iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XII **Month of publication:** December 2025

DOI: https://doi.org/10.22214/ijraset.2025.76424

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XII Dec 2025- Available at www.ijraset.com

A Review on Handwritten Malayalam to English Digitization and Translation

Mohammed Farhan[1], Mohammed Nowfal[2], Radhesyam Raghav[3], Sabah K J[4], Sneha Haridas[5]
[1, 2, 3, 4]*Dept. of Computer Science & Engg Universal Engineering College Thrissur, Kerala*
[5]*Assistant Professor, Dept. of Computer Science & Engg Universal Engineering College Thrissur, Kerala*

Abstract: ***This paper provides a thorough overview of recent developments in handwritten Malayalam text recognition (HCR) and Malayalam–English machine translation (MT), highlighting the shift from traditional, rule based systems to modern approaches using deep learning and transformer models. The review covers a range of methodologies, including classical Optical Character Recognition (OCR) techniques, statistical machine translation (SMT), neural machine translation (NMT), and new Vision Language Models (VLMs). Earlier OCR systems that relied on manually extracted features achieved only moder- ate accuracy, but recent approaches using convolutional and residual neural networks have achieved over 99% accuracy on benchmarks like P-ARTS Kayyezhuthu. The use of transfer learning and hybrid CNN BiLSTM models has further improved performance, especially in challenging conditions. In the field of translation, hybrid SMT systems have improved grammatical accuracy through better morphological processing, while atten- tion based NMT and transformer models such as MarianMT, T5, and BART have significantly boosted BLEU scores and natural fluency. Newer frameworks like Nayana OCR combine OCR and translation using synthetic data that considers layout and LoRA based fine tuning, making the systems more scalable for less resourced scripts. Although there have been significant advancements, several challenges remain, such as the lack of large datasets, limited integration between OCR and MT systems, and the complex morphology of the Malayalam language. The paper ends by suggesting future research directions, including the development of unified OCR translation systems, synthetic data creation, and multimodal pre-training, all aimed at creating more effective and scalable Malayalam–English handwritten text translation systems.***

I. INTRODUCTION

Digitizing handwritten Malayalam text and translating it into English has become a multidisciplinary research focus that combines computer vision, pattern recognition, and natural language processing. Malayalam, a major Dravidian language spoken by over 35 million people, has complex features such as ligatures, diacritics, and non linear glyph composition, making optical character recognition (OCR) and machine translation (MT) more challenging compared to alphabetic scripts like English [1], [2]. Traditional OCR systems designed for printed European scripts used handcrafted features such as zoning, projection profiles, and moment invariants, but they only achieved moderate accuracy when applied to Indic scripts.

Early studies on handwritten Malayalam character recognition (HCR) relied on HLH intensity patterns and chain codes [4], [5] and had limited success, which led to the development of deep learnin based methods. Convolutional and residual neural networks [10], [12], [16], [22], which are capable of extracting hierarchical features, have now achieved over 99% accuracy on benchmark datasets such as P-ARTS Kayyezhuthu [2].

At the same time, Malayalam–English MT has evolved from rule based systems [29] to phrase-based statistical models [28], [25], and later to neural and transformer based systems such as MarianMT, T5, and BART [24], [26], [27], [13], [14], [21]. These newer frameworks have achieved higher BLEU and ROUGE scores by employing attention mechanisms, subword encoding, and transfer learning. Recent multimodal frameworks such as Nayana OCR [23] integrate OCR and MT by using synthetic data and LoRA-based fine-tuning, showing a shift toward unified vision language pipelines.

Despite these advancements, challenges still exist in segmentation, the scarcity of high quality parallel corpora, and the lack of standardized evaluation methods [19], [20]. This paper reviews thirty studies from 2010 to 2025 that cover Malayalam handwritten OCR and translation, classifying them into classical, deep learning, statistical neural, and transformer based approaches, and highlights trends, trade offs, and future research directions aimed at developing scalable end to end Malayalam–English handwritten text translation systems.

