IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, April 2026
Impact Factor: 8.2

Deep Learning-Driven OCR for Complex and Classical Scripts

[1]Dr. M. Rama Chandra, [2]SK. Naimisha, [3]V. Vishal, [4]Ch. Rohith Kumar
[1]Assistant Professor, Dept of CSE, Sreenidhi Institute of Science and Technology, Hyderabad.
[2]Dept of CSE, UG Student, Sreenidhi Institute of Science and Technology, Hyderabad.
[3]Dept of CSE, UG Student, Sreenidhi Institute of Science and Technology, Hyderabad.
[4]Dept of CSE, UG Student, Sreenidhi Institute of Science and Technology, Hyderabad.
[1]ramchandra.m@sreenidhi.edu.in,[2]22311a05q8@sreenidhi.edu.in,
[3]22311a05n6@sreenidhi.edu.in, [4]22311a05r2@sreenidhi.edu.in

Abstract: *Optical Character Recognition (OCR) is a revolutionary technology that enables documents to be digitized, preserved and accessed by converting images of textual materials into machine readable text. Latin scripts have reached the level of maturity in OCR systems, but Indian scripts, including Telugu and Sanskrit, are quite challenging because of their complicated character structures, compound symbols, and high incidence of diacritical marks. The proposed work is an OCR system based on deep learning with Convolutional Neural Networks (CNNs) to extract features that aid in the classification of both printed and handwritten Telugu and Sanskrit characters. The wide range of samples will make the data set robust and flexible. Binarization and morphological processing are used as preprocessing methods in the system to enhance image quality. The accuracy metrics of experimental assessment show that the proposed CNN-based model is more effective than the traditional machine learning techniques. The study is related to digital preservation, transcription of manuscripts, and contemporary OCR to Indian languages*

Keywords: Optical Character Recognition (OCR), Deep Learning, Convolutional Neural Networks (CNN), Telugu Script Recognition, Sanskrit Script Recognition, Handwritten Character Recognition, Printed Text Recognition, Image Preprocessing, Binarization, Morphological Operations, Character Segmentation, Pattern Recognition, Document Digitization

I. INTRODUCTION

Optical Character Recognition (OCR) is a high technology that helps computers to read and transcribe the textual data in images, scanned documents and photographs and open it as edible, searchable and machine readable text. This has changed how documents are viewed, handled and accessed in many areas like libraries, government offices, institutions, bank systems and in digital archives. OCR is of special importance in document digitization, historical record archiving, data entry automational processes, and language translation and access assistance software. OCR could reduce the amount of manual work, enhance data accuracy, and allow saving of valuable information in long-term format by transforming printed or hand-written text into cyberspace.

Though, OCR systems have attained high level of maturity in English and other latest Latin scripts, the performance of OCR systems on Indian scripts is a major research concern. The scripts of Indian languages are visually complex, structurally rich and very varied in representing them. Of these, Telugu and Sanskrit are the most special because of the complex nature of characters and the wide application in the historical and other literary books.

The Telugu script is marked by rounded forms, loops and curves and this makes borders of characters less prominent than the Latin characters. It is made up of a blend of the vowels, consonants and compound characters created by combining symbols with each other. Further complexity is added by the use of conjunct consonants (ligatures), and multiple vowel modifiers in the form of diacritical marks in Sanskrit, which was frequently written in Telugu script in

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, April 2026 Impact Factor: 8.2

ancient manuscripts. These modifiers may come before, after, over or under the base character with change in its pronunciation and meaning. Moreover, most of the characters in these scripts cannot be differentiated even by the human reader since they appear to be similar to the eyes.

The other difficulty is the availability of printed as well as handwritten versions of these scripts. Handwritten text adds in the diversity of stroke thickness, character spacing, orientation and writing styles among various people. Historical manuscripts are usually destroyed, discoloured, torn paper, and noises that are caused during the scanning. All these combine together to render the task of accurately character recognition next to impossible and hard to execute by standard OCR systems.

The traditional OCR solutions are heavily based on the manual feature extraction procedures, template matching, and rule-based recognition of pattern. These methods involve fixed rules and precise character shapes, that do not scale to the differences in handwriting, font type and document quality. Consequently they do not correctly identify complex scripts (such as Telugu, Syanskrit, etc.); in particular at the ligature and diacritical levels.

The recent development with Deep Learning and especially with Convolutional Neural Networks (CNNs) has completely changed the image recognition and the pattern analysis field. CNNs can automatically extract hierarchical features directly out of raw image data, avoiding the laborious process of manually engineered features. They have the ability to detect edges, curves, textures and structural patterns in characters and are therefore very effective in the recognition of complex scripts. CNNs are also resistant to differences in writing style, noise and distortions, that are prevalent in handwritten and historical documents.

This project is based on the strengths of CNNs to produce a strong and effective OCR system specifically geared towards the identification of Telugu and Sanskrit characters printed and handwritten. The system combines developed image preprocessing methods and deep learning-based feature extraction to overcome the shortcomings of the conventional OCR method and provide a higher recognition rate. The suggested solution will help to digitalize the documents of regional languages, preserve Indian culture, and create contemporary OCR products of Indian language.

II. LITERATURE SURVEY

Optical Character Recognition (OCR) has developed a lot during the last several decades. Overall, earlier methods were mainly based on template matching, whereby the input character images were matched to a bank of pre-stored templates to find the most similar. These techniques work when character shapes are very close to templates, but do not work so well in practice where font styles, font sizes, and font distortions are highly diverse [1]. Prohibitively rigid nature of the template matching renders it ineffective under conditions that characters are distorted in shape, size, or orientation.

To address these drawbacks, scientists proposed feature-based machine learning algorithms that are based on handcrafted features derived out of the images. In this category:

The K-Nearest Neighbors (KNN) uses the nearest examples in the feature space to classify characters.

The Support Vector Machines (SVM) identifies the optimal decision boundaries among the types of characters.

Decision Trees are maps taught on training data, labeled, and learned in order.

The techniques are used in features like edges, stroke direction, contours, and geometric properties extracted manually. Despite offering a better performance than basic template matching, they are still not applicable to the sophisticated Indian scripts. Writing systems of Indians such as Telugu, Sanskrit feature curved and looped character shapes, many ligatures, and various forms of diacritical marks leading to extremely challenging feature selection [2]. Besides, these models have a poor generalization with regard to processing handwritten characters because of a high variance in individual styles of writing.

The current popular engines such as Tesseract used in OCR provide less support to the Indian languages. Though they do quite a good job with clean printed text in Latin scripts, their performance deteriorates considerably when we look at Indian scripts, when reading handwritten or damaged documents. Recognition accuracy of Tesseract has been

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, April 2026
Impact Factor: 8.2

demonstrated to decrease significantly with scripts composed of complicated ligatures and modifiers, which further demonstrates the shortcomings of rule-based and template-based systems [3].

Over the past several years, the Deep Learning methods, specifically Convolutional Neural Networks (CNNs) have transformed image and pattern recognition. CNNs automatically extract hierarchical features that are naturally learned using only input pixels, without using human created features. Since CNNs are able to simultaneously extract both low-level structures (e.g., edges) and high-level structures (e.g., character shapes) across spatial hierarchies, they have been positively proven superior to conventional machine learning algorithms in a wide spectrum of uses in OCR [4]. According to several comparative studies, CNN-based models can significantly enhance the recognition accuracy of Indian scripts by enhancing it against noise, distortions and intra-class variation [5].

In spite of these improvements, the majority of the current literature investigates individual Indian scripts (e.g., Devanagari, Kannada, Tamil), and little is available on the joint recognitions of Telugu and Sanskrit signatures of mixed printed and hand written data. The lack of publicly available datasets, which combine both scripts with handwritten samples, also forces development and benchmarking, and is easily noticeable. These shortcomings of existing studies suggest a definite gap, which leads to the development and launch of a deep learning-based OCR system that can effectively identify printed and handwritten Telugu and Sanskrit characters.

III. METHODOLOGY

The way of doing things in this work is based on the systematic pipeline beginning with the pre-processing of datasets, model training, and evaluation. All the stages are thoroughly developed to address the complexity of the Telugu and Sanskrit scripts and achieve high recognition of printed or handwritten characters.

