IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DEGRADED DOCUMENT ANALYSIS

[1]Sanika Santaji Khot, [2]Dhanashri Suryakant Jadhav, [3]Sneha Santosh Koli,

[4] Asst. Prof. Roopa Rahul Gaur

[1234]Student, [4]Assistant Professor,

[1234]Computer Science And Engineering,

[1]Nanasaheb Mahadik College Of Engineering, Peth.

Abstract: In this era of digitization, most hard copy documents are being transformed into digital formats. In the process of transformation, large quantities of documents are stored and preserved through electronic scanning. These documents are available from various sources, such as ancient documentation, old legal records, medical reports, music scores, palm leaves, and reports on security-related issues. In particular, ancient and historical documents are hard to read due to their degradation in terms of low contrast and the existence of corrupted artifacts [1]. In recent times, degraded document analysis has been studied widely, and several approaches have been developed to deal with issues and challenges in document analysis. In this project, a comprehensive review is conducted on the issues and challenges faced during the document analysis process, followed by insights on various methods used for document analysis. This project also discusses the advanced methods used for the enhancement of degraded documents that improve the quality of documents during the analysis process. Further discussions are made on the effectiveness and robustness of existing methods, and there is still scope to develop a hybrid approach that can deal with degraded document analysis more effectively [5].

Index Terms - image quality, accuracy, OCR, historical document, image enhancement, stain, noise, motion picture, old medical reports, old legal records, missing character.

1. INTRODUCTION

The first technique, called image binarization, is suitable for tasks such as moving object detection and finding regions of interest in text-like images in corrupted files. For some applications, binary images require less memory, data layout analysis, data skew detection, and speed calculation. The most commonly used method to binarize images is the use of a thresholding algorithm, which divides the image into background and foreground based on whether the pixel is determined globally or locally using the threshold value. At the same time, manuscripts and historical documents are preserved in archives and libraries. Their quality deteriorates over time due to various environmental conditions such as dust, moisture, human handling errors, and old paper. It is difficult to convert files into binary and save them digitally when there is bad corruption. The binarization of image data has evolved rapidly due to increasing interest in historical data analysis. Damage to historical documents can take many forms, including fading or yellowing of ink, bleeding, uneven lighting, alterations, and stains. Since there is no way to control all these effects and the possibility of improving the quality of binarization always exists, new ideas are constantly being developed. According to Sezgin and Sankur, there are six types of tests: the histogram-based method, the geographical method, the clustering-based method, the entropy-based method, the object attribute based method, and the local method. Depending on the decision, we can classify the process as a global process or a local process. The general method divides the image into two groups, foreground (text) and background, and then performs the calculation of the image based on the entropy, histogram, or clustering algorithm [3].

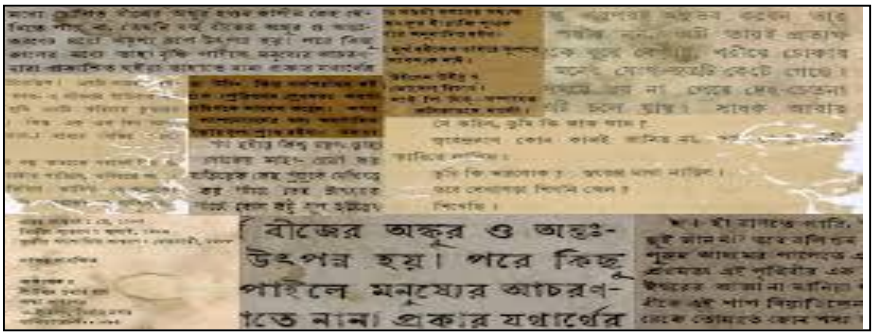

Figure 1.1: Degraded Document.

In contrast, the local measurement technique divides the image into smaller pieces and estimates the threshold of each piece separately, using information from pixels closer to increasingly smaller parts of the image. Use the international review process to get the best value for your paper from quality data. At the same time, many technologies emerged for converting ancient images to text, including international methods (like Otsu, Kittler, etc.) and local (like Niblack, Sauvola) methods in binary form. Although Otsu's method works well for images with dual-mode histograms, it cannot solve problems such as uneven lighting, color bleeding, or blurry text. To find the segmentation boundary, Kittler's theory assumes that the pixel values at each pixel level exhibit a Gaussian distribution. This method works very well for good data.

1.2 PROBLEM STATEMENT

The majority of the handwritten and map-formatted writings from the old and vanished human culture are written on regular paper. The purpose of this study is to improve degraded papers caused by adverse environmental conditions and other harmful elements. In this project, we create a system that can accurately analyze documents that have been damaged, including historic manuscripts or scanned pages with fading ink or uneven lighting conditions. In this system, we remove all noise from the document, such as ink stains, shadows, motion of the document, blurring, wrinkles, etc., and we also extract the text from the document and generate a summary of that text.

