Language Recognization From Handwritig Based On Machine Learning And Deep Learning

S. Satyanarayana[1], M.B.V.Gangadararao[2], S. Ahmed[3], R. Jyothi Babu[4], Mr. P. Aditya Shiva Shankar[5]
[1,2,3,4]*Department of Computer Science and Engineering (Data Science), Raghu Engineering College (Autonomous), Dakamarri, Visakhapatnam, Affiliated to JNTU Gurajda, Vizianagaram*
[5]*Guide, Assistant Professor, Department of Computer Science and Engineering (Data Science), Raghu Engineering College (Autonomous), Dakamarri, Visakhapatnam,Affiliated to JNTU, Gurajda,Vizianagaram*

***Abstract*—Handwritten text recognition plays a vital role in document digitization, multilingual communication, and intelligent human–computer interaction. With the growing diversity of handwritten scripts and languages, automatic language recognition from handwriting has become a challenging research problem. This project presents a language recognition system from handwritten text using machine learning and deep learning techniques. The proposed approach involves preprocessing handwritten input images through noise removal, normalization, and segmentation, followed by feature extraction to capture structural and statistical characteristics of handwriting. Traditional machine learning classifiers such as Support Vector Machines (SVM) and Random Forests are explored, along with deep learning models like Convolutional Neural Networks (CNNs) for automatic feature learning. The system is trained and evaluated on handwritten samples from multiple languages to accurately identify the language without prior knowledge of the script. Experimental results demonstrate that deep learning–based models outperform conventional machine learning methods in terms of accuracy and robustness. This work highlights the effectiveness of combining image processing and deep learning techniques for reliable handwritten language recognition and its potential applications in document analysis, translation systems, and smart archival solutions. Keywords: Handwritten Language Recognition, Machine Learning, Deep Learning, Convolutional Neural Networks, Image Processing, Multilingual Handwriting.**

I. INTRODUCTION

Handwritten text recognition is an important research domain in the fields of pattern recognition, artificial intelligence, and computer vision. With the rapid growth of digital technologies, there is an increasing need to convert handwritten documents into machine-readable formats. Many sectors such as banking, education, healthcare, government institutions, and historical archives require automatic processing of handwritten documents for efficient storage, retrieval, and analysis.

Unlike printed text, handwritten text varies significantly from person to person due to differences in writing styles, stroke patterns, character spacing, writing speed, and cultural influences. These variations make handwritten language recognition a challenging task.

Additionally, multilingual handwritten recognition becomes more complex due to the structural differences between scripts such as Latin, Devanagari, Tamil, and Telugu. Therefore, developing an efficient and robust handwritten language recognition system remains an active research problem.

Traditional handwritten recognition systems mainly relied on manual feature extraction techniques followed by classical machine learning algorithms such as Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Decision Trees, and Random Forest classifiers. These approaches require domain expertise for feature engineering and often fail when dealing with large variations in handwriting styles. Furthermore, their performance heavily depends on the quality of extracted features.

Recent advancements in deep learning have significantly improved the performance of handwriting recognition systems. Deep learning models, particularly Convolutional Neural Networks (CNNs), automatically learn hierarchical feature

representations directly from input images without the need for manual feature extraction. CNN models have demonstrated remarkable success in various computer vision tasks such as image classification, object detection, face recognition, and optical character recognition (OCR).

Figure.1 Overview of Handwritten Language Recognition System

II. LITERATURE SURVEY

Handwritten text and language recognition has been an active research area in pattern recognition and computer vision for several decades. Many researchers have proposed different techniques based on machine learning and deep learning algorithms to improve recognition accuracy and robustness. Early research mainly focused on optical character recognition (OCR) using traditional machine learning algorithms. Support Vector Machine (SVM) became one of the most widely used classifiers due to its strong generalization capability and effectiveness in high-dimensional data. Several studies demonstrated that SVM provides good accuracy when combined with structural and statistical feature extraction methods. However, these approaches required careful feature engineering and were sensitive to variations in handwriting styles.

Artificial Neural Networks (ANN) were also widely used for handwritten character recognition. Neural networks mimic the human brain by learning patterns from data through interconnected neurons. Multi-Layer Perceptron (MLP) models were commonly applied for character classification tasks. Although ANN models improved recognition performance compared to traditional classifiers, they required significant computational resources and careful parameter tuning.

Naïve Bayes classifiers were also explored for handwritten recognition due to their simplicity and fast computation. These probabilistic classifiers work based on Bayes theorem and assume independence between features. While they performed well for simple classification tasks, their performance was limited when dealing with complex handwriting variations. With the advancement of deep learning, Convolutional Neural Networks (CNNs) have become the most dominant approach for handwritten recognition problems. CNN models automatically extract important features directly from images, eliminating the need for manual feature extraction. Researchers have demonstrated that CNN architectures such as LeNet, AlexNet, VGGNet, and ResNet significantly improve recognition accuracy compared to traditional machine learning approaches.

Recent studies have shown that CNN-based recognition systems achieve accuracy above 95% for handwritten character recognition tasks across multiple languages. The success of CNN is mainly due to its ability to learn hierarchical features such as edges, shapes, textures, and complex patterns from handwritten images. Pooling layers further improve robustness by reducing spatial variations.

Some researchers have also explored hybrid approaches combining machine learning and deep learning techniques. These systems use traditional preprocessing and segmentation techniques along with deep learning classifiers to achieve better performance. The integration of image processing techniques such as noise filtering, binarization, and normalization has also been shown to significantly improve recognition accuracy.

Despite these advancements, several challenges still exist in handwritten language recognition. These include large variations in handwriting styles, limited availability of multilingual datasets, noise in scanned documents, and similarities between characters of different languages. These challenges require more robust and generalized recognition models.

This research aims to address these limitations by developing a multilingual handwritten language recognition system using both machine learning and deep learning techniques. The study also performs a comparative analysis to demonstrate the superiority of CNN models over traditional classifiers in terms of accuracy and robustness.

Another important direction in handwritten language recognition research focuses on feature extraction techniques. Researchers have proposed various feature extraction approaches such as Histogram of Oriented Gradients (HOG), Scale Invariant Feature Transform (SIFT), zoning methods, and projection profiles. These techniques help in identifying unique patterns in handwritten characters. However, the major limitation of these methods is that they depend heavily on handcrafted features, which may not generalize well across different datasets and writing styles.

In recent years, deep learning approaches have addressed this limitation by introducing automatic feature learning.

CNN models learn low-level features such as edges and curves in the initial layers and high-level semantic features in deeper layers. This hierarchical learning structure makes CNNs highly effective for handwritten language recognition. Researchers have reported that deep learning models significantly reduce preprocessing dependency and improve classification performance. Some studies have also explored Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks for sequence-based handwriting recognition. These models are particularly useful when recognizing continuous handwritten text because they can capture sequential dependencies between characters. However, these models require large training datasets and high computational power, which limits their practical deployment in some applications.

Transfer learning is another emerging technique used in handwritten recognition research. Pretrained deep learning models such as VGG16, ResNet, and Inception are fine-tuned on handwriting datasets to improve recognition performance. This approach reduces training time and improves accuracy, especially when datasets are limited. Many recent studies have shown that transfer learning improves model generalization and reduces overfitting problems. Researchers have also investigated multilingual handwritten recognition systems to support multiple languages within a single framework. These systems aim to identify the language before performing character recognition. Language identification helps improve the efficiency of recognition systems by selecting appropriate models for specific scripts. However, multilingual recognition remains challenging due to similarities between characters across languages and limited availability of benchmark datasets. Another important research focus is improving robustness against noise and distortions. Handwritten documents often contain noise due to scanning errors, paper quality, and ink variations. To address this issue, researchers have proposed preprocessing techniques such as median filtering, morphological operations, adaptive thresholding, and contrast enhancement. These techniques improve image quality and help classification models perform better.

Recent developments also include the use of attention mechanisms and transformer-based architectures in handwriting recognition. These models have shown promising results in natural language processing and are now being explored for computer vision tasks. Transformer-based models can capture global relationships in images and may further improve recognition performance in future handwritten recognition systems.

Although significant progress has been made, there is still scope for improvement in areas such as real-time recognition, low-resource language support, and lightweight models for mobile applications. Future research is expected to focus on improving computational efficiency, reducing training data requirements, and developing more generalized recognition frameworks.

III. PROPOSED METHODOLOGY

The proposed handwritten language recognition system is designed to automatically identify the language from handwritten text images using machine learning and deep learning techniques. The system consists of multiple stages including image acquisition, preprocessing, segmentation, feature extraction, classification, and language prediction. The overall framework is designed to improve recognition accuracy while maintaining robustness against variations in handwriting styles.

The workflow of the proposed system is shown in Fig. 2. The process starts with collecting handwritten samples from different languages, followed by image preprocessing to improve quality. The processed images are then segmented and important features are extracted. Finally, classification models are applied to identify the language.

The main stages of the proposed methodology are

described below.

1. Image Acquisition: Image acquisition is the first step of the proposed system. In this stage, handwritten text samples are collected and converted into digital image format. The images may be obtained using scanners, mobile cameras, or publicly available datasets.The quality of input images plays an important role in system performance. Poor quality images containing blur, noise, or distortions can negatively affect recognition accuracy. Therefore, high-resolution images are preferred for better performance.The dataset used in this work contains multilingual handwritten samples collected from different writers to ensure variability and robustness of the system.
2. Image Preprocessing : Noise present in scanned images is removed using Gaussian filtering and median filtering techniques. This helps in improving image clarity. Input RGB images are converted into grayscale images to reduce computational complexity. Thresholding techniques such as Otsu's method are used to convert grayscale images into binary images. This simplifies feature extraction. Images are resized into fixed dimensions (for example 64×64 or 128×128 pixels) to ensure uniform input to the classification model. Handwritten text may be slightly rotated during scanning. Skew correction techniques align the text properly. These preprocessing steps significantly improve the quality of input data and increase classification accuracy.
3. Classification: The classification stage identifies the language from extracted features. In this research, both machine learning and deep learning classifiers are used to compare performance.The following classifiers are implemented: SVM is used due to its effectiveness in high dimensional data classification. It works by finding an optimal separating hyperplane. Character level segmentation improves feature extraction efficiency and classification accuracy.

Figure.2 System Architecture of Handwritten Language Identification

IV. RESULTS

The performance of the proposed handwritten language recognition system was evaluated using multilingual handwritten datasets consisting of English, Hindi, and Telugu scripts. The evaluation was performed using both machine learning models and deep learning CNN models. The results demonstrate that the CNN model provides better accuracy compared to traditional machine learning approaches. The dataset consisted of handwritten samples collected from different writers to ensure variability in writing styles. Sample handwritten inputs used for testing are shown in Fig. 3. These samples demonstrate variations in stroke thickness, character spacing, and writing patterns, which makes the recognition task challenging.

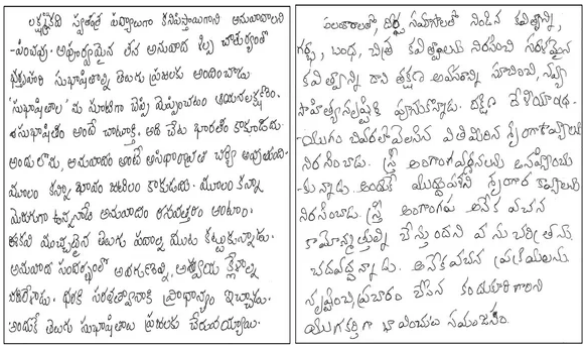

Figure.3 Handwritten Input Sample

The system was also tested on noisy and degraded handwritten documents. Some common challenges observed in handwritten documents include ink smearing, touching characters, uneven character height, and background noise. These challenges are

illustrated in Fig. 4.

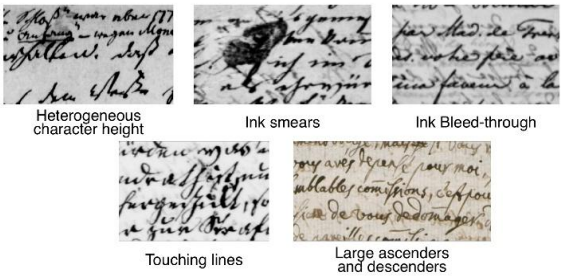

Fig. 4. Challenges in handwritten document recognition.

The CNN model was trained using training and validation datasets. The training performance was evaluated using accuracy and loss curves. The training results are shown in Fig. 5 . The accuracy graph shows that the CNN model gradually improves performance with increasing epochs and stabilizes after sufficient training iterations. Similarly, the loss graph shows a decreasing trend, indicating effective learning of features.

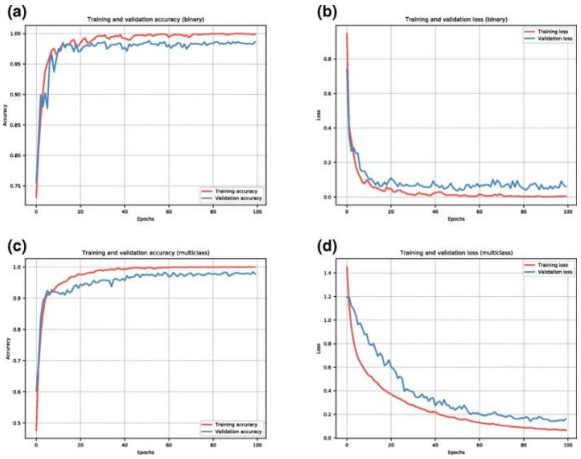

Figure.5 Accuracy graph

The final stage of the proposed handwritten language recognition system is the classification of the input handwritten image into the corresponding language category. After feature extraction and model training, the Convolutional Neural Network (CNN) produces an output vector representing the probability distribution of each language class. The Softmax activation function is used in the output layer to convert the raw prediction scores into probability values. The classification output provides confidence scores for each language class such as English, Hindi, and Telugu. The language with the highest probability score is selected as the final predicted output. This probability-based classification improves the reliability of the recognition system and helps in understanding the confidence level of predictions.

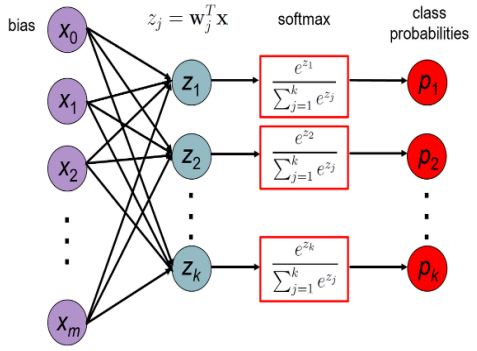

Figure.7 Classification output

Fig. 7 illustrates the classification output of the proposed CNN model, where the Softmax layer assigns probability values to each class. The model successfully identifies the correct language with the highest probability value, demonstrating the effectiveness of deep learning based classification. The experimental results show that the CNN model produces highly accurate predictions due to its ability to learn discriminative features automatically from handwritten images. Compared to traditional machine learning classifiers, CNN achieves better generalization and robustness against handwriting variations. These outputs confirm that the proposed system can effectively recognize multilingual handwritten text with high accuracy. The classification results also demonstrate that deep learning approaches are more suitable for complex handwriting recognition problems compared to conventional machine learning methods.

REFERENCES

[1] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P. "Gradient-based learning applied to document recognition". Proc. IEEE Vol. 86 No. 11, pp. 2278–2324, 1998.

[2] J. Schmidhuber, "Deep learning in neural networks: an overview", Neural Networks,Vol. 61, pp. 85–117, 2015.

[3] Y. LeCun, Y. Bengio, G.E. Hinton, "Deep learning", Nature Vol. 521,No. 7553, pp. 436–444, 2015.

[4] Sharma, R., Kaushik, B., and Gondhi, N. K.

"Devanagari and Gurmukhi Script Recognition in the Context of Machine Learning Classifiers." Journal of Artificial Intelligence Vol.11, No. 2, pp: 65- 70, 2018.

[5] Dongre, V. J., Mankar, V. H. "Development of comprehensive Devnagari numeral and character database for offline handwritten character recognition",Applied Computational Intelligenceand Soft Computing, pp. 1-5, 2012.

[6] Hong, Y., Kwong, S., Wang, H. "Decision-based median filter using k-nearest noise-free pixels". In International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, pp. 1193-1196, April, 2009.

[7] Sahare, P. and Dhok, S.B. "Multilingual character segmentation and recognition schemes for Indian document images". IEEE Access, Vol. 6, pp.10603-10617, 2018.

[8] Boulid, Y., Souhar, A. and Ouagague, M.M. "Spatial and textural aspects for Arabic handwritten characters recognition". IJIMAI, Vol. 5, No. 1, pp.86-91, 2018.

[9] Vapnik, V. "The nature of statistical learning theory". Springer science & business media, 2013.

[10] Simonyan, K., Zisserman, A. "Very deep convolutional networks for largescale image recognition", CoRR abs/, pp. 1409-1556, 2014.

[11] Pramanik, R., Bag, S. "Shape decomposition-based handwritten compound character recognition for bangla ocr". Journal of Visual Communication and Image Representation Vol. 50, pp: 123–134, 2018.

[12] Amarappa, S. and Sathyanarayana, S.V. "Kannada named entity recognition and classification (nerc) based on multinomial na\" ive bayes (mnb) classifier". IJNLC, Vol. 4, No.4, 2015.

[13] Zeiler, M. D., Fergus, R.: "Visualizing and understanding convolutional networks", CoRRabs/, pp. 1311-2901, 2013.

[14] Shalini Puri, Satya Prakash Singh. "An efficient Devanagari character classification in printed and handwritten documents using SVM". Procedia Computer Science, Vol. 152, pp: 111-121, 2019.

[15] Zhang, X.-Y., Bengio, Y., Liu, C.-L.: "Online and offline handwritten Chinese character recognition: a comprehensive study and new benchmark". Pattern Recognition., Vol. 61, pp. 348˜360, 2017.

[16] Xiao, X., Jin, L., Yang, Y., et al.: "Building fast and compact convolutional neural networks for offline handwritten Chinese character recognition". Pattern Recognition., Vol. 72, pp. 72˚81, 2017.

[17] Sandhya, N., and R. Krishnan. "Broken Kannada character recognition—A neural network based approach". In Electrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on, pp. 2047-2050. IEEE, 2016.

[18] Khémiri, A., Echi, A. K., Belaïd, A., et al. "A system for off-line Arabic handwritten word recognition based on Bayesian approach". Int. Conf. Frontiers in Handwriting Recognition (ICFHR), Shenzhen, China, October, pp. 560˚565, 2016.