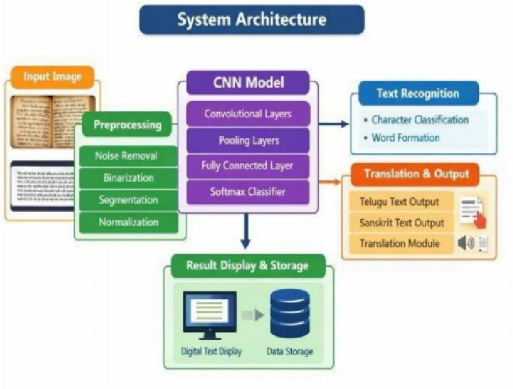

Fig 1:System Architecture

Dataset Collection

An effective OCR system is based on a credible and ecumenical data. In this project, the dataset is built based on various sources to bring as much variability as there is among Telugu and Sanskrit scripts.

Printed Character Samples

Cut-outs contained in texts, storybooks, newspapers, magazines, PDFs.

Diverse font styles and font sizes are dictated to portray the reality of the world.

The character clarity is preserved with the help of high-resolution scans.

Handwritten Character Samples

Gathered and sampled among a group of people of various ages and styles of writing.

The samples are drawn by pens and pencils on plain white papers.

This guarantees stroke variability with thickness, slant, spacing and style.

Diversity of fonts and styles used.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, April 2026
Impact Factor: 8.2

Bold, italic and regular font types are included.
Change in character spacing and alignment.
Samples are isolated characters as well as those words extracted characters.
The labeling of each character image is done manually to produce a supervised learning dataset. The data is then sorted into:
Training set (70%)
Validation set (15%)
Testing set (15%)
This separation guarantees objective consideration of the model.

Data Augmentation

In order to enhance the model to generalize, and to avert overfitting, data augmentation methods are used. Such methods artificially expand the size of the dataset with the formation of altered copies of preexisting images.

Rotation

Rotations of characters are applied in very small angles (between -10 and +15 degrees).
Assists the model in dealing with slanted or distorted scans.

Scaling

Pictures are enlarged and reduced to a small extent.
Enables the model to be aware of the characters of various sizes.

Flipping

Minor horizontal/vertical transformations where necessary.
Brings in diversity in spatial orientation.

Noise Addition

Noise is introduced randomly to represent the actual scanning case.
Trains the model to be robust to image degradation.

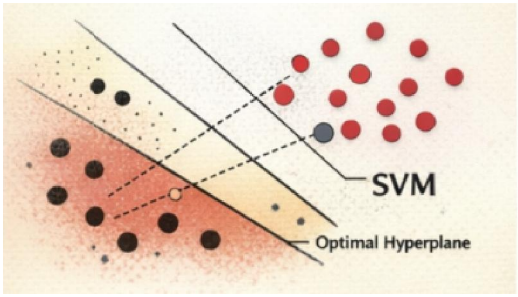

Fig 2:Support Vector Machine

Such augmentation methods make sure that the CNN is fit to invariant features and works well even on unseen data.

Image Preprocessing Techniques

Prior to feeding images to the CNN, a series of preprocessing operation is used to improve image quality and eliminate any undesired noise.

Binarization

Applicants convert grayscale pictures to black and white pictures, with thresholding (e.g. the Otsu technique).
Enhances readability of text and background.

Morphological Operations

Small things that are undesired but happen to be white noises are taken away through erosion.
Broken character strokes are repaired by dilation.
These operations aid in the cleaning of the picture and keeping structure of the character.

DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, April 2026
Impact Factor: 8.2

Edge Detection

Identifies edges of characters with edge operators.

Helps in proper segmentation of words or lines into characters.

Normalization

The images are reduced to a constant size (e.g. 32 by 32 pixels).

To train more quickly, pixel values are scaled to the set of [0,1].

CNN Model Design

Convolutional Neural Network (CNN), is constructed in such a way that it automated character features and classification.

Convolution Layers

Do a series of filters (kernels) to the image.

Extract simple features in low-level layers (edges and curves).

The deeper layers are able to reveal intricate patterns such as loops and ligatures.