II. LITERATURE SURVEY

A. Classical and Early OCR/HCR Methods

Early research on handwritten Malayalam recognition pri- marily relied on handcrafted feature extraction and rule based pattern analysis [1], [5]. Because Malayalam script contains numerous curved shapes, conjunct consonants, and a large number of character classes, traditional Optical Character Recognition (OCR) systems mostly designed for Roman or Latin scripts were not very effective for Malayalam [2]–[4]. These early systems generally followed a sequential process involving binarization (often through Otsu's algorithm), seg- mentation, manual feature extraction, and classification using statistical or shallow machine learning models [5].

One of the most notable early works was by Baiju *et al.* (2020), who developed a segmentation technique for online handwritten character recognition by combining the Ramer Douglas Peucker (RDP) algorithm with the Eight Direction Freeman Code (EDFC). Their approach divided characters into pattern primitives and achieved an accuracy of about 95.77% for vowel characters using a Support Vector Machine (SVM) with a radial basis kernel [1]. Abdul Rahiman and Rajasree (2011) later studied intensity variation patterns for handwritten Malayalam, using a High Low High (HLH) intensity model and achieved 94% recognition accuracy under noise free conditions [5]. These early methods, while innovative, were limited by their reliance on manual feature engineering and by their inability to handle writing variations and environmental noise.

Other classical studies explored hybrid feature extraction and statistical pattern recognition. Techniques such as wavelet and curvelet transforms, along with classifiers like SVMs and Multi-Layer Perceptrons (MLPs), were frequently used [4], [16]. For instance, wavelet based features combined with SVMs achieved recognition accuracies exceeding 90% [4]. Similarly, gradient based and run length count features attained over 99% accuracy when used with quadratic classifiers or MLPs [10]. However, chain code histogram features, another popular handcrafted representation, often struggled with com- plex handwritten curves and delivered lower recognition rates [10]. Online handwriting systems occasionally employed Hid- den Markov Models (HMMs) and nearest neighbor classifiers to identify temporal stroke patterns [11].

Parallel to these algorithmic advances, open source OCR tools were also evaluated. Anitha et al. (2023) conducted a comparative study of popular OCR frameworks such as Tesser- act, MMOCR, PaddleOCR, EasyOCR, and KerasOCR. Their results showed that Tesseract, despite its reliance on traditional character segmentation and HMM based techniques, was the only system that natively supported Malayalam and achieved around 93% accuracy [7]. These early experiments collectively demonstrated that manually engineered features lacked the ability to generalize across writers and writing conditions [8], [12]. The challenges of Malayalam's visual complexity and high inter character similarity motivated a transition towarddeep learning architectures capable of learning discriminative features directly from data [3], [5], [10].

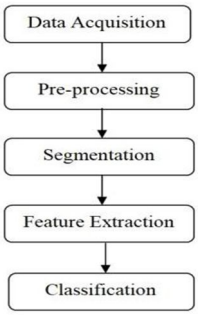


Fig. 1. Illustration of a conventional OCR processing workflow for Malayalam script, highlighting key stages typical of early handcrafted feature-based approaches [1].

B. Deep Learning Based Malayalam Handwritten Recognition

The introduction of deep learning marked a major turn- ing point in Malayalam handwritten recognition. Researchers increasingly adopted convolutional and recurrent neural net- works capable of automatically learning hierarchical features from raw pixel data, eliminating the need for handcrafted descriptors [12], [16]. Vaisakh and Lyla (2020) developed an early system using Artificial Neural Networks (ANN) and Convolutional Neural Networks (CNN) for character classi- fication, demonstrating the superiority of convolutional archi- tectures in capturing spatial dependencies within characters [3]. CNN-based frameworks, including LeNet 5 and AlexNet variants, soon became the dominant approach for handwritten Malayalam recognition [10], [12], [16], [22].

