Doument Extractor

1. Problem Statement

Organizations frequently handle documents such as **PAN cards, Aadhaar cards, GST certificates, and Bank passbooks** during onboarding, KYC, and vendor management processes.
Manually extracting information from these documents is **time-consuming, error-prone, and inefficient**, especially when dealing with large volumes.

The goal of this project is to **automate the extraction of structured data** from scanned documents and PDFs using OCR and rule-based logic, and to present the extracted information in a **reliable, confidence-scored format** suitable for ERP systems and automation workflows.

2. Architecture / Flow

1. User uploads an image or PDF document through the web interface
2. The document is preprocessed (resize, grayscale, noise removal) to improve OCR accuracy
3. OCR engine extracts raw text from the document
4. Rule-based and pattern-based logic identifies required fields
5. Confidence scores are calculated for each extracted field
6. The structured output is displayed in the UI and returned via API

3. Tools & Libraries

- **Python** – Core programming language
- **FastAPI** – Backend REST API framework
- **Tesseract OCR** – Text extraction from images
- **OpenCV** – Image preprocessing and enhancement
- **PyMuPDF** – PDF to image conversion
- **HTML / CSS / JavaScript** – Frontend user interface
- **pytest** – Unit testing framework

4. Key Logic Explanation

- Regular expressions are used to detect structured patterns such as **PAN numbers, Aadhaar numbers, GSTINs, bank account numbers, and IFSC codes**
- Context-based heuristics are applied to extract **names and addresses** from nearby text
- Each extracted field is assigned a **confidence score** based on OCR clarity and pattern reliability
- The system handles low-quality or unreadable documents gracefully by returning low confidence values instead of incorrect data