# AI-BASED DIGITIZATION OF HANDWRITTEN HISTORICAL DOCUMENTS IN REGIONAL LANGUAGES

[1]Mr.M.Devendran, [2]Antony Sebastin X, [3]Ashik B, [4]Barathvaj S, [5]Gurumoorthy M S
[1]Assistant Professor, Department of Computer Science and Engineering,
Hindusthan Institute of Technology, Coimbatore
[2,3,4,5] UG student, Department of Computer Science and Engineering,
Hindusthan Institute of Technology, Coimbatore
[1] md.devendran@gmail.com, [2] antonysebastin1521@gmail.com ,[3]shivaviswa1424@gmail.com,
[4] barathvaj726@gmail.com, [5]gurumoorthy2212005@gmail.com

**ABSTRACT:** The preservation of historical documents is essential for maintaining cultural heritage and enabling future research. Many valuable historical records exist only in handwritten form and are often written in regional languages, making their preservation and accessibility challenging. Traditional digitization methods mainly focus on scanning documents as images, which does not allow efficient searching, editing, or analysis of the content. To address these challenges, this study explores the use of Artificial Intelligence (AI) for the digitization of handwritten historical documents in regional languages. The proposed approach uses advanced AI techniques such as Optical Character Recognition (OCR), Deep Learning, and Natural Language Processing (NLP) to automatically recognize and convert handwritten text into machine-readable digital formats. The system is designed to handle variations in handwriting styles, ink quality, and aging effects commonly found in historical manuscripts. Special attention is given to regional language scripts, which often have complex characters and limited digital resources. The AI-based digitization process involves document image preprocessing, handwritten text recognition, language modeling, and digital archiving. By training machine learning models on regional language datasets, the system improves accuracy in recognizing handwritten characters and words. The digitized content can then be stored in structured databases, enabling efficient search, translation, and analysis. This approach not only helps in preserving historical documents but also makes them accessible to researchers, historians, and the general public. The implementation of AI-driven digitization can significantly reduce manual effort, improve accuracy, and ensure long-term preservation of valuable historical records in regional languages.

**Keywords:** Autonomous Medical Robot, Contactless Health Testing, Smart Medical Assistant, Healthcare Robotics, Arduino-based Robot, Hospital Automation

**Corresponding Author:** *Mr.M.Devendran*
*Assistant Professor / CSE, Hindusthan Institute of Technology*
*Coimbatore, Tamil Nadu, India*
*Mail: md.devendran@gmail.com*

## INTRODUCTION

Historical documents play a vital role in preserving the cultural, social, and political history of a region. Many of these valuable records exist only in handwritten form and are often written in regional languages. These documents include manuscripts, letters, government records, temple records, literature, and historical archives that contain important information about past civilizations and traditions. However, due to aging, environmental damage, and lack of proper preservation techniques, many of these documents are at risk of deterioration and permanent loss.

Traditional digitization methods mainly involve scanning documents and storing them as image files. While this method helps in preserving the visual form of the document, it does not make the content easily searchable, editable, or analyzable. In addition, handwritten documents present several challenges such as variations in handwriting styles, faded ink, damaged pages, and complex scripts, especially in regional languages like Tamil, Telugu, Kannada, Hindi, and others. These challenges make manual transcription time-consuming and prone to human errors. With the advancement of Artificial Intelligence (AI), new techniques have emerged that can automatically recognize and convert handwritten text into machine-readable formats. AI technologies such as Optical Character Recognition (OCR), Deep Learning, and Natural Language Processing (NLP) can analyze scanned images of handwritten documents and extract textual information with high accuracy. These technologies are capable of learning different handwriting patterns and adapting to various regional scripts.

AI-based digitization systems also include processes such as image preprocessing, feature extraction, text recognition, and language modeling. These processes help in improving the quality of scanned documents and accurately identifying handwritten characters and words. By applying AI techniques to historical manuscripts written in regional languages, it becomes possible to convert large collections of handwritten documents into structured digital data.

The digitization of handwritten historical documents using AI not only helps preserve cultural heritage but also improves accessibility for researchers, historians, and the general public. Digitized documents can be easily stored, searched, translated, and shared through digital platforms. Therefore, AI-based digitization plays a significant role in protecting historical

knowledge and ensuring that valuable information written in regional languages is preserved for future generations.

## Literature Survey

Ahmed et al. (2021) explored the application of deep learning techniques for handwritten text recognition in historical manuscripts. Their system combines Convolutional Neural Networks (CNN) with Long Short-Term Memory (LSTM) networks to recognize handwritten character sequences from scanned document images.

Chen et al. (2022) introduced a transformer-based handwritten text recognition model designed for complex document transcription tasks. The system uses attention mechanisms to capture contextual relationships between characters and words, improving the accuracy of recognition in historical texts. However, the system requires a large amount of labeled training data and high computational resources to achieve optimal performance.

Garcia et al. (2023) developed an AI-driven digital archival framework for preserving historical manuscripts. Their system integrates image enhancement, document segmentation, and machine learning-based OCR techniques to convert handwritten manuscripts into searchable digital text.

Patel et al. (2023) proposed a machine learning-based OCR system specifically designed for regional language documents. The system focuses on recognizing handwritten scripts used in Indian regional languages and converts them into machine-readable text. It also incorporates language modeling techniques to improve spelling and grammar correction. However, the system faces challenges in recognizing complex ligatures and overlapping characters.

Kumar et al. (2024) introduced an AI-based document digitization platform for preserving historical records written in regional scripts. The system utilizes image preprocessing, noise removal, and deep neural networks for character recognition. It enhances the readability of faded manuscripts through advanced image restoration techniques. Despite these improvements, the system's accuracy decreases when processing documents with irregular handwriting or damaged text.

Zhang et al. (2024) presented a deep learning-based handwritten text recognition system that combines convolutional neural networks with recurrent neural networks to capture both spatial and sequential handwriting patterns. This approach significantly improves transcription accuracy

for historical manuscripts. However, the system requires extensive training datasets and may struggle with stylistic variations in ancient calligraphy.

Liu et al. (2025) developed a multilingual handwritten text recognition framework designed for historical archives. The system employs transformer-based neural networks capable of recognizing multiple regional language scripts. It enables automatic transcription, indexing, and retrieval of historical documents in digital libraries. However, the framework requires high computational power and may face difficulties when processing severely degraded documents.

Smith et al. (2025) proposed an advanced handwritten document digitization system using deep learning-based Optical Character Recognition (OCR). Their model integrates convolutional neural networks with attention-based sequence learning to recognize complex handwritten scripts from historical documents. The system improves recognition accuracy through advanced image preprocessing techniques, but its performance depends heavily on the availability of large annotated datasets.

Overall, existing research highlights the significant role of artificial intelligence, deep learning, and OCR technologies in the digitization of handwritten historical documents. These approaches improve document preservation, accessibility, and searchability in digital archives. However, challenges such as handwriting variability, document degradation, and limited training datasets for regional languages remain key areas for future research.

## Proposed System

The proposed system aims to develop an AI-based framework for digitizing handwritten historical documents written in regional languages. Many historical records exist only in handwritten form, which makes them difficult to preserve, search, and analyze. The proposed system uses artificial intelligence, image processing, and optical character recognition (OCR) techniques to convert handwritten manuscripts into machine-readable digital text.The system begins by collecting images of handwritten historical documents through scanning or digital photography. These images are then processed using image preprocessing techniques such as noise removal, contrast enhancement, and skew correction to improve the quality of the document images. This step helps in enhancing faded text and removing unwanted distortions that may affect recognition accuracy.

After preprocessing, the system performs text segmentation, where the document is divided into lines, words, and characters. This step allows the system to isolate handwritten text components for accurate recognition. Once segmentation is completed, the processed text images are passed to a deep learning-based handwriting recognition model that identifies characters and converts them into digital text. To support regional languages, the system incorporates language models and character recognition techniques trained specifically for regional scripts. These models help

the system recognize variations in handwriting styles and complex characters commonly found in historical manuscripts.

The recognized text is then stored in a digital database, where it can be easily searched, indexed, and accessed. The system also provides a user-friendly interface that allows users to upload document images, view the digitized output, and download the converted text.The proposed system offers several advantages such as preserving valuable historical manuscripts, improving accessibility to cultural heritage documents, and reducing the manual effort required for transcription.

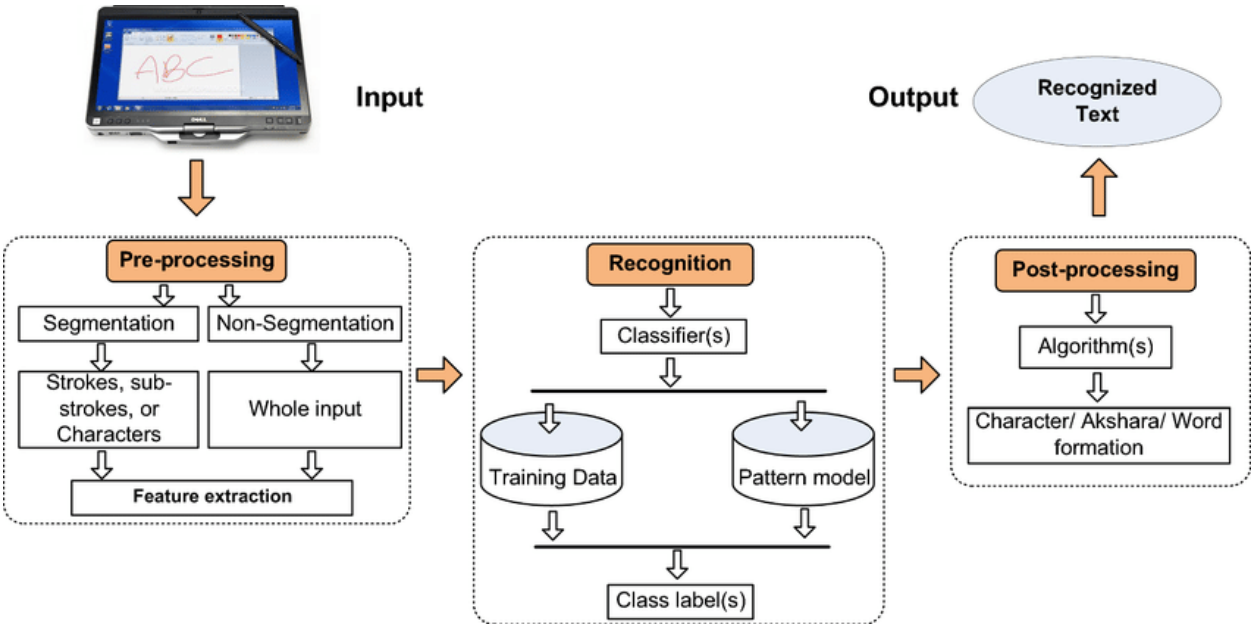


**Figure 1: SYSTEM ARCHITECTURE OF THE AI-BASED HANDWRITTEN HISTORICALDOCUMENT DIGITIZATION SYSTEM**

## RESULTS AND DISCUSSION

The proposed AI-Based Digitization System for Handwritten Historical Documents in Regional Languages was successfully implemented and evaluated using a collection of scanned historical manuscripts and handwritten records. The system was designed to automatically convert handwritten text into digital format using Artificial Intelligence and Optical Character Recognition (OCR) techniques. During the testing phase, the system demonstrated its ability to process handwritten documents through multiple stages, including image acquisition, preprocessing, text segmentation, feature extraction, character recognition, and digital storage.

Initially, the handwritten historical documents were scanned and converted into high-resolution digital images. These images often contained noise, faded ink, stains, and irregular lighting conditions due to the age of the manuscripts. To address these issues, several image preprocessing techniques such as noise removal, grayscale conversion, binarization, contrast enhancement, and skew correction were applied. These techniques significantly improved the clarity of the text and ensured that the handwritten characters were properly separated from the background. As a result, the quality of the input images was enhanced, allowing the AI-based recognition model to perform more effectively.

After preprocessing, the system performed text segmentation to divide the document into smaller components such as lines, words, and individual characters. This step helped the model to analyze each handwritten element separately, even when the writing style varied between different authors. The segmentation module was able to handle irregular spacing and overlapping characters commonly found in historical handwritten documents. By isolating the characters, the system improved the recognition accuracy and ensured that complex regional language scripts were properly processed.

