Major Project Report On

**AN AUTOMATED ANNOTATION SYSTEM FOR DOCUMENT TEXT IMAGES**

Submitted in partial fulfillment of the requirements for the award of the

## **Bachelor of Technology**

In

**Department of Computer Science and Engineering**

By

**B. JHANSI LAKSHMI 20241A0566**

**B. SRIJA 20241A0564**

**MARIA JABEEN 20241A0593**

**K. KUSHI REDDY 20241A0583**

Under the Esteemed guidance of

## **Dr. ASHLIN DEEPA R N Associate Professor**



**Department of Computer Science and Engineering**

**GOKARAJU RANGARAJU INSTITUTE OF ENGINEERING AND TECHNOLOGY**

**(Autonomous)**

Ⅰ



GOKARAJU RANGARAJU

INSTITUTE OF ENGINEERING AND TECHNOLOGY

**(Autonomous)**

**CERTIFICATE**

This is to certify that the major project entitled “**An Automated Annotation System for Document Text Images**” is submitted by **B. Jhansi Lakshmi (20241A0566), B. Srija (20241A0564), Maria Jabeen(20241A0593), K Kushi Reddy(20241A0583)** in partial fulfillment of the award of degree in BACHELOR OF TECHNOLOGY in Computer Science and Engineering during academic year 2023- 2024.

INTERNAL GUIDE HEAD OF THE DEPARTMENT

**Dr. ASHLIN DEEPA R N Dr. B. SANKARA BABU**

**Associate Professor Professor and HOD**

EXTERNAL EXAMINER

Ⅱ

**ACKNOWLEDGEMENT**

There are many people who helped us directly and indirectly to complete our project successfully. We would like to take this opportunity to thank one and all. First we wish to express our deep gratitude towards our internal guide **Dr. Ashlin Deepa R N, Associate Professor,** Department of CSE for her/his support in the completion of our project report. We wish to express our honest and sincere thanks to **Dr. B. Sankara Babu, HOD, Department of CSE** and to our principal **Dr. J. Praveen** for providing the facilities to complete the major project. We would like to thank all our faculty and friends for their help and constructive criticism during the project completion phase. Finally, we are very much indebted to our parents for their moral support and encouragement to achieve goals.

**B. Jhansi Lakshmi (20241A0566)**

**B. Srija (20241A0564)**

**Maria Jabeen(20241A0593)**

**K Kushi Reddy(20241A0583)**

Ⅲ

**DECLARATION**

We hereby declare that the major project entitled **“An Automated Annotation System for Document Text Images**” is the work done during the period from **19th July 2023 to 23rd Dec 2023** and is submitted in the partial fulfillment of the requirements for the award of degree of Bachelor of Technology in Computer Science and Engineering from Gokaraju Rangaraju Institute of Engineering and Technology (Autonomous under Jawaharlal Nehru Technology University, Hyderabad).The results embodied in this project have not been submitted to any other university or Institution for the award of any degree or diploma.

**B. Jhansi Lakshmi (20241A0566)**

**B. Srija (20241A0564)**

**Maria Jabeen(20241A0593)**

**K Kushi Reddy(20241A0583)**

Ⅳ

|  | **Table of Contents** |  |
| --- | --- | --- |
| **Chapter** | **TITLE** | **Page No** |
|  | **Abstract** | **1** |
| **1** | **Introduction** | **2** |
| **2** | **Literature Survey** | **3** |
| **3** | **Modules & UML diagram** | **5** |
| **4** | **Implementation** | **9** |
| **5** | **Test cases** | **13** |
| **6** | **Conclusion & Scope** | **19** |
|  | **References** | **20** |
|  | **Appendix**  **i)Review Paper**  **ii)Result** | **22** |

Ⅴ

**Abstract**

Annotation involves enriching text with additional details or labels to provide context, improve comprehension, or aid in refining Natural Language Processing (NLP) models. The central aim of this project is to develop a robust system capable of annotating text images in documents with minimal human intervention. This involves delineating distinct regions within an image and generating descriptive text that effectively captures the characteristics of these outlined areas. While manual annotation is both time-consuming and error-prone, automatic image annotation tools step in to address these challenges, significantly boosting efficiency and reliability. A primary emphasis is placed on achieving word-level annotations for handwritten document images. The resulting system not only alleviates the time and resource demands associated with manual annotation but also elevates the reliability and accuracy of annotations. This project is poised to make a substantial contribution to the advancement of NLP applications, particularly in scenarios that involve the analysis and understanding of handwritten documents.

**Chapter 1**

**Introduction**

Optical Character Recognition (OCR) stands as a transformative technology crucial for converting printed or handwritten text from physical documents or images into editable and machine-readable formats. This technology plays a pivotal role in digitizing books, articles, historical records, and various textual materials. The field of OCR has witnessed significant advancements[16].

While OCR primarily focuses on printed text[15], Handwritten Text Recognition (HTR) takes text recognition a step further by specializing in deciphering handwritten content. An optimized OCR system tailored for handwritten Marathi text document classification and recognition is introduced in [17]. Handwritten materials, such as historical documents and manuscripts, pose unique challenges due to variations in writing styles, ink fading, and paper degradation. HTR technology utilizes neural networks, machine learning algorithms, and pattern recognition techniques to transcribe handwritten content, thereby preserving the authenticity and historical significance of such documents with the primary goal of converting them into a digital format[4].

Although handwritten documents provide valuable historical and cultural information, manually annotating and transcribing them is a time-consuming and error-prone task. Automated Annotation, characterized by superior efficiency and user-friendly interfaces, surpasses manual annotation. Automated Annotation Systems[5] streamline and enhance tasks related to text analysis, information retrieval, and data organization. They leverage techniques from Natural Language Processing (NLP), machine learning, and data mining to analyze and understand text, facilitating the automated generation of tags for large volumes of textual content.

An innovative automatic annotation-based approach for digitized text recognition is proposed in [10]. Manual annotation is labor-intensive, time-consuming, and error-prone. Automated annotation systems provide an efficient and user-friendly alternative, expediting the annotation process while improving accuracy. These systems find diverse applications, including content recommendation, information retrieval, sentiment analysis, and content categorization.

**Chapter 2**

**Literature Survey**

In this digital age, the need for efficient document digitization and text extraction has led to the development of Optical Character Recognition (OCR) systems. These systems aim to automatically convert printed or handwritten text from document images into machine-readable and searchable text. One crucial aspect of enhancing OCR accuracy and usability is the automatic annotation of document text images.

Choi and Kim (2012) delve into automatic image annotation using semantic text analysis[3]. Their work highlights the intersection of image analysis and natural language processing, demonstrating how semantic understanding of text can aid in annotating images.

Hu et al. (2015) introduce DocRicher[2] an automatic annotation system that leverages social media to annotate text documents. This innovative approach showcases the significance of incorporating external contextual information, such as social media trends or user-generated content, to annotate and enrich document text.

In the study by[1] Mondal et al. (2023), the authors focus on handwritten document images and propose a method for automatic annotation at the word level. Automatic annotation systems play a pivotal role in enhancing the accuracy and usability of OCR technology. They contribute to the enrichment of document text images with metadata, semantic context, and linguistic information.

[6,7] uses a hybrid CNN and RNN architecture that tackles the challenges of recognizing handwritten and printed text.

In the realm of Optical Character Recognition (OCR), recognizing handwritten paragraphs poses a significant challenge.Language models like BERT and GPT for Indian languages, offering insights into their adaptability and performance in multilingual NLP tasks. [8] addresses this challenge by introducing an innovative approach using a Vertical Attention Network (VAN). VAN employs an attention mechanism focused on vertical text regions, enabling it to capture the intricate layout and structure of handwritten paragraphs seamlessly, thereby eliminating the need for explicit segmentation.

Traditional metrics like WER and CER are widely used in OCR experiments for assessing recognition accuracy[13].The paper introduces the concept of bWER[14], a metric designed to assess word recognition independently of reading order.

Dealing with limited labeled data, especially for symbols or alphabets in regional languages, presents a common challenge in HTR. [9] introduces a few-shot learning approach for HTR, particularly beneficial when data availability is scarce.This approach starts with a minimal number of labeled examples for each symbol and progressively refines its understanding as more labeled data becomes accessible. To overcome the data limitation, the paper introduces the technique of Pseudo-labeling.

Automatic annotation systems have direct applications in OCR technology. By associating meaningful annotations with text images, these systems improve various aspects of OCR, including text recognition accuracy, context preservation tasks.

**Chapter 3**

**Modules and UML diagram**

**Modules:**

There are three main modules in our project.

**1) De-GAN Module:**

● De-GAN utilizes Generative Adversarial Networks (GANs) for denoising handwritten documents.

● It identifies and removes various types of noise, such as background artifacts and smudges.

● It enhances the overall quality of scanned documents, making them more legible and suitable for annotation.

● It employs deep learning techniques to learn and generate denoised versions of input documents.

● It significantly improves dataset readiness for subsequent annotation and analysis.

**2) CRAFT Module:**

● After binarizing the image using De-GAN, it is sent to the CRAFT model.

● CRAFT is a text detection model that uses a Convolutional Neural Network(CNN) model to accurately extract text regions from the document images.

● It analyzes pixel-level information to determine character and word boundaries.

