# InvoiceSight

# Invoice Extraction and Verification

# Kaviya N S

# 22PD16

# ****InvoiceSight**** is a robust, intelligent document processing system designed to extract structured information from scanned invoice PDFs. The application not only identifies key fields such as invoice number, dates, GST numbers, and purchase order references, but also detects signatures/seals and verifies tabular line items for completeness and accuracy.

### **Approach Adopted**

1. **Text Extraction**:
   * **Primary**: Text is extracted from PDF using **pdfplumber**, which is efficient for machine-readable PDFs.
   * **Fallback OCR**: For scanned or low-quality PDFs, **PaddleOCR** is used to perform Optical Character Recognition (OCR) on images.
2. **Regex-based Field Extraction**:  
   A dictionary-driven regex system is used to extract fields such as:
   * Invoice number
   * Date
   * GSTINs (Supplier & Buyer)
   * Purchase Order Number
   * Shipping Address
   * Total Amount
3. **Signature/Seal Detection**:
   * The page is processed using **OpenCV** and **scikit-image** to:
     + Convert to grayscale.
     + Apply thresholding and connected component labeling.
     + Remove small noise and isolate large contour regions likely to be signatures.
   * The largest contour is extracted as a potential seal/signature.
4. **Table Parsing and Line Item Verification**:
   * Tabular data is extracted using pdfplumber.extract\_tables().
   * Supports multiple formats:
     + HSN/SAC-based line items.
     + Service descriptions with quantity embedded.
   * Line items are parsed into structured JSON with quantity, unit price, and total amount fields.
5. **Validation & Confidence Scoring**:
   * Confidence scores are assigned to each field (high for regex matches, low if not found).
   * Verification reports are generated to indicate completeness and potential issues.