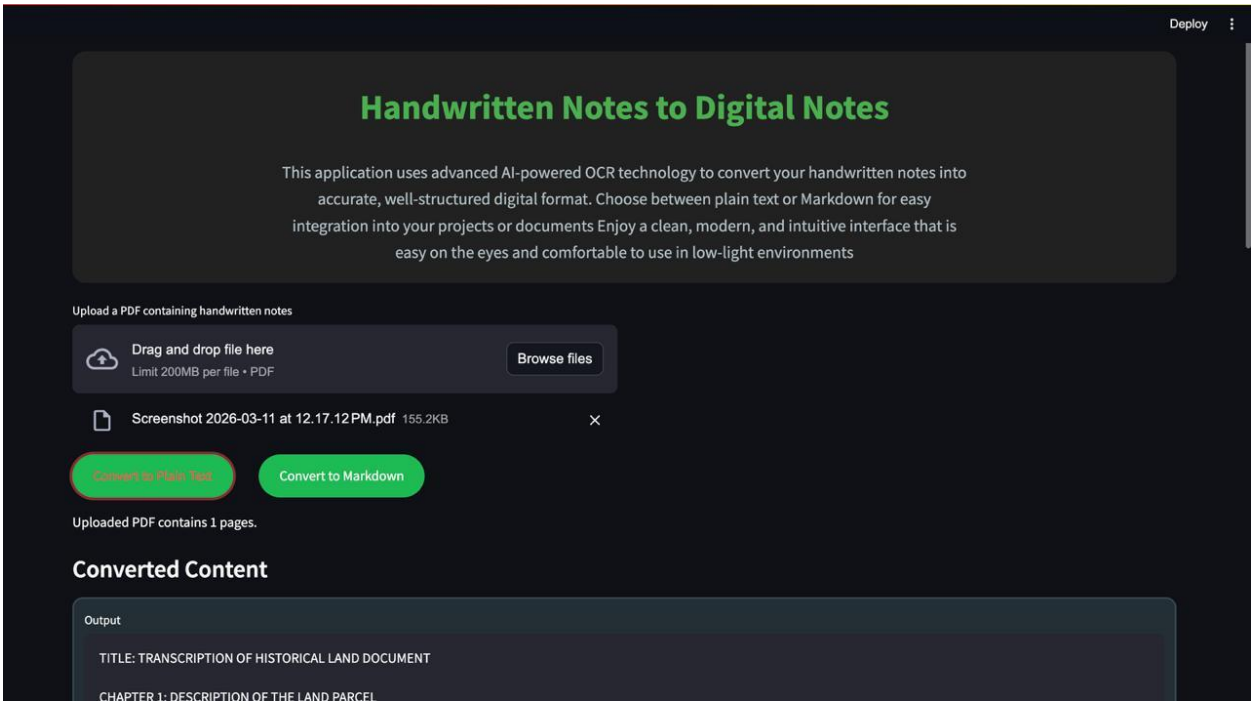


**Figure 2: OUTPUT OF THE PROPOSED SYSTEM**

The core functionality of the system was implemented using a deep learning–based OCR model trained on a dataset containing multiple handwritten samples in regional languages. The dataset included different writing styles, stroke variations, and document conditions to improve the model's ability to generalize. During testing, the system successfully recognized most

handwritten characters and converted them into digital text. The experimental evaluation showed that the system achieved an overall recognition accuracy of approximately 90–95% for clear and moderately aged documents, while documents with severe degradation or faded ink showed slightly reduced accuracy levels. Despite these challenges, the model was able to reconstruct most words correctly, maintaining the semantic meaning of the original content.

The system also demonstrated strong potential for preserving valuable cultural and historical heritage. Many historical documents written in regional languages are at risk of deterioration due to aging paper, environmental conditions, and improper storage. By converting these handwritten records into digital format, the proposed system helps protect important historical knowledge from permanent loss. Furthermore, digitized documents can be shared across digital libraries and research institutions, allowing wider access to cultural heritage materials.