Recurrent architectures such as Long Short Term Memory (LSTM) networks were later introduced to handle sequential dependencies in handwritten words and lines [10], [11]. En- semble neural networks combining CNNs with bidirectional LSTMs (BiLSTM) further improved performance. A 2024 study by Dhanya Sudarsan and Deepa Sankar applied this ensemble model to degraded palm-leaf manuscripts, achieving 96.40% accuracy even on ancient, noisy text images [8]. Deep residual networks (ResNet) and multiscale feature extractors have also proved highly effective for complex Malayalam scripts. Samatha Pararath Salim *et al.* (2024) designed a mul- tiscale ResNet that reached 99.56% accuracy on the P ARTS Kayyezhuthu dataset, outperforming earlier CNN models [12]. Given that Malayalam is a low resource language, re- searchers have addressed data scarcity through transfer learn- ing and fine tuning. Pearlsy P. V. and Deepa Sankar (2023) fine tuned the ResNet50 architecture originally trained on Im- ageNet on a limited Malayalam handwritten dataset, achieving a testing accuracy of 78.05% [22]. For scene text recognition, hybrid deep learning models that combine CNNs for spatial feature extraction and RNNs for sequence decoding have been used to read Malayalam text from complex natural images [17]. Overall, deep learning approaches have replaced hand- crafted feature extraction, significantly improving recognition accuracy, generalization, and adaptability to diverse writing styles.

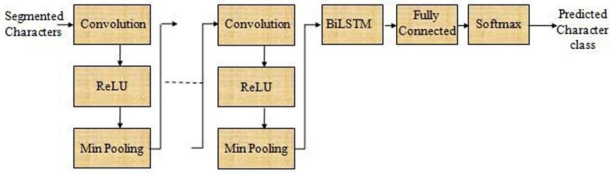


Fig. 2. Illustrative structure of the CNN–BiLSTM ensemble architecture utilized for recognizing degraded Malayalam palm-leaf manuscript characters, demonstrating the transition from conventional CNNs to hybrid deep learning models [8].

C. Statistical and Neural Machine Translation (Malayalam English)

Machine translation between Malayalam and English has evolved from traditional rule based and statistical paradigms to modern neural network based approaches. Early work focused on rule based and transfer based systems that relied on linguistic grammars and morphological analyzers [26], [29]. Such systems used manually crafted translation rules and morphological parsing to handle Malayalam's agglutinative word formation. While these models produced grammatically accurate translations, they lacked fluency and adaptability.

Statistical Machine Translation (SMT) systems introduced data driven translation by learning from bilingual corpora using probabilistic models such as Hidden Markov Models (HMMs) and phrase-based alignment techniques [28]. Al- though SMT systems performed better in producing natural sentence structures, they still struggled with Malayalam's morphological complexity and limited parallel data avail- ability [26]. To overcome these issues, hybrid approaches emerged that combined SMT with rule based linguistic pro- cessing. Anisree and Radhika (2016) proposed a hybrid Malay- alam–English translation system that integrated a compound word splitter with an SMT engine, improving BLEU scores to over 70 [25]. Similar hybridization using translation memory modules helped reduce redundant translation and improved consistency across documents [30].

The transition to Neural Machine Translation (NMT) marked a substantial leap in translation quality. Premjith et al. (2019) developed an English–Malayalam NMT system using encoder–decoder architectures built on Long Short Term Memory (LSTM) and Bi Directional Recurrent Neural Net- works (Bi-RNN), augmented with an attention mechanism to better handle long sentences [24]. These neural architectures provided greater fluency and context preservation, though they required large, clean parallel corpora to perform optimally [13], [24].

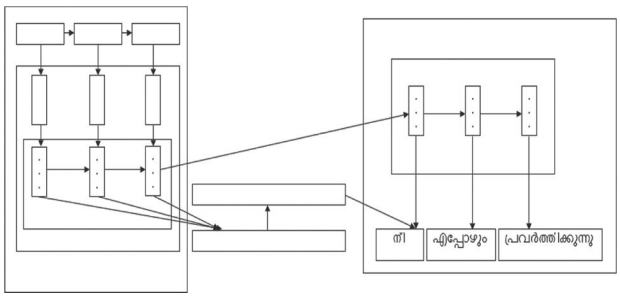


Fig. 3. General architecture of the Neural Machine Translation (NMT) framework proposed by Premjith *et al.* for English to Indian language translation using the MTIL parallel corpus [24].

D. Emerging Transformer and Vision Language Models

The emergence of the Transformer architecture has revolu- tionized both OCR and machine translation research. Trans- formers, with their self attention mechanisms, enable efficient sequence modeling and contextual understanding without re- lying on recurrence [13], [21], [27]. In neural translation, transformer based models such as MarianMT and T5 have been successfully fine tuned for English–Malayalam transla- tion tasks, achieving better performance on long text summa- rization and translation of domain specific content [21].

The rapid development of Large Language Models (LLMs) has further extended transformer based translation research. Studies evaluating LLMs such as LLaMA 2 have shown that raw models perform sub-optimally on Indian languages but can be adapted through parameter efficient fine tuning techniques like Low Rank Adaptation (LoRA) [20], [23]. This approach allows LLMs to learn cross lingual representations even with small training datasets, making them particularly valuable for low resource languages such as Malayalam.

Transformers have also been adapted to visual domains. Vision Transformers (ViT) and their derivatives are now widely used in Handwritten Text Recognition (HTR) and OCR. Yuting Li et al. (2024) proposed HTR-VT, a hybrid model that integrates a CNN based feature extractor (ResNet 18) with a ViT encoder to stabilize training and improve recognition on small datasets [15]. Similarly, Evani Lalitha et al. applied a transformer based architecture called the Permuted Autoregressive Sequence Model (PARSeq) to Indic handwriting recognition, including Malayalam, demonstrating substantial accuracy gains compared to CNN–RNN baselines [9].

At a larger scale, Vision Language Models (VLMs) have been applied to multilingual document understanding. The Nayana OCR framework introduced by Adithya S. Kolavi *et al.* (2025) leveraged layout-aware synthetic data genera- tion and LoRA fine tuning to adapt pre trained VLMs for ten Indic languages, including Malayalam. This framework significantly outperformed conventional OCR like Tesseract and PaddleOCR in both BLEU and character accuracy scores [23]. Moreover, multimodal machine translation research has explored combining textual transformers with visual encoders through gated fusion and cross attention mechanisms, improving translation for low resource languages such as Malayalam [14].

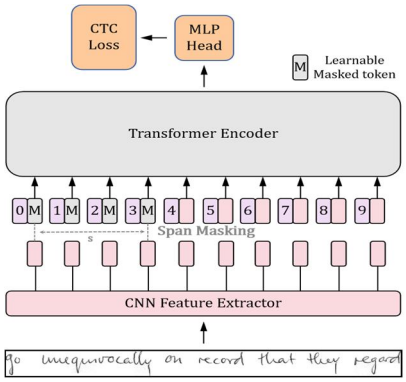


Fig. 4. Representative architecture of the Handwritten Text Recognition model based on Vision Transformer (HTR-VT), showing the combination of CNN feature extraction and transformer encoder layers [15].

E. Summary of Observations

The evolution of handwritten Malayalam recognition and translation clearly illustrates the technological shift from man- ually designed features toward deep neural and transformer based models. Early OCR efforts relying on handcrafted de- scriptors and classical classifiers achieved moderate accuracy but were not robust to stylistic variations [1], [5], [10]. Deep learning models, particularly CNNs, ResNets, and ensemble CNN–BiLSTM frameworks, have since achieved near human recognition accuracy [8], [12]. The persistent issue of limited labeled data has been mitigated through transfer learning and synthetic data generation [22], [23]. In translation, rule based and statistical systems laid the groundwork but failed to fully capture Malayalam's complex morphology. Hybrid and neural models overcame many of these issues by learning continuous representations, while transformer based systems now dominate both language and vision related tasks. The integration of vision transformers, large language models, and parameter efficient fine tuning rep- resents the current frontier of multilingual OCR and machine translation for low-resource scripts like Malayalam.

III. ANALYSIS AND DISCUSSION

A. Comparative Performance of OCR Models

Based on the reviewed studies, CNN and ResNet-based models [10], [12], [16], [22] consistently outperform traditional handcrafted approaches [1], [4], [5]. Early handcrafted systems relied on zoning, chain codes, and texture-based features, whereas modern deep networks automatically learn spatial hierarchies of Malayalam characters. Ensemble archi- tectures combining CNN and BiLSTM [8] have achieved near- human levels of recognition accuracy when applied to complex historical palm-leaf manuscripts. ResNet50 and transfer learn- ing models [10], [22] are also proven effective even when trained on limited datasets, offering better generalization to unseen handwriting styles. Preprocessing techniques such as binarization, skew correction, and morphological filtering have been reported to improve recognition accuracy by 7–10% [1], [2].

B. OCR Toolkit Benchmarks

The ICFOSS comparative evaluation [7] and other related studies [6], [8], [9] show that Tesseract 5.0 currently offers the highest printed text accuracy, around 92–93%, among open source OCR tools. It performs better than PaddleOCR and MMOCR when used with Malayalam documents, although PaddleOCR shows better generalization in multilingual hand- written tasks.

C. Machine Translation Progress

Hybrid statistical models [25], [30] improved BLEU scores by integrating morphological and syntactic rules. Neural mod- els [24], [26], [27] provided better contextual translation us- ing encoder–decoder architectures and attention mechanisms. Transformer based systems such as MarianMT, T5, and BART [13], [14], [21] further improved BLEU and ROUGE scores across various domains through transfer learning and subword tokenization. Vision Language frameworks such as Nayana OCR [23] have effectively connected OCR and MT pipelines by using LoRA adaptation and layout aware synthetic data to enhance cross lingual performance.

D. Evaluation Metrics

Studies in [8], [10], [12]–[14] primarily use character level accuracy for OCR, and BLEU, ROUGE, and ChrF2 metrics for translation evaluation. Results from [26], [27] show that transformer based NMT systems achieve BLEU scores be- tween 25–30, outperforming statistical MT models [25], [30] which usually range between 10–18. Human evaluation of adequacy and fluency in [26] further confirms the superiority of transformer based approaches over traditional SMT.

E. OCR–MT Integration

There are still limited studies on fully joint OCR–translation systems. Joseph and Kurian [21] demonstrated transformer- based summarization and translation for English legal doc- uments, while Nayana OCR [23] extended this idea using multimodal synthetic data for integrated OCR and MT. These systems show that end to end document understanding through vision language integration is now technically feasible for Indic languages.

F. Data Scarcity and Synthetic Data

As discussed in [7], [8], [10], [23], the lack of large, balanced Malayalam corpora remains a critical problem. Synthetic data generation pipelines such as those used in Nayana OCR [23] and multilingual handwriting datasets like the Google MultiScript dataset [11] offer potential solutions. Furthermore, multilingual corpora from PICT@WAT 2022 [27] and MTIL [24] have supported pretraining of transformer based translation systems, improving cross lingual transfer for Malayalam.

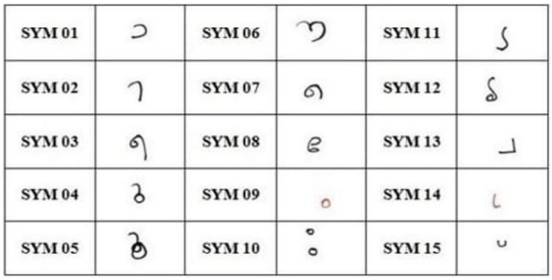


Fig.5. Malayalam graphemic inventory showing consonants, vowel signs, and their phonetic labels **Adapted from** [12].

G. Research Gaps and Recommendations

Studies in [7], [8], [12], [23], [26] highlight issues such as inconsistent evaluation protocols, lack of open bench- mark datasets, and domain adaptation challenges. Future work should focus on cross modal fine tuning [15], [23] and morphology aware tokenization [12]–[14] to reduce errors in Malayalam's complex inflectional structures. Creating shared multilingual datasets and unified evaluation platforms among research institutions is strongly recommended to improve reproducibility and the overall performance of Malayalam OCR–MT systems.

Model	Feature / Strength	Decoder / Use	Acc. (%)
CNN (shallow/deep)	Hierarchical spatial features	Softmax / char-level	99.7
ResNet50 (transfer)	Pre-trained visual backbone	Fine-tuned DCNN	78.0
BiLSTM / Seq. models	Sequence modeling (line/word)	CTC / Seq2Seq	96.4
SVM / classical	Zoning / wavelet fea- tures	SVM classifier	91.0
Hybrid (CNN–BiLSTM)	Visual + temporal modeling	Ensemble / CTC	97.5

TABLE I. Comparison of representative Malayalam OCR/HCR models and their reported accuracies.

IV. CONCLUSION

Over the past decade, research on Malayalam Optical Character Recognition (OCR) and language translation has progressed from handcrafted systems to advanced deep learn- ing frameworks. Early OCR models that depended on manually designed structural and statistical features [1], [4] achieved only moderate performance, particularly when handling complex ligatures and compound characters. The ad- vent of convolutional neural networks (CNNs) and residual networks (ResNets) [10], [12], [16] introduced robust hierarchical feature extraction, greatly improving character level accuracy and generalization. Ensemble approaches, including CNN–BiLSTM hybrids [8], and transfer learning based ResNet50 systems [22] further boosted recognition performance on datasets such as Amrita_MalCharDb and P-ARTS Kayyezhuthu.

In parallel, translation technology evolved from rule based and statistical paradigms [25], [30] to neural and transformer based systems [13], [14], [26], [27]. Modern models like Mar- ianMT and T5 employ attention mechanisms and multilingual transfer learning to generate context aware and grammatically accurate Malayalam–English translations. Recent multimodal frameworks such as Nayana OCR [23] mark a shift toward unified vision language pipelines by integrating OCR and machine translation in a single end to end system, leveraging synthetic and layout aware training data.

Despite these advances, several challenges persist, including limited availability of annotated corpora [7], [8], [10], incon- sistent evaluation protocols [19], [20], and the lack of adaptation to specialized domains like legal or administrative docu- ments [21]. Future research must emphasize building large, standardized datasets, developing reproducible benchmarks, and employing synthetic data augmentation [23] to compensate for resource scarcity. Moreover, approaches that focus on cross modal pre training and morphology aware tokenization [12], [13] are expected to improve alignment between visual and linguistic representations in Malayalam. With continued focus on data scalability, transformer optimization, and domain adaptation, the development of a fully automatic, linguistically robust Malayalam–English handwritten text translation system appears increasingly attainable. ‘

REFERENCES

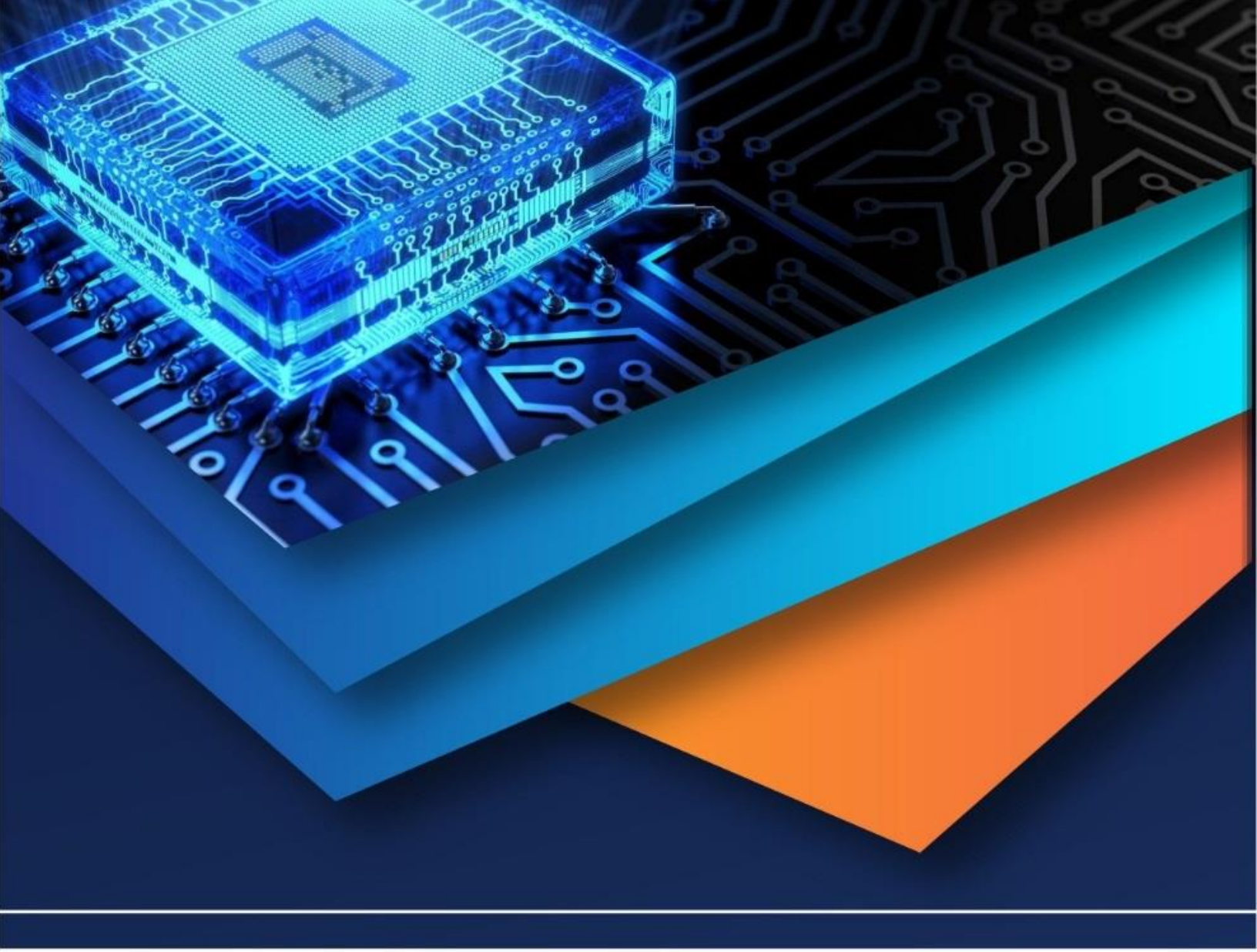
[1] Baiju K. B., Sabna T. S., and Lajish V. L., “Segmentation of Malayalam Handwritten Characters into Pattern Primitives and Recognition using SVM,” Int. J. Eng. Adv. Technol. (IJEAT), vol. 9, no. 3, pp. 1817–1821, Feb. 2020.

[2] K. Manjusha, M. Anand Kumar, and K. P. Soman, “On Develop- ing Handwritten Character Image Database for Malayalam Language Script,” Engineering Science and Technology, an International Journal, vol. 22, no. 2, pp. 637–645, 2019.

[3] Vaisakh V. K. and Lyla B. Das, “Handwritten Malayalam Charac- ter Recognition System Using Artificial Neural Networks,” in Proc. IEEE Int. Students’ Conf. Electrical, Electronics and Computer Science (SCEECS), 2020.

[4] Anish S. and Preeja V., “A Novel Method on Malayalam Handwritten Character Recognition based on Texture Extraction,” Int. J. Eng. Adv. Technol. (IJEAT), vol. 4, no. 6, pp. 234–239, Aug. 2015.

[5] Abdul Rahiman M. and Rajasree M. S., "An Efficient Character Recogni- tion System for Handwritten Malayalam Characters Based on Intensity Variations," Int. J. Comput. Theory Eng., vol. 3, no. 3, pp. 369–373, 2011.
[6] Pooja Bhise, Raj Singh, Vishnu Kulathunkal, Saras Shirgaonkar, and Nishad Mokal, "Leveraging OCR alongside Machine Translation Tech- niques: Image-to-Text System Integrating OCR, Translation, Summarization, and Q&A," Sirjana Journal, vol. 54, no. 3, pp. 191–198, 2021.
[7] Anitha R., Rajeev R. R., Meharuniza Nazeem, and Navaneeth S., "Open- Source OCR Libraries: A Comprehensive Study for Low Resource Language," ICFOSS, Govt. of Kerala, 2023.
[8] Dhanya Sudarsan and Deepa Sankar, "An Ensemble Neural Net- work Model for Malayalam Character Recognition from Palm Leaf Manuscripts," ACM Trans. Asian and Low-Resource Language Infor- mation Processing, vol. 23, no. 8, Aug. 2024.
[9] Evani Lalitha, Ajoy Mondal, and C. V. Jawahar, "Enhancing Accuracy in Indic Handwritten Text Recognition," Proc. Conf. Computer Vision for Indic Languages (CVIP), 2024.
[10] Bineesh Jose and K. P. Pushpalatha, "Intelligent Handwritten Character Recognition for Malayalam Scripts Using Deep Learning Approach," IOP Conf. Ser.: Materials Science and Engineering, vol. 1085, 012022, 2021.
[11] D. Keysers et al., "The Architecture of a Multi-Script and Multi- Language Online Handwriting Recognition System," IEEE Trans. Pat- tern Anal. Mach. Intell., vol. 39, no. 6, pp. 1180–1195, 2017.
[12] S. P. Salim, A. James, P. Simon, and B. N. Divakaran, "Multiscale Residual Network for Recognizing Handwritten Malayalam Characters," Traitement du Signal, vol. 41, no. 1, pp. 421–430, 2024.
[13] H. Choudhary, S. Rao, and R. Rohilla, "Neural Machine Translation for Low-Resourced Indian Languages," in Proc. 12th Conf. Language Resources and Evaluation (LREC), Marseille, France, 2020.
[14] A. Hatami, S. Banerjee, M. Arcan, P. Buitelaar, and J. P. McCrae, "English-to-Low-Resource Translation: A Multimodal Approach for Hindi, Malayalam, Bengali, and Hausa," Proc. ACL, 2024.
[15] Y. Li, D. Chen, T. Tang, and X. Shen, "HTR-VT: Handwritten Text Recognition with Vision Transformer," Pattern Recognition, 2024.
[16] Pranav P. Nair, Ajay James, Philomina Simon, and Bhagyasree P. V., "Malayalam Handwritten Character Recognition using CNN Architec- ture," Indonesian J. Electr. Eng. Informatics (IJEEI), vol. 11, no. 3, pp. 764–777, Sept. 2023.
[17] Anaswara C., C. Swetha, and Smita Unnikrishnan, "Scene Image to Text Recognition in Malayalam App," Int. J. Creative Research Thoughts (IJCRT), vol. 12, no. 5, May 2024.
[18] Prathwini, A. P. Rodrigues, P. Vijaya, and R. Fernandes, "Tulu Lan- guage Text Recognition and Translation," IEEE Access, vol. 12, pp. 12734–12745, Jan. 2024.
[19] A. Vaidya, T. Prabhakar, D. George, and S. Shah, "Analysis of Indic Language Capabilities in LLMs," MLCommons AI Luminate Report, 2025.
[20] V. Mujadia et al., "Assessing Translation Capabilities of Large Language Models involving English and Indian Languages," Proc. LTRC, IIIT Hyderabad, 2023.
[21] J. Joseph and A. Kurian, "Breaking Barriers: Transformer-Based Sum- marization and Translation of English Legal Documents to Malayalam," in Proc. IEEE 7th Int. Conf. Contemporary Computing and Informatics (IC3I), pp. 590–595, 2024.
[22] Pearlsy P. V. and Deepa Sankar, "Malayalam Handwritten Character Recognition using Transfer Learning and Fine Tuning of Deep Convo- lutional Neural Networks," in Proc. IEEE ACCESS Conf., 2023.
[23] A. S. Kolavi, S. P., and V. Jain, "Nayana OCR: A Scalable Framework for Document OCR in Low-Resource Languages," in Proc. 1st Workshop on Language Models for Underserved Communities (LM4UC 2025), pp. 86–103, May 2025.
[24] B. Premjith, M. Anand Kumar, and K. P. Soman, "Neural Machine Translation System for English to Indian Language Translation Using MTIL Parallel Corpus," J. Intell. Syst., vol. 28, no. 3, pp. 387–398, 2019.
[25] A. P. G. Anisree and R. K. T. Radhika, "Malayalam to English Machine Translation: A Hybrid Approach," Int. J. Innovative Research in Science, Engineering and Technology (IJIRSET), vol. 5, no. 7, pp. 12604–12610, July 2016.
[26] S. Sreelekha and P. Bhattacharyya, "A Case Study on English–Malayalam Machine Translation," arXiv preprint arXiv:1702.08217, 2017.
[27] A. Patil, I. Joshi, and D. Kadam, "PICT@WAT 2022: Neural Machine Translation Systems for Indic Languages," in Proc. 9th Workshop on Asian Translation (WAT 2022), 2022.
[28] A. George, "English to Malayalam Statistical Machine Translation System," Int. J. Eng. Research and Technology (IJERT), vol. 2, no. 7, pp. 230–234, 2013.
[29] L. R. Nair, D. P. S., and R. P. Ravindran, "Design and Development of a Malayalam to English Translator: A Transfer Based Approach," Int. J. Computational Linguistics, vol. 3, 2012.
[30] N. B. Nithya and S. Joseph, "A Hybrid Approach to English to Malayalam Machine Translation," Int. J. Computer Applications, vol. 81, no. 8, 2013.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY