MEDISCAN – INTELLIGENT MEDICAL DOCUMENT PARSER

OVERVIEW

- Automates extraction of structured data from unstructured PDF documents.
- Targets use cases like medical prescriptions and patient records.
- Built using Python, FastAPI, Tesseract OCR, and PDF2Image.

PROBLEM STATEMENT

- Manual data entry from medical PDFs is error-prone and time-consuming.

- Unstructured medical documents make it challenging to extract key details like patient info, medicines, or dosage without specialized tools.

TECH STACK

Languages & Tools:

- 🐍 Python (Backend logic and OCR)
- 🔎 Tesseract OCR (Text extraction)
- 📄 pdf2image + Poppler (PDF to image conversion)
- ⚙️ FastAPI (REST API)
- 🧪 Streamlit (Testing / frontend)
- 📂 JSON (structured output)

PROJECT EXECUTION STEPS

Fast API Backend

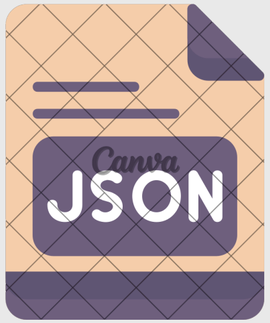
Upload Pdf

pdf to img

Extract txt

python™

Regex
(frontend)



Final output

OCR & PARSING LOGIC

OCR Phase:

- Uses Tesseract with preprocessed images.
- Converts entire PDF into plain text.

Parsing Phase:

- Regex and keyword-based logic to extract:

👤 Patient Name

🏠 Address

💊 Medicines

📋 Directions

🔁 Refill count

JSON OUTPUT SAMPLE


```
{
   "patient_name": "John Doe",
   "patient_address": "123 Street Name",
  "medicines": ["Paracetamol 500mg", "Amoxicillin 250mg"],
   "directions": "Take twice daily after meals",
    "refill": "3 times"
   }
```

BENEFITS

- Easily extract text from scanned or printed PDF documents.
- No technical skills needed – just upload and get results.
- Saves time compared to manual data entry.
- Keeps user data private with local processing.
- Supports structured output for easy understanding and reuse

FUTURE SCOPE

Achievements:

- Fully functional backend for PDF parsing
- Accurate extraction with handwritten/typed docs (if clear)

Future Enhancements:

- Multilingual support
- Image-only inputs
- Mobile app integration
- AI-based parsing using NLP (for better accuracy)

THANK YOU