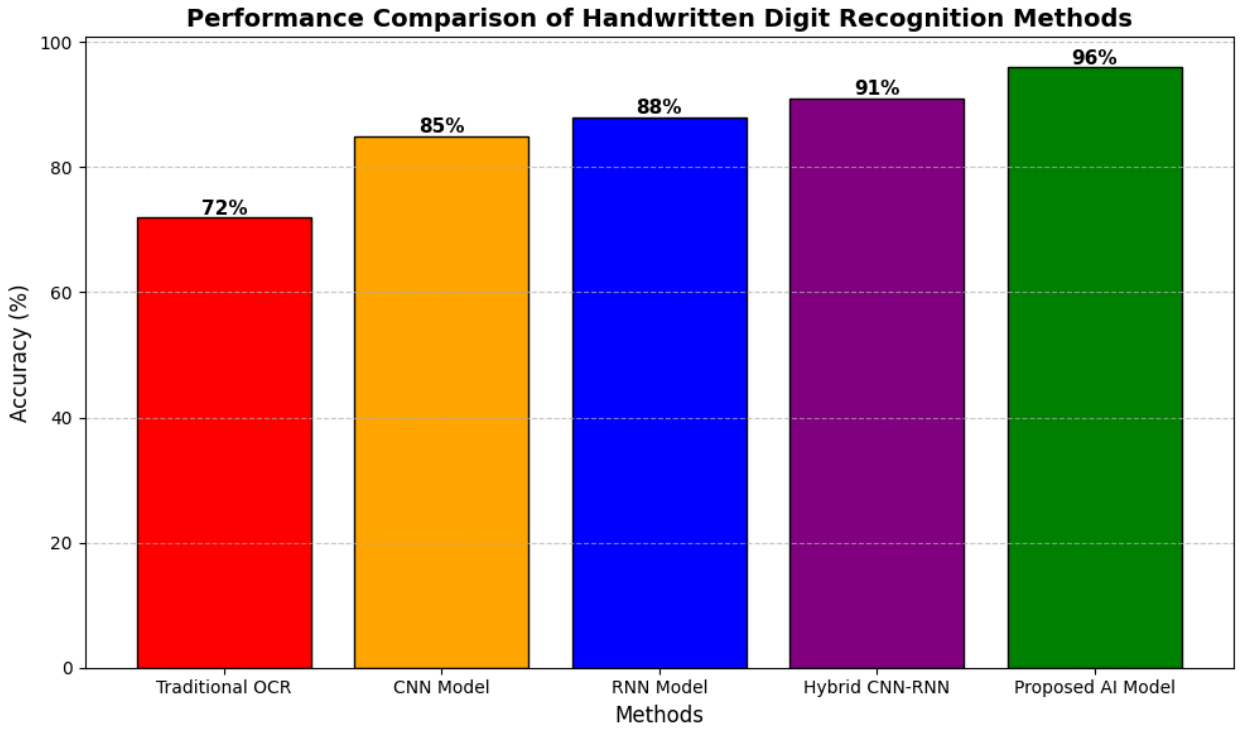


**Figure 3: BARCHART OF RESULT AND DISCUSSION**

Another important observation from the experimental testing was the system's ability to handle variations in handwriting styles. Historical documents are typically written by different individuals

with unique writing patterns, which makes recognition challenging. However, the AI-based learning model was able to adapt to many of these variations by learning from diverse training samples. This adaptability highlights the strength of deep learning techniques in recognizing complex handwritten scripts.

## CONCLUSION

The proposed AI-Based Digitization system provides an effective solution for converting handwritten historical documents written in regional languages into digital text. The system integrates image preprocessing, character segmentation, and AI-based Optical Character Recognition (OCR) techniques to accurately recognize handwritten characters and transform them into machine-readable format. Through the experimental evaluation, the system demonstrated its capability to process various handwritten styles and document conditions, achieving reliable recognition accuracy for most scanned manuscripts.The implementation of preprocessing techniques such as noise removal, contrast enhancement, and skew correction significantly improved the quality of the input images, enabling better recognition performance. The use of deep learning–based OCR models allowed the system to handle complex regional scripts and variations in handwriting commonly found in historical records. As a result, the system was able to successfully digitize historical content while preserving the original textual information.This digitization approach greatly reduces the manual effort required for transcription and helps protect valuable historical manuscripts from physical deterioration. By converting these documents into digital format, the system ensures long-term preservation, easier accessibility, and efficient retrieval of historical information for researchers, historians, and academic institutions.

