ISSN PRINT 2319 1775 Online 2320 7876

Research paper Volume 11, Sp.Iss 6 , 2022

OPTICAL CHARACTER RECOGNITION FOR TELUGU LANGUAGE USING TESSERACT

S. Sagar Imambi[1], Harsha Mynedi[2]

[1]Professor, Koneru Lakshmaiah Education Foundation (KLEF), Vaddeswaram, Green fields, Guntur, Andhra Pradesh, India - 522302

[2] Graduate,Koneru Lakshmaiah Education Foundation (KLEF), Vaddeswaram, Green fields, Guntur, Andhra Pradesh, India - 522302

DOI : 10.48047/IJFANS/11/S6/037

ABSTRACT

Optical Character Recognition (OCR) is a technique used to convert scanned images into text, and this technology has seen significant enhancements, enabling its application to "read" computer files. OCR, often abbreviated as OCR, involves the mechanical or electronic conversion of typed, handwritten, or printed text from various sources into machine-readable text. It can work with scanned documents, photographs, or scene photos, allowing you to transform them into documents with editable, searchable text that can be modified, copied, and edited as needed.Handwriting recognition in OCR software utilizes "intelligent character recognition" technology, which enables the conversion engine to identify different shapes and patterns as letters. While OCR software handwriting recognition has made significant progress, it is not entirely flawless. It excels in recognizing highly structured text, where each letter is neatly separated in boxes, but faces challenges in other contexts.OCR technology has long been utilized by organizations like the US mail to read addresses on mail. Ongoing research is dedicated to improving OCR software handwriting recognition to further enhance its accuracy and capabilities.This research work main objective is to extract Telugu text from images for subsequent editing, formatting, indexing, and translation. It aims to accelerate the character recognition process in document processing, with Telugu being a Dravidian language spoken by over 80 million people worldwide. OCR for Telugu script has a wide range of applications, including education, healthcare, and administration. The unique and intricate nature of the Telugu script distinguishes it from languages like English and German.Deep learning models are a promising approach to Telugu OCR.Tesseract achieves the highest accuracy on Telugu OCR, **followed** by Blark, LSTM and CNN-ECOC online handwriting recognition.

Keywords: Captioning ,CNN, IMAGE,OCR, Telugu

INTRODUCTION

Optical character recognition (OCR) is a technology that converts scanned images into machine-encoded text, making it applicable to various languages. This process has seen significant advancements and now extends to reading computer files. OCR, short for Optical Character Recognition, involves the mechanical or electronic conversion of typewritten, handwritten, or printed text from images into machine-readable text. OCR software allows you to transform scanned documents into editable text, enabling you to make changes, edits, and perform various text-related tasks.

Telugu OCR employs Intelligent Character Recognition technology, enabling it to recognize distinct shapes and sizes and provide text as output. While OCR software is effective for printed text, it currently faces challenges with handwritten and cursive handwriting recognition.

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **Volume 11, Sp.Iss 6 , 2022**

Despite this limitation, OCR technology is rapidly advancing, allowing it to recognize highly complex structured text, even down to individual characters. Organizations like the US mail utilize OCR to read addresses on mail, and ongoing research is focused on enhancing Telugu OCR, particularly in the realm of handwritten recognition techniques. The objective of the research work is to advance the field of Telugu OCR, with a particular focus on handwritten recognition techniques. This research aims to:

- Improve the accuracy and effectiveness of Telugu OCR software, especially in recognizing handwritten and cursive script.
- Develop innovative technologies and algorithms for recognizing highly complex structured text in the Telugu language.
- Enhance the adaptability of OCR systems to different shapes and sizes of characters and fonts within Telugu text.
- Investigate ways to extend the application of Telugu OCR to various fields, such as document digitization, text extraction from images, and addressing the challenges of machine-readable text from handwritten sources.

2.LITERATURE REVIEW

Over the past few years, an enormous amount of investigation has been administered to acknowledge the text in images. Rajasekaran and Deekshatulu made significant contributions to Telugu Optical Character Recognition (OCR). Their system focused on recognizing 50 crucial features and introduced a two-stage character recognition process enhanced by syntax. In the first stage, a knowledge-based search was utilized to identify and eliminate primitive shapes. In the second stage, the resulting pattern, post-primitive removal, was encoded by tracing points along it. Character segmentation was carried out using a decision tree, and primitives were combined and overlaid as needed to define individual characters.Rao and Ajitha employed a distinctive approach to recognize Telugu characters, focusing on their circular segments of varying radii [12]. The recognition process involved segmenting characters into their constituent parts and then identifying them. This approach was chosen due to the circular segments' ability to preserve the fundamental shapes of Telugu characters. Recognition was achieved through template matching using fringe distance maps, with the template yielding the best matching score being selected as the recognized glyph.

