# Project Report: Automated Lab Report Digitization

Ishaan Gupta 2022A7PS1180P

Date: October 2, 2025

### Abstract

The manual extraction of data from medical lab reports is a time-consuming, error-prone bottleneck in clinical workflows. This project solves this by developing an intelligent automation system to digitize and structure information from PDF lab reports. The system uses a hybrid method combining OCR, a rule-based engine, and a custom-trained Named Entity Recognition (NER) model, along with a human-in-the-loop web interface for data verification and continuous model improvement. The core NER model achieved a high precision of 88.89%, while the end-to-end system demonstrated an F1-score of 39.63% on a real-world test set, highlighting its potential to significantly improve efficiency in clinical data management.

### 1. Introduction

Manually transcribing medical lab reports into electronic health records (EHRs) is slow and susceptible to errors. This project creates a robust, automated pipeline to extract data from varied lab report layouts. The solution is centered on a hybrid model that uses a rule-based system for simple cases and a sophisticated NER model for complex ones. A web UI allows for human review, ensuring data accuracy and providing a stream of corrected data for retraining the model.

### 2. Methodology

The system is a modular pipeline that processes uploaded lab reports (PDF, PNG, JPG). Files are preprocessed using PyMuPDF and OpenCV to create clean images for the Tesseract OCR engine. The extracted text is then fed into a hybrid extraction model. A custom spaCy NER model, trained on human-corrected data, identifies key entities. If the model's confidence is low, the system falls back on a rule-based module that uses heuristics and regular expressions. A critical component is the "human-in-the-loop" web interface, which allows users to correct inaccuracies, with saved corrections used to create an ever-growing dataset for retraining the NER model.

### 3. Results and Evaluation

The system was evaluated at two levels: the performance of the core NER model during training and the performance of the complete end-to-end pipeline against a manually verified test set.

NER Model Performance

The spaCy NER model was evaluated during its training phase against a validation set. It reached a peak F1-score of 62.02%, demonstrating high precision in identifying entities correctly.

* **Precision**: 88.89%
* **Recall**: 47.62%
* **F1-Score**: 62.02%

End-to-End System Performance

The entire hybrid pipeline was evaluated against 9 manually corrected "ground truth" documents. This score reflects the system's real-world performance, including OCR errors and fallback logic.

* **Precision**: 63.37%
* **Recall**: 28.83%
* **F1-Score**: 39.63%

### 4. Discussion and Conclusion

The results show a precise NER model, meaning the data it identifies is very reliable. The lower recall and F1-score for the end-to-end system highlight real-world challenges like OCR errors and diverse document layouts (only 3 layouts used here) that the model has yet to learn. The hybrid approach ensures a result is always returned, and the human-in-the-loop mechanism is crucial for generating the data needed to improve recall over time.

This project successfully created an automated system for digitizing lab reports.

Attached herewith are images for reference

