IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 7, July 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207102|

Investigation of the Best OCR

Gahana.S[1], Manasa M R[2], Thejashwini B L[3]

Department of Studies in Computer Science University of Mysore, Manasa Gangotri, Mysuru, Karnataka, India[1, 2, 3]

ABSTRACT: This paper examines the effectiveness of many OCR (Optical Character Recognition) technologies with an emphasis on how well they work with various kinds of text images. Even with the enormous advancements in OCR technology, it is still quite difficult to reliably recognize text from images that contain difficult-to-read fonts, such handwritten prescriptions from doctors. This study assesses the benefits and drawbacks of several OCR options, including open-source libraries and industry standards. One noteworthy discovery is that, even with its widespread use and strong performance on printed text, EasyOCR frequently performs poorly when processing images of handwritten prescriptions from doctors. By performing a comparative analysis based on accuracy, processing speed, and simplicity of integration, this study seeks to determine the most dependable OCR tool for such difficult scenarios, assisting practitioners in choosing the best OCR solution.

KEYWORDS: Doctor's handwritten and printed prescription recognition, Machine learning, Optic Character Recognition(OCR), Tesseract OCR, EasyOCR.

I. INTRODUCTION

Optical Character Recognition (OCR) technology has emerged as a key instrument for text digitization from images, bringing significant advantages to a range of sectors including banking, healthcare, and education. OCR transforms various document formats such as scanned paper documents, PDF files, or digital camera photos into editable and searchable data. However, the quality of the image and the type of text can have a substantial impact on how accurate OCR technologies are. Identifying handwritten text is a particularly difficult situation, especially when it comes to prescription drugs, which are infamous for having illegible handwriting and differing formats.

In order to determine which OCR tool is best for processing handwritten prescriptions from doctors, this study examines the functionality of a number of OCR programs. The study covers a broad spectrum of OCR solutions, from well-known market leaders to recently developed open-source substitutes. One among them, EasyOCR has drawn praise for its easy-to-use interface and reliable output of printed text. Nevertheless, its effectiveness falls when handling the intricate and frequently irregular handwriting included in prescription drugs. This disparity emphasizes the necessity for OCR technologies to take a more specialized approach.

The main goal of the study is to compare different OCR technologies and evaluate each one's accuracy, speed of processing, and simplicity of integration. This research attempts to determine the advantages and disadvantages of each solution in managing handwritten doctor prescriptions by utilizing an extensive evaluation framework. The investigation will shed light on why some OCR tools perform better in these difficult situations and provide useful suggestions for healthcare professionals.

It is essential to comprehend the capabilities and constraints of existing OCR technology if one hopes to enhance document digitization procedures. The ultimate goal of this research is to improve the effectiveness and dependability of text recognition in medical documentation by bridging the gap between current OCR technologies and the specific requirements of healthcare practitioners.

II. DATASET INFORMATION

An organized collection of data is called a data set. Data, as we all know, is a collection of information gathered via measurements, observations, research, or analysis. Information like names, numbers, Images, figures, facts, and even brief descriptions of items could be included.

The collection consists of real-time images of handwritten prescriptions from doctors collected from various sources. I collected both handwritten and printed medical prescription dataset from many sources, such as pharmacies, hospitals, clinics, and even directly from patients. I used my handheld cameras to capture images. In the dataset, there are two distinct kinds of prescriptions: one type contains only handwritten text from physicians, while the other kind only contains printed text.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207102|

A. Handwritten prescriptions

The dataset is made up of a number of handwritten prescriptions from doctors, which reflect a wide range of medical scripts with different degrees of complexity and legibility. Prescriptions from several healthcare professionals are included in this dataset, assuring a broad range of handwriting styles from moderately readable to completely unreadable. Standard medical data, including patient information, medication names, dosages, and directions, are included with every prescription, offering a thorough foundation for evaluating the effectiveness of OCR technologies.

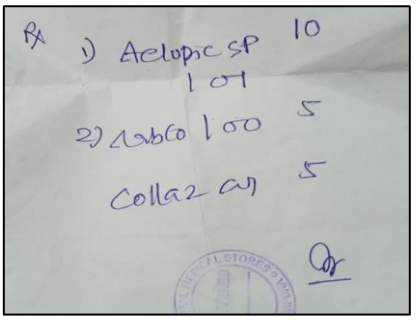

Fig 1. Handwritten prescription

B. Printed prescriptions:

Prescriptions written by doctors with unified, machine-printed text that includes patient information, medicine names, dosages, and instructions make up the dataset. Since most of these prescriptions are readable and straightforward, OCR technologies have an easier time bench marking against them. With the use of this dataset, OCR performance on more standardized and organized medical documents can be assessed, guaranteeing precise text recognition and data extraction for applications in the healthcare industry.

Fig 2. Printed prescription

III. PROPOSED METHODOLOGY

A complete methodology including dataset preparation, tool selection, performance measures, and in-depth analysis is proposed to evaluate the optimal OCR tool for text recognition from handwritten prescriptions issued by doctors.

3.1 Preparation of the dataset:

Considering two types of datasets, one with printed prescriptions and the other with handwritten prescriptions. The printed prescription dataset will offer standardized, readable text for a baseline comparison, while the handwritten prescription dataset will showcase a variety of handwriting styles with readability. To standardize image quality, prepossessing techniques such as scaling, gray-scale conversion, and noise reduction will be applied to every dataset.

3.2 Tool Selection:

A number of OCR tools will be chosen for assessment, such as: EasyOCR: It is an open-source OCR program that works well with printed text and it is well-known for its ease of use.

Tesseract OCR: An open-source OCR engine that is widely utilized. This OCR is efficient in-line and character pattern recognition enable users to extract text from images.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207102|

3.3 Measures of Performance:
Several important performance criteria will be used to assess the OCR tools: Accuracy: Based on the accurate identification of medical terms and dosages, accuracy is measured by character and word recognition rates.

Error Rate: The frequency and kinds of errors, such misinterpreted characters or text whose presence is missing.

METHOD 1: EasyOCR

A. Applying EasyOCR for Handwritten Prescription:
1. Dataset image: Considering one of the original image from the dataset for implementation EasyOCR.

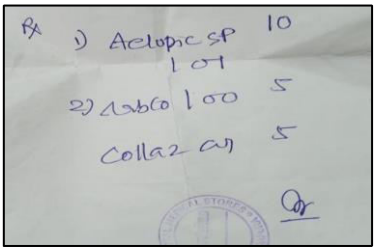

Fig 3. Original handwritten prescription image

2. Line segmentation: Line segmentation in an image entails splitting the text into distinct lines in order to help in correct EasyOCR processing. This phase is critical for accurately identifying and reading each line of text, particularly in handwritten texts. Effective line segmentation increases EasyOCR accuracy by distinguishing text lines and minimizing recognition errors.

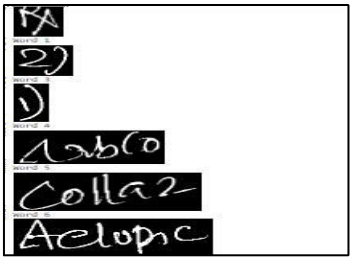

Fig 4. Line Segmented handwritten image

3. Word detection: Word detection in images includes identifying and separating individual words from lines of text for EasyOCR processing. This phase is critical for accurate text recognition because it allowsthe EasyOCR engine to discriminate between different words and interpret them correctly. Effective word detection improves both EasyOCR's accuracy and readability.

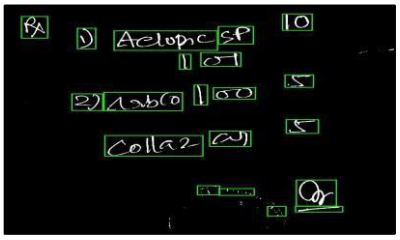

Fig 5. Word Detected handwritten image

4. OCR implementation: OCR implementation on an image includes preprocessessing` it, segmenting it into lines and words, and then using an EasyOCR algorithm to recognize and extract text. This procedure transforms visual text data into machine-readable text, allowing for digital storage, searchability, and analysis of the content within the image.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207102|

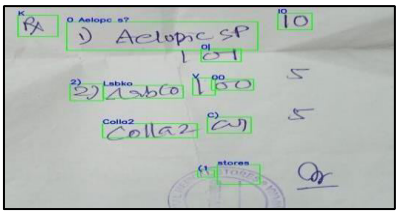

Fig 6. EasyOCR implemented handwritten image

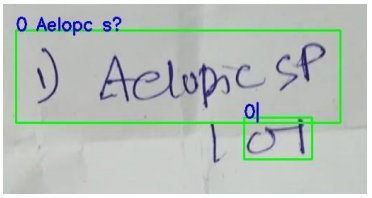

Fig 7. Partially classified handwritten image

B. Applying EasyOCR for Printed Prescription:

1. Dataset image: Considering one of the original image from the dataset for implementation.

Fig 8. original image

2. Line segmentation: Line segmentation in printed images involves discovering and separating specific lines of text within a document. This procedure is critical for accurate EasyOCR because it guarantees that each line of text is handled independently, enhancing recognition accuracy and reducing errors caused by overlapping or misaligned text lines.

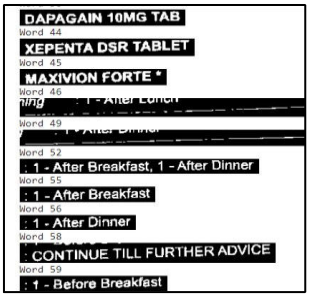

Fig 9. Line Segmented printed image

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207102|

3. Word detection: Word recognition in a printed image includes recognizing and extracting specific wordswithin the image. The method uses EasyOCR technology to scan and analyse the text, identifying word boundaries based on spacing and font properties, ensuring accurate extraction and conversion of printed text into digital format.

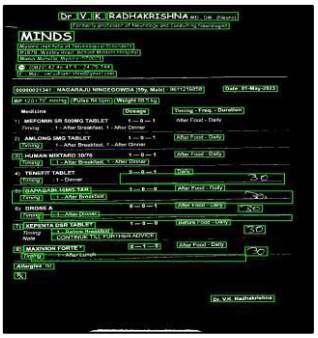

Fig 10. Word Detected printed image

4. OCR implementation:4. To implement OCR on a printed image, first it must be preprocessed. Next, it must be divided into words and lines. Finally, text may be recognized and extracted using the EasyOCR algorithm. Through this process, printed text data is converted into machine-readable text, enabling digital storage, content analysis, and searchability.

Fig 11. EasyOCR implemented printed image

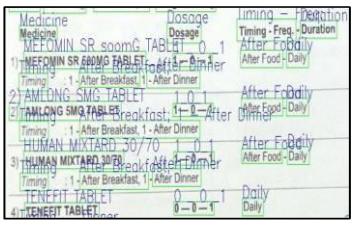

Fig 12. classified printed image.

METHOD 2: Tesseract OCR

A. Applying Tesseract OCR for Handwritten Prescription:

1. Dataset image: Considering one of the original image from the dataset for implementation of TesseractOCR.

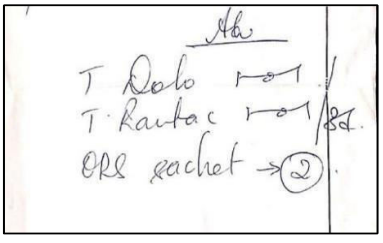

Fig 13. Original handwritten prescription image.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207102|

2. Line segmentation: Line segmentation in an image entails splitting the text into distinct lines in order to help in correct Tesseract OCR processing. This phase is critical for accurately identifying and reading each line of text, particularly in handwritten texts. Effective line segmentation increases Tesseract OCR accuracy by distinguishing text lines and minimizing recognition errors.

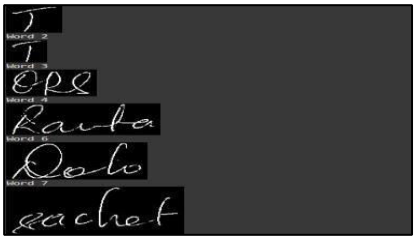

Fig 14. Line Segmented handwritten image.

3. Word detection: Word detection in images includes identifying and separating individual words from lines of text for Tesseract OCR processing. This phase is critical for accurate text recognition because it allows the Tesseract OCR engine to discriminate between different words and interpret them correctly. Effective word detection improves both Tesseract OCR's accuracy and readability.

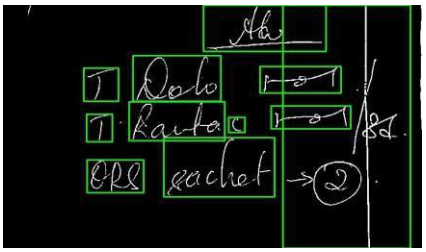

Fig 15. Word detected handwritten image.

4. OCR implementation: OCR implementation on an image includes preprocessessing` it, segmenting it into lines and words, and then using an Tesseract OCR algorithm to recognize and extract text. This procedure transforms visual text data into machine-readable text, allowing for digital storage, searchability, and analysis of the content within the image.

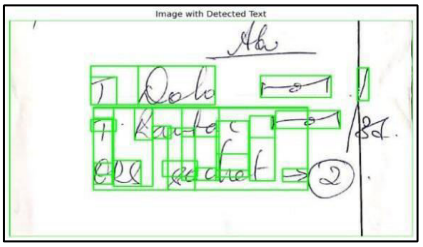

Fig 16.Tesseract OCR implemented handwritten image

Fig 17. Partially extracted text fromthe image using Tesseract OCR

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207102|

B. Applying EasyOCR for Printed Prescription:

1. Dataset image: Considering one of the original Printed image from the dataset for implementation .

Fig 18. Original printed prescription image.

2. Line segmentation: Line segmentation in printed images involves discovering and separating specific lines of text within a document. This procedure is critical for accurate Tesseract OCR because it guarantees that each line of text is handled independently, enhancing recognition accuracy and reducing errors causedby overlapping or misaligned text lines.

Fig 19. Line segmented printed image.

3. Word detection: Word recognition in a printed image includes recognizing and extracting specific words within the image. The method uses Tesseract OCR technology to scan and analyze the text, identifying word boundaries based on spacing and font properties, ensuring accurate extraction and conversion of printed text into digital format.

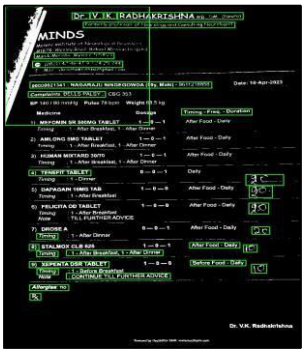

Fig 20. Word detected printed image.

4. OCR implementation: To implement Tesseract OCR on a printed image, first it must be preprocessed. Next, it must be divided into words and lines. Finally, text may be recognized and extracted using the Tesseract OCR algorithm. Through this process, printed text data is converted into machine-readable text, enabling digital storage, content analysis, and searchability.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207102|

Fig 21. Tesseract OCR implemented printed image.

Fig 22. Extracted text from the printed image.

IV. CONCLUSION

This study examines the variable performance of OCR tools when applied to various types of text images, with a special emphasis on handwritten medical prescriptions. While techniques like EasyOCR work well with printed text, their effectiveness decreases with sophisticated handwriting. After a thorough investigation, it was discovered that no single OCR program excelled across all image types. Instead, the finest OCR solution for handwritten medical prescriptions frequently use specialized or hybrid approaches, which combine numerous tools to improve accuracy and dependability. By identifying the strengths and limitations of each OCR tool, this study helps healthcare professionals and developers choose the best OCR technology for digitising handwritten medical documents, improving efficiency and accuracy in medical data management.

REFERENCES

1. Comparison of Various Machine Learning Algorithms For Recognizing Text On The Medical Prescriptions. Sandhya P1 , Rama Prabha K.P2 , Jayanthi.R3 , V. Sujatha4 , Asha N5 , M B Benjula anbu malar
2. Medical Prescription Recognition using Machine Learning. Authors: Esraa Hassan, Mai hazem, Habiba Tarek, Shaza Bahnacy, Lobna Shaaheen, Walaa H. Elashmwai
3. Machine learning application in epilepsy. Authors: Bardia Abbasi, Daniel M. Goldenholz
4. Medical Imaging using machine Learning and Deep Learning Algorithms: A Review. Author: Jahaanzaib Latif , Chuangbai Xiao, Azhar Imran, Shanshan Tu
5. Interpreting Doctor notes using Handwriting Recognition and Deep Learning Techniques: Authors: Nithin Kumar, Amrutha Chikmath

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details