● It utilizes computer vision algorithms to identify regions containing textual content.

● The purpose of this model is to minimize errors and enhance annotation efficiency by focusing solely on relevant text.

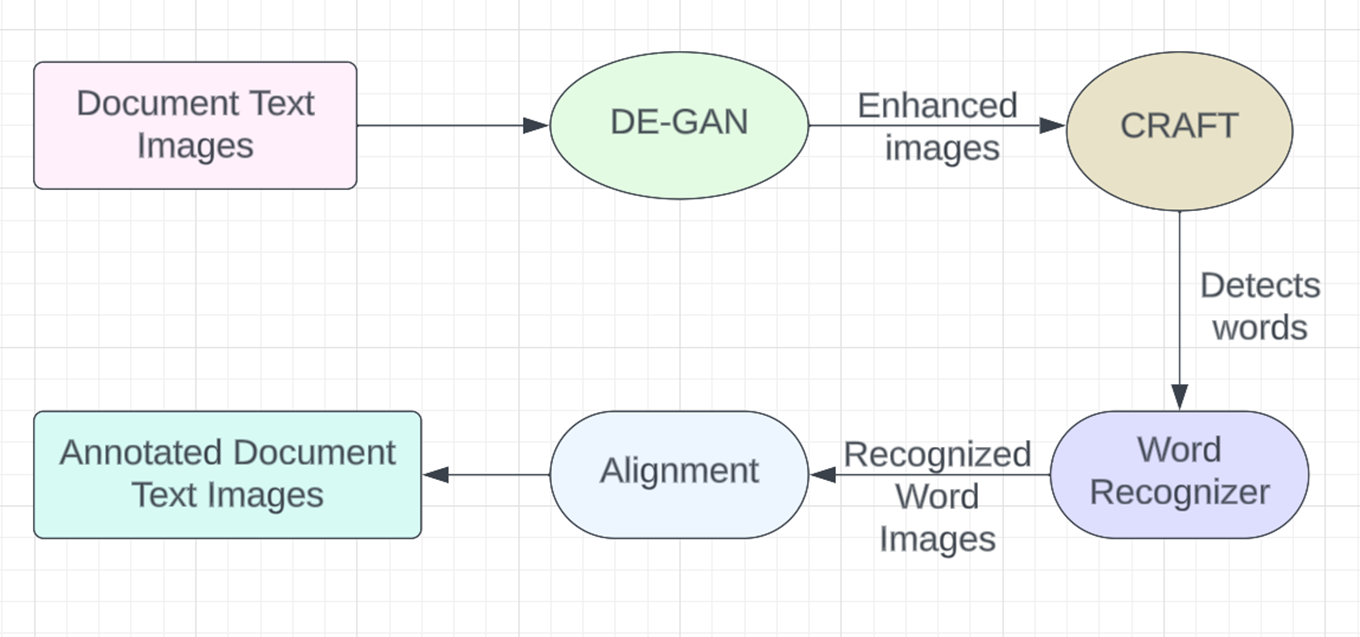
**3) Annotation Module:**

● Annotation involves the automated labeling of pre-processed documents to make it recognizable and usable for various applications.

● It employs machine learning models like CNNs, GCN, LSTM-based and encoder-decoder with attention models for automated text recognition.

● It involves manual review to rectify discrepancies and ensure high accuracy.

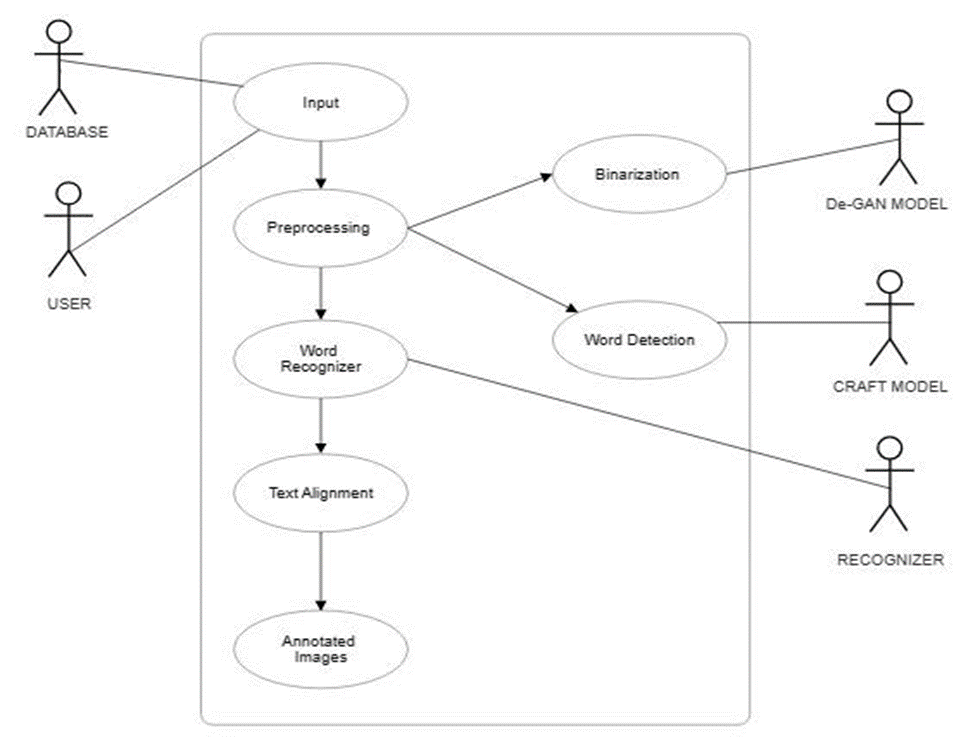
**Architecture:**



**Fig 1: Architecture of Automated Annotation System**

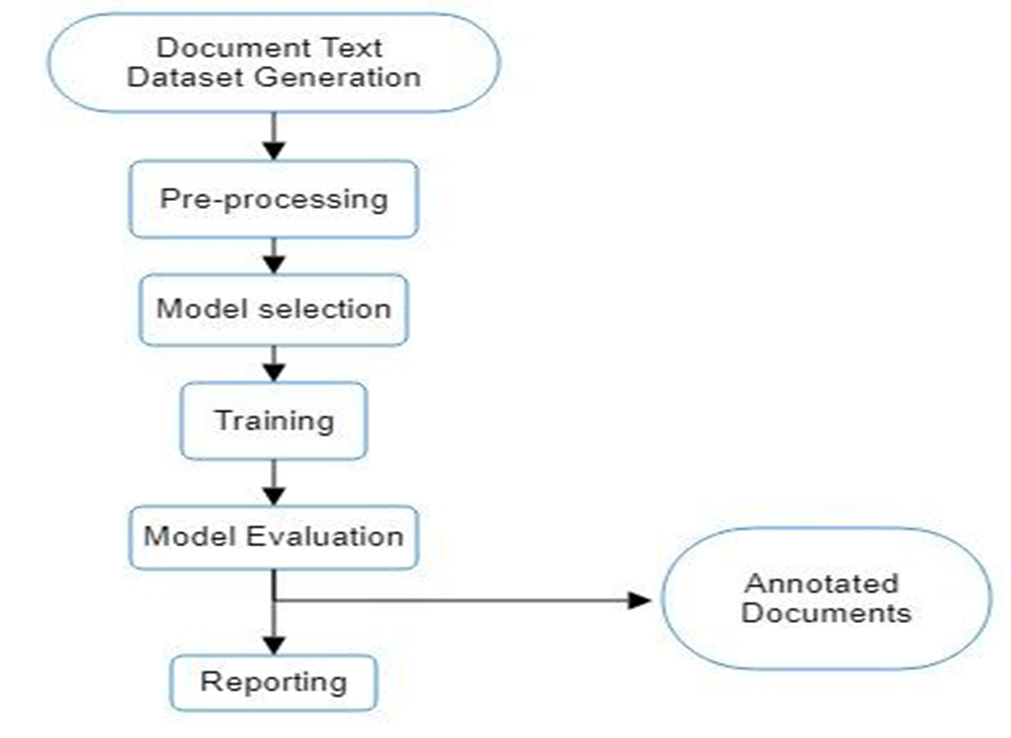
**UML Diagrams:**

**Use Case Diagram:**



**Fig 2: Use Case Diagram of Automated Annotation System**

**Data Flow Diagram:**



**Fig 3: Data Flow Diagram of Automated Annotation System**

**Chapter 4**

**Implementation**

**Implementation steps:**

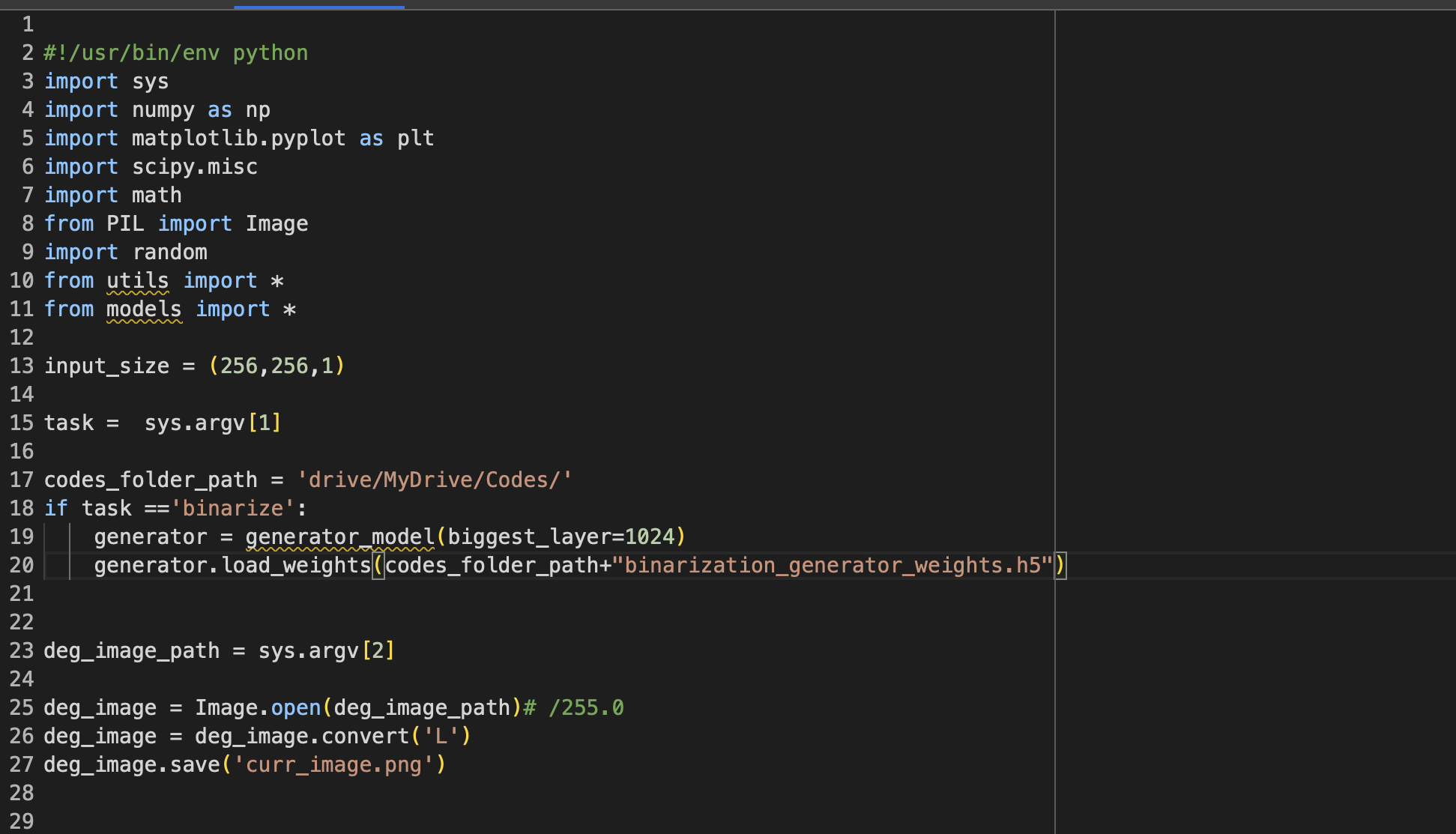
●The process begins with the acquisition of an image, ensuring it is taken without the influence of any filters.

●The unfiltered image is fed into the pretrained De-GAN (Denoising GAN) model.

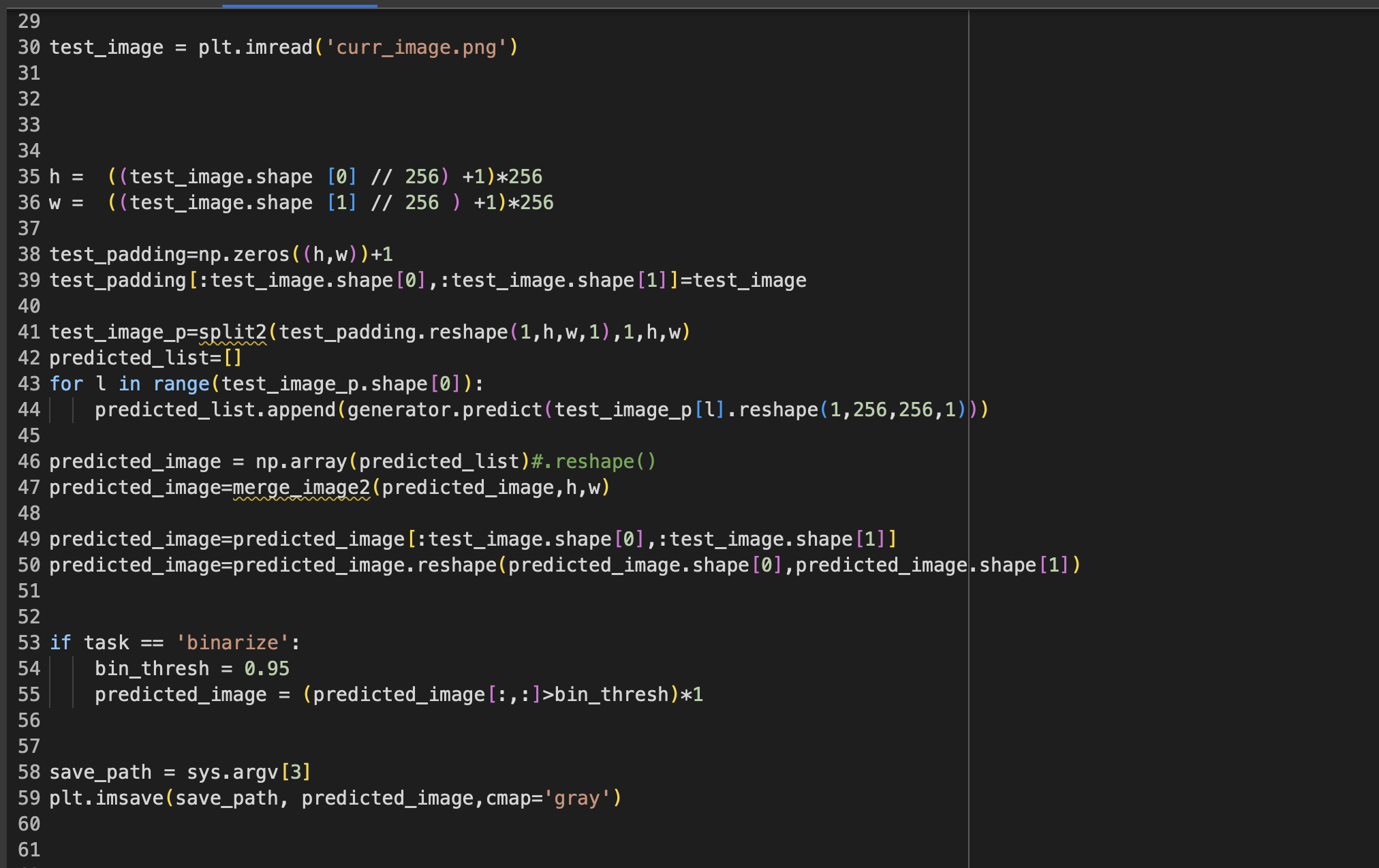
●De-GAN specializes in denoising and enhances the image quality, generating a binarized version that eliminates background noises.

●The binarized image produced from De-GAN serves as input for the CRAFT model, which generates a bounding box image and a mask image highlighting detected text regions. Additionally, a text file is created, containing coordinates detailing the position of bounding boxes.

**De-GAN snapshots :**

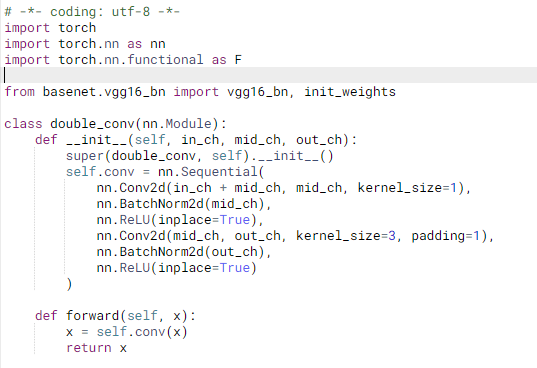


**Fig 4**

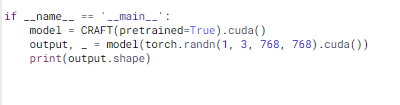


**Fig 5**

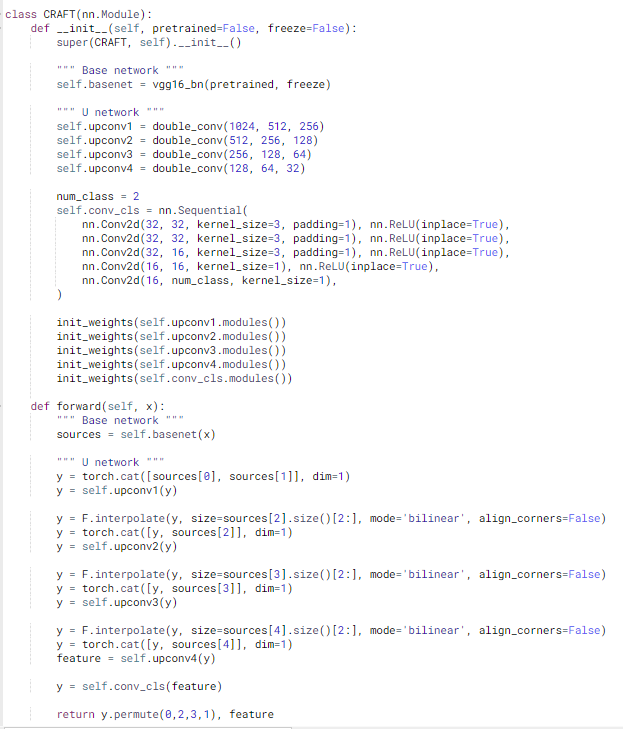
**Craft code snapshots:**



**Fig 6**



**Fig 7**



**Fig 8**

**Chapter 5**

**Test cases**

**De-GAN Module:**

| **ID** | **Test Data** | **Output** | **Status(pass/Fail)** |
| --- | --- | --- | --- |
| 1 |  |  | Pass |
| 2 |  |  | Pass |
| 3 |  |  | Fail |
| 4 |  |  | Pass |
| 5 |  |  | Pass |

**CRAFT Module:**

| **ID** | **Test Data** | **Output** | **Status(pass/Fail)** |
| --- | --- | --- | --- |
| 1 |  |  | Pass |
| 2 |  |  | Pass |
| 3 |  |  | Fail |
| 4 |  |  | Pass |
| 5 |  |  | Pass |

**Chapter 6**

**Conclusion and scope**

In conclusion, an Automated Annotation System designed for document text images brings transformative benefits to various industries. By leveraging advanced image processing techniques, such as Optical Character Recognition (OCR) and deep learning algorithms, this system significantly enhances document analysis, data extraction, and information retrieval processes. It optimizes efficiency, accuracy, and productivity by automating the time-consuming task of manual annotation. This system finds applications in areas ranging from digitizing historical archives and cataloging large document databases to aiding in content indexing and semantic understanding. As technology continues to advance, the development and adoption of Automated Annotation Systems for document text images hold the promise of revolutionizing the way we interact with and extract knowledge from textual documents.

**References**

1. Ajoy Mondal, Krishna Tulsyan, and C.V. Jawahar. "Automatic Annotation of Handwritten Document Images at Word Level." ICVGIP '22, ACM, 2023.

2. Qiang Hu, Qi Liu, Xiaoli Wang, Anthony K.H. Tung, Shubham Goyal, and Jisong Yang. "DocRicher: An Automatic Annotation System for Text Documents Using Social Media." SIGMOD '15, ACM, 2015.

3. Choi, D., Kim, P. "Automatic Image Annotation Using Semantic Text Analysis." CD-ARES 2012, Springer, 2012.

4. Balci, Batuhan, Dan Saadati, and Dan Shiferaw. "Handwritten text recognition using deep learning." CS231n: Convolutional Neural Networks for Visual Recognition, Stanford University, Course Project Report, Spring (2017): 752-759.

5. Mondal, Ajoy & Tulsyan, Krishna & Jawahar, C.. (2022). Automatic Annotation of Handwritten Documents in Word Level. 10.1145/3571600.3571645.

6. Jain, Mohit & Mathew, Minesh & Jawahar, C.V.. (2017). Unconstrained OCR for Urdu Using Deep CNN-RNN Hybrid Networks. 747-752. 10.1109/ACPR.2017.5.

7. Gunna, Sanjana, Rohit Saluja, and C. V. Jawahar. "Transfer learning for scene text recognition in Indian languages." International Conference on Document Analysis and Recognition. Cham: Springer International Publishing, 2021.

8. Coquenet, Denis & Chatelain, Clement & Paquet, Thierry. (2020). End-to-end Handwritten Paragraph Text Recognition Using a Vertical Attention Network.

9. Souibgui, M. A., Fornés, A., Kessentini, Y., and Megyesi, B., “Few shots are all you need: A progressive learning approach for low resource handwritten text recognition”, <i>Pattern Recognition Letters</i>, vol. 160, pp. 43–49, 2022. doi:10.1016/j.patrec.2022.06.003.

10. Sankar, K. Pramod and C. V. Jawahar. “Enabling Search over Large Collections of Telugu Document Images - An Automatic Annotation Based Approach.” Indian Conference on Computer Vision, Graphics & Image Processing (2006).

13. Bansal, Siddhant, Praveen Krishnan, and C. V. Jawahar. "Fused Text Recogniser and Deep Embeddings Improve Word Recognition and Retrieval." International Workshop on Document Analysis Systems. Cham: Springer International Publishing, 2020.

14. Vidal, Enrique and Toselli, Alejandro Héctor and Rios-Vila, Antonio and Calvo-Zaragoza, Jorge, End-to-End Page-Level Assessment Of Handwritten Text Recognition. Available at SSRN: https://ssrn.com/abstract=4347108 or<http://dx.doi.org/10.2139/ssrn.4347108>

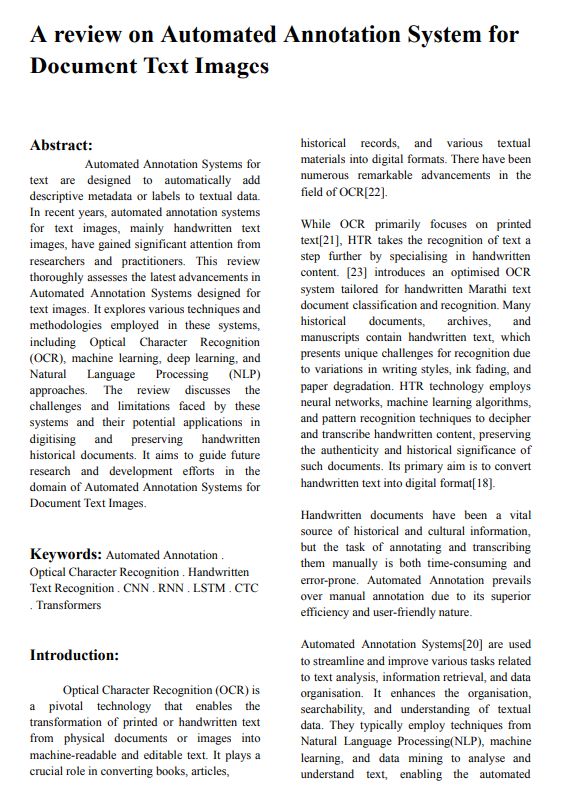
15. B M, Sagar & Shobha, G. & P., Ramakanth. (2008). OCR for printed Kannada text to machine editable format using database approach. WSEAS Transactions on Computers. 7. 766-769.

16. Sharma, Parikshit. (2023). Advancements in OCR: A Deep Learning Algorithm for Enhanced Text Recognition. International Journal of Inventive Engineering and Sciences. 10. 1-7. 10.35940/ijies.F4263.0810823.

17. Ramteke, Surendra & Gurjar, Ajay & deshmukh, Dhiraj. (2018). A streamlined OCR system for handwritten MARATHI text document classification and recognition using SVM-ACS algorithm. International Journal of Intelligent Engineering and Systems. 11. 186-195. 10.22266/ijies2018.0630.20.

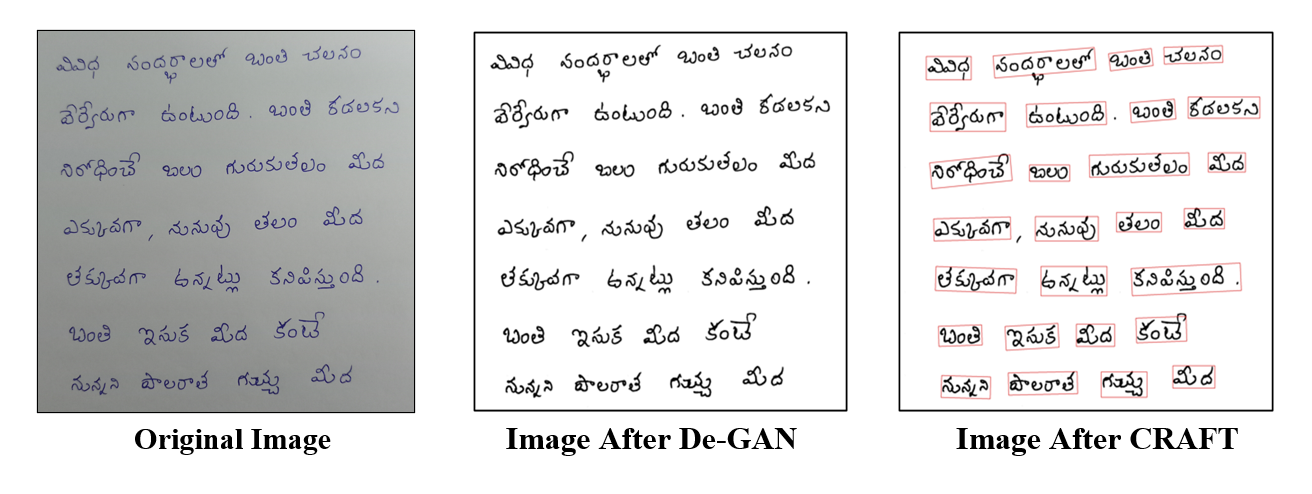
**Appendix**

**i)Review Paper:**

****

**ii)Result:**

**De-GAN and CRAFT Modules:**

****

**Fig 9: Results generated**