



# International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

*(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)*



**Impact Factor: 8.699**

**Volume 14, Issue 3, March 2025**

# Handwritten Language Detection using Deep Learning

Prof. Vishal Patil<sup>1</sup>, Prof. Vivek Patil<sup>1</sup>, Prassana Naikade<sup>2</sup>, Aayush Pandhare<sup>2</sup>, Atharva Jangade<sup>2</sup>

Professor, Department Computer Engineering, D. Y. Patil University, Ambi, Pune, India<sup>1</sup>

Student, Department Computer Engineering, D. Y. Patil University, Ambi, Pune, India<sup>2</sup>

**ABSTRACT:** The written character recognition (HCR) methodology still includes a big variety of applications. Across any union government like an Asian nation, reading communicating addresses in many states contains distinct languages. One vital application of HCR within the automatic industry altogether developed nations is the verification of order of payment amounts and signatures. The documents' optical character recognition is compared to written documents by an individual. Characters from many sorts of files, as well as image and word document files, square measure translated victimization this OCR. This study article's primary goal is to supply an answer for numerous handwriting recognition ways, as well as bit input via a mobile screen and picture files. Artificial neural networks applied math techniques, and different techniques square measure utilized by popular methodologies to handle issues that don't seem to be linearly dissociative. This analysis study uses a spread of comparison and recognition techniques to spot handwriting characters in image documents in addition, the study contrasts the analytical way to support vector machine (SVM) heuristic network methodology with applied math, templet similarity, well-organized pattern identification, and comprehensive ways. The technique, depiction, and style of the Written Character Recognition System, as a system assessing and developing outcomes, also are coated within the article. The objective is to point out how well neural networks recognize handwriting characters. Offline Handwritten Recognition which works on scanned images. Offline handwritten is the hardest to find the handwriting. MNIST data set is used for this handwritten digit recognition process and it has 70000 handwritten digits. Many Machine Learning Algorithms are developed which can be used for this digit classification. This paper performs the analysis of accuracies and performance measures of algorithm Convolutional Neural Networks (CNN). Handwritten language detection has garnered significant attention in recent years due to its wide range of applications, including document digitization, forensic analysis, and historical document preservation. This paper presents a comprehensive review and analysis of various techniques employed in handwritten language detection.

**KEYWORDS:** SVM, CNN, handwriting, MNIST, Machine learning

## I. INTRODUCTION

Handwritten Recognition (HR) is a challenging task in the field of pattern recognition. Because every person in this world has their own way of writing. In real-time applications like the conversion of handwritten information into digital format, verification of signatures, number plate recognition, this kind recognition is required. This recognition was implemented using many Machine Learning techniques like Random Forest, Naïve Bayes and Support Vector Machine, Convolution Neural Network(CNN), etc. The accuracy and correctness are mostly crucial in handwritten digit recognition applications. Even 1% error may lead to inappropriate results in real-time applications. The MNIST (Modified National Institute of Standards and Technology database) data set is used for this recognition process. The MNIST data set has 70,000 handwritten digits. Here 60,000 images are used training examples and 10,000 images are used testing examples. MNIST database is a combination of Special Database 1 and Special Database 3. Special Database1 consists of digits written by school students. Special Database 3 consists of digits written by employees. Half of the training and test set have taken from NIST's training dataset, and the remaining training set and test set were taken from NIST's testing dataset. Artificial intelligence (AI) in simple words is basically making a computer do the work that traditionally requires the human brain. AI has the ability to take in large amounts of data unlike the human and uses that data to recognize patterns, make decisions, and give judgment. In this AI we have a subset which is called Machine learning. ML is used to make computers to learn to behave as humans. This is done by two ways, supervised learning in which the computer is given a set of input data and the required output for it. Now it uses ML to learn the algorithm to understand how that particular input gives this particular output. Now the unsupervised learning is when

input data is provided but with no output. ML (Machine Learning) is the branch of computer science which helps the machines to learn without being programmed. A program learns from data given. It performs tasks and measures accuracy and tells if its performance at doing tasks improved with the experience or not. We use algorithms and other techniques in Machine Learning instead of doing the programming part. Machines learn from past experiences and examples. A model can be built based on their past experiences, so that it can be used to predict the new values. If the question or the problem is too large and difficult to solve, then it can be used. It can also help finding the answers to questions based on the analysis of data. It requires very less time to find important things when a large amount of data is given. It can also solve very complex problems as machines can learn faster than humans and they may even exceed humans in some fields. As a result, its demand is rising continuously. Machine Learning is catching up with cloud computing and big data as it can solve many difficult problems with ease. It's used in many applications. Handwritten language detection is a fascinating and essential field of research and technology with profound implications for our digital age. In a world where written communication spans numerous languages and scripts, the ability to automatically identify the language or script of handwritten text is of utmost importance. This introduction provides an overview of what handwritten language detection is, why it is significant, and the challenges and applications associated with it. Handwritten language detection, often referred to as script or language identification, is the process of determining the language or script in which a piece of handwritten text or document is composed. This technology aims to bridge the gap between the analog and digital worlds, making it possible to understand, analyze, and interact with handwritten content in a more automated and intelligent manner.

## II. LITERATURE SURVEY

[1] Andrew Smith have introduced Offline handwritten text recognition is the process of extracting text from a photograph and translating it to a digital image. The past approaches were dependent on lexical segmentation, complex feature extractions along with some knowledge related to the language. The proposed approach combines a convolutional recurrent neural network with connectionist temporal classification, which is a sort of neural network output and will be connected with a score function, to recognize offline handwritten texts. The applied methods are character independent which makes the model globally trainable and can be applied for various languages. Optical character recognition is the process of converting printed texts, whether they are printed or handwritten, into digital, machine-readable language. To recognize texts, a convolutional recurrent neural network with three neural network blocks was used: a convolutional block for image feature extraction, a recurrent block for sequence learning, and a final block for labelling with connectionist temporal classification linked to the scoring function.

[2]Thomas Deselaers developed the handwritten text Recognition (HTR) system is implemented in Optical Character Recognition (OCR) model. Three techniques that were used to build a scalable handwritten text recognition were HTR system, OCR model and a Dual head which is the integration of HTR into OCR. When compared to handwritten text recognition systems and optical character recognition models, the dual head model was able to categorize both printed and handwritten text, yielding the highest accuracy. Firstly there is a need of a good amount of training data with good quality and then a line recognition model based on the neural network and without any recurrent connections.

[3] Anshul Gupta have introduced the method of identifying offline handwritten recognition of English words in which individual characters are recognized first, followed by the complete word. The primary approaches to handwritten text recognition are divided into two categories: holistic and segmentation. For the recognition of limited size vocabulary holistic based approach is used where from the entire word image the global features that are extracted are considered. When the vocabulary grows larger, the complexity of this holistic approach grows as well, resulting in a decrease in the rate of character recognition. The other is segmentation-based, which uses a bottom-up approach that starts at the character level and works its way up to generating a meaningful word. After the segmentation is done the problem will be reduced from recognizing an entire word into recognizing a single character and thereby increasing the size of the vocabulary. To identify the individual characters neural networks will be used

[4] Peng Ren Using deep learning techniques, they have developed a handwritten text recognition programme. The proposed method is based on a detection algorithm for objects. Here the recognition of offline handwritten texts takes place in two steps that is first by preprocessing then by the recognition of the characters. Preprocessing is done first with a quicker R-CNN, and then the character is recognized with a convolutional neural network. The problem of the segmentation of characters have been reduced as the detection of objects. There will be a preliminary processing where the segmentation of sentence into words takes place and then in the meticulous processing there involves segmentation

of words into characters and later the characters will be recognized.