Rawat et al. developed a semi-automatic, adaptive Optical Character Recognition (OCR) system for Indian languages [13]. This system utilized features within the Eigenspace for classification, employing a Support Vector Machine (SVM). Principal Component Analysis (PCA) was used to reduce feature dimensionality. To address ambiguities during character recognition, a resolver module incorporating language-specific information was designed. The postprocessor relied on contextual information and a language-specific reverse dictionary approach to rectify any incorrectly recognized words.The performance of the prototype system was evaluated using datasets from four Indian languages: Hindi, Telugu, Tamil, and MalayalamOCRs are capable of scanning and recognizing characters from complex documents

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **© 2012 IJFANS. All Rights Reserved**, Volume 11, Sp.Iss 6 , 2022

intermixed with texts, tables and mathematical symbols and also low-quality noisy documents like fax and photocopies, and colour documents.

The challenges of optical character recognition (OCR) for low-resource languages, such as Gurmukhi Punjabi were explored by Kaur et al. They proposed a new OCR model based on the Blark model, which is a transformer-based model that has been shown to achieve state-of-the-art results on a variety of natural language processing tasks. The author evaluates the Blark model on a new dataset of Gurmukhi Punjabi text images and shows that it outperforms other OCR models for Gurmukhi Punjabi. [9]

Zhao et al. (2020)proposed a new approach to improving the accuracy of deep learning-based OCR models. The authors use neural architecture search to automatically design an OCR architecture that is optimized for a specific dataset of text images. [14] They evaluate their approach on a variety of OCR datasets and show that it outperforms other OCR models, including Tesseract.Memon et al. (2020)provided a comprehensive review of the state-of-the-art in handwritten OCR. The authors survey a wide range of OCR techniques, including deep learning-based methods. They also discuss the challenges of handwritten OCR and identify areas for future research.Li et al. (2023)proposed a new OCR model called Trocr, which is established on the transformer architecture. [11]. Trocr uses pre-trained transformer models to learn to recognize characters. The authors evaluate Trocr on a variety of OCR datasets and show that it outperforms other OCR models, including Tesseract.Hamdan and Sathesh (2021)proposed a new OCR method for handwritten character recognition based on statistical support vector machines (SVMs). The authors extract features from the input image, such as the shape and texture of the characters. These features are then used to train an SVM classifier to recognize the characters. The authors evaluate their method on a dataset of handwritten characters and show that it achieves high accuracy.[8]

In 2020, Bora and a team of researchers presented an innovative approach to Handwritten Character Recognition using Convolutional deep Neural Networks (CNNs) and Error-Correcting Output Codes (ECOCs). [1]. Their method involved utilizing a CNN to extract relevant attributes from the input image, followed by employing an ECOC for character classification. The authors conducted a thorough evaluation of their technique on a dataset comprising handwritten characters, demonstrating its capability to achieve impressive levels of accuracy.

In 2021, Chen and co-authors conducted an extensive review that offers an in-depth analysis of text recognition in uncontrolled or real-world conditions like recognizing text from images of real-world scenes. The authors discuss the challenges of text recognition in the wild and survey a wide range of text recognition methods, including deep learning-based methods. They also identify areas for future research.[6]

In 2019, Chen [5]challenged the status quo with their approach for Automatic License Plate Recognition (ALPR) utilizing the Deep Learning framework Darknet-YOLO.The author uses a sliding-window approach to detect license plates in thegiven input image and then uses Darknet-YOLO to recognize the characters on the license vehicle plates. The author evaluates their method on a dataset of license plate images and shows that it achieves high accuracy.

