MedDocParser: Automated OCR, Medical NER, and ICD-10 Coding

A Project Report Submitted by

Aatif Ahmad

B.Tech in Artificial Intelligence and Data Science

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

Bachelor of Technology

Indian Institute of Technology Jodhpur
Department of Computer Science and Engineering

November, 2025

Declaration

I hereby declare that the work presented in this Project Report titled **“MedDocParser: Automated OCR, Medical NER, and ICD-10 Coding”**, submitted to the Indian Institute of Technology Jodhpur in partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology**, is a bonafide record of the research work carried out by me under the supervision of **Dr. Sumit Kalra**.

The contents of this Project Report, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Signature:
Name: Aatif Ahmad
Roll Number: B22AI002

Certificate

This is to certify that the Project Report titled **"MedDocParser: Automated OCR, Medical NER, and ICD-10 Coding"**, submitted by **Aatif Ahmad (Roll Number: YOUR ROLL)** to the Indian Institute of Technology Jodhpur for the award of the degree of **Bachelor of Technology**, is a bonafide record of the research work done by him under my supervision.

To the best of my knowledge, the contents of this report have not been submitted to any other Institute or University for the award of any degree or diploma.

Supervisor:
Dr. Sumit Kalra
Assistant Professor
Department of Computer Science and Engineering
IIT Jodhpur

Acknowledgements

I would like to express my sincere gratitude to my supervisor, **Dr. Sumit Kalra**, for his continuous guidance, support, and encouragement throughout this project. I also thank **Dr. Manik Sejwal** for providing valuable medical and technical insights that shaped the direction of this work.

I am grateful to IIT Jodhpur for providing the computational resources and academic environment necessary for completing this project.

Abstract

MedDocParser is a prototype system designed for parsing medical documents such as prescriptions and reports. It integrates OCR for text extraction, NLP methods for recognizing clinical entities (diseases, symptoms, medications), and maps these entities to ICD-10 codes. The system includes a Streamlit web interface for fast upload and interpretation of results.

The project addresses the challenges of healthcare data digitization and reduces manual errors during clinical documentation. Experiments conducted on sample prescriptions demonstrate effective entity extraction and ICD mapping. Limitations include OCR difficulty on handwritten text and reliance on external APIs. Future work includes integrating local models, increasing ICD coverage, and enhancing UI with entity highlighting.

Contents

Chapter 1

Introduction and Background

Healthcare generates large amounts of unstructured text in prescriptions, lab reports, and discharge summaries. Manual extraction of medical entities is time-consuming and error-prone, which complicates EHR workflows and ICD-10 standardization.

This project automates document parsing using PaddleOCR for text extraction, GPT-based NER for identifying clinical concepts, and a fuzzy-matching ICD-10 lookup system. The goal is to reduce manual workload and improve healthcare documentation efficiency.

Objectives

- Develop an end-to-end pipeline for parsing medical PDFs/images.
- Use GPT-based NER to extract diseases, symptoms, medications, and findings.
- Map extracted entities to ICD-10 codes using multi-level matching.
- Build an interactive Streamlit interface for testing and demonstration.

Chapter 2

Literature Survey

OCR tools such as Google Vision API and Tesseract are widely used, while PaddleOCR offers strong accuracy with multilingual support. Medical NER is commonly performed using models like BioBERT, cTAKES, or transformer-based architectures.

ICD coding systems typically rely on keyword rules or similarity-based matching. In this project, GPT-4o-mini is used for high-flexibility NER, and fuzzy matching improves recall in ICD lookup.

Chapter 3

Problem Definition and Objective

Problem Definition

Hospitals rely heavily on manual interpretation of prescriptions for identifying diseases, symptoms, and clinical terms. There is no easy automated pipeline that can extract these entities and map them to ICD-10 codes in a reliable manner.

Objective

To design and implement an automated system that:

- extracts text from medical documents,
- identifies key medical entities,
- assigns ICD-10 codes,
- and presents results via a user-friendly interface.

Chapter 4

Methodology

System Architecture

The pipeline consists of three modules:

1. **OCR Module:** Extracts raw text using PaddleOCR.
2. **NER Module:** Uses GPT-4o-mini to identify diseases, symptoms, medications, and findings.
3. **ICD Mapping Module:** Performs exact, substring, and fuzzy matching against an ICD-10 dataset.

Implementation Details

- Streamlit for user interface.
- Python backend with modular code structure (`app.py`, `app_ui.py`).
- ICD dataset stored in `icd10.csv`.
- Environment variables used for API keys.

Chapter 5

Theoretical/Numerical/Experimental Findings

Evaluation

The system was tested on multiple sample prescriptions. PaddleOCR effectively extracted printed text, while GPT-based NER achieved high recall across clinical entities.

Example mapping:

- "Hypertension" → ICD-10: I10
- "Diabetes Mellitus Type-2" → ICD-10: E11

Discussion

The model is modular and scalable. Compared to traditional rule-based systems, GPT-based NER adapts better to varied document formats.

Chapter 6

Summary and Future Plan of Work

Summary

The project successfully demonstrates an end-to-end automated pipeline for medical document parsing, combining OCR, NER, and ICD mapping within a simple web interface.

Future Work

- Use fully offline transformer-based medical NER models.
- Integrate complete ICD-10 hierarchy.
- Improve UI to highlight extracted entities in the document.
- Evaluate system against real clinical datasets.

Publications (if any)

None as of now.

Appendix (if any)

Bibliography

[1] PaddleOCR Documentation. https://github.com/PaddlePaddle/PaddleOCR.

[2] OpenAI API Documentation. https://platform.openai.com/docs.

[3] World Health Organization. ICD-10 Coding Guidelines. https://icd.who.int.

[4] J. Lee et al., “BioBERT: a pre-trained biomedical language representation model,” *Bioinformatics*, 2019.