[5] Francisco Zamora-Martinez introduced a new technique for removing slants from handwritten text has been developed, and the size of the text images has been balanced using artificial neural networks. Normalization of the handwritten text from the scanned image includes image cleaning, page skew correction and line detection. Skew-corrected lines database have been used so pages skew correction and line detection are skipped. There are various preprocessing steps to reduce the variations in handwritings. The line that has been scanned from the image is first cleaned before preprocessing. The next step is the slope removal and then the slant removal. When the entire image is free from slopes and the slants normalization is done for the size of the text line so as to minimize the changes in size and the position. The ANN system have been used for recognizing the offline handwritten text lines.

[6] Martin Rajnoha introduced the method for recognition for handwritten text which is of comenia font. The proposed method has preprocessing of data, normalization of data and based on the Support Vector Machine there is an optical character recognition. The comenia script is simple and modern and is same as block letters. The proposed model made use of more than one classification model for the purpose of character recognition which showed an increase in the accuracy when compared with the single model approach. The entire method of recognizing handwritten texts involves feature extraction and it uses support vector machine for the classification.

[7] Batuhan Balsi developed method to classify the individual handwritten word so that it can be translated to a digital form. To approaches have been used in this method first is to classify the words directly then then the segmentation of the characters. To train a model for word classification, a convolutional neural network with several topologies was utilized. To construct the bounding boxes for a single character long short term memory networks with convolution is used. The segmented characters are then transferred to a convolutional neural network for classification, and each word is then reconstructed based on the findings of character classification and segmentation. To improve the results of direct classification for a word made an alternate move by segmenting the characters and then the word is reconstructed by classifying each character individually.

[8] Rohan Vaidya introduced the method for the detection of offline handwritten characters using deep neural networks. The handwritten character recognition system based on image segmentation have been designed in this proposed method for the development of this system python programming language have been used. Various tools like android opencv and tensor flow have been used to design the offline handwritten character recognition system. The built Android application must allow the user to take a photo of handwritten text that must be identified using the Android phone's camera. In order to serve the predictions there is a use of pretrained neural network model then the image processing operation is performed. The neural network model is trained using tensor flow. The processing of image involves preprocessing where the noise from the image will be removed. After the preprocessing there will be conversion of image into grayscale. After the image has been converted to grayscale, thresholding will be used to separate the image's darker and lighter regions. [9] Vu Pham introduces a recurrent neural network to recognize the handwritten texts where the image which is taken as an image is divided into blocks and is then sent to the long short term memory layers which will scan the input and the output of this long short term memory layer will be sent to a convolutional layer. The output of the convolutional layer will be added from vertically and then will be given for softmaxing where in the softmax layer and then the output of the softmax layer is processed by connectionist temporal classification. Here dropout involves removal of some hidden units from the training randomly and then using it for the purpose of testing. Another method which is same as that of dropout is Drop connect where instead of dropping the hidden units' values here the connection is dropped. Both convolution and recurrent layers are compatible with the dropout. First the dropout will be applied for the top layer of the long short term memory. If the size of the model is large and if there is an over fitting found then the dropout is helpful

### III. PROPOSED SYSTEM

Handwritten language detection involves determining the language of text written by hand. In below figure 1 is a high-level algorithm for handwritten language detection: Algorithm for Handwritten Language Detection: **Data Collection and Preprocessing:** Gather a dataset of handwritten text samples for different languages. Preprocess the handwritten images to improve their quality (e.g., grayscale conversion, denoising, deskewing). **Feature Extraction:** Extract relevant features from the preprocessed images. Common features include: Histogram of Oriented Gradients (HOG): Captures gradient information. Scale-Invariant Feature Transform (SIFT): Extracts key points. Convolutional Neural

Network (CNN) features: If using deep learning approaches. **Data Labeling:** Label each handwritten text sample with the corresponding language. This step is crucial for supervised learning. **Split Data:** Split the dataset into training and testing sets. Typically, 70-80% for training and the rest for testing.

**Model Selection:** Choose a machine learning algorithm or model for language detection. Common choices include: Support Vector Machines (SVM) Random Forest Convolutional Neural Networks (CNN) Recurrent Neural Networks (RNN) - Transformer-based models (e.g., BERT, GPT) **Training:** Train the selected model using the training dataset. The model learns to recognize the features and patterns associated with each language. **Testing:** Evaluate the model's performance on the testing dataset to assess its accuracy. **Inference:** Use the trained model to make predictions on new handwritten text samples.

**Post-processing:** Implement post-processing steps to further refine the language detection result. This might include voting mechanisms or language model integration for context-based decisions.

**Deployment:** Deploy the trained model in a suitable application or platform. Set up an API or interface for language detection. **Monitoring and Maintenance:** Regularly update the model with new data for better accuracy. Monitor its performance in the real-world application and make necessary adjustments.

**User Interface:** Design a user-friendly interface for users to interact with the system.

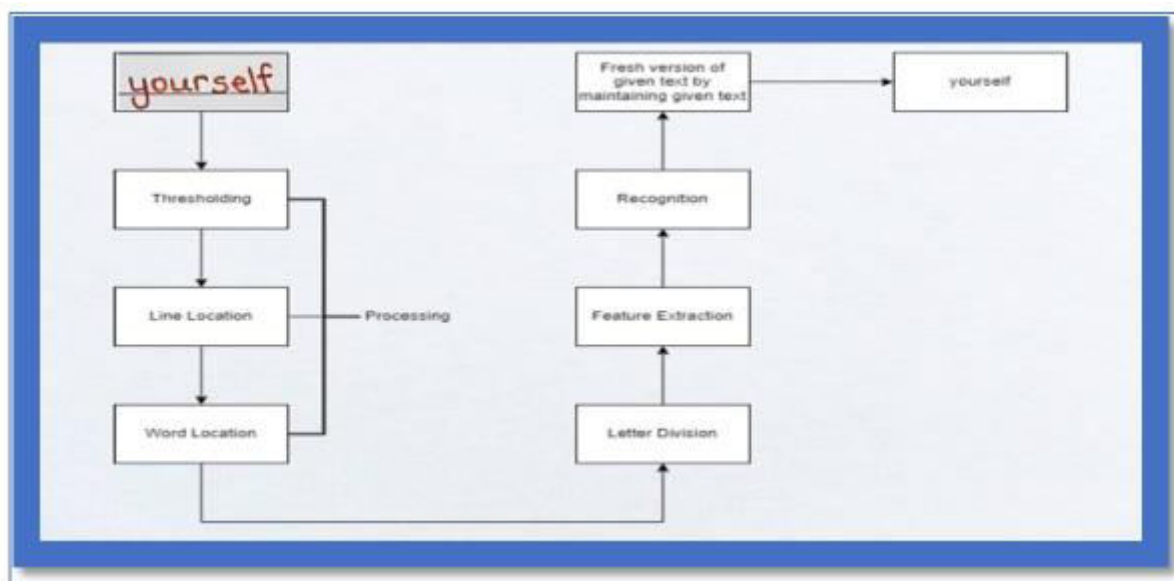


Figure 1. System Architecture diagram of the setup

#### IV. SCOPE

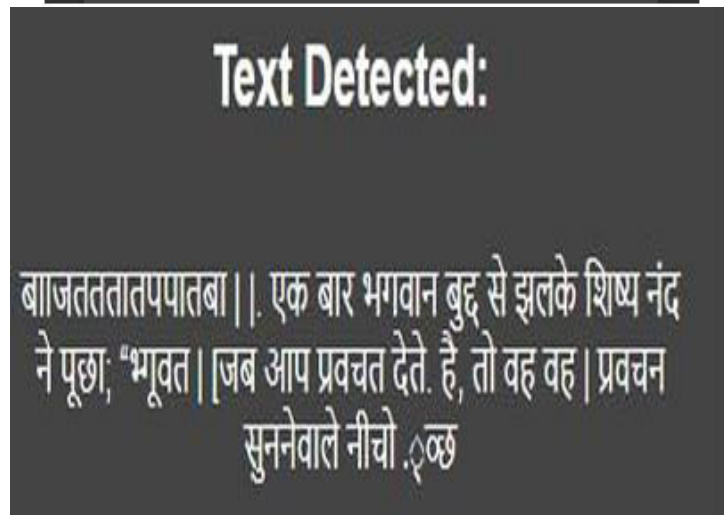
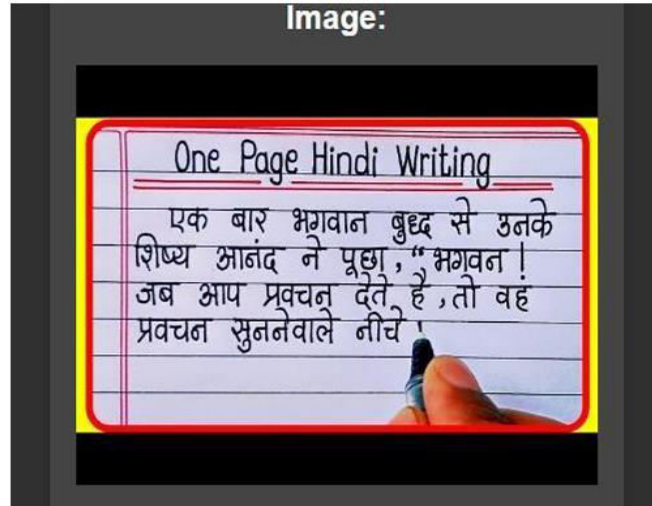
##### 1. Academic & Education Sector

- **Automated Grading:** Recognizing handwritten answers for digital evaluation.
- **Digital Note Conversion:** Converting handwritten lecture notes into editable digital text.
- **Smart Learning Applications:** AI-powered handwriting analysis for personalized learning.

##### 2. Business & Documentation

- **Paperless Offices:** Digitizing handwritten forms, contracts, and invoices.
- **Banking & Financial Services:** Automating check and signature verification.
- **Medical Records Management:** Converting doctors' prescriptions and patient records into structured data.

## V. RESULT



## VI. CONCLUSION

In Conclusion handwritten language detection using machine learning is a valuable and versatile application with numerous real-world use cases. It involves the development of models and systems that can automatically identify and classify handwritten text into different languages or scripts. Here are some key points to summarize the significance and implications of this technology:

- **Multilingual Recognition:** Handwriting language detection enables the automatic recognition of multiple languages or scripts within handwritten documents. This is particularly useful in diverse linguistic and cultural contexts.
- **Automated Data Entry:** It facilitates the automation of data entry tasks, making it more efficient and reducing the risk of human error when dealing with handwritten text.

## REFERENCES

- [1] Hassan, L, "Accessibility of games and game-based applications: A systematic literature review and mapping of future directions," New Media & Society, 2024, 26(4), 2336-2384.
- [2] Ammour N, Bazi Y, Alajlan N, "Multimodal Approach for Enhancing Biometric Authentication," Journal of Imaging. 2023, 9(9):168.

- [3] Nguyen, T. T. H., Jatowt, A., Coustaty, M., & Doucet, A, “Survey of post-OCR processing approaches,” ACM Computing Surveys (CSUR), 2021, 54(6), 1-37
- [4] Graves, A, “Generating sequences with recurrent neural networks,” arXiv preprint arXiv:1308.0850.
- [5] Manogaran, G., & Lopez, D, “A survey of big data architectures and machine learning algorithms in healthcare,” International Journal of Biomedical Engineering and Technology, 2017, 5(2-4), 182-211.
- [6] Schenk, J., Lenz, J., & Rigoll, G, “ Novel script line identification method for script normalization and feature extraction in on-line handwritten whiteboard note recognition,” Pattern recognition, 42(12), 2009, 3383-3393.
- [7] A. Antonacopoulos, B. Gatos and N. Stamatopoulos, “Handwriting Segmentation Contest, in 2007,” 9th International Conference on Document Analysis and Recognition, Parana, 2007, pp. 1284-1288



INTERNATIONAL  
STANDARD  
SERIAL  
NUMBER  
INDIA



# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  [ijirset@gmail.com](mailto:ijirset@gmail.com)



[www.ijirset.com](http://www.ijirset.com)

Scan to save the contact details