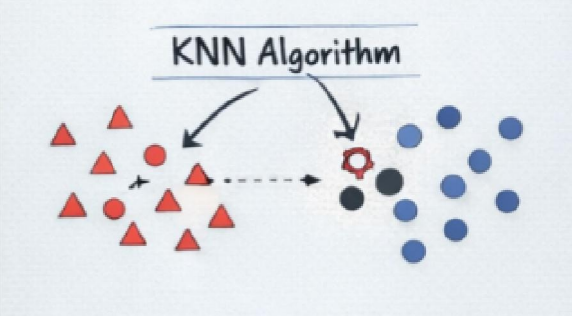

Fig 3: KNN-Algorithm

Pooling Layers

Multiplication of maximum-pooling (2×2) downsizes space dimensions.

Assists in lessening computational complexity and overfitting.

Keeps the most significant attributes.

Flatten Layer

Maps 2D feature maps to 1D.

Fully Connected (Dense) Layers.

High level associations among features extracted.

Deliver decision making capacities in classification.

Softmax Output Layer

Generates probability distribution of all character classes.

The most probable class is picked as the output.

Through this architecture, hierarchical representations of Telugu and Sanskrit characters in the network can be learnt effectively.

Training Parameters

The model performance is much dependent on the training structure.

Optimizer: Adam

Effective optimization algorithm to optimize the learning rates of individual parameters and accelerate the learning process.

Loss Function: Categorical Cross-Entropy.

Applicable in the preferences of multi-class classification tasks such as character recognition.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, April 2026
Impact Factor: 8.2

Epochs: 30–50

The model is trained in a series of iterations to obtain learning stability without overfitting.

Batch Size: 32

A moderate size of a batch, which results in stability of gradient updates and optimization of memory.

In training, both validation accuracy and loss are tracked to prevent overfitting. Best-performing model can also be saved by use of early stopping and model checkpointing method.

IV. RESULTS

It was established and contrasted with the performance of proposed CNN based OCR system to the traditional machine learning models, such as KNN and SVM. This was measured using both printed text and hand written text to know the strength of each of the models under different input conditions.

Evaluation Metrics

The assessment is done mainly on Recognition Accuracy (%), which is computed as:

Accuracy = (Correct Predictions / Total Test Samples)×100

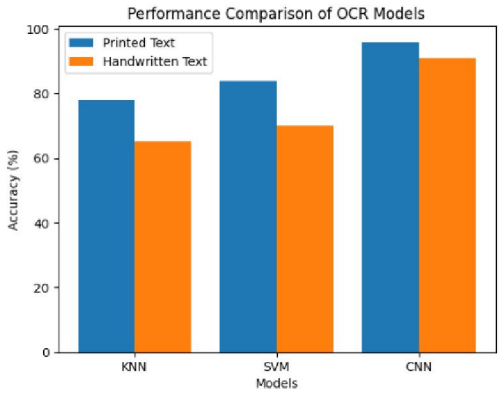

Fig 4:Performance Evaluation Graph

According to the performance evaluation graph, recognition accuracy of various models on both printed and handwritten text is very clear and compares very well. Based on the graph, it is clear that accuracy of the three methods, KNN, SVM, and CNN are dramatically different, which demonstrates the influence of a fundamentally based learning strategy on OCR results.

K-Nearest Neighbors (KNN) uses a relatively low accuracy model especially when it comes to a handwritten text. This is mainly because of the great variation in personal writing styles, the thickness of strokes, the spacing and orientation. Because KNN places a great deal of emphasis on distance-based similarity in feature space, and because it uses manually-extracted features, KNN does not generalize well in the presence of such variations. Consequently, it becomes significantly less effective in handling the handwritten samples when the characters shapes are not similar to its training data.

SVM model reveals a better improvement as compared to KNN in both printed and handwritten texts. SVM proves to be more effective in forming decision boundaries between the classes based on the extracted features. Nevertheless, in spite of this, SVM continues to experience challenges related to the recognition of complex ligatures and diacritical marks typical of Telugu and Sanskrit scripts. Its dependency on hand-crafted features constrains its capability of capturing the complex curves, loops and modifiers in such characters.

The Convolutional Neural Network (CNN) model, on the other hand, performance is significantly high when compared to both KNN and SVM with printed and handwritten text. This is because the CNN can automatically learn the hierarchical features which it uses straight on the input images. CNN layers do not require much of preset functions

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, April 2026
Impact Factor: 8.2

instead they learn to identify edges, curves, stroke patterns and structural properties of characters at various levels. This allows the model to comfortably deal with differences in writing style, noise, and distortion of real world information.

