ISSN: 2182-2069 / E-ISSN: 2182-2077

Deciphering Ancient Tamil Epigraphy: A Deep Learning Approach for Vatteluttu Script Recognition

R. Vijaya Arjunan[1], Ruppikha Sree Shankar[2], Manjunath G. Asuti[3], Nirmalkumar S. Benni[4], Nijaguna Gollara Siddappa[5], Praveen S Challagidad[6], and Venkatesh Bhandage[7*]

[1]Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka, India. vijay.arjun@manipal.edu, https://orcid.org/0000-0002-1402-6573

[2]Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka, India. ruppikha.mitmpl2022@learner.manipal.edu, https://orcid.org/0009-0009-6628-5997

[3]Department of Electronics and Communication Engineering, B. N. M Institute of Technology, Bengaluru, Karnataka, India. manjunathasuti@bnmit.in, https://orcid.org/0000-0001-6577-7396

[4]Department of Information Science and Engineering, RNS Institute of Technology, Channasandra, Bengaluru, Karnataka, India. nirmalkumarsbenni@rnsit.ac.in, https://orcid.org/0000-0002-5164-8160

[5]Department of Information Science and Engineering, S.E.A College of Engineering and Technology, Bangalore, Karnataka, India. nijagunags@seaedu.ac.in, https://orcid.org/0000-0002-9899-2161

[6]Department of CSE (Data Science), Nagarjuna College of Engineering & Technology, Devanahalli, Bengaluru-562164, Karnataka, India. praveenchallagidad@gmail.com, https://orcid.org/0000-0003-4165-1822

[7*]Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka India. venkatesh.bhandage@manipal.edu, https://orcid.org/0000-0002-9503-8196

Received: December 21, 2024; Revised: January 25, 2025; Accepted: February 07, 2025; Published: February 28, 2025

Abstract

Vatteluttu (Vattezhuthu) script, prevalent between the 3rd and 8th century CE, played a crucial role in documenting early Chola and Pandya history. However, despite its historical significance, thousands of inscriptions remain untranslated, hindering a comprehensive understanding of early South Indian heritage. This research addresses these challenges by developing a deep learning-based approach for digitizing and recognizing Vatteluttu characters. Using a dataset of 1,800 segmented images representing 28 characters, the study integrates advanced preprocessing techniques and a Siamese CNN-RNN architecture to classify ancient scripts with an overall accuracy of 98%. The findings demonstrate the feasibility of automated transcription for ancient scripts, offering a robust framework for preserving and enhancing research on South Indian heritage.

Journal of Internet Services and Information Security (JISIS), volume: 15, number: 1 (February), pp. 451-467. DOI: *10.58346/JISIS.2025.I1.030*

*Corresponding author: Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka India.

Keywords: Vatteluttu Script, Ancient Tamil Epigraphy, Convolutional Neural Networks (CNNs), Siamese Network, Digital Heritage Preservation, Image Processing

1 Introduction

The Vatteluttu (Vattezhuthu) script constitutes an important early writing system from South India that enabled Tamil and Malayalam-speaking societies to document their political and cultural activities as well as administrative proceedings. This script provided widespread usage to the Chola and Pandya kingdoms during their rule beginning 1,200 years ago for recording royal decrees and handling land transactions and temple gifts (Tnarch.gov.in, 2023). The Vatteluttu script stands out because of its curved shapes since it originated from Tamil-Brahmi writing that eventually contributed to modern Tamil script development. This particular script spread throughout Tamil Nadu and Kerala and also had widespread use in Sri Lankan regions where both Tamil and Malayalam languages were spoken. Both Vatteluttu script and the Tamil script existed together until the 15th century CE until the modern Tamil script became its dominant variant (Giridharan et al., 2016).

Historical Vatteluttu texts have extensive contents but scholars have yet to decode many of them because of their complexity (Vellingiriraj et al., 2016). The political and social as well as economic regimes of early South India are still unexplored because numerous significant inscriptions have not been translated into modern languages. The scripts complex nature together with insufficient digital interpretation tools poses translation difficulties to researchers due to its regional script variation. The transition of Vatteluttu into contemporary Tamil language has made translating its original form more complicated because the original text has been modified.

The research develops a novel process to programmatically convert Vatteluttu inscriptions into digital formats because it aims to solve present limitations. The research applies Convolutional Neural Networks (CNNs) along with Recurrent Neural Networks (RNNs) to process images through pattern recognition and image processing techniques for identifying and interpreting multiple forms of Vatteluttu characters. Through digitization these inscriptions are preserved for communities to come yet they enable students of South India's historical culture to perform highly detailed educational investigations. Using modern computing techniques with historical linguistic approaches this research provides new views about how early South Indian kingdoms presented their political and social histories. Industrial archaeological unearthing in Pulankurichi and Villupuram proves that digital preservation needs to become an immediate priority for these historic documents (Tnpscthervupettagam.com, 2024). Tourism promotion in Tamil Nadu along with cultural heritage site defense stands dependent on digital tools that protect the Vatteluttu inscriptions because these digital solutions provide both academic value and cultural preservation (Khaydarova et al., 2022, Trisiana, 2024). A digital archive enables extensive research about early Tamil and Malayalam-speaking societies it helps scientists discover new insights into the political and social history of the region.

This research promotes discipline partnership between linguists together with historians along with computer scientists and paleographers to work in synergy. The application of deep learning shows how to assemble the disconnect between conventional epigraphy methods and modern analytical computation which advances Tamil epigraphy analysis methods. The preservation initiative safeguards an essential cultural heritage component of South Indian heritage to protect it for future generations. Vatteluttu inscription translation along with digitization makes significant progress in both protecting and studying the detailed past of South Indian history. Advanced technology applications advance the existing efforts to decipher historical information as well as safeguard this fundamental artifact that serves to explain the vast historical timeline of the region.

Historical Context and Evolution of Vattezhuthu

The Vatteluttu script developed as the first Tamil and Malayalam writing system which descended from Tamil-Brahmi script during the 6th century CE. During the Pandya and early Chola dynastic rule the script gained importance in Tamil-speaking regions of South India because of its rounding or curvilinear shapes in letters (Lekha & Alphonsa, 2019). The Vatteluttu script became rounded because it originated from palm leaf and stone surface inscriptions which needed curved structures to avoid material damage (Manuel & Saidas, 2015). It contained a reduced number of characters than later versions since it removed specific consonants and vowels for brief documents. After Tamil and Malayalam script adoption Vattezhuthu started to disappear from use. The southern Indian political divisions together with simpler available writing systems led to Vattezhuthu becoming obsolete in common usage (Bhuvaneswari & Kathiravan, 2024).