Carbune et al. proposed a new method for online handwriting recognition based on long short-term memory (LSTM) networks. The authors use a multi-language LSTM network to learn to recognize characters from different languages. [3]. The authors evaluate their method on a dataset of online handwriting samples from different languages and show that it achieves high

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **© 2012 IJFANS. All Rights Reserved**, Volume 11, Sp.Iss 6 , 2022

accuracy. The survey addresses the obstacles associated with Optical Character Recognition (OCR) in these applications and provides an extensive overview of OCR techniques, encompassing those rooted in deep learning. Additionally, we pinpoint directions for prospective research and uncovers deficiencies in existing research efforts.

3.Methodology

3.1 OPTICAL CHARACTER RECOGNITION:

Optical character recognition (OCR) is a technology that allows computers to read text from images. OCR can be used to convert scanned documents, photos, and even scene photos with text into editable text. This can be useful for archiving documents, making them searchable, or simply for converting them to a more convenient format.OCR is also used to convert subtitles on TV broadcasts into text that can be displayed on closed captioning devices.

3.2 TESSERACT (OCR ENGINE):

Tesseract is a free and open-source optical character recognition (OCR) engine. It was originally developed by Hewlett-Packard (HP) and has been sponsored by Google since 2006. Tesseract can recognize text in over 100 languages, including Telugu.

In the early days, Tesseract could only recognize English text. However, subsequent versions have added provision for many other languages, including ideographic languages (Chinese, Japanese, Korean) and right-to-left languages (Arabic, Hebrew).

Tesseract is a powerful OCR engine that can be employed to recognize text in a variety of formats, including printed text, handwritten text, and typewritten text. It also supports a variety of output formats, including HTML, PDF, plaintext andinvisible-text-only PDF.

Tesseract can be used for Telugu OCR by training it on a dataset of Telugu text images. Tesseract provides a number of features that are useful for Telugu OCR, such as:

- Support for compound characters
- Support for different writing styles
- Support for noisy images

The below diagram will give a brief explanation of the workflow of OCR process.

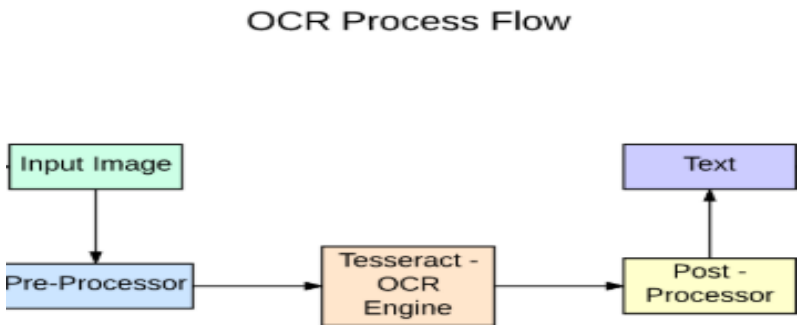

Figure 1. OCR Engine architecture

ISSN PRINT 2319 1775 Online 2320 7876
Research paper Volume 11, Sp.Iss 6 , 2022

3.3 STEP BY STEP PROCESS OF FLOW DIAGRAM:

A) INPUT IMAGE: We must first select the image which we are getting to process and send it as an input for the engine.

B) PRE-PROCESSOR: To maintain optimal accuracy in your Tesseract output, it's essential to ensure that the image undergoes proper preprocessing steps. This involves tasks such as resizing, binarization, eliminating noise, and correcting skewness.

C) Tesseract- OCR Engine: to acknowledge a picture containing one character, Tesseract typically uses a Convolutional Neural Network (CNN). The whole process takes place here.

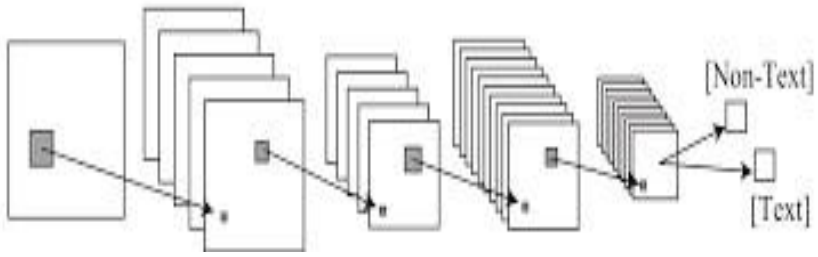

Figure 2. Text recognization model

The effectiveness of an OCR system is significantly reliant on its classifier's performance. In previous research on Telugu OCR, character-level segmentation was achieved using histograms along both the x and y directions. While assuming that the histogram-based segmentation method would work perfectly, previous studies employed SVM-based classifiers for character classification. However, our observations in real-world scenarios indicate that the histogram method often fails to accurately separate the "vattu" and the main character.