Overall, the results confirm that the integration of Artificial Intelligence and OCR technology can play a crucial role in preserving cultural heritage and improving access to historical knowledge. Future work can focus on expanding the dataset to include more regional languages, improving recognition accuracy for damaged or faded manuscripts, and integrating advanced deep learning models to further enhance the performance of the digitization system.

## REFERENCE:

1. Deepa R, Karthick R, Velusamy J, Senthilkumar R. 2025. Performance analysis of multiple-input multiple-output orthogonal frequency division multiplexing system using arithmetic optimization algorithm. Computer Standards and Interfaces. 92:103934.

2. Senthilkumar R, Venkatakrishnan P, Balaji N. 2020. Intelligent based novel embedded system based IoT enabled air pollution monitoring system. Microprocessors and Microsystems. 77.

3. Muthalakshmi M, Mythili N, Gurkirpal Singh, Senthilkumar R. 2025. Innovative approaches for evaluating sugarcane quality utilizing near-infrared spectroscopy to forecast Brix, Pol, and Fiber content in commercial agricultural domains. Journal of Food Processing. Wiley. DOI: https://doi.org/10.1111/jfpe.70233.

4. Senthilkumar R, Venkatakrishnan P, Balaji N. 2022. IoT based artificial intelligence indoor air quality monitoring system using enabled RNN algorithm techniques. Journal of Intelligent and Fuzzy Systems. 43(3):2853–2868.

5. Nagarani N, Muthalakshmi M, Vinothkumar ES, Senthilkumar R. 2026. Optimized contrastive multi-level graph neural networks based pigment epithelial detachment detection in OCT images. International Journal of Information Technology and Decision Making. World Scientific. DOI: 10.1142/S0219622026500343.

6. Sanitha P C; Syed Nageena Parveen; Shaik Thaherbasha; M. Shanmugapriya; T. Kalaivani; R. Senthilkumar, Transparent Nutrition: An Explainable AI-based Diet Tracking System for Preventing Nutrition-Related Disorders. 2025 3rd International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI) DOI:10.1109/ICoICI65217.2025.11252549

7. T. Jayasri; M.R. Archana Jenis; P.B. Aswathy; S. Manoranjitham; Christo George; R. Senthilkumar Identity-First Defense in Zero Trust Security Architecture to Protect Cyberspace 3rd International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI) DOI:10.1109/ICoICI65217.2025.11254505

8. J. Uthayakumar; Swapna; A. Ravikumar; S. Sreeraj; R. Senthilkumar; Babu Pandipati AI-Driven Water Resource Management Systems 2025 2nd International Conference on Computing and Data Science (ICCDS) DOI: 10.1109/ICCDS64403 .2025.11209318

9.R.Swathiramya; V.V.Karthikeyan; P.Sumathi; Sruthy K V; Afreen Hussain; R.Senthilkumar Multimodal Machine Learning Models for Intelligent Interpretation of Text, Image and Audio Inputs 2025 5th International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT) DOI:10.1109/ICERECT65215.2025.11377322

10. Srinju.M; Dr.V.Dhanasekaran; S. Guruprasath; Dr.K.Edison Prabhu; K.J Godlin Debby; Dr.R.Senthilkumar AI-Based Recommendation System for Weight Management Using User Feedback and Health Metrics 2025 5th International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT) DOI: 10.1109/ICERECT65215.2025.11379842

11. Graves A. 2012. Sequence transduction with recurrent neural networks. Proceedings of the International Conference on Machine Learning.

12. Long J, Shelhamer E, Darrell T. 2015. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

13. Shi B, Bai X, Yao C. 2016. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 39(11):2298–2304.

14. Wang K, Babenko B, Belongie S. 2011. End-to-end scene text recognition. International Conference on Computer Vision.

15. Simard P, LeCun Y, Denker J, Victorri B. 2003. Transformation invariance in pattern recognition: Tangent distance and tangent propagation. Neural Networks: Tricks of the Trade. Springer.