Model	Printed Accuracy	Handwritten Accuracy
KNN	78%	65%
SVM	84%	70%
CNN	**96%**	**91%**

Considering the above, it is evident that the CNN-based method is extremely efficient in OCR of Telugu and Sanskrit scripts. This allows it to learn more complex visual patterns, so it is especially well adapted to recognising characters with ligatures, diacritical marks and curved ones, and its recognition accuracy is far higher than with classic machine learning approaches.

V. DISCUSSION

Structural richness and visual complexity of Telugu and Sanskrit scripts makes the task of developing an OCR system of these scripts challenging. In contrast to Latin writing, in which the characters are mainly straight and distinctly separated, the Telugu and Sanskrit characters are shaped like curves, loops, a combination of formations, and various marks of diacritical symbols fastened in different parts. The features render character segmentation, feature extraction and classification much more challenging. This section of discussion examines the way the proposed CNN-based OCR system tackles these challenges and its performance compared to the traditional methods.

Character variability was one of the major issues that were experienced during the implementation. Writing samples taken by hand of various people were found to have vast variation in the thickness of strokes, the angle of the strokes, distance between them and the style of writing. Conventional methods, including KNN and SVM, were poor at generalizing to these variations as these models are based on manually created features, which cannot adequately model the variability in character shapes. Conversely, CNN model was highly adaptable to learn features directly on the raw images making it to identify patterns despite the fact that the characters were written in various styles.

The other significant challenge was that there were ligatures and diacritical marks. In Telugu and Sanskrit scripts, the vowel modifiers may be located above, below or before or after the base consonant and modify the physical appearance of the character. Similar to several symbols formed by combination of many characters, separating individuals into parts is complicated. The layer design of the CNN was also able to resolve this issue by capturing the local features (edges and curves) and global features (whole character shapes) resulting in a better recognition of these more complex structures.

The quality of the image was also a contributing factor especially with scanned documents and historical writings. Segmentation and recognition were impaired by noise, ink fading and paper degradation. Preprocessing stages of binarization, morphological operations and normalization of images, were necessary in improving clarity of images prior to inputting data into the network. These measures minimised background noise, but increased the visibility of character boundaries, which contributed directly to increased classification performance.

Comparing the results of performance, it can be concluded that CNN model can substantially outperform the traditional machine learning approaches. The accuracy of KNN and SVM was decent with printed text, but their accuracy declined significantly with handwritten samples. This validates the fact that the handcrafted features do not-sufficiently deal with variability in the Indian scripts in the real world. Conversely, the CNN model was highly accurate in both printed and handwritten text and thus showed that it is robust and can be scaled.

The significance of data augmentation is also mentioned in the discussion. The dataset was made more varied through rotation, scaling, flipping, and noise, which means the CNN could learn more about the features of the image that were invariant. This minimized overfitting and enhanced denial to unknown data.

Although the proposed system was successful, there were a few limitations that were experienced. Character segmentation errors were found where the characters were too much linked in terms of words. Also, in rare cases,

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, April 2026 Impact Factor: 8.2

misclassification arose due to scarcity of the training samples. These problems indicate that future research can potentially improve performance by adding the word-level recognition model, i. e. Recurrent Neural Network (RNNs) or Long Short-Term Memory (LSTM) networks.

Altogether, that the discussion supports the idea that deep learning, specifically CNNs, is very effective in OCR-ing Telugu and Sanskrit scripts. CNNs are much better suited to solving the complexities of Indian languages than traditional methods because of the possibility of automatic acquisition of complex visual patterns. The system created within the scope of the current project has great potential to have applications in real-life scenarios like digitization of manuscripts, automated transcription, and preservation of cultural heritage, in a digital form.

VI. CONCLUSION

The project was an optical character recognition (OCR) application that was developed with deep learning with support for Telugu and Sanskrit scripts and employs Convolutional Neural Networks (CNNs). The paper has considered the significant issues relating to these scripts, such as curved and looped character designs, conjunct consonants (ligatures), multiple diacritical characters, and much greater difference between printed and handwritten versions.

The system was able to produce high recognition rates of both printed and handwritten characters through meticulous dataset gathering, preprocessing, data enrichment, and a well-crafted CNN model. The experimental findings showed that CNN model was far better than the independent stand of the traditional machine learning strategies like KNN and SVM, especially with handwritten text and intricate character combinations. The capability of CNNs to extract features (hierarchy) automatically and images has been found to be very effective in identifying complex patterns in Telugu and Sanskrit scripts.

The suggested system will add to the process of digitization of documents, historical manuscripts preservation, and enhanced access to regional language resources. It is also used as a basis to come up with practical OCR applications in education, research and digital libraries using the Indian languages.

VII. FUTURE ENHANCEMENT

Despite the promising outcomes of the system, it is possible to make several improvements to enhance its performance and its applicability:

Sentence-Level Sentence-Word Recognition.

By incorporating sequence models, like RNN or LSTM, it is possible to identify whole words, sentences, rather than individual characters.

Improved Segmentation Techniques

The sophisticated methods of segmentation may be used in order to deal with characters that might touch and overlap better.

Greater and a More varied Dataset.

Increasing samples, particularly, rare ligatures and worn-out manuscripts will make the model more robust.

Extensure to Other Indian Scripts.

It is possible to expand the model to identify the scripts, like Kannada, Tamil, Hindi (Devanagari), and Malayalam.

Real Time OCR and Mobile Application.

Installing the system as a mobile application or real-time camera OCR utility to use it practically.

Post-Processing using Language Models.

Making corrections on misclassified characters through the application of NLP techniques and dictionaries to correct them by their situational context.

Cloud-Based OCR API

Creating an API service to be integrated into digital libraries, archives, and learning platforms.

Manuscript Restoration Integration

Using image improvement and restoration methods and the OCR to enhance recognition of ancient documents.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, April 2026
Impact Factor: 8.2

Such improvements of the future can change the system into an all-encompassing OCR system of Indian languages, facilitating the digital preservation and access to the cultural and literary heritage.

REFERENCES

[1]. Shaik Moinuddin Ahmed and Abdul Wahid (2023). Handwritten OCR for Indic Scripts: A Comprehensive Overview of Machine Learning and Deep Learning Techniques. International Journal on Recent and Innovation Trends in Computing and Communication. — This review examines state-of-the-art approaches for handwritten OCR in Indic scripts, including Deep Neural Networks and hybrid methods.

[2]. M. Gnaneswari and T. Chaitanya Kumar (2023). Application of Deep Convolutional Neural Networks to Telugu Scripts for Optical Character Recognition. International Journal of Computer Trends and Technology (IJCTT), Vol. 71, No. 1, pp. 50–55. — A recent research article applying CNNs to OCR specifically for the Telugu script.

[3]. M. V. Vijaya Saradhi, K. Rakesh, D. Ravi Prasanna, K. Swetha, and B. Prawin (2023). Comprehensive Study of Deep Learning Based Telugu OCR: A Survey. International Journal of Science and Research Archive, Vol. 08(01), pp. 353–356. — Provides a survey of deep learning techniques used for Telugu OCR, covering preprocessing, segmentation, and classification methods.

[4]. Nikitha Nayak, Ms. Rakshitha P., and Mr. Hareesh B. (2026). A Comparative Study of Traditional, Deep Learning, and Multimodal OCR Systems for Document Digitalization and Information. International Journal of Engineering Research & Technology (IJERT), Volume 14, Issue 01. — A recent comparative analysis of conventional and deep learning OCR systems, highlighting modern approaches for document digitization.

[5]. Shivraj Gaikwad, Renu Kachhoria, and Gitanjali Yadav (2025). AI-Based OCR for Digitizing Ancient Indian Texts: Preserving Linguistic Heritage and Overcoming Script Challenges. International Journal of Linguistics Applied Psychology and Technology (IJLAPT), 2(03):47–50. — Discusses AI-powered OCR for ancient Indian scripts, emphasizing integration of deep learning for preservation and accessibility.

[6]. ·Deep Learning, Yoshua Bengio, and Aaron Courville (2016). Deep Learning. MIT Press. — Standard reference for understanding deep learning fundamentals, especially CNN architectures used in modern OCR systems.

[7]. ·Google (2024). Tesseract Open Source OCR Engine Documentation. — Documentation and technical details of a widely used OCR engine for baseline comparison with deep learning methods.