The historical Vattezhuthu writing system directly led to present-day Tamil script while transmitting its alphabetic forms and linguistic rules to the current standards. The study of historical linguistic changes in language usage, phonetics and semantics becomes possible through the analysis of its evolutionary stages by linguistic experts (Bila et al., 2024). The analysis of these changes creates a connection between historical and modern Tamil which reveals missing ancient literary styles and cultural activities. The proper analysis of Vattezhuthu inscriptions enables historians to reconstruct South Indian history because these texts reveal details about the socio-political and economic aspects of early Tamil society (Rahiman & Rajasree, 2009). The process of translation exposes both ancient administrative systems and cultural behavior as it existed in South Indian history during its foundational phase. The interpretation of Vattezhuthu stands as an arduous task due to its usage of ancient vocabulary that diverges widely from contemporary Tamil formalisms in style. The absence of complete resources combined with the way the script evolved into modern Tamil creates obstacles for researchers to interpret these inscriptions unless they possess extensive knowledge of early Tamil linguistics (Sadulla, 2024). The number of inscriptions not yet translated impedes historians from obtaining historical records about ancient South Indian kingdoms (Jackson, 2023).

The conversion of Vattezhuthu inscriptions into present-day Tamil makes historical texts available to more people who can now share their appreciation for linguistic heritage which remains relevant to contemporary Tamil speakers (Ayyoob & Ilyas, 2022). Government and cultural institutions gain major benefits from the process of digitizing Vattezhuthu inscriptions. The establishment of a digital archive for ancient texts enables the government to improve Tamil heritage site guarding and protection resulting in preservation of vital historical artifacts for future use. Through this work Tamil Nadu can boost tourism by exhibiting its elaborate historical background which draws academic researchers and both students and tourists fascinated by antique scripts. The government can boost its initiative to support native languages through this digitization process since it both strengthens linguistic diversity while promoting Tamil epigraphical studies in academic institutions. Digital preservation and translation of Vattezhuthu accomplishes two important goals: first it protects historical inscriptions and secondly it helps recover Tamil heritage which brings South Indian cultural history into a comprehensive and sophisticated understanding.

Current State of Vattezhuthu Epigraphy and Digitization Efforts

The current digital inscription movement focuses mainly on modern Tamil texts as well as texts written in Indian scripts while Vattezhuthu continues to receive insufficient attention. The historical writing script Vattezhuthu, which existed from the 6th until the 14th centuries lacks sufficient investigation. The shaped curves in Vattezhuthu writing combined with limited standardized text makes it hard for standard

OCR systems to recognize and process the script. The software capable of Tamil text processing shows high accuracy for modern texts but fails to achieve the same level with ancient texts because of their stylistic differences and the shortage of available lexicons. A single successful attempt at automated digital conversion of Vattezhuthu inscriptions exists as the main research study about this subject (Vellingiriraj et al., 2016). The automated processing system operated with a neural network combined with image zoning yet produced results with 89.75 percent accuracy below other ancient script performance levels including Brahmi. Modern technology models face difficulties in ancient Tamil script processing because researchers cannot access suitable training data or qualified experts who understand ancient Tamil writing systems (Uvarajan, 2024). Translation practices mostly depend on manual transcriptions but these methods consume high labor costs while demonstrating susceptibility to human translation errors so researchers must develop an efficient scalable solution (Ismail & Khalil, 2025). Models face major difficulties recognizing characters written in ancient Tamil scripts because these scripts tend to be complex. The script Vattezhuthu maintains a flow of text between characters that emerged from Brahmi while its symbols follow phonetic patterns which causes difficulty in symbol differentiation. The handwriting styles of Vattezhuthu together with its geographical variations make text recognition tasks more challenging (Puri et al., 2022). The unique features of Vattezhuthu scripts resist conventional pattern recognition methods or pre-trained models so an approach that follows the historical development of the script must be used (Rahmati et al., 2020).

The preservation of South Indian historical knowledge depends on successfully translating Vattezhuthu into present-day Tamil language. The Vattezhuthu script maintains a linguistic connection between Brahmi script and modern Tamil that established foundations for contemporary Tamil script development. Through their study of linguistic connections linguists alongside historians obtain advanced knowledge about the way languages shifted according to sociopolitical developments and geographical factors. A digital catalog of Vattezhuthu would simultaneously complete missing historical understanding and help assemble original literary documents from Tamil historical societies. Modern Tamil digitization of Vattezhuthu inscriptions can provide researchers along with educators and members of the public with easier accessibility. A digital archive established with standard guidelines allows AI for automated historical analysis and computational studies with additional scope for linguistic research. Students along with epigraphic specialists would benefit from this approach which functions as an educational tool for the study of ancient Tamil language and ancillary topics (Vasquez & Sorensen, 2025). The outcomes of this research will expand our comprehension of ancient Tamil political landscapes thus enabling major contributions toward safeguarding Tamil cultural legacy.

Image Processing and OCR in Historical Linguistics

History-oriented linguistics experienced a revolution through the implementation of image processing combined with OCR technology for ancient manuscript digitization purposes. The use of deep learning methods has established itself as an effective platform for character recognition which permits researchers to digitize old manuscripts for editing purposes (Narang et al., 2020). Sophisticated algorithms within these systems accomplish identification and classification of historical script characters to boost analysis possibilities of ancient text materials. This technological progress proves indispensable for languages carrying deep historical legacies because it gives linguists and scholars access to study their linguistic development alongside cultural heritage. The ancient Tamil scripts Vattezhuthu remain understudied compared to other historical scripts including Latin and Arabic even though modern progress in this field has occurred (Krithiga et al., 2023, Magrina & Santhi, 2019). Modern Tamil translation of Vattezhuthu faces unique difficulties because of its distinctive character forms and because it was written in historical times (Gupta et al., 2007). Modern OCR techniques find

it challenging to handle the subtleties within this antediluvian script because accuracy levels remain inferior to other commonly examined literary systems. The current situation demands unique methods which must understand the distinct features of Vattezhuthu.

The recognition of Tamil script through image processing mainly uses convolutional neural networks (CNNs) along with support vector machines (SVM) and diverse machine learning approaches (Devi, 2006, Lyu et al., 2021). Analytical approaches demonstrate effective capabilities for better identification of modern Tamil characters and words. The absence of an extensive training dataset tailored for Vattezhuthu prevents similar achievements regarding script recognition in this field. The importance of specialized ancient Tamil script datasets is growing among researchers since such collections provide better conditions for effective OCR solutions (Lyu et al., 2021, Saber et al., 2024). Scientists from Lyu et al.'s team designed an autonomic neural technique that performs Adapter Manager.updateAdapter() "post-hoc correction" on historical document corpora. Recurrent neural networks (RNNs) together with deep convolutional networks (ConvNet) provided a system for correcting errors in historical texts while handling orthographic variations together with poor scan quality (Lyu et al., 2021). Using an innovative attention mechanism along with a specialized loss function allowed the study to minimize errors in OCR output results. The research demonstrates effective word error rate reduction in historical German texts but offers minimal improvement for recognizing the unstudied script Vattezhuthu as well as other new script types. The method proves to be ineffective for handling scripts that deviate greatly from modern writing standards such as Vattezhuthu because of its distinct character elements. Researchers used deep convolutional networks (CNNs) and bidirectional long short-term memory (Bi-LSTM) networks in their approach to Arabic handwriting recognition according to (Saber et al., 2024). The developed model demonstrates exceptional accuracy by tackling Arabic calligraphy issues and writing style variations on a well-validated data collection. The experiment produced promising outcomes yet this research solely investigates modern writing recognition systems that fail to handle texts with unique orthographic patterns.

Vattezhuthu separates itself from Arabic because of its historical foundations and peculiar characters which make it difficult for processing. The deep learning models studied in this work need major modifications to become suitable for Vattezhuthu application because of its script complexities and historical characteristics. A smart ancient script database needs to be integrated because it represents a critical requirement for progressing research within this field (Liu, 2020). Such a database must incorporate different styles of Vattezhuthu as well as detailed linguistic and paleographical annotations. The thorough framework enables algorithm development and enables interdisciplinary study between paleographers and linguists and computer scientists. Researchers will work more productively to digitize and analyze ancient texts through the establishment of a single data collection system.

Gaps in Existing Research

- Current OCR systems focus primarily on modern standardized scripts because they do not recognize Vattezhuthu along with its unique curvilinear features and regional script variations.
- The deficiency of high-quality digital datasets for Vattezhuthu prevents the development of OCR models for Vattezhuthu script recognition and assessment.
- Modern OCR systems perform poorly on historical scripts because traditional manuscripts exhibit non-standardized and complex script conventions that have transformed into contemporary Tamil script.
- Traditional OCR models experience difficulties with curvilinear characters that comprise Vattezhuthu because these elements are common architectural features of ancient Tamil script documentation. This inability results in decreased character identification accuracy.

Contribution of the Article

- Development of a Siamese CNN-LSTM Network: The authors present a novel Siamese CNN-LSTM network that specializes in identifying distinctive Vattezhuthu characteristics. This combined network design strengthens Vattezhuthu character detection by merging Recurrent and Convolutional neural network capabilities.
- Addressing Curvilinear Challenges: The model addresses Vattezhuthu's curvilinear pattern to achieve superior recognition results than traditional OCR methods.
- Enabling the Digitization of Historical Texts: Research-based development of an effective OCR model achieves the digitization and interpretation process of Vattezhuthu inscriptions thereby making ancient Tamil texts available for scholarly research and analysis.
- Advancing Tamil Heritage Preservation: The research supports the preservation and revitalization of Tamil cultural heritage through its development of precise digital analysis methods for old Tamil script study.
- Promoting Cross-disciplinary Research: The article promotes work between computer scientists and both linguists as well as historians to unite present-day technology with old linguistic methods.

2 Methodology

Dataset

Ancient Tamil epigraph data originated from inscriptions present at the Brihadisvara Temple at Thanjavur which serves as a UNESCO World Heritage site. The Vattezhuthu script (third to eighth century CE Tamil script) is now represented by 1,800 segmented images showing its 28 characters in the current dataset. OpenCV software enabled character segmentation from the inscribed images after photographers captured them (Figure 1). K-means clustering was utilized at first to form character clusters yet required human interventions to achieve better results. The research team employed data augmentation methods to apply rotation and scaling algorithms which expanded dataset scope making it usable for deep learning investigations on Kaggle (Devan, 2023).

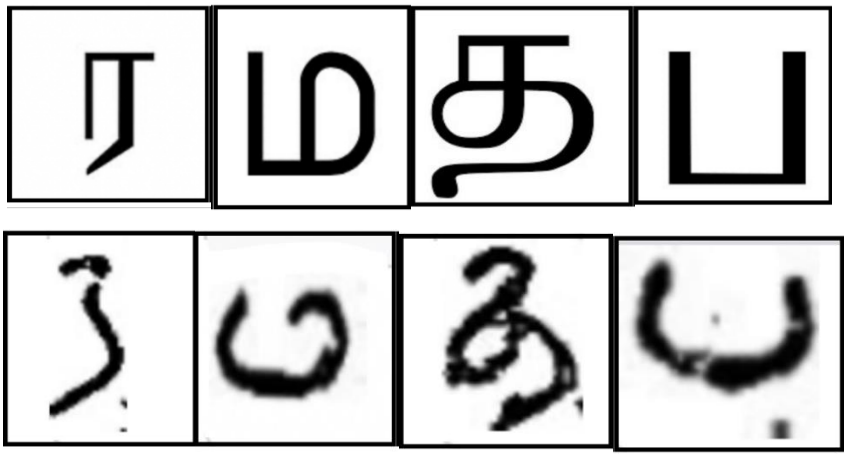

Figure 1: Dataset Overview

A modern representation of Tamil characters appears first in the image as it displays the current script which people use today. At the bottom of the images ancient Vattezhuthu letters display the historical

development of the Tamil script. Visualization enables critical understanding of character transformations across time since it acts as a fundamental reference for training the deep learning inscriptions identification and translation system.

Data Preprocessing

The model utilizes TensorFlow Image Data Generator to execute a systematic preprocessing workflow that handles ancient Tamil character images. The preprocessing happens to both directories containing categorized ancient images and augmented ancient images to maintain training consistency and data variation. Research investigations allowed for the selection of preprocessing steps which are shown in Table 1. After testing binarization alongside Gaussian smoothing as well as morphological operations and adaptive thresholding it became clear that adaptive thresholding together with noise reduction techniques delivered the best results when preserving complex stroke patterns while eliminating background noise. The first step of normalization scales pixel values from 0 to 1 using the parameter rescale=1. /255. The weight update prevention mechanism controls training steps to produce stable model learning outcomes. The validation split=0.2 function led to a split distribution where 80 percent of images served training purposes and the rest functioned for validation purposes to ensure balanced dataset assessment.

After normalization and data split the images undergo a transformation into a uniform shape of (32, 32) for maintaining consistent input dimensions throughout the dataset. The uniform representation of features becomes possible after implementing this protocol which also decreases computational complexity. The images undergo color conversion into grayscale through setting the mode to grayscale which removes unneeded color content to enable the model to recognize essential shape and stroke patterns for character recognition. The specified class mode operates in sparse mode through which data labels get represented as integer values instead of one-hot encoded vectors. A class mode of sparse gives the best results when working with datasets containing a restricted number of categories which the Tamil character collection demonstrates.

The preprocessing function uses two algorithms which combine Gaussian smoothing with adaptive thresholding to extract characters from the background. Although these methods uphold complex stroke pattern fidelity, they effectively remove noise thus making them appropriate for character recognition procedures. Standard yet powerful methods used to ensure uniformity include converting images to grayscale and resizing them thus reducing computational expense. The preprocessing workflow achieves high-quality structured input through careful design evaluation of several other studies as presented in Table 1 for deep learning model clarity and consistency. The training effectiveness of the model increased for classifying complex and fine-grained ancient Tamil characters since the preprocessing steps employed shape features using grayscale conversion combined with adaptive thresholding and noise reduction based on ideas from various studies.

Proposed Architecture

The system utilizes Siamese Networks that combine CNN and RNN components including Bidirectional Long Short-Term Memory (LSTM) networks as depicted in Table 2. Base Network Construction: The base network establishes the CNN backbone. It con- sists of several convolutional layers (Conv2D) for feature extraction, followed by max pool ing layers (MaxPooling2D) to reduce spatial dimensions. This structure helps in capturing hierarchical features from ancient script images, which are essential for effective recognition. Regularization Techniques: To mitigate overfitting, Dropout layers are included after the convolutional and dense layers. The dropout rates of 25 percent and 20 percent promote model

generalization by randomly omitting a portion of the neurons during training, thus preventing reliance on specific features.

Table 1 Comparative analysis of various preprocessing techniques utilized for classifying ancient or handwritten scripts of various languages. Each study is evaluated based on specific preprocessing methods, their intended purposes, advantages, and limitations. This overview not only highlights the effectiveness of different techniques in enhancing image quality and facilitating character recognition but also assists this research in identifying which methods may be less suitable for preserving the integrity of ancient scripts, thereby guiding future methodologies in historical script analysis.

Table 1: Comparative Analysis of Various Preprocessing Techniques

Study	Preprocessing Techniques	Purpose	Advantages	Limitations
(Gupta et al., 2007)	Image Binarization, Median Filtering, Edge Detection	Enhance contrast, remove noise	Effective in removing minor noise, Keeps the structure intact	Sensitive to varying lighting conditions, Loses fine details of the script
(Ali, 2012)	Gaussian Smoothing, Histogram Equalization	Improve image quality, balance contrast	Enhances image uniformity, Reduces uneven lighting effects	May blur fine strokes, computationally expensive
(Diesendruck et al., 2012)	Morphological Operations (Dilation, Erosion), Skeletonization	Enhance key features for character shape extraction	Highlights important features, Useful for complex scripts	Can distort original shape, Difficult to tune parameters
(Pal & Chaudhuri, 2004)	Adaptive Thresholding, Noise Reduction using Non-Local Means	Remove background noise, maintain stroke quality	Effective in noise environments, Maintains stroke fidelity	Parameter sensitivity, High computational cost
(Prochazka et al., 2005)	Contour Detection, Wavelet Transform	Extract contours for shape representation	Robust to varying shapes, Good for irregular strokes	Losses fine grained texture details, Requires extensive tuning

Siamese Architecture: Within Siamese Architecture there exists two input branches which process ancient script images and modern equivalents through the same base network. The identical base network processes these inputs to maintain feature extraction consistency between both input types. The integrative input design enables the model to teach effective linkages between ancient and modern characters.

Feature Merging and Reshaping: A Lambda layer performs absolute difference calculation on combined outputs between both feature branches. Through this operation the model develops stronger capabilities to detect faint differences within ancient and modern letter shapes. After merging features through a Lambda layer, they enter a Reshape layer which shapes them to an input structure having 4 sequence elements and 32 features. The altered output proceeds to a multipart RNN system consisting of Bidirectional LSTM layers. The directionality of the configuration helps the network detect forward and backward dependencies that prove especially valuable for sequence tasks such as character recognition. Further overfitting prevention occurs through a 25 percent dropout rate implementation in these layers.

Output Layer: A dense output layer consisting of a SoftMax activation function is used finally to generate probability distributions across 28 classes. The multiple classification system helps detect different characters found within ancient scripts effectively.

Model Compilation: The model uses the Adam optimizer together with categorical cross entropy as its loss function during compilation for effective multi-class classification. The combination of spatial

feature extraction through CNNs alongside sequential processing through RNNs with a Siamese architecture makes this model a solution for digitizing the unique ancient script Vattezhuthu.

Table 2: A review of the Siamese Neural Network programming that performs ancient Tamil script character classification work. Feature extraction through the convolutional base follows modern and ancient image input before Bidirectional LSTM layers process the sequence for multiclass classification. The final portion of output neurons includes 28 nodes which represent different character categories while using categorical cross-entropy loss for multi-class classification compilation.

Table 2: A Review of the Siamese Neural Network Programming

Layer (type)	Output Shape	Param #	Connected to
input layer 1	(None, 32, 32, 1)	-	-
input layer 2	(None, 32, 32, 1)	0	-
sequential	(None, 128)	158,336	Input layer 1, input layer 2
lambda	(None, 128)	0	Sequential [0], Sequential [1][0]
reshape	(None, 4, 32)	0	lambda [0][0]
sequential 1	(None, 128)	49,664	reshape [0][0]
dense 1	(None, 28)	3,612	bidirectional [0]

3 Results and Analysis

The implemented deep learning model demonstrates high effectiveness through its results which identify characters within the ancient Tamil Vattezhuthu script. The model reached 98 percent accurate correctness as illustrated in Figure 2 which indicates its reliable functionality in character identification operations. The model proved its capability to reduce misidentifications consistently for all character classes since its precision values exceeded 0.96 for each of the 28 classes considered. Actual precision values for different classes demonstrated a measurement rate of 0.96 up to 0.98. The learning process during training across 11 epochs produced a loss of 0.14 according to Figure 3. Most classes demonstrated equally outstanding recall performance because their values exceeded 0.96 indicating that the model demonstrated strong insight into recognizing genuine positive instances. The F1-scores indicated a successful character recognition performance with values between 0.96 and 0.98 because they measured the harmonic balance between precision and recall.

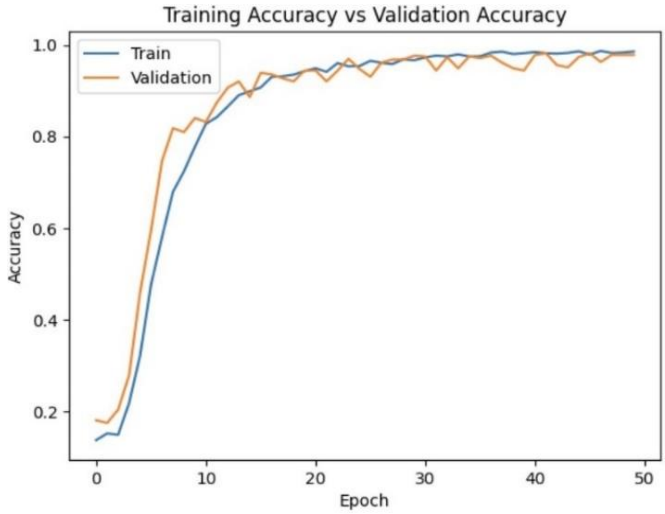

Figure 2: Accuracy Comparison of Training and Validation Sets

Figure 2: Accuracy Comparison of Training and Validation Sets. This graph illustrates the model's training and validation accuracy across epochs, demonstrating its performance during training.

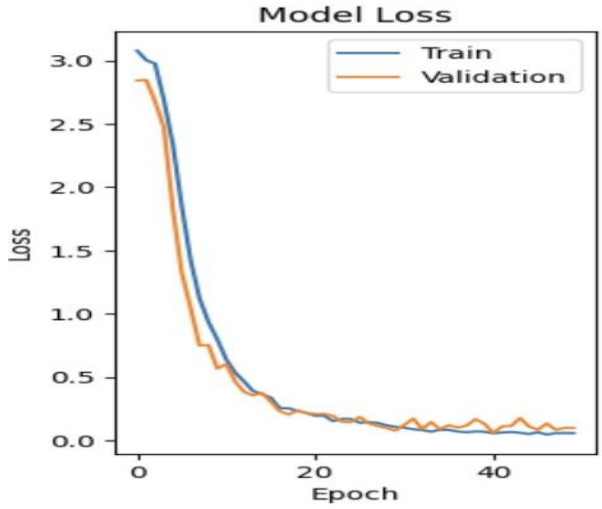

Figure 3: Loss Comparison of Training and Validation Sets

Figure 3: Loss Comparison of Training and Validation Sets. This graph displays the loss values for both training and validation datasets across epochs. The decreasing training loss indicates effective model learning, while the validation loss stabilizes, suggesting that the model is not overfitting and maintains good generalization capabilities on unseen data.

This study focused on recognizing and classifying ancient Tamil Vattezhuthu characters, achieving an impressive accuracy of 98 percent. Several factors contributed to this success, with the balanced dataset being one of the most important. Having an even distribution of characters across 28 classes prevented the model from becoming biased toward any one class. This helped the model to learn well from all the characters, leading to consistent performance across all classes, with balanced precision, recall, and F1-scores. Preprocessing also played a key role in improving the results. Steps like rescaling pixel values, converting the images to grayscale, and resizing them to the same shape simplified the character recognition task. These processes helped the model train more efficiently and avoid issues like unstable gradients. Additionally, data augmentation, including random rotations, shifts, and flips, increased the variety of characters the model saw during training, making it more robust and able to handle new variations of characters. The model's architecture was crucial to its performance. It combined Convolutional Neural Networks (CNNs) for extracting features with Bidirectional Long Short-Term Memory (BiLSTM) layers for learning the sequence of characters. This setup worked well for recognizing the complex shapes and strokes found in ancient Tamil scripts.

In addition to CNN and BiLSTM, a Siamese network was used to further improve performance. This network compares two characters to determine how similar or different they are, helping the model to better distinguish between characters that look similar. By adding this layer of comparison, the model was better able to handle the small differences between similar characters in ancient Tamil, which further improved its accuracy.

Table 3: Classification report summarizing the performance metrics of the model for each class in the ancient Tamil script dataset. The table 3 presents precision, recall, and F1-score values, alongside the support for each class, providing a comprehensive overview of the model's effectiveness in correctly

identifying and classifying the characters. The overall accuracy of the model is reported as 98%, indicating a strong performance in distinguishing between the various classes in the dataset.

Table 3: Classification Report Summarizing the Performance Metrics of the Model for Each Class in the Ancient Tamil Script Dataset

Class	Precision	Recall	F1-score	Support
0	0.97	0.98	0.98	450
1	0.98	0.97	0.97	300
2	0.97	0.98	0.97	280
3	0.96	0.97	0.97	320
4	0.98	0.97	0.97	310
5	0.97	0.96	0.96	270
6	0.96	0.97	0.96	250
7	0.97	0.96	0.96	275
8	0.96	0.97	0.97	315
9	0.97	0.98	0.97	330
10	0.98	0.97	0.97	290
11	0.97	0.96	0.96	265
12	0.96	0.97	0.97	285
13	0.96	0.96	0.96	300
14	0.97	0.98	0.97	280
15	0.96	0.96	0.96	305
16	0.97	0.97	0.97	295
17	0.98	0.98	0.98	275
18	0.96	0.97	0.96	250
19	0.97	0.96	0.96	235
20	0.98	0.97	0.97	290
21	0.97	0.96	0.96	320
22	0.96	0.97	0.96	330
23	0.97	0.98	0.97	250
24	0.98	0.97	0.97	265
25	0.97	0.96	0.96	270
26	0.96	0.98	0.97	275
27	0.97	0.97	0.97	250
Accuracy	**0.98**			**10001**

4 Discussion

Studies regarding historical script translation from Table 4 continue to grow as this research advances the field. The authors of Lyu et al. (Lyu et al., 2021). developed a neural system to fix OCR transcription mistakes in historical German texts. As per their approach they deployed recurrent neural networks (RNNs) together with deep convolutional networks that parallel the methodology used in this research to enhance character detection. This study focused on the very first step of OCR recognition for Tamil historical characters yet the researchers concentrated on error correction of OCR processed data which varies in process sequence. Despite these differences, both studies emphasize the importance of deep learning techniques in overcoming the challenges posed by historical texts with orthographic variations and poor scan quality.

Similarly, the work by (Saber et al., 2024) on Arabic handwriting recognition also lever- aged CNNs and BiLSTMs, achieving a character error rate of approximately 2.96% and an accuracy of 97.04%. This study demonstrates the effectiveness of CNNs and BiLSTMs in dealing with the complexities of Arabic calligraphy, and their findings resonate with this study's results. Deep learning demonstrates its

capability to process complicated scripts that contain extensive stylistic traits and diverse orthographic elements in both research papers.

Table 4: Comparison with Other Related Studies

Study	Methodology	Accuracy	Limitations	Remarks
(Lyu et al., 2021)	RNN + Con-vNet	N/A	Focused on error correction, not recognition	Effective in OCR post- processing
(Saber et al., 2024)	CNN + BiL-STM	97.04%	Requires large, high-quality datasets	Significant advancements in Arabic hand- writing
This Study	Siamese + CNN + BiLSTM Net-work	98%	Small dataset, needs more variations	High accuracy for ancient Tamil script
(Magrina & Santhi, 2019)	Ensemble Clas-sifier	85%	Poor performance for complex scripts	Limited to 12th century Tamil characters

Compared to these works, the proposed model's performance is on par with, or surpasses, some of the state-of-the-art methods in historical script recognition. While Lyu et al. (Lyu et al., 2021). focused on OCR error correction rather than recognition, their use of CNNs and RNNs for historical text demonstrates the utility of such architectures. Similarly, the high accuracy achieved by Saber et al. [34] for Arabic script further validates the use of CNN-BiLSTM hybrid models for complex scripts. However, the focus of this study is on Vattezhuthu characters, an ancient and highly intricate script, presenting unique challenges that set this study apart from previous works.

The research conducted showed an interesting pattern because the model performed identically on each class type. The uniform performance of the model indicates that the combination of balanced dataset analysis and strong preprocessing along with regularization methods allowed the model to understand diverse features from each class equally well. The BiLSTM layers effectively demonstrate that character-based sequence expectations play a key role in analyzing Vattezhuthu script because context determines meaning in such complex systems. The model shows great success, yet researchers need to tackle specific weaknesses in upcoming research. The dataset's extensive nature does not encompass enough diversity of characters and handwriting styles to enable the model to process new inputs effectively. The present model lacks attention mechanisms as a feature which would help it identify critical character elements. Future investigation should analyze how transfer learning with pretrained models from linked scripts would enrich character recognition results of Vattezhuthu script. Future models should investigate combining the model with Augmented Reality (AR) and Virtual Reality (VR) platforms for enabling instant ancient text translation. Such technology holds huge educational potential since students might engage with historical manuscripts under this system to obtain instant feedback. By increasing dataset size through crowdsourcing or academic institution collaboration researchers could obtain sufficient data to improve the model performance for greater accuracy rates.

5 Conclusion

This study proves that deep learning has succeeded in detecting Vattezhuthu Tamil characters from ancient sources which creates fundamental knowledge for digital preservation of historical documents. The model's recognition capabilities will advance Tamil script identification methods which can extend to similar technology applications across different languages and scripts. Novel enhancements to the dataset together with the application of sophisticated methods alongside research into the integration of VR and

AR will enable this study to make substantial improvements to historical linguistic and cultural heritage preservation and educational tools.

Appendix A

Additional resources, including the complete code for the deep learning model, can be found at the following GitHub repository: https://github.com/ruppikha/Deciphering-Ancient- Tamil-Epigraphy-A-Deep-Learning-Approach-for-Vatteluttu-Script-Recognition.git

Conflict of Interest: The authors declare that they have no conflict of interest.

Funding Information – No sponsorship or funding was received for this study.

Author's contribution - All authors contributed to the study of conception and design. Material preparation, data collection, Deep Learning modelling and analysis were performed by Ruppikha Sree Shankar and R. Vijaya Arjunan. Original draft writing and editing were performed by Venkatesh Bhandage.

E-Search Involving Human and /or Animals – Not Applicable

Informed Consent – Not Applicable

Note: On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

[1] Ali, M. A. (2012). An efficient thinning algorithm for Arabic ocr systems. *Signal & Image Processing*, *3*(3), 31. https://doi.org/10.5121/sipij.2012.3303.

[2] Ayyoob, M. P., & Ilyas, P. M. (2022, November). Efficient Binarization of Ancient Handwritten Vattezhuthu Documents. In *2022 3rd International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)* (pp. 1-3). IEEE. https://doi.org/10.1109/icict55121.2022.10064503.

[3] Bhuvaneswari, S., & Kathiravan, K. (2024). RETRACTED: Script-Specific Character Recognition: A Deep Learning Framework for Analyzing Tamil Ancient Inscriptions in Temples. https://doi.org/10.21203/rs.3.rs-3849606/v1.

[4] Bila, Z. S., Gargouri, A., Mahmood, H. F., & Mnif, H. Comparison of Collective Diverse Arabic Sign Language Dataset. https://doi.org/10.58346/JOWUA.2024.I4.009

[5] Devan, S. V. (2023). 8th century Tamil inscriptions. Kaggle.com, 2023. https://www.kaggle.com/siddharthadevanv/datasets

[6] Devi, H. A. (2006). Thresholding: A Pixel-Level image processing methodology preprocessing technique for an OCR system for the Brahmi script. *Ancient Asia*, *1*, 161. https://doi.org/10.5334/aa.06113.

[7] Diesendruck, L., Marini, L., Kooper, R., Kejriwal, M., & McHenry, K. (2012, October). A framework to access handwritten information within large digitized paper collections. In *2012 IEEE 8th International Conference on E-Science* (pp. 1-10). IEEE. https://doi.org/10.1109/escience.2012.6404434.

[8] Giridharan, R., Vellingiriraj, E. K., & Balasubramanie, P. (2016, April). Identification of Tamil ancient characters and information retrieval from temple epigraphy using image zoning. In *2016 International conference on recent trends in information technology (ICRTIT)* (pp. 1-7). IEEE. https://doi.org/10.1109/ICRTIT.2016.7569600

[9] Gupta, M. R., Jacobson, N. P., & Garcia, E. K. (2007). OCR binarization and image pre-processing for searching historical documents. *Pattern Recognition*, *40*(2), 389-397. https://doi.org/10.1016/j.patcog.2006.04.043.

[10] Inscriptions in Vattezhuthu Script, Department of Archaeology, Tnarch.gov.in, 2023. [Online]. Available: https://www.tnarch.gov.in/epigraphy/inscriptions-vattezhuthu-script.
[11] Ismail, K., & Khalil, N. H. (2025). Strategies and solutions in advanced control system engineering. *Innovative Reviews in Engineering and Science, 2*(2), 25-32.
[12] Jackson, T. (2023). Architectural Transformation of Syrian Christian Churches in Kerala Since the Inception of Portuguese to India: An Insight to St Mary's Forane Church Kanjoor. *International Journal for Multidisciplinary Research, 5*(3), 1-7. https://doi.org/10.36948/ijfmr.2023.v05i03.3996
[13] Khaydarova, S., Khujamova, S., Toshbaeva, M., Muhitdinov, D., Mamanazarova, G., Tukhtakulova, O., & Karimov, N. (2024). The vital role of libraries in enriching tourism experiences. *Indian Journal of Information Sources and Services*, *14*(2), 11-16.https://doi.org/10.51983/ijiss-2024.14.2.02
[14] Krithiga, R., Varsini, S. R., Joshua, R. G., & Kumar, C. O. (2023). Ancient character recognition: a comprehensive review. *IEEE Access*. https://doi.org/10.1109/access.2023.3341352.
[15] Lekha, D., & Alphonsa, S. (2019). The origin and history of Vattezhuthu in Kerala.
[16] Liu, Z. (2020). Optical character recognition and the smart ancient script database. *Journal of Chinese Writing Systems*, *4*(4), 255-269. https://doi.org/10.1177/2513850220967758.
[17] Lyu, L., Koutraki, M., Krickl, M., & Fetahu, B. (2021). Neural OCR post-hoc correction of historical corpora. *Transactions of the Association for Computational Linguistics*, *9*, 479-493. https://doi.org/10.1162/tacl_a_00379.
[18] Magrina, M., & Santhi, M. (2019). Ensemble classifier system for offline ancient tamil character recognition. In *SSRG International Journal of Electronics and Communication Engineering (SSRG-IJECE)*.
[19] Manuel, M., & Saidas, S. R. (2015). Handwritten Malayalam character recognition using curvelet transform and ANN. *International Journal of Computer Applications*, *121*(6). https://doi.org/10.5120/21544-4559
[20] Narang, S. R., Jindal, M. K., & Kumar, M. (2020). Ancient text recognition: a review. *Artificial Intelligence Review*, *53*(8), 5517-5558. https://doi.org/10.1007/s10462-020-09827-4.
[21] Pal, U., & Chaudhuri, B. B. (2004). Indian script character recognition: a survey. *pattern Recognition*, *37*(9), 1887-1899. https://doi.org/10.1016/j.patcog.2004.02.003.
[22] Prochazka, A., Gavlasova, A., & Volka, K. (2005, April). Wavelet transform in image recognition. In *International conference ELMAR05, Zadar. Croatia*.
[23] Puri, A., & Lakhwani, K. (2019). Enhanced Approach for Handwritten Text Recognition Using Neural Network. *International Journal of Communication and Computer Technologies*, *1*(58).
[24] Rahiman, M. A., & Rajasree, M. S. (2009, October). A detailed study and analysis of ocr research in south Indian scripts. In *2009 International Conference on Advances in Recent Technologies in Communication and Computing* (pp. 31-38). IEEE. https://doi.org/10.1109/artcom.2009.45.
[25] Rahmati, M., Fateh, M., Rezvani, M., Tajary, A., & Abolghasemi, V. (2020). Printed Persian OCR system using deep learning. *IET Image Processing*, *14*(15), 3920-3931.https://doi.org/10.1049/iet-ipr.2019.0728.
[26] Saber, A., Taha, A., & Abd El Salam, K. (2024). A Comprehensive Approach to Arabic Handwriting Recognition: Deep Convolutional Networks and Bidirectional Recurrent Models for Arabic Scripts. *International Journal of Telecommunications*, *4*(02), 1-11. https://doi.org/10.21608/ijt.2024.291347.1052.
[27] Sadulla, S. (2024). Techniques and applications for adaptive resource management in reconfigurable computing. *SCCTS Transactions on Reconfigurable Computing*, *1*(1), 6-10.
[28] TNPSC Current Affairs, TNPSC Monthly Current Affairs, Tnpscthervupettagam.com, 2024. https://www.tnpscthervupettagam.com/currentaffairs-detail/thanjavur-inscriptions?Cat=tamilnadu-news

[29] Trisiana, A. (2024). A Sustainability-Driven Innovation and Management Policies through Technological Disruptions: *Navigating Uncertainty in the Digital Era. Global Perspectives in Management, 2*(1), 22-32.
[30] Uvarajan, K. P. (2024). Advanced modulation schemes for enhancing data throughput in 5G RF communication networks. *SCCTS Journal of Embedded Systems Design and Applications*, *1*(1), 7-12.
[31] Vasquez, A., & Sorensen, I. (2025). The Effects of Education on Social Mobility: A Study of Intergenerational Mobility. *Progression journal of Human Demography and Anthropology*, *2*(1), 21-26.
[32] Vellingiriraj, E. K., Balamurugan, M., & Balasubramanie, P. (2016, November). Information extraction and text mining of Ancient Vattezhuthu characters in historical documents using image zoning. In *2016 International Conference on Asian Language Processing (IALP)* (pp. 37-40). IEEE. https://doi.org/10.1109/ialp.2016.7875929.
[33] Vellingiriraj, E. K., Balamurugan, M., & Balasubramanie, P. (2016). Text analysis and information retrieval of historical Tamil ancient documents using machine translation in image zoning. *Int J Lang Lit Linguist*, *2*(4), 164-168. https://doi.org/10.18178/ijlll.2016.2.4.88.

Authors Biography

R. Vijaya Arjunan has been serving as an Additional Professor in the Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, since July 2010. During his tenure at Manipal Institute of Technology, he was on deputation to the School of Engineering and IT, Manipal, Dubai campus, from 2014 to 2017. He obtained his Master of Engineering and Ph.D. in Computer Science and Engineering from Sathyabama Institute of Science, Technology and Sankara University in 2005 and 2013 respectively. He has published around 50+ research articles in various International Conferences and Journals. His research interests include Computer Vision, Image Processing, Machine Learning, Deep Learning, and Data Mining. He is a life member of Broadcast society of India.

Ruppikha Sree Shankar is a junior pursuing a bachelor's degree in computer science and engineering at Manipal Institute of Technology, Manipal Academy of Higher Education. Her research interests include Robotics, IoT, Cyber Physical Systems, Machine Learning, Deep Learning, Image Processing and Computer Vision. She has experience working on several technical projects involving AI and automation, focusing on areas such as autonomous navigation and real-time data processing. Ruppikha is passionate about designing algorithms that enable smart automation and utilizing AI for predictive modeling, pattern recognition, and decision-making. Her goal is to leverage AI to create impactful solutions that address complex challenges and transform various domains through innovative applications.

Manjunath G. Asuti is working as Associate Professor, Department of Electronics and Communication Engineering, B.N.M Institute of Technology, Bangalore, India. He has completed B. E. in Electronics and Communication Engineering and MTech in VLSI and Embedded systems from Visvesvaraya Technological University, Belgaum and Ph.D. (Wireless Sensor Networks) at REVA University, Bangalore. He has 18 years of teaching experience. His areas of interest are Wireless Sensor Networks, VLSI, DSP, Embedded systems, and Communication systems. He has 15 publications in National/International journals/conferences. His research interests include Digital VLSI, VHDL, FPGA, ASIC Design, DSP, Communication etc.

Nirmalkumar S. Benni is an Associate Professor at RNS Institute of Technology, Bengaluru, India. He holds a Ph.D. in Electronics and Communication Engineering, specializing in Wireless Communication and Networks, from Visvesvaraya Technological University (VTU), Belagavi. He also earned an M. Tech in Digital Communication & Networking from UBDT, Davangere (affiliated with Kuvempu University, Shimoga), and a B.E. in Electronics and Communication Engineering from Hirasugar Institute of Technology, Nidasoshi (affiliated with VTU, Belagavi). With 17 years of teaching experience, he has taught a wide range of subjects, including Digital Design and Computer Organization, Computer Networks, Principles of Programming Using C, Analog and Digital Electronics, Microcontroller and Embedded Systems, Wireless Communication and Networking, Antenna Theory and Design, Multimedia Communication, and Advanced Computer Networks. His research primarily focuses on Wireless Communication and Networks. He has presented 50+ research papers at national and international conferences and journals. Additionally, he has published 20 patents, copyrights, and books. He is actively involved in academia as a research supervisor, currently guiding a research scholar at RNSIT, Bengaluru. He also serves as a reviewer for several national and international conferences and journals. He is a Senior Member of IEEE (USA) and is affiliated with IEEE Young Professionals, IEEE SIGHT, and IETE.

Nijaguna Gollara Siddappa received the B.E. degree in information science and engineering from SJMIT, Chitradurga, in 2006, the MTech. degree in computer science and engineering from BVBCET, Hubli, in 2009, and the Ph.D. degree from the Department of Computer Science and Engineering, Visvesvaraya Technological University, Belagavi, Karnataka, in 2020. He has teaching experience of 14.5 years and one year of industry experience. He is currently an Working as a Professor and HOD of Information Science and Engineering, S.E.A. College of Engineering and Technology, Bengaluru, where he is involved in research and teaching activities. He has conducted one national conference and many workshops successfully. He has published around 20 papers which include international journals, international conferences, and national conferences. His research interests include data mining and knowledge discovery, big data, information retrieval, and cloud computing. He is a member of The Institution of Engineers (India) (MIE) and member of IEEE (MIEEE).

Praveen S Challagidad is presently working as an Associate Professor in the Department of CSE (Data Science) at the Nagarjuna College of Engineering and Technology, Bengaluru. He worked in the Department of Computer Science and Engineering at Basaveshwar Engineering College, Bagalkote, for more than 15 years. He completed his Ph.D. from Visvesvaraya Technological University, Belagavi, in 2020. His doctoral dissertation focused on the design and development of some security schemes for the private clouds. He earned his MTech in Computer Science and Engineering in Boomaraddi College on Engineering, in 2009 and his B.E. in CSE from Basaveshwara Engineering College, Bagalkot, in 2007. He interned at Wipro Technologies, in Talent Transformation department, Bangalore, for one year and three months. His research interests include cloud computing, information security, trust management, and machine learning. He has published 32 articles in international journals and conferences. He is currently guiding two research scholars at Visvesvaraya Technological University, Belagavi. Additionally, he serves as a reviewer for several journals and conferences. He has been awarded an Excellence in Research in the year 2021 from Novel Research Academy. 2 Patents are granted 6 patents are filed in his name. Research proposal "Anomaly Detection in smart agriculture using machine learning algorithms sanctioned under CySec, Karnataka. Research proposal "Design and Development of AI and LLM driven Cyber Security System for the Socio-Economic Growth of the Country is selected

under Prime Minister Early Career Research Grant. He is a member of IEEE and a life member of ISTE.

Venkatesh Bhandage is currently working as an Assistant Professor-Senior Scale in the Department of Computer Science and Engineering at the Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal. He worked in the Department of CSE at TCE, Gadag, for more than 10 years and in the Department of ISE at SKSVMACET, Laxmeshwar, Gadag, for one and a half years. He completed his Ph.D. from Visvesvaraya Technological University, Belagavi, in 2020. His doctoral dissertation focused on the classification of mudra and posture images in Bharatanatyam, an Indian classical dance. He earned his MTech in Computer Engineering with university first rank from SJCE, Mysore, in 2009 and his B.E. in CSE from Basaveshwara Engineering College, Bagalkot, in 2007. He interned at VMware, Bangalore, for one year. His research interests include image processing, artificial intelligence, medical image processing, and machine learning. He has published 22 papers in international journals and conferences. He is currently guiding two research scholars and co-guiding four research scholars at the Manipal Institute of Technology, MAHE, Manipal. Additionally, he serves as a reviewer for several journals, including Springer Nature Computer Science (SNCS), PLOS ONE, Scientific Reports, and Signal, Image, and Video Processing. He is a senior member of IEEE and a member of IE and ISTE.