1.3 OBJECTIVES

1.3.1 Document Digitization: This refers to the conversion of a physical document into a digital format, typically an image file, through scanning.

1.3.2 Document Enhancement: This stage focuses on improving the quality and readability of the digitized document by removing various types of noise. This may include:

- Ink stains
- Shadows
- Motion blur (caused by camera movement or document shake)
- Blurs (from other factors)
- Wrinkles

Document enhancement techniques can significantly enhance the visual clarity of the digital document.

1.3.3 Character Recognition (OCR): Optical Character Recognition, or OCR, aims to recover lost or damaged characters from the document. This may involve:

- Analyzing the surrounding context of the missing character(s)
- Utilizing pattern recognition algorithms trained on large datasets of characters.

Character recognition helps restore the integrity of the digitized documents.

2. ANALYTICAL MODEL

2.1 System Architecture

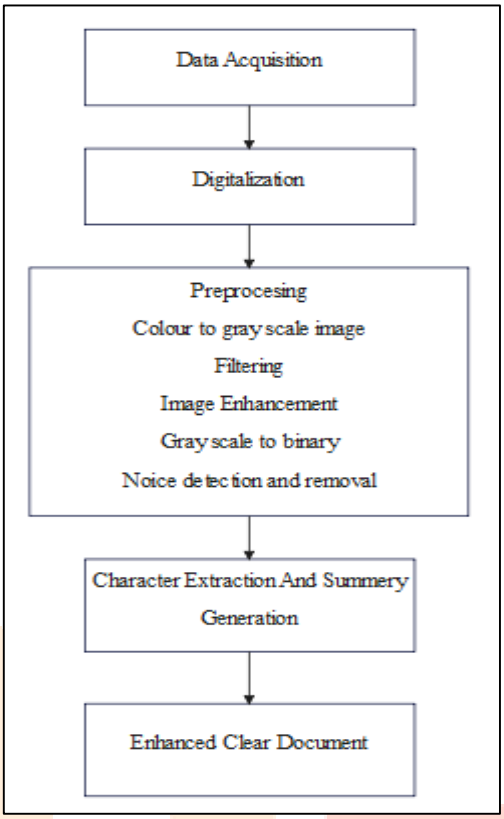

Fig. 2.1: System Architecture.

i) Data Acquisition: In this system, we first give one image as input.

ii) Digitization: Document digitization is the process of creating digital versions of historical documents. Digital copies can serve as backups to the originals or as more accessible file formats for users to view in multiple locations and adapt as needed. The digitization process is more than scanning and creating images of documents.

iii) Pre-processing: Document pre-processing is an important step to ensure high-quality extraction. It refers to cleaning, organizing, and transforming raw data to fit what machine learning models need or desire. It is basically a data mining process to improve the quality of data. Data pre-processing uses some of the previous steps to clean and prepare documents for subsequent analysis and comparison with other documents.

- Color to grayscale image: It is an image conversion technology in digital photography. It removes all information from the color data, leaving only different shades of gray. The lightest is white and the darkest is black.
- Filtering: Here, we can filter the document.
- Grayscale to binary: Here we convert a grayscale image to a binary image. Thresholding is the simplest image segmentation method and the most common way to convert grayscale images to binary images.
- Noise detection and removal: Information can be transmitted through noise during the process of converting, scanning, or converting to digital form. We can classify noise by analyzing its features and looking for similar patterns in image data to choose appropriate methods to eliminate noise.

iv) Restoration and Character Recognition: Here we recover and identify missing characters from the file. Identifying and recognizing text from documents and other images, commonly known as optical character recognition (OCR), is a widely used method of data processing.

v) Enhanced Clear Document: Clear Document: After processing from above ages stages above, the perfect enhanced clear document.

2.2 Data Flow Diagram

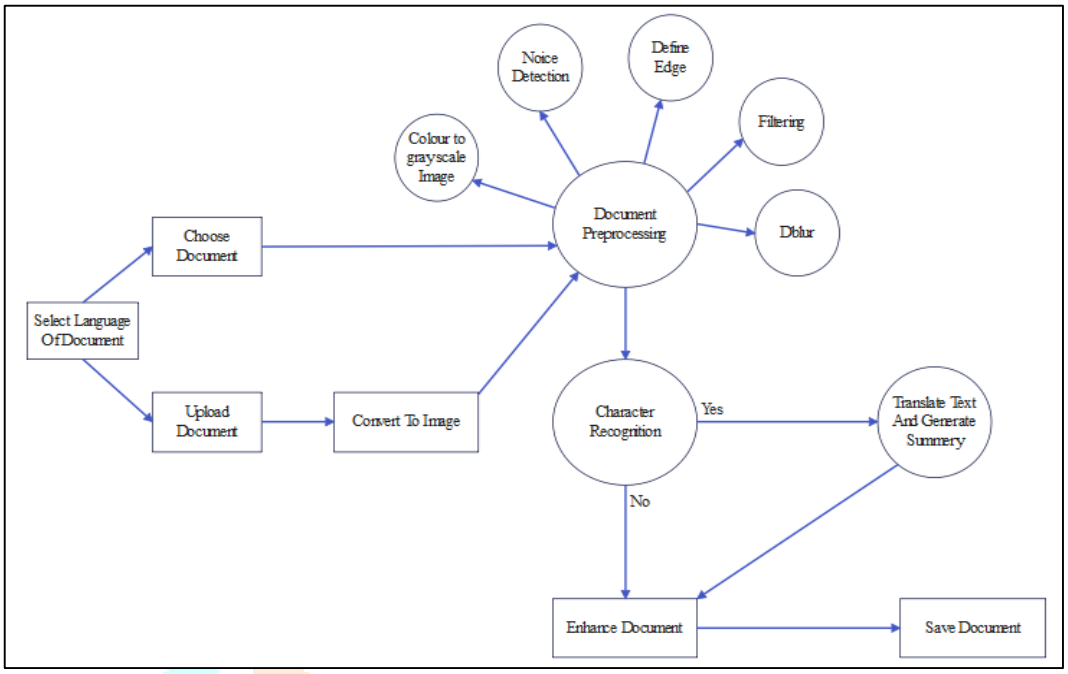

Fig. 2.2: Data Flow Diagram.

The data flow diagram (DFD) depicts a degraded document analysis that can process different kinds of documents, including text and images. Here a breakdown of the process:

• Users can choose to upload a document or select an existing one.

• The system can then perform document preprocessing, which includes converting the document to a gray-scale image an filtering it.

• The DFD then shows two possible paths for the document, depending on whether it's a text document or an image document.

• If its a text document, the system performs optical character recognition (OCR) to convert the image into text and then translates and generates a summary of the text.

• If its an image document, the system performs noise, edge detection, and filtering.

• Finally, the processed document is saved.

3 EXPERIMENTAL WORK

3.1 Algorithms used

3.1.1 OCR Technology: OCR (Optical Character Recognition) plays a vital role in degraded document analysis by acting as a bridge between the damaged text and a digital, editable format. Even though some approaches attempt to reconstruct the original document without OCR, it remains a powerful tool for extracting the remaining information. Here's how OCR tackles degraded documents:

• Image Preprocessing: Before diving into character recognition, OCR technology often cleans up the image through techniques like filtering to reduce noise and improve clarity. This preprocessed image becomes the basis for OCR to analyze the text.

• Character Recognition: OCR engines then segment the image into individual characters. Degraded documents may have faded characters, smudges, or even tears, making this segmentation a challenge. However, OCR algorithms are trained to recognize these damaged characters despite the imperfections.

3.1.2 2D Filtration: In order to prepare damaged documents for optical character recognition (OCR) and additional analysis, 2D filtration methods are essential. These algorithms function as digital filters, effectively eliminating noise and improving image quality to facilitate accurate OCR. Documents that are old frequently have different kinds of noise, like speckle that comes from age or flaws in the scanner. This noise can be reduced with 2D filters, resulting in a sharper image with more legible text.

Averaging filters, Gaussian filters, and median filtering are common filtering techniques. Each has advantages over the other when it comes to tackling particular noise patterns. 2D filtration methods enhance the overall quality of the image by lowering noise, which facilitates character segmentation by OCR engines. This is particularly crucial.

3.1.3 Median Filter: The median filter is a specific type of 2D filtration algorithm particularly useful in degraded document analysis for its ability to tackle a common problem: salt-and-pepper noise. Here's a closer look at how it helps:

- Salt-and-pepper noise manifests as isolated white and black pixels scattered randomly across the image. This noise can significantly disrupt the integrity of characters in degraded documents.
- Median Filter's Strength: The median filter works by analyzing a small window around each pixel in the image. It replaces the center pixel's value with the median value of all the pixels within the window. This approach effectively removes isolated outliers like the white and black pixels of salt and pepper noise.

3.2 Technology Used

To develop this project, we used different types of technology, which are as follows:

3.2.1 Python: Python is a powerful language for degraded document analysis projects due to its extensive libraries and functionalities. We also use Python to create the front end of the project. Python provides a set of specialized libraries designed for image processing, computer vision, and machine learning tasks that are important for degraded document analysis. Here are some notable points about the players:

- poppler: In a degraded document analysis project, you can upload images as well as PDFs. This PDF might have a mix of clean and degraded text or even embedded images. Poppler can be used to render these PDFs into image formats (like JPEG or PNG) suitable for further processing.
- OpenCV: OpenCV (CV2) is a powerful Python library for degraded document analysis due to its extensive image processing capabilities. Defective files are often affected by noise caused by obsolescence, browser defects, or background noise. cv2 provides various filters, such as a median filter, a Gaussian filter, and a binary filter, to remove noise and improve image clarity. Converting gray scale images to black and white (binarization) is an important step in the OCR engine. cv2 provides methods such as adaptive transformation and Otsu thresholding to achieve optimal binarization of corrupted data. This improves the separation between text and background. Poor text or low contrast can make OCR difficult. CV2 provides techniques such as histogram equalization and contrast expansion to improve the visibility of text in images. Technologies such as canny edge detection can be used to define the boundaries of characters and data, aiding segmentation and layout analysis.

3.2.2 MySQL: MySQL offers a secure and scalable solution for managing user accounts in a degraded document analysis project that involves login and registration functionalities.

3.2.3 Machine Learning: Machine learning can be used to learn without explicit programming. It finds patterns in the data. Noise reduction is tackled in degraded document analysis. Even with broken fonts or touching characters, it refines character classification for accurate text extraction. Information extraction and organization can be made more efficient by analyzing document layout and document types. By using machine learning, degraded documentscan be improved, and valuable information from damaged documents can be found.

4. RESULT

4.1 Select Document

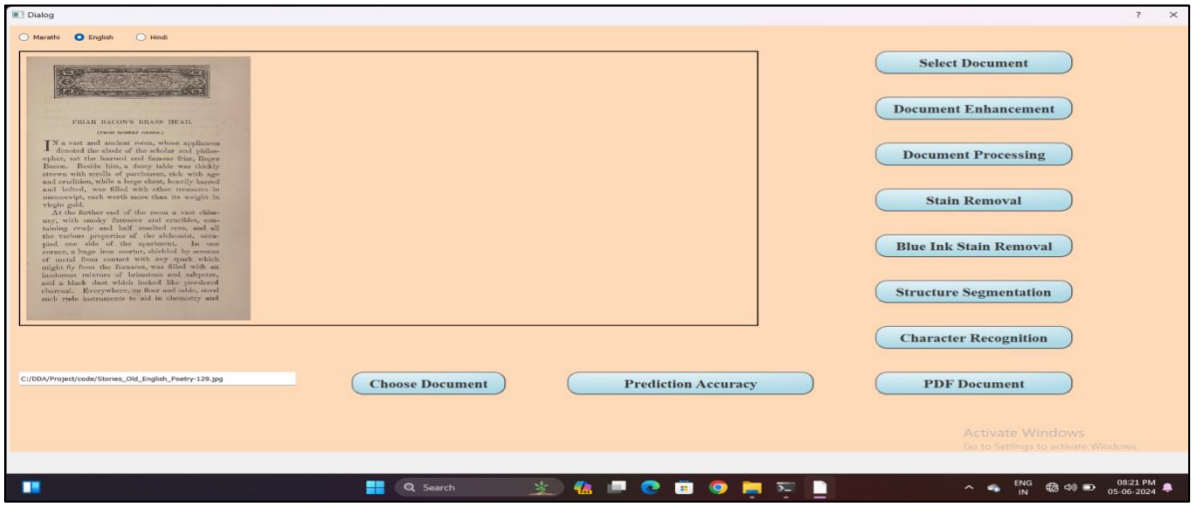

Fig 4.1: Select Document.

4.2 Document Enhancement

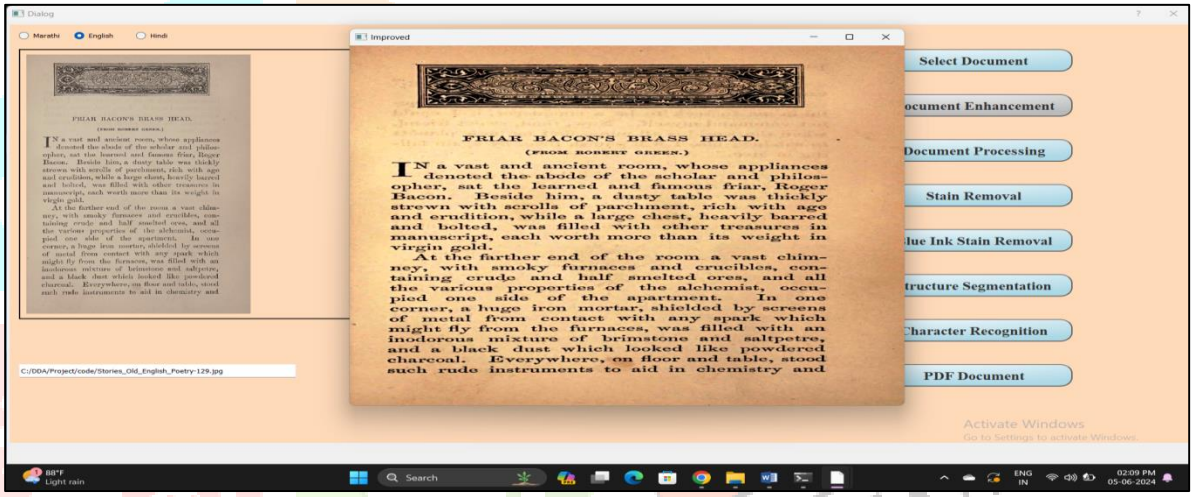

Fig. 4.2: Document Enhancement.

4.3 Noise remove

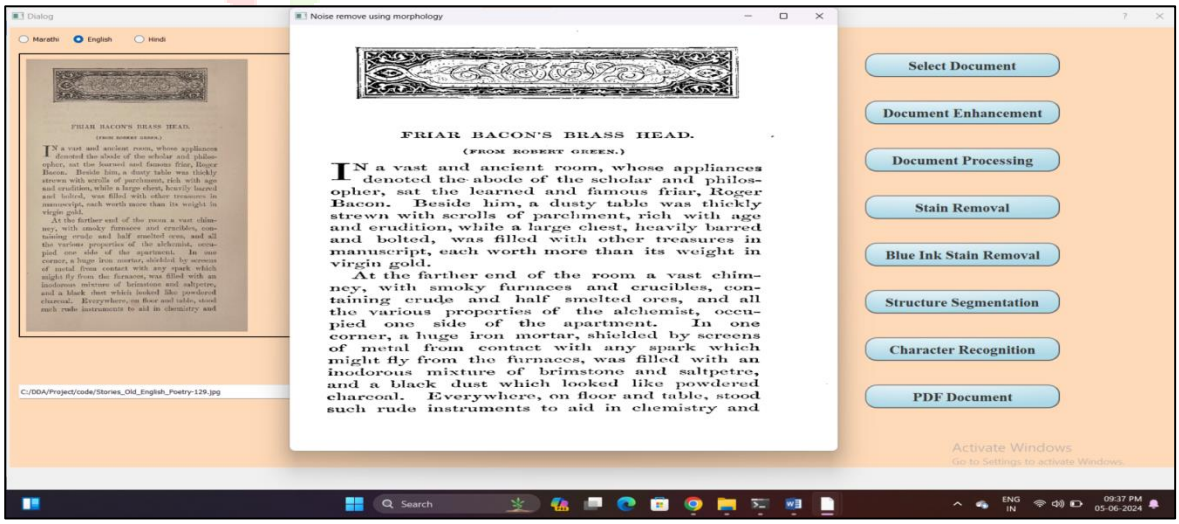

Fig 4.3: Noise remove.

4.4 Document Dblur

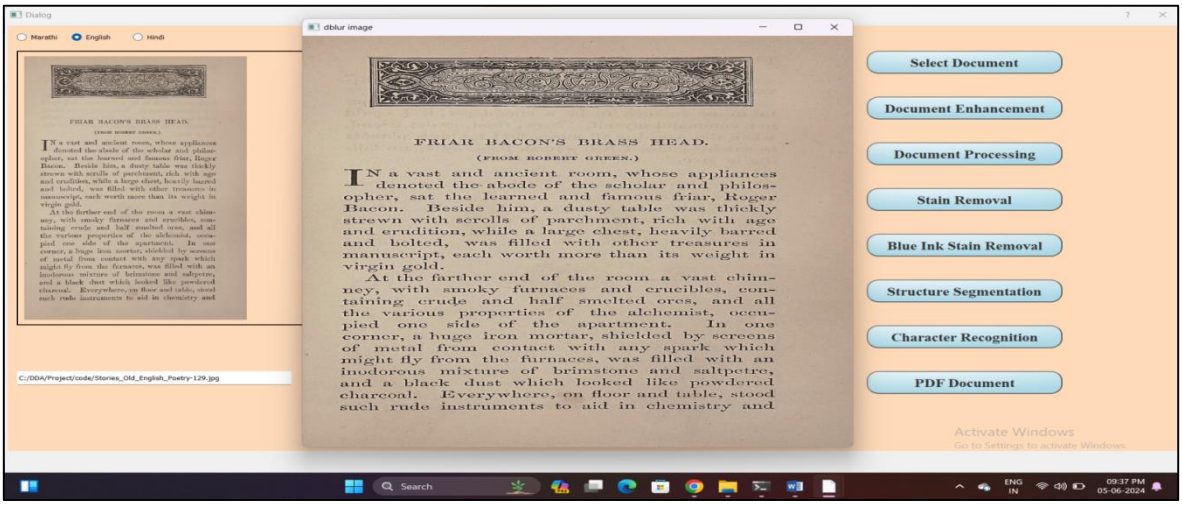

Fig 4.4: Document Dblur.

4.5 Blue Ink Stain Removal

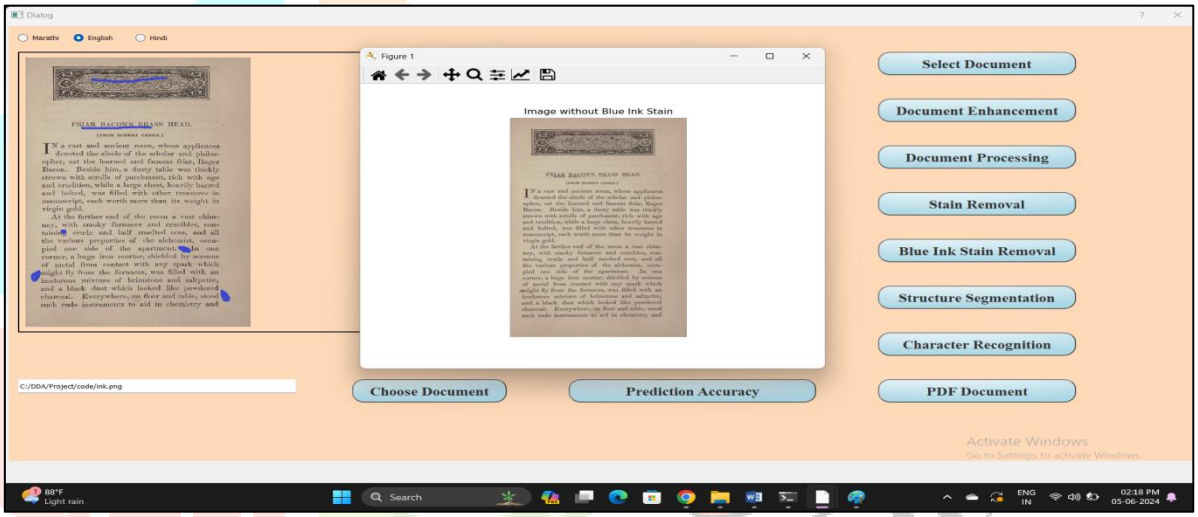

Fig. 4.5: Blue Ink Stain Removal.

4.6 Accuracy Prediction

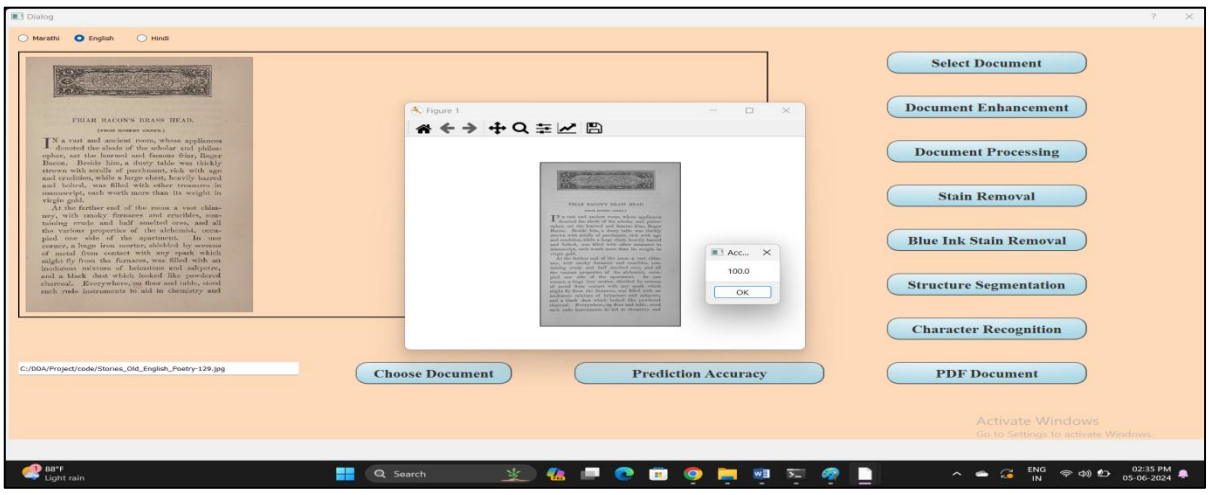

Fig. 4.6: Accuracy Prediction.

4.7 Translate Text

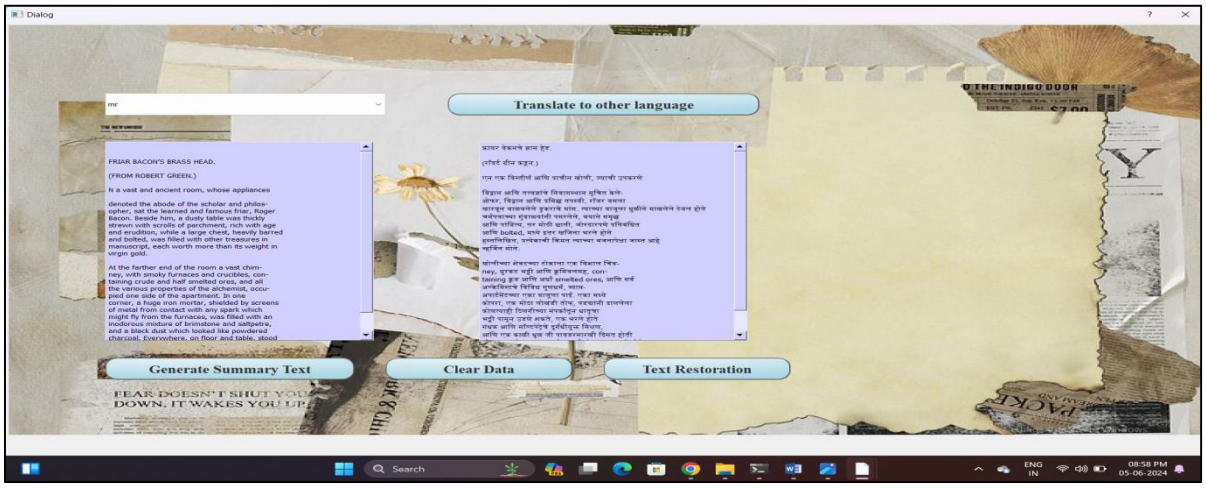

Fig. 4.7: Translate text.

4.8 Generate Summery

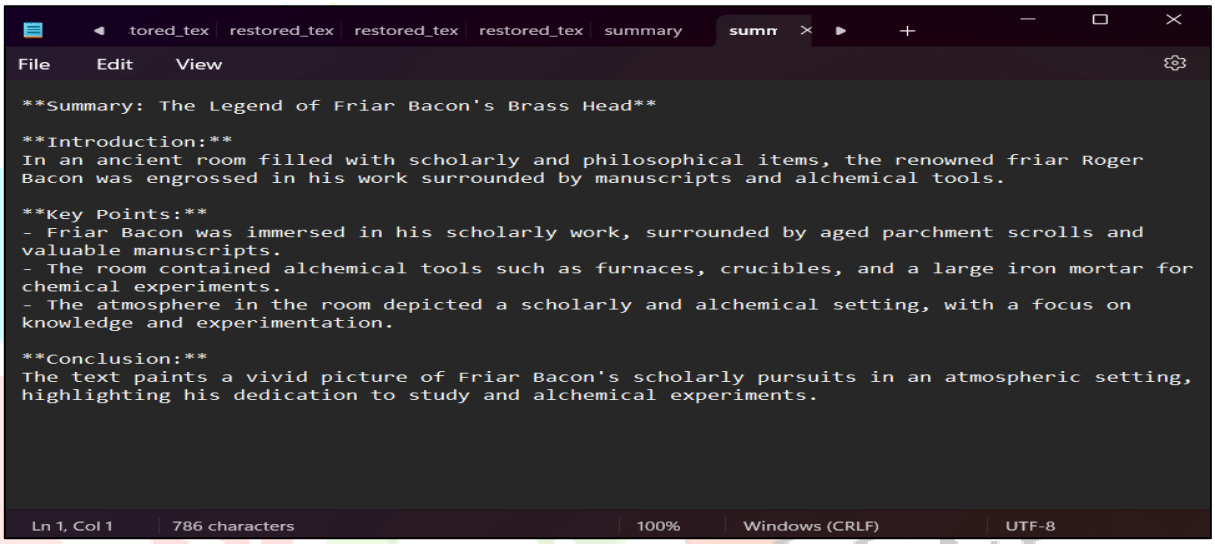

Fig. 4.8: Generate Summery

4.9 Text Restoration

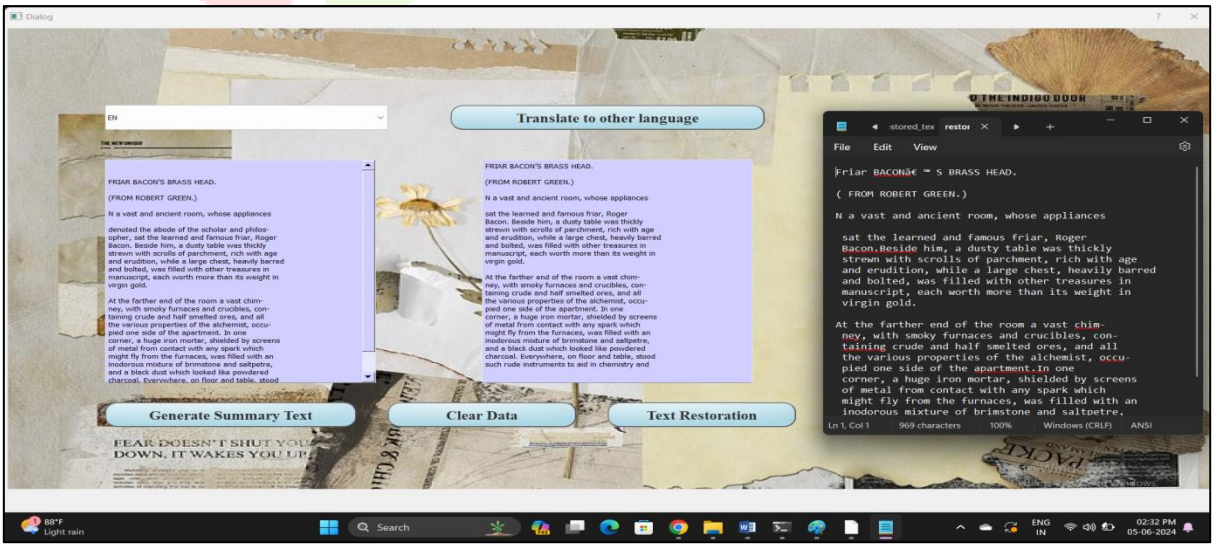

Fig. 4.9: Text Restoration.

5. CONCLUSION

This paper discusses various challenges related to document degradation and offers a solution. These problems pertain to historical materials that are printed or handwritten. This system classifies degradation issues such as noise detection and removal, blurring, ink stain removal, edge detection, character recognition and restoration, summery generation, and determining the accuracy of the text extraction automatically from the document using machine learning and image processing based techniques.

6. REFERENCES

[1] Vaishali Patil and Dr. Tripti Arjariya, "Implementation of Document Image Binarization for Removing Noise from Degraded Document", International Journal for Research in Engineering Application & Management (IJREAM), vol. 7, pp. 1-?, July 2021.

[2] Wei Xiong and Song Wang, "An enhanced binarization framework for degraded historical document images", EURASIP Journal on Image and Video Processing, 2021.

[3] Jeba Shain and Divesh Kumar, "A Review Analysis on Development of Historical Manuscript Images", Saudi Journal of Engineering and Technology, December 2021.

[4] Harshit Jindal and Manoj Kumar, "Degraded Document Image Binarization using Novel Background Estimation Technique", in Proceedings of the 2021 6th International Conference for Convergence in Technology (I2CT), pp. 1-6, Pune, India, April 2-4, 2021.

[5] Rohithkumar A and Rajath T, "Document Image Binarization", Journal of Seybold Report, vol. 15, no. September, pp. 1-?, 2020.

[6] Johan Tordsson and Li Wu, "MicroRAS: Automatic Recovery in the Absence of Historical Failure Data for Microservice Systems", in Proceedings of the 2020 IEEE/ACM 13th International Conference on Utility and Cloud Computing (UCC), pp. 1-10, 2020.

[7] Alaa Sulaiman and Khairuddin Omar, "Degraded Historical Document Binarization: A Review on Issues, Challenges, Techniques, and Future Directions", Journal of Imaging, vol. 5, no. 2, pp. 1-?, February 2019.

[8] Younes Akbari and Somaya Al-Maadeed, "Binarization of Degraded Document Images Using Convolutional Neural Networks and Wavelet Based Multichannel Images", in Proceedings of 2019 IEEE Region 10 Symposium, pp. 1-6, 2019.

[9] Poulami Raha and Bhabatosh Chanda, "Restoration of Historical Document Images Using Convolutional Neural Networks", in Proceedings of 2019 IEEE Region 10 Symposium, pp. 1-6, 2019.

[10] Tien-Ying Kuo and Yu-JenWei, "Automatic Damage Recovery of Old Photos Based on Convolutional Neural Network", in Proceedings of the 2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), pp. 1-6, 2019.