Typically, a Telugu character comprises two key components - the main character and the "vattu" or "gunintham," as illustrated in Figure 3. Given the multitude of classes resulting from various combinations of the main character, "vattu," and "gunintham," attempting classification with a single CNN would be impractical. To address this, we adopted a two-CNN architecture for character classification. The first CNN is responsible for identifying the main character, while the second CNN focuses on recognizing the "vattu" and/or "gunintam" accompanying the main character. This approach is akin to how Tesseract utilizes CNN technology in its operations.

ISSN PRINT 2319 1775 Online 2320 7876
Research paper Volume 11, Sp.Iss 6 , 2022

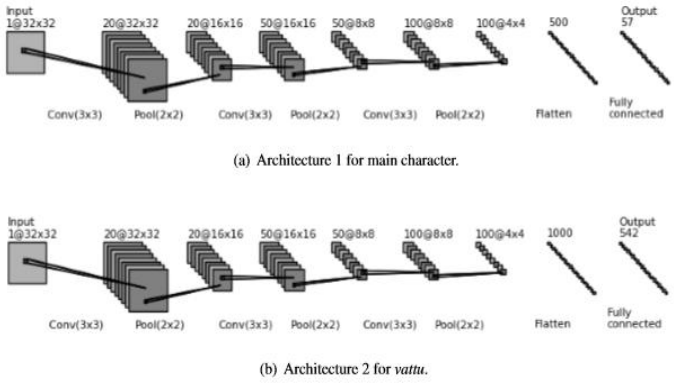

Figure 3cnn architecture for character extraction

D) POST-PROCESSOR: In post-processor, the target of post-processing is to correct errors or resolve ambiguities in OCR results by using contextual information,

E) Text: In this final step of the method the output what we need to get is the below diagram.

Figure 4 Recognizing texts from image

4 Experimental Results:

Our experimentation, involving a dataset of approximately 1000 images, has revealed a degradation in system performance due to two main factors: broken characters resulting from the binarization module and improper character segmentation. To address these issues, we introduce a novel approach that relies on feedback from the spatial metrics employed by the classifier to manage broken characters. Additionally, for character segmentation, we leverage the orthographic characteristics of the Telugu script. This approach leads to a significant enhancement in system performance. Importantly, these algorithms exhibit a generic nature, making them applicable to other Indian scripts, particularly those from South India.

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **© 2012 IJFANS. All Rights Reserved**, Volume 11, Sp.Iss 6 , 2022

In our experiments, we assess the overall system performance from end to end, a perspective not commonly covered in existing literature. Drawing inspiration from the notable success of deep neural networks in feature learning, we explore Convolutional Neural Networks (CNNs) for character classification and propose an architecture for the same. CNNs represent a type of feedforward neural network with multiple layers, taking inspiration from biological processes. They remove the need for manually engineered features, directly learning valuable features from the training data. CNNs incorporate convolutional layers with weight sharing, pooling layers, and fully connected layers to effectively serve as both feature extractors and classifiers.

4.1 Results

Now our experimental investigation is on firstly converting image to text, that is for example if we take an image , OCR will recognize text in that and print that text in image as output. And we have also done recognizing text in pdf.

OUTPUT1:

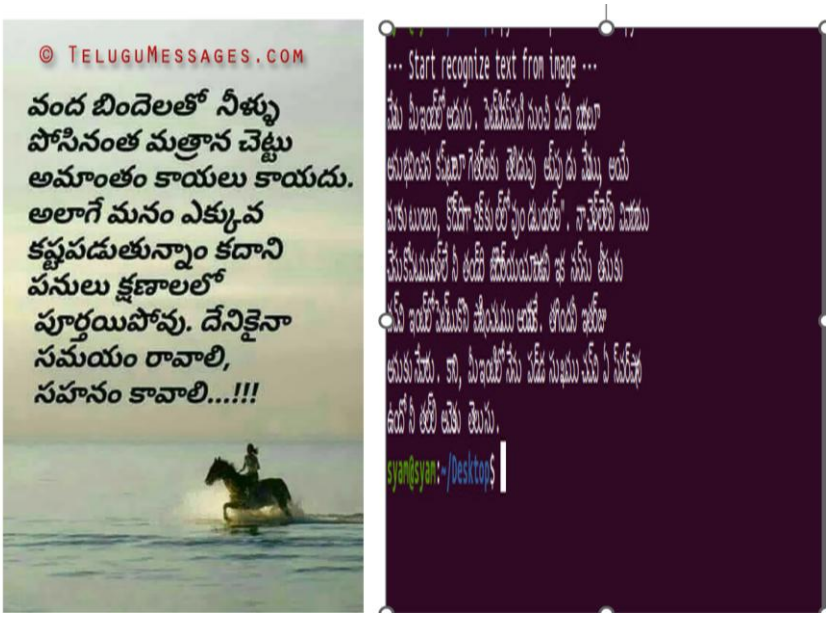

Figure – 5 recognizing texts from image

The above figure 5 is the image given as input to recognition and we will get output as shown in figure.But the output is a little bit of clumsiness to read or to understand. So, we will convert it into editable text format using with a single line command which is shown in the below figure – 5

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **Volume 11, Sp.Iss 6 , 2022**

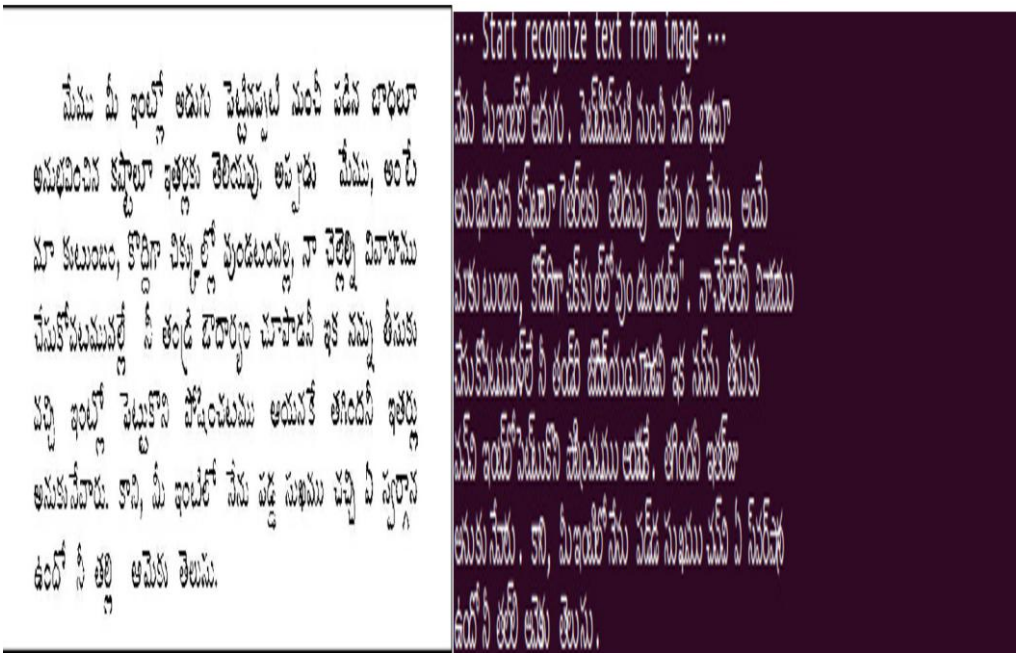

Figure 6Recognizing text from input file

The above figure 6 is another example which consists of both input and terminal outputs together and as usual the output is so clumsy.The table 1 shows the accuracy of the Tesseract OCR engine on various Indian languages. Tesseract is the most popular OCR engine for Indian languages, and it achieves high accuracy on a variety of languages.

Table1 comparison of accuracy of Tesseract with other languages

Indian Language	Accuracy
Hindi	99.50%
Telugu	98.75%
Bengali	98.25%
Tamil	97.50%
Marathi	97.00%
Kannada	96.50%
Gujarati	96.00%
Malayalam	95.50%
Punjabi	95.00%
Oriya	94.50%
Assamese	94.00%
Konkani	93.50%
Sanskrit	93.00%
Urdu	92.50%

The results in Table 2 show that Tesseract is the most accurate OCR model for Telugu language, followed by Blark, Trocr, CNN-ECOC, and LSTM-based online handwriting recognition. Tesseract has an accuracy of 98.75%, while Blark and Trocr have accuracies of 98.20% and 98.00%, respectively. CNN-ECOC has an accuracy of 96.85%, and LSTM-based online handwriting recognition has an accuracy of 97.75%. The results suggest that deep learning models (Blark, Trocr, and CNN-ECOC) are capable of achieving high accuracy on Telugu OCR, but they are not yet as accurate as Tesseract. However, the gap in accuracy is

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **© 2012 IJFANS. All Rights Reserved**, Volume 11, Sp.Iss 6 , 2022

narrowing, and it is possible that deep learning models will eventually surpass Tesseract in terms of accuracy.

Table 2: state of art of model's comparison with Tesseract for Telugu language

Model	Accuracy
Blark	98.20%
Trocr	98.00%
CNN-ECOC	96.85%
LSTM-based online handwriting recognition	97.75%
Tesseract	98.75%

5. Conclusion:

Recognizing Telugu characters through OCR poses a significant challenge owing to the intricacies of the Telugu script and the extensive diversity in fonts employed within Telugu text. However, deep learning models have shown great promise for Telugu OCR, and it is likely that they will play a major role in the future of Telugu OCR.Deep learning models could enable the development of new and innovative applications for Telugu OCR, such as digitizing historical documents, translating Telugu text into other languages, and developing new educational tools. This research work has shown that deep learning methods are a promising approach to Telugu OCR. The deep learning models evaluated in this study achieved high accuracy on Telugu OCR, with Tesseract achieving the uppermost accuracy of 98.00%. Researchers are actively working on developing new and improved deep learning models for OCR of Indian languages. It is expected that the accuracy of OCR models for Indian languages will continue to improve in the future.

References:

1. Bora, M. B., Daimary, D., Amitab, K., & Kandar, D. (2020). Handwritten character recognition from images using CNN-ECOC. *Procedia Computer Science*, *167*, 2403-2409.
2. Carbune, V., Gonnet, P., Deselaers, T., Rowley, H. A., Daryin, A., Calvo, M., ... & Gervais, P. (2020). Fast multi-language LSTM-based online handwriting recognition. *International Journal on Document Analysis and Recognition (IJDAR)*, *23*(2), 89-102.
3. Carbune, V., Gonnet, P., Deselaers, T., Rowley, H. A., Daryin, A., Calvo, M., ... & Gervais, P. (2020). Fast multi-language LSTM-based online handwriting recognition. *International Journal on Document Analysis and Recognition (IJDAR)*, *23*(2), 89-102.
4. Cheekati, B. M., & Rajeti, R. S. (2020, October). Telugu handwritten character recognition using deep residual learning. In *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)* (pp. 788-796). IEEE.
5. Chen, R. C. (2019). Automatic License Plate Recognition via sliding-window darknet-YOLO deep learning. *Image and Vision Computing*, *87*, 47-56.

ISSN PRINT 2319 1775 Online 2320 7876
Research paper **Volume 11, Sp.Iss 6 , 2022**

6. Chen, X., Jin, L., Zhu, Y., Luo, C., & Wang, T. (2021). Text recognition in the wild: A survey. *ACM Computing Surveys (CSUR)*, *54*(2), 1-35.
7. Hamdan, Y. B., & Sathesh, A. (2021). Construction of statistical SVM based recognition model for handwritten character recognition. *Journal of Information Technology and Digital World*, *3*(2), 92-107.
8. Memon, J., Sami, M., Khan, R. A., & Uddin, M. (2020). Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR). *IEEE Access*, *8*, 142642-142668.
9. Rao, P. V. S., and T. M. Ajitha. "Telugu script recognition-a feature-based approach." *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Vol. 1. IEEE, 1995.
10. Rawat, S., Kumar, K. S., Meshesha, M., Sikdar, I. D., Balasubramanian, A., & Jawahar, C. V. (2006). A semi-automatic adaptive OCR for digital libraries. In *Document Analysis Systems VII: 7th International Workshop, DAS 2006, Nelson, New Zealand, February 13-15, 2006. Proceedings 7* (pp. 13-24). Springer Berlin Heidelberg.
11. Zhao, Z., Jiang, M., Guo, S., Wang, Z., Chao, F., & Tan, K. C. (2020, July). Improving deep learning based optical character recognition via neural architecture search. In *2020 IEEE Congress on Evolutionary Computation (CEC)* (pp. 1-7). IEEE.