Decognize: Prescription Digitization Using Knowledge Graphs

Group Members:
Muhammad Sherjeel Akhtar (P20-0101)
Mahad Ashraf(P20-0563)

Supervisor:
Mr. Muhammad Shoaib Khan

Table of contents

1. Literature Review
2. Problem Statement
3. System Diagram
4. UML Diagrams
5. Objectives
6. Expected Output
7. Gantt Chart

1. Problem Statement

Problem Statement

- **Problem**: Inefficient healthcare data management for prescriptions.
- **Challenge**: Illegible handwriting , medical jargon and Knowledge Graph
- **Consequence**: Errors in healthcare due to traditional OCR systems.
- **Goal**: Develop NLP-based system for accurate prescription transcription

2. Literature Review

Literature Review

Sr. no	Year	Basic Idea	Methodologies	Results	Limitations
[1]	2023	OCR with Open CV and tesseract	Implemented Tesseract OCR with Open CV in python. Focusing on image pre-processing for optimal results integrated text detection and recognition components	Achieved satisfactory OCR accuracy with well-preprocessed images. However, Tesseract struggled with complex backgrounds and artifacts, yielding suboptimal outputs.	Tesseracts accuracy is hindered by poor image quality. Requiring meticulous preprocessing . Challenges arise in handling artifacts handwriting and diverse languages .
[2]	2021	Optical Character Recognition Using TensorFlow	Implemented OCR with TensorFlow Enhanced model robustness with data augmentation technique. Implemented a custom ResNet architecture for OCR	These results showcase the effectiveness of the OCR model, particularly in accurately recognizing characters within the test set, demonstrating its robustness and suitability for the specified task.	Our Model can fail if the image is complex . E.g cursive writing images or images with Continous Characters Currently our model is trained only on digits and English language
[3]	2021	Construct a Bio Medical Knowledge Graph with NLP	Extracted text from biomedical document using OCR and applied BERN and utilized zero relation extractor.	Successfully established a Neo4j knowledge graph, showcasing versatility through demonstrated applications such as search engine, co-occurrence analysis and author expertise inspection. While emphasizing its utility for diverse biomedical machine learning applications.	Limitations include persistent NER challenges with BERN, potential inaccuracies in the zero shot relation extractor and the need for expert validation with external database enrichment reliant on data consistency
[4]	2018	Build a Handwritten Text Recognition System using TensorFlow	Implemented HTR using TensorFlow, with NN trained on IAM word-images, including CNN, RNN and CTC layers. Preprocessed data with resizing normalization and potential augmentation. Utilized RMSProp for training and explored enhancements like data augmentation, input size adjustments and decoding strategies.	Implemented successful HTR on IAM word -images, enabling flexible NN customization and identifying areas for accuracy improvements.	Limited Diversity due to reliance on IAM dataset Potential recognition errors especially for non-dictionary words CPU based training may be slower : GPU recommended
[5]	2022	Doctor Handwritten Prescription recognition system in multilanguage using deep learning	Implemented a system employing machine learning techniques such as CNNs,RNNs,LSTMs for recognizing and translating handwritten prescription notes in diverse language	Successful recognition and translation of handwritten prescriptions in various languages. Demonstrated the efficiency of CNNs, RNNs, and LSTMS in multilingual handwritten text processing.	Sensitivity to variations in handwriting styles. Reliance on quality and diversity of training data for optimal performance
[6]	2022	A Comparison of various Machine learning Algorithms for recognizing Text on Medical prescriptions	Proposed approach involves image scanning pre-processing and CNN-based feature extraction for recognizing handwritten medical prescriptions. Results are compared with drug name database using OCR for medicinal name identification	Successful implementation of CNN-based recognition for medical prescription. Need for further investigation into alternative machine learning algorithms for comprehensive comparision.	Limited Exploration of alternative machine learning algorithms Identification challenges with low accuracy medical names in OCR
[7]	2020	Online Cursive Handwritten Medical Words Recognition System	Implemented an online cursive handwritten medical word recognition system using a bidirectional LSTM network. Employed data augmentation techniques to enhance recognition efficiency.	Successful Utilization of bidirectional LSTM for cursive medical word recognition Recognition efficiency improvements achieved Through data augmentation	The system is restricted to providing output only for the trained data Inability to generate output for the new unseen data due to lack of adaptability
[8]	2021	Medical Prescription Recognition Using Machine Learning	Developed a Medical Prescription Recognition System employing image processing techniques and machine learning algorithms to identify handwritten medicine names from prescription note images.	Successful integration of image processing and machine learning for medical prescription recognition . Acknowledged limitations include reliance on small dataset and lower accuracy	Limited dataset usage in the system The system exhibits low accuracy levels
[9]	2021	Medical Prescription Identification Solution	Implemented a Medical Prescription Identification Solution employing a neural network for character recognition and knowledge-based maching for accurate results.	Utilized neural network approach and knowledge based matching for effective prescription identification	Restricted to reading only one line at a time

3. System Diagram

System Diagram

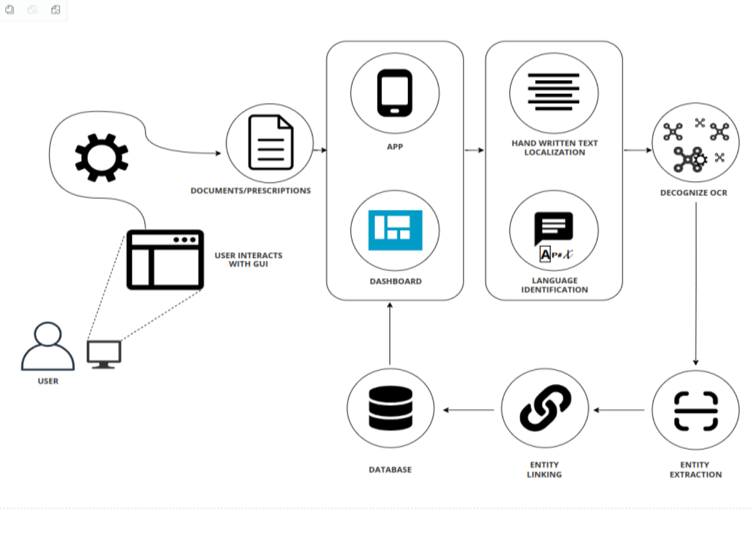

Figure 1: Architecture Diagram of DeCognize

4. UML Diagrams

Use Case Diagram

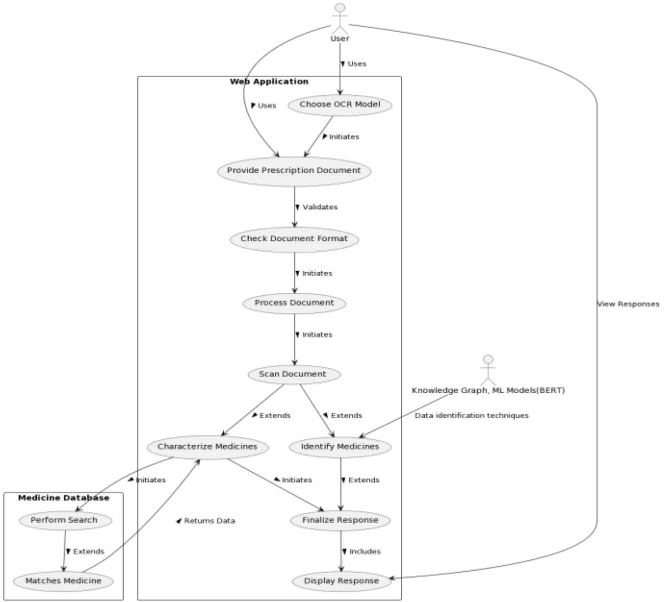

Figure 2: Use Case Diagram of DeCognize

Activity Diagram

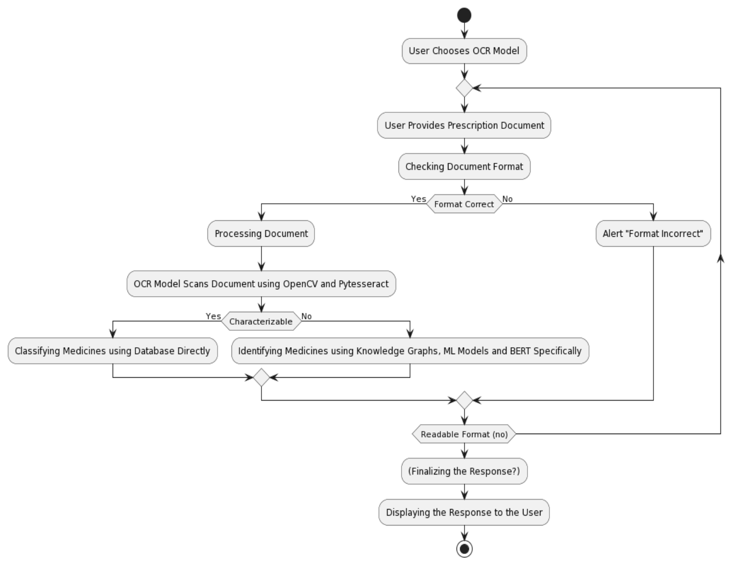

Figure 3: Activity Diagram of DeCognize

Swimline Diagram

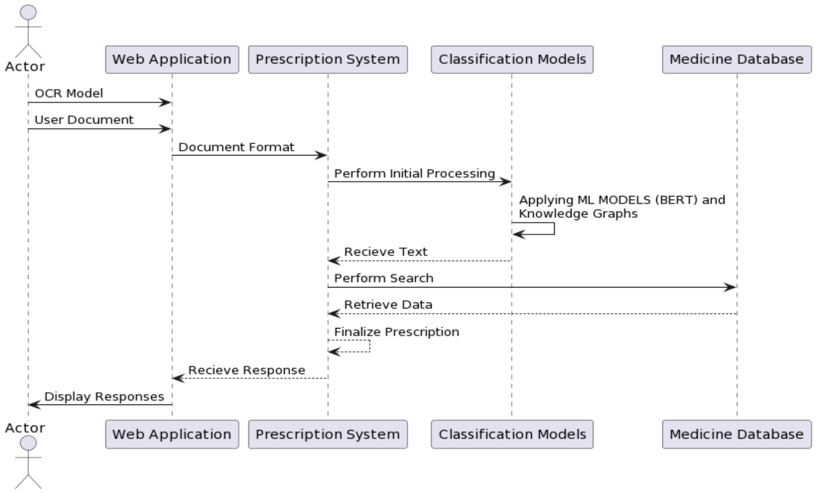

Figure 3: Swimline Diagram of DeCognize

Flow Diagram

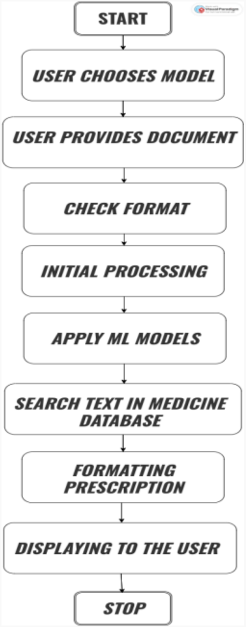

Figure 3: Flow Diagram of DeCognize

5. Objectives

Objectives

- To reduce error percentage in prescriptions readability.
- To create an improved OCR system which could later on deployed on other real-life-domains as well.
- To allow user to save and access their prescription data conveniently.

6. Sample Demo

Sample Code

```
import cv2
import pytesseract

pytesseract.pytesseract.tesseract_cmd = r"C:\Program Files\Tesseract-
OCR\tesseract.exe"

# Reading image
img = cv2.imread("sample.png")

# Convert to RGB
img_rgb = cv2.cvtColor(img, cv2.COLOR_BGR2RGB)

# Use pytesseract to detect and print text
custom_config = r'--oem 3 --psm 6'
texts = pytesseract.image_to_string(img_rgb, config=custom_config)
print("Texts:", texts)

# Save the text to a file
output_file_path = "output.txt"
with open(output_file_path, "w", encoding="utf-8") as text_file:
   text_file.write(texts)

# Use pytesseract to get bounding boxes
boxes = pytesseract.image_to_boxes(img_rgb,
config=custom_config)

# Draw bounding boxes on the image
for b in boxes.splitlines():
   b = b.split()
   x, y, w, h = int(b[1]), int(b[2]), int(b[3]), int(b[4])
   img_rgb = cv2.rectangle(img_rgb, (x,
img_rgb.shape[0] - y), (w, img_rgb.shape[0] - h), (0,
255, 0), 2)

# Show the image with bounding boxes
cv2.imshow("Output", img_rgb)
cv2.waitKey(0)
cv2.destroyAllWindows()

print(f"Texts saved to {output_file_path}")
```

Figure 4: Sample Code

```
Texts: DD tony 1289
1 NOV 71
DOD PRESCRIPTION
Ne 22 CoS dhe tes nl)
FOR (Full name, address, & phone number) (it under 12, give age)
John R Doe, HH3, VSN
Pe eee ee ee cree eee ee
US.S. Never forgotten (00 178)
MEDICAL FACILITY DATE
US.S. Neverforgotten (00 178) I Sand
BR (Superscription) gm or ml.
(nscription)
ta (lliden ra 15 | nl
Amphege geek 120\me
(Subscription)
IN + JL Polar
(Signe)
Se Sm tid ac
MEGA: th
LOT NO: 39K /06
{ack R. Frost
- LCDR. WD. USKR
BR NUMBER SIGNATURE RANK AND DEGREE
EDITION OF 1 JAN 60 MAY BE USED FOR
S/N 0102-LF-012-6201
```

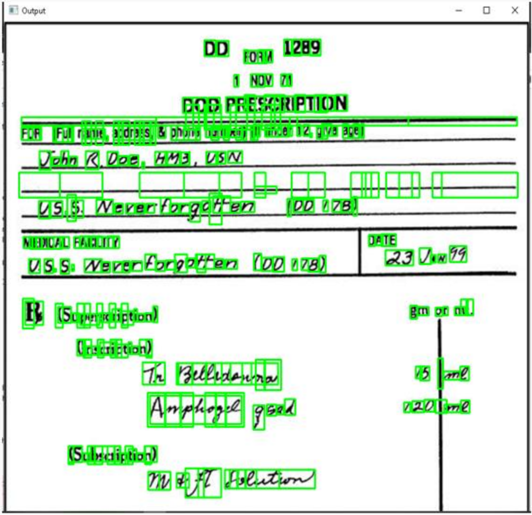

Figure 4: Sample Output

7. Expected Output Using Knowledge Graph

Sample Knowledge Graph

8. Gantt Chart

Gantt Chart

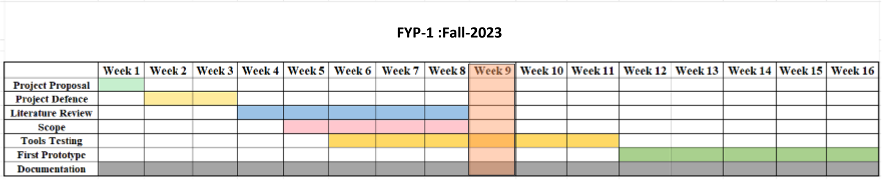

Figure 5: Gantt Chart

9. References

1. Filip Zelic and Anuj Sable. A review on on OCR with Tesseract OpenCV and Python. Nanonets, 2023.

2. Kamlesh Solanki . A review on optical character recognition using tensor flow. Medium, 10:39154–39176, 2021.

3. Tomaz Bratanic , D. Kim *et al.*, "A Neural Named Entity Recognition and Multi-Type Normalization Tool for Biomedical Text Mining," in *IEEE Access*,Medium vol. 7, pp. 73729–73740, 2019, doi: 10.1109/ACCESS.2019.2920708. , 2021

References [3]

4. Harald Scheidl. Article on Build handwritten text recognition using tensorflow. Medium, 9:87643–87662, 2018.

5. Kamalanaban, E., M. Gopinath, and S. Premkumar. "Medicine box: Doctor’s prescription recognition using deep machine learning." International Journal of Engineering and Technology (UAE) 7 (2018): 114-117.

6. Sandhya, P., and K. P. Rama Prabha. "Comparison Of Various Machine Learning Algorithms For Recognizing Text On The Medical Prescriptions." Journal of Pharmaceutical Negative Results (2022): 2083-2091.

References [2]

7. Tabassum, Shaira, Nuren Abedin, Md Mahmudur Rahman, Md Moshiur Rahman, Mostafa Taufiq Ahmed, Rafiqul Islam, and Ashir Ahmed. "An online cursive handwritten medical words recognition system for busy doctors in developing countries for ensuring efficient healthcare service delivery." Scientific reports 12, no. 1 (2022): 1-13

8. Hassan, Esraa, Habiba Tarek, Mai Hazem, Shaza Bahnacy, Lobna Shaheen, and Walaa H. Elashmwai. "Medical prescription recognition using machine learning." In 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), pp. 0973-0979. IEEE, 2021.

9. Wijewardena, W. R. A. D. "Medical Prescription Identification Solution." PhD diss., 2021