Project Title:

Automated Personal Loan Document Processing System

1. Methodologies Used

- **Image Preprocessing (OpenCV):**
  Uploaded loan application images (JPG/PNG) are first converted to grayscale, denoised, and thresholded to improve clarity and contrast — optimizing them for OCR.
- **Text Extraction (Tesseract OCR):**
  The preprocessed image is passed to Tesseract OCR to extract raw textual data from the scanned document.
- **Information Extraction (Regex):**
  Regular expressions are applied to parse and extract key fields:
  - Full Name
  - Date of Birth (DOB)
  - PAN Number
  - Income
  - Mobile Number
- **Data Validation:**
  Extracted fields are validated for completeness and correct formatting (e.g., valid PAN pattern, DOB format). Any missing or incorrect fields are flagged.
- **User Interface (Streamlit):**
  A web-based interface displays all extracted fields in editable input boxes, allowing users to review and correct them.
- **Data Storage (Pandas → CSV):**
  Once validated, the final data is appended to a CSV file, simulating storage in the bank's loan processing system. A timestamp is included for audit tracking.

2. Results

- Successfully processed scanned application forms with over **90% OCR accuracy** after preprocessing.
- Key information like PAN, DOB, and mobile number were reliably extracted using pattern-based parsing.
- Editable field interface allowed easy correction of occasional OCR misreads.
- Generated structured CSV files with time-stamped, validated data entries — ready for integration into backend systems.

3. Conclusion

This project effectively automates the extraction and validation of data from scanned personal loan documents. By combining OCR, rule-based parsing, and a user