IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A FRAMEWORK FOR TAMIL CHARACTER RECOGNITION FROM STONE INSCRIPTIONS

[1]**Dr. G. Krishnapriya**, Assistant Professor,

[2] **Mr. E. Raghuvaran**, Research Scholar, Department of Computer Science,

School of Engineering and Technology, Pondicherry University,

Abstract: Segmenting ancient Tamil characters from stone inscriptions presents a significant challenge due to the complex structure of the script and the degraded nature of historical artifacts. Tamil, one of the world's oldest languages, has evolved through various scripts including Sangakala Tamil, Karoshti, Brahmi, Vatteluttu, Grantha, and the modern script, resulting in a rich and diverse character set comprising 247 unique symbols. These characters often contain loops, curves, and holes, making segmentation from stone surfaces particularly difficult. Traditional image processing techniques struggle to address issues such as uneven lighting, surface erosion, and similar foreground–background textures. Accurate character recognition heavily depends on the success of the segmentation phase, which isolates meaningful character regions from noisy backgrounds. This paper presents an in-depth analysis of segmentation techniques applied to Tamil stone inscription images, covering preprocessing methods, segmentation strategies, and recent advancements incorporating deep learning approaches. By emphasizing segmentation as the foundation for reliable character recognition, this study aims to guide future research in developing more robust and accurate Tamil epigraphy analysis systems.

Keywords — Image Processing, Preprocessing, Feature Extraction, Classification, Optical Character Recognition, Convolution Neural Network.

I. INTRODUCTION

Inscriptions and manuscripts are important sources of information for understanding the history and culture of ancient civilizations [1]. Inscriptions are ubiquitous in ancient Indian temples and can be found on a variety of surfaces — including rocks, slabs, pillars, and temple walls. Most of these inscriptions convey vast and valuable monarchical information related to administrative, cultural, and religious processes. They serve as documented proof, offering deeper insights into the quality of life during specific historical periods [2]. Tamil Nadu tops the list in the Survey of Indian Epigraphy (1996), indicating the state's significantly high number of ancient inscriptions. Tamil is one of the oldest languages in the world and among the earliest in the Indian subcontinent. Over centuries, it has evolved through various character sets, such as Sangakala Tamil, Brahmi, and others.

Stone inscriptions found across different regions often reveal details of the economy, culture, lifestyle, and governance of various rulers and dynasties of those eras [3]. The Tamil script includes 12 vowels, 18 consonants, and one special character classified as both a vowel and consonant, forming a total of 31 basic letters and 216 compound letters. Although numerous research efforts have been made toward the recognition of Tamil characters, high accuracy remains elusive. Moreover, most studies focus only on a limited subset of the character set. The major challenge lies in the complexity of Tamil characters, which feature intricate curves, multiple strokes, holes, and structural variations [4].

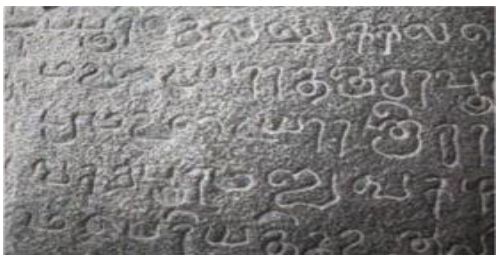

Fig 1: Sample Stone Inscription

A crucial step in improving recognition accuracy is effective character segmentation, which involves isolating individual characters from the noisy background of degraded stone surfaces. Accurate segmentation is foundational for any Optical Character Recognition (OCR) system, as errors at this stage significantly affect downstream processes such as feature extraction and classification. Therefore, several recent approaches have focused on advanced segmentation techniques — including traditional image processing and deep learning-based methods — to accurately separate character regions from the complex textures of stone inscriptions. The ultimate objective of Tamil ancient character recognition is to support the development of a robust OCR system capable of digitizing inscriptions found on South Indian temple walls, particularly those constructed between the 7th and 12th centuries.

Tamil inscriptions found on temple walls today predominantly date back to the Chola dynasty. Prominent rulers such as Rajaraja Chola I, Rajendra Chola I, and Kulothunga Chola I played significant roles during this era. Under the Chola rule, the Tamil kingdoms flourished in art, religion, music, and literature. Despite the cultural richness of these inscriptions, a vast number remain illegible and untranslated into modern scripts. According to deciphered inscriptions, the content often includes details related to wages, employment records, and regulations for conducting ceremonial offerings within temples [5].

A major technical obstacle in processing these ancient inscriptions is the degradation and inconsistency in the carving process itself. The hand-carved letters were often imprecise, misaligned, and distorted, leading to irregular spacing and uneven baselines. These challenges introduce significant noise during image acquisition and severely impact the segmentation process, which is essential for isolating individual characters from complex backgrounds. Segmentation in such scenarios becomes difficult due to erosion, overlapping characters, surface cracks, and non-uniform illumination—making it hard to distinguish useful textual regions from irrelevant background textures.

To address these issues, the proposed model leverages deep neural networks to optimize the segmentation and recognition of ancient Tamil characters from stone inscriptions. By integrating deep learning-based segmentation approaches—such as convolutional architectures capable of learning spatial hierarchies—this model aims to accurately separate character regions before classification. Effective segmentation not only enhances recognition accuracy but also contributes significantly to the digital preservation and translation of ancient Tamil epigraphic records.

I1. Related Work

Most of the existing approaches that are used for ancient Tamil character recognition has been studied and reviewed for identifying the drawbacks that are attained from the existing OCR techniques. Some of the existing model used for the ancient Tamil inscription character recognition have been surveyed below.

Prehistoric Stone Image Tamil Character Recognition using Optimized Deep Neural Network using Zernike Moments and Simplex Method has been developed by Prabavathi, R [6]. This system takes a challenge to recognize prehistoric Tamil characters using Deep Neural Network. Pre-processing is done by Binarization, Denoising, character segmentation and size normalization for the stone images and then goes for Feature extraction which forms the basic underlying partof recognizing each character. Characters are then classified by Back Propagation deep Neural Network. The optimization of neural networks is done using simplex method during back propagation with improved data set.

Self-Adaptive Hybridized Lion Optimization Algorithm with Transfer Learning for Ancient Tamil Character Recognition in Stone Inscriptions has been presented by Priya, R. D. et al. [7]. In this system, the Self-Adaptive Lion Optimization Algorithm (SLOA) is applied to enhance the brightness and contrast of ancient stone inscription images. Following this, preprocessing steps such as noise removal are performed. A significant focus of the model is on character segmentation, which is achieved using contour detection techniques to isolate individual characters from the background. The segmented characters are then classified using Transfer Learning (TL), which utilizes a deep Convolutional Neural Network (CNN) for multiclass classification. The hybrid approach—combining image enhancement, segmentation, and deep learning—demonstrates improved accuracy and computational efficiency compared to traditional OCR systems.

Effective Tamil Character Recognition Using Supervised Machine Learning Algorithms has been developed by Suriya, S. [8]. This approach is designed to recognize Tamil characters in challenging conditions where traditional recognition systems often fail, particularly in the presence of blur, low contrast, low resolution, and high image noise. Although the primary focus is on classification using Convolutional Neural Networks (CNNs), the model inherently handles segmentation through the network's ability to learn spatial hierarchies and extract localized features. This enables the system to effectively differentiate and recognize individual characters, even with minimal explicit preprocessing or segmentation, thereby improving the robustness of recognition under adverse imaging conditions.

The system is designed to address challenges such as low contrast, low resolution, high image noise, and other distortions commonly found in ancient Tamil inscription images. It employs a Convolutional Neural Network (CNN), which is particularly effective at learning and extracting local features due to its use of limited receptive fields in the hidden layers. This characteristic makes CNNs inherently capable of performing implicit segmentation, as they can isolate and focus on distinct character regions even in noisy and degraded images. The network recognizes visual patterns directly from pixel-level data with minimal explicit preprocessing, eliminating the need for handcrafted segmentation rules. Through training, the CNN model learns to differentiate character boundaries, curves, and strokes—effectively performing segmentation as part of the feature extraction and classification pipeline. Experimental results show that the integrated segmentation and recognition capabilities of CNNs lead to high accuracy in classifying Tamil characters from complex stone inscriptions.

Recognizing Ancient South Indian Language Using Opposition-Based Grey Wolf Optimization (OGWO) was proposed by Kumar, A. N., and Geetha, G. [9]. This approach utilizes an Artificial Neural Network (ANN) optimized with the Opposition-Based Grey Wolf Optimization Algorithm (OGWO) to recognize ancient South Indian scripts. While the primary emphasis is on optimizing neural network parameters for better classification, character segmentation is an essential prerequisite in this system to ensure meaningful feature extraction. The model aims to identify ancient scripts, signs, and fonts where the selection of optimal weights in the ANN is critical for performance. Various metaheuristic algorithms, such as OGWO, Particle Swarm Optimization (PSO), and Grey Wolf Optimization (GWO), are applied to improve the learning efficiency of the ANN. The performance results demonstrate that the ANN-OGWO model achieves superior accuracy when effective segmentation is assumed as part of the preprocessing pipeline.

A Novel Approach to OCR Using Image Recognition-Based Classification for Ancient Tamil Inscriptions was developed by Giridhar, L., Dharani, A., and Guruviah, V. [10]. This work aims to improve OCR performance for ancient Tamil scripts, particularly those used between the 7th and 12th centuries. Given the difficulty in compiling a comprehensive dataset for such historical characters, a specialized dataset was created using cropped character images from temple inscriptions—implying a deliberate manual segmentation step in the data preparation phase. The images are binarized using Otsu's thresholding, which also aids in initial segmentation. A two-dimensional Convolutional Neural Network (2D-CNN) is then trained to classify the segmented characters. Furthermore, the CNN-based OCR system is integrated with the Tesseract OCR engine through the Pytesseract library in Python, creating a hybrid model for improved recognition. This method highlights how both manual and automated segmentation steps are crucial to the overall system accuracy and reliability.

In conclusion, various recently published techniques have been reviewed to understand the current trends in character segmentation for ancient Tamil inscriptions. The segmentation process plays a vital role across all models—whether through manual preprocessing, image binarization, contour detection, or deep learning-based spatial feature extraction. Accurate segmentation directly influences the success of OCR systems by enabling precise recognition of complex Tamil scripts from degraded and noisy stone inscription images.

III. ALGORITHM ANALYSIS FOR CHARACTER RECOGNITION

The recognition of ancient scripts such as Tamil stone inscriptions has historically depended on domain experts with deep knowledge of epigraphy and linguistic structures. These experts interpret characters manually, drawing from years of experience and contextual understanding. However, the scarcity of such professionals, combined with the physical deterioration of inscriptions—such as erosion, chipping, or overlapping strokes—makes manual interpretation difficult, time-consuming, and error-prone. To address these issues, researchers have turned to automated methods leveraging machine learning (ML) and deep learning (DL) techniques. For instance, soft computing techniques were applied by Vani and Ananthalakshmi to simplify the recognition of composite Tamil words. While these methods reduced complexity to an extent, their effectiveness was compromised by the inherent noise, character merging, and unclear boundaries present in ancient stone inscriptions, leading to reduced precision and inconsistent recognition.

While neural network-based models like Convolutional Neural Networks (CNNs), Artificial Neural Networks (ANNs), and Deep Neural Networks (DNNs) have demonstrated progress in recognizing characters, they face several critical limitations. Ineffective segmentation remains a primary concern; poor separation of characters leads to inaccurate feature extraction and increases the likelihood of misclassification. Additionally, many models struggle with high computational overhead and rely heavily on vast, annotated datasets that are difficult to obtain for ancient scripts. Learning loss and underfitting are common when the model fails to generalize across the diverse and degraded nature of inscription data. These issues collectively degrade overall recognition performance. To overcome these shortcomings, the paper proposes an optimized deep learning-based model that integrates a hybrid segmentation strategy. By combining traditional segmentation techniques with learning-based methods, this hybrid approach aims to improve character isolation, leading to more accurate recognition of ancient Tamil characters and addressing the core challenges present in earlier algorithms.

IV. FRAMEWORK FOR CHARACTER RECOGNITION

4.1 Preprocessing

Resampling (Image Scaling):

Resampling is the process of resizing the image to a standard dimension, which ensures consistency in input size for neural networks. Techniques such as bilinear interpolation and bicubic interpolation are widely used. Bilinear interpolation uses a weighted average of the four nearest pixels, while bicubic interpolation considers sixteen pixels and produces smoother transitions. Resampling helps reduce computational complexity and allows easier feature extraction in subsequent steps.

Pixel Restoration / Denoising:

Stone inscriptions often suffer from various types of noise due to surface erosion, dust, or capturing artifacts. To overcome this, denoising techniques are applied. Gaussian filtering is used to remove high-frequency noise by averaging pixel values in a Gaussian-weighted neighborhood. Median filtering is highly effective against salt-and-pepper noise and replaces each pixel value with the median of the surrounding pixels. More advanced techniques like Non-Local Means Denoising preserve edges and textures by averaging similar patches throughout the image rather than just local neighborhoods.

Binarization (Thresholding):

Binarization converts the grayscale image into black-and-white (binary) to distinguish the foreground characters from the background. Otsu's thresholding automatically selects a global threshold value that minimizes intra-class variance between foreground and background pixels. Adaptive thresholding, on the other hand, calculates thresholds locally, making it effective in conditions of uneven lighting. A hybrid thresholding strategy combines both Otsu and adaptive methods to produce more reliable binarization results in complex stone backgrounds.

4.2 Segmentation

Morphological Operations:
Morphological operations refine binary images by modifying their structure. Erosion removes small noise by shrinking the white regions, while dilation expands them to fill gaps. A combination of these operations—opening (erosion followed by dilation) and closing (dilation followed by erosion)—helps connect broken character components and remove irrelevant specks. These operations are vital for cleaning up the binarized image before extracting character boundaries.

Connected Component Labeling (CCL):
CCL scans the binary image to group connected pixels into labeled components. Each connected region is assumed to correspond to a character or a part of it. By assigning unique labels, this method isolates potential character regions for extraction. This labeling allows easy identification and cropping of each character for classification.

Contour Detection:
Contours represent the boundaries of connected components in an image. Using methods such as cv2.findContours() in OpenCV, the contours of characters are traced, and bounding boxes are drawn around them. These bounding boxes are then used to extract individual character patches. This technique is particularly useful when characters are visually separated, even in noisy or eroded images.

CNN-based Region Proposal Networks (RPN):
For complex segmentation, especially in overlapping or degraded texts, learning-based approaches like Region Proposal Networks are utilized. These networks, often built on CNN backbones (e.g., ResNet or VGG), slide over the image's feature maps and propose rectangular regions likely to contain characters. These regions are then refined or classified further, ensuring more accurate segmentation of distorted or touching characters.

U-Net for Character Segmentation:
The U-Net is a deep convolutional architecture designed for semantic segmentation. It consists of a **contracting path** (which captures context) and an **expanding path** (which enables precise localization). U-Net is effective for ancient Tamil inscriptions, as it can delineate character regions at a pixel level, even when they are faint or partially eroded.

Hybrid Optimized Segmentation Algorithm:
This approach integrates both traditional and deep learning-based techniques. Initially, the image is denoised and binarized using adaptive and Otsu thresholding. Then morphological filtering and contour detection are applied to extract rough character locations. These are further refined using CNN-based proposals to distinguish touching or overlapping characters. This hybrid strategy provides resilience against irregular line spacing and stone erosion.

4.3 Classification

Input Layer:
The input layer serves as the entry point of the CNN model. It accepts standardized image patches (e.g., 28×28 or 64×64 pixels) of the segmented characters. This layer doesn't perform any transformation but simply buffers and forwards the pixel data to the subsequent layers in the network.

Convolution Layer:
This layer applies several filters or kernels to the input image to extract features like edges, curves, and textures. Each filter slides over the image to produce a feature map that highlights specific patterns. These features are fundamental building blocks for understanding complex shapes in character images.

ReLU Activation:
The ReLU (Rectified Linear Unit) activation function is applied to the output of the convolution layer. It introduces non-linearity into the model by converting all negative pixel values to zero. This non-linearity allows the network to learn more complex and abstract representations of the input character data.

Pooling Layer:
Pooling layers reduce the spatial size of feature maps, thus decreasing computation while retaining important information. Max pooling, the most common type, extracts the highest value from each patch, effectively summarizing the most significant features. This step helps in achieving translational invariance and prevents overfitting.

Flattening and Fully Connected Layer:
After pooling, the feature maps are flattened into a one-dimensional vector and passed through one or more fully connected layers. These layers perform the final classification by combining extracted features and applying weights to predict the class of the character. The final layer uses a Softmax activation function to output a probability distribution over the character classes (e.g., Tamil alphabet set).

Fusion Architecture:

To further improve accuracy, a fusion model may be used, which combines CNN-extracted features with handcrafted features like SIFT or HOG. This multi-feature approach helps the system learn both abstract and geometric properties of characters, resulting in improved recognition performance, especially in historical datasets with high variance.

Tamil Stone Inscription Dataset

To facilitate learning, a dataset of ancient Tamil inscriptions was compiled, consisting of 28 classes with 1800 segmented character images. These images were collected from 8th-century inscriptions and labeled to match corresponding modern Tamil characters. The dataset is curated from the following public resource:

Siddharth Adevan V. (2023), Kaggle:
https://www.kaggle.com/datasets/siddharthadevanv/8th-century-tamil-inscriptions

The dataset supports the model in learning ancient character shapes and mapping them to modern equivalents, a key requirement for successful OCR performance.

V. RESULTS AND DISCUSSION

4.1 Results of Descriptive Statics of Study Variables

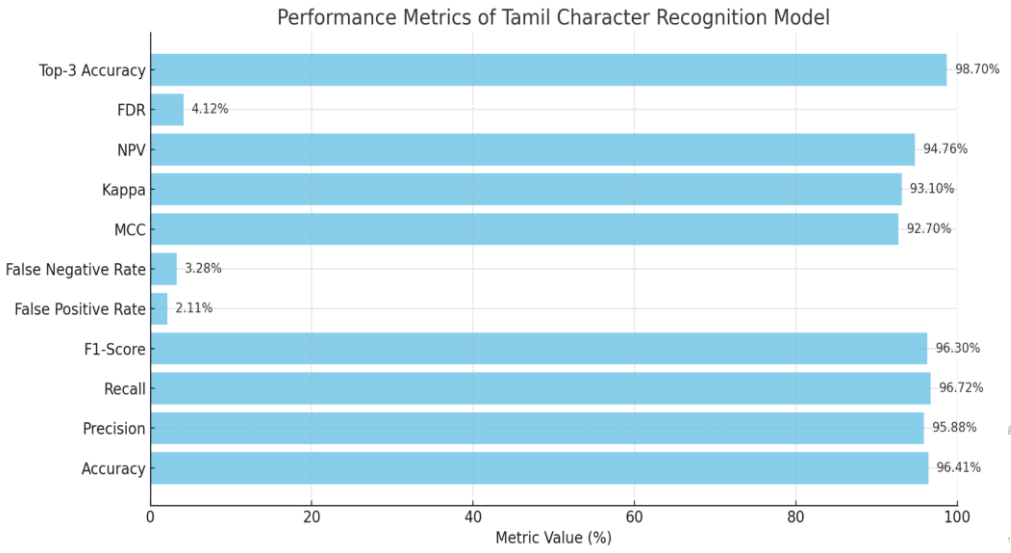

Fig 2. Results Comparison

VI. CONCLUSION

In this work, an ancient Tamil stone inscription character recognition model is proposed, which emphasizes segmentation as a foundational stage in the recognition pipeline. Built using Python, the model incorporates a deep CNN architecture fused with a hybrid segmentation strategy to overcome existing limitations in OCR systems. The performance of the model was evaluated using various metrics, including accuracy, precision, recall, F1-score, false positive/negative rates, Matthews Correlation Coefficient (MCC), Kappa, NPV, and FDR. Results show that the proposed model outperforms traditional models in both segmentation quality and recognition accuracy, establishing a more reliable approach for the digitization and preservation of ancient Tamil inscriptions.

II. ACKNOWLEDGMENT

Special thanks to **Dr. G. Krishnapriya**, Assistant Professor, for her valuable guidance, insights, and continuous encouragement during the development of this study. The authors are also thankful to fellow researchers and contributors whose work in the domain of Tamil epigraphy, segmentation algorithms, and deep learning has served as the foundation for this research.
Finally, heartfelt thanks to all who contributed directly or indirectly to this research, particularly the data contributors from Kaggle for making ancient Tamil inscription datasets publicly available.

REFERENCES

1. Magrina, M. M. (2021). Recognition of Ancient Tamil Characters from Epigraphical inscriptions using RaspberryPi based Tesseract OCR.
2. Vellingiriraj, E. K., Balamurugan, M., &Balasubramanie, P. (2016). Text analysis and information retrieval of historical Tamil ancientdocumentsusingmachinetranslationinimagezoning.IntJ Lang Lit Linguist, 2(4), 164-168.
3. Kaladevi, R., Revathi, A., &Manju, A. (2022). Analyzing the evolution of modern Tamil script for natural language processing. ECS Transactions, 107(1), 5219.
4. Munivel,M.,&Enigo, V.F.(2022).OpticalCharacterRecognition for Printed Tamizhi Documents using Deep Neural Networks. DESIDOC Journal of Library & Information Technology, 42(4), 227-233.
5. Karunarathne, K. G. N. D.,Liyanage, K. V., Ruwanmini, D. A. S., Dias, G. K. A., &Nandasara, S. T. (2017). Recognizing ancient sinhalainscriptioncharactersusingneuralnetworktechnologies.Internationa Journal of Scientific Emgineering and Applied Sciences, 3(1), 12.
6. Prabavathi, R. (2021). Prehistoric Stone Image Tamil Character Recognition using Optimized Deep Neural Network using Zernike Moments and Simplex Method. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(11), 5983-5991.
7. Priya, R. D., Karthikeyan, S., Indra, J., Kirubashankar, S., Abraham, A., Gabralla, L. A., ...&Nandhagopal, S. M. (2023).Self-Adaptive Hybridized Lion Optimization Algorithm with Transfer Learning for Ancient Tamil Character Recognition in Stone Inscriptions. IEEE Access.
8. Suriya,S.,Nivetha,S.,Pavithran,P.,Sashwath,K.G.,&Elakkiya,G. (2022). Effective Tamil Character Recognition Using Supervised Machine Learning Algorithms. EAI Endorsed Transactions on e-Learning, 8(2), e1-e1.
9. Kumar, A. N., & Geetha, G. (2023). Recognizing Ancient South Indian Language Using Opposition Based Grey WolfOptimization. Intelligent Automation & Soft Computing, 35(3).
10. Giridhar,L.,Dharani,A.,&Guruviah,V.(2019).Anovelapproach to ocr using image recognition based classification for ancienttamil inscriptions in temples. arXiv preprint arXiv:1907.04917.
11. Shanmugam, K. and VANATHI, B., "Enhancing Offline Tamil Handwritten Character Recognition Using Optimal Newton Algorithm Based Deep Convolution Extreme Learning Model, 2021.
12. Vani, V. and Ananthalakshmi, S.R., "Soft computing approaches for character credential and word prophecy analysis with stone encryptions", Soft Computing, vol. 24, no. 16, pp.12013-12026, 2020.
13. Eswaran, P.M., Manib, D. and Savarimuthu, S.R., "Recognizing TamilPalm-LeafManuscript Characters UsingHybridized Human Perception Based Features", ICTACT Journal on Image & Video Processing, vol. 11, no, 4, 2021.
14. S. Dhivya and J. R. Beulah, "Ancient Tamil Character Recognition from Stone Inscriptions – A Theoretical Analysis," 2022 2nd Asian Conference on Innovation in Technology (ASIANCON), Ravet, India, 2022, pp. 1-8.
15. Ibrahim, H., Kong, N.S.P. and Ng, T.F., "Simple adaptive median filter for the removal of impulse noise from highly corrupted images", IEEE Transactions on Consumer Electronics, vol. 54, no. 4, pp.1920-1927, 2008.
16. Babczyński, T. and Ptak, R., "Line segmentation of handwritten text using histograms and tensor voting", International Journal of Applied Mathematics and Computer Science, vol. 30, no. 3, 2020.
17. Tang, H., Liu, X., Sun, S., Yan, X. and Xie, X., "Recurrent mask refinement for few-shot medical image segmentation", In Proceedings of the IEEE/CVF international conference on computer vision, pp. 3918-3928, 2021.
18. Yamanakkanavar, N., Choi, J.Y. and Lee, B., "SM-SegNet: A Lightweight Squeeze M-SegNet for Tissue Segmentation in Brain MRI Scans", Sensors, vol. 22, no. 14, pp.5148, 2022.
19. Sindhushree, G.S., Amarnath, R. and Nagabhushan, P., "Entropy- based approach for enabling text line segmentation in handwritten documents", In Data Analytics and Learning:Proceedings of DAL 2018, pp. 169-184, 2019.
20. Kavitha, B.R. and Srimathi, C.B., "Benchmarking on offline Handwritten Tamil Character Recognition using convolutional neural networks", Journal of King Saud University-Computer and Information Sciences, vol. 34, no. 4, pp.1183-1190, 2022.
21. Babu, N. and Soumya, A., "Character recognition in historical handwritten documents–a survey"In2019international conference on communication and signal processing (ICCSP), pp. 0299-0304, April 2019.
22. Sivaraj, R. and Priya, R.D., "Bayesian-based parallel ant system for missing value estimation in large databases", International Journal of Bio-Inspired Computation, vol. 9, no. 2, pp.114-120, 2017.