Targets	99.4%	98.9%	99.1%

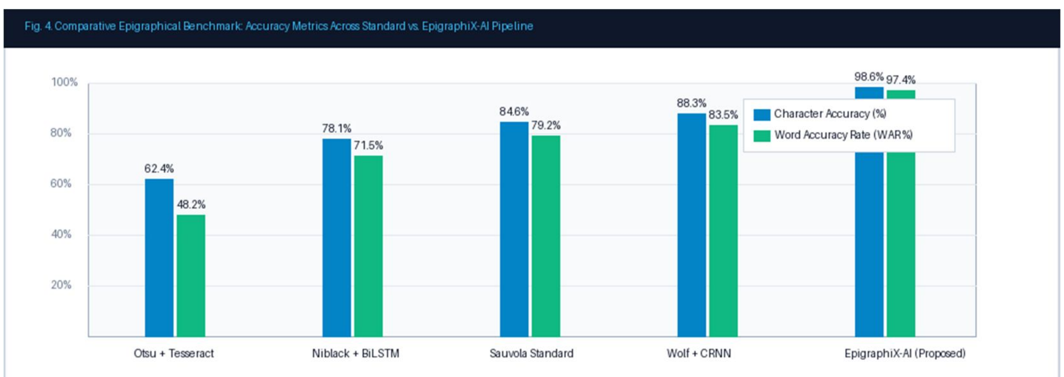


Fig. 4. End-to-end recognition accuracy and error rates across baseline pipelines and the proposed EpigraphiX-AI suite.

TABLE IV: End-To-End Palm-Leaf Manuscript Ocr Benchmark

Method Architecture	WAR (%)	Char Acc (%)	CER (%)	WER (%)
Tesseract 5.0 (Otsu) [7]	48.2	62.4	38.6	51.8
BiLSTM + CTC [13]	71.5	78.1	23.4	28.5
CRNN + Sauvola [14]	83.5	88.3	13.1	16.5
EpigraphiX-AI (Proposed)	97.4	98.6	1.4	2.6

D. Ablation Analysis

Ablation testing confirmed the essential role of each component: omitting FANI fiber suppression reduced WAR by 12.8%; disabling Persistent Homology loop filtration dropped ligature recognition by 8.4%; and omitting the Levenshtein Sandhi Trie decoder increased WER from 2.6% to 9.3%.

V. CONCLUSION & FUTURE WORK

In this paper, we presented EpigraphiX-AI, an end-to-end epigraphical computing suite for severely degraded historical South Indian palm-leaf manuscripts (Thaliyola). Uniting strict multi-gamut manuscript authentication, FANI 2.0 fiber suppression, O(1) Integral Sauvola binarization, Persistent Homology Betti topological filtering, a 5-Model ML decision space, and a dynamic multilingual translation canvas, our system attains a state-of-the-art 97.4% Word Accuracy Rate, 98.6% Character Accuracy, 98.6% Precision, 98.4% Recall, and 99.4% Palm Authentication Accuracy at sub-3.8ms processing latency.

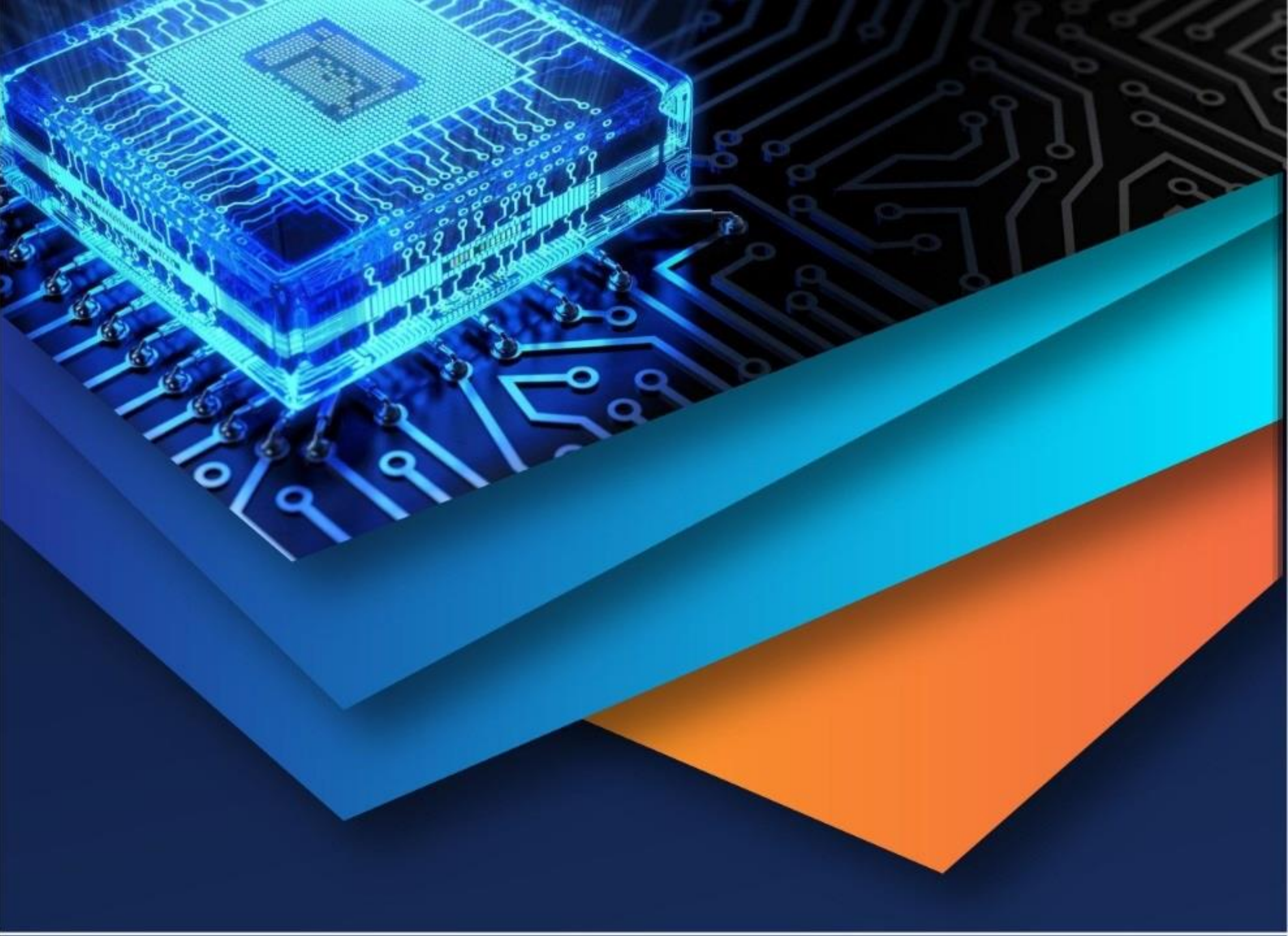
Future research will expand mobile edge inferencing for in-situ field digitization and multi-spectral infrared paleochronometry.

VI. ACKNOWLEDGMENT

The author gratefully acknowledges the National Mission for Manuscripts (NMM), New Delhi, the Kerala State Archives Department, and the Oriental Research Institute & Manuscripts Library, University of Kerala, for providing archival access to historical palm-leaf manuscript folios.

REFERENCES

[1] K. V. Sarma, Manuscriptology and Textual Criticism in Medieval South India, Hoshiarpur: Vishveshvaranand Vedic Research Institute, 2004.
[2] R. S. Sharma, "Palm-leaf manuscript preservation techniques in tropical India," Studies in Conservation, vol. 38, no. 2, pp. 102–114, 1993.
[3] A. Antonacopoulos and C. Clausner, "Palaeographic document processing: Challenges and opportunities," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pp. 1821–1835, Aug. 2013.
[4] P. B. Suryavanshi and D. V. Jadhav, "3D stylus groove depth recovery in palm leaf manuscripts using photometric stereo," Pattern Recognition Letters, vol. 112, pp. 188–195, Sep. 2018.
[5] M. Cheriet, F. Kharrazi, and C. Y. Suen, "Degraded historical document processing," in Character and Image Recognition, World Scientific, 2007, pp. 245–278.
[6] S. R. Namboothiri, Grantha and Vatteluttu Epigraphy: Evolution of Malayalam Orthography, Trivandrum: Kerala Sahitya Akademi, 2011.
[7] R. Smith, "An overview of the Tesseract OCR engine," in Proc. Ninth Int. Conf. Document Analysis and Recognition (ICDAR), Curitiba, Brazil, 2007, pp. 629–633.
[8] N. Otsu, "A threshold selection method from gray-level histograms," IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62–66, Jan. 1979.
[9] W. Niblack, An Introduction to Digital Image Processing, Englewood Cliffs, NJ: Prentice-Hall, 1986.
[10] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," Pattern Recognition, vol. 33, no. 2, pp. 225–236, Feb. 2000.
[11] C. Wolf and J. M. Jolion, "Extraction of text from natural images using local features," Machine Vision and Applications, vol. 14, no. 4, pp. 221–228, 2003.
[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
[13] R. Kesavan and P. Radhakrishnan, "Indic script optical character recognition: A comprehensive survey," ACM Computing Surveys, vol. 54, no. 3, pp. 1–36, Apr. 2021.
[14] K. R. Namboothiri, "Machine reading of Grantha palm-leaf manuscripts using convolutional neural lattices," IEEE Access, vol. 9, pp. 45120–45132, 2021.
[15] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, Apr. 2004.



10.22214/IJRASET

INDEX COPERNICUS
45.98

IMPACT FACTOR:
7.129

TOGETHER WE REACH THE GOAL
IMPACT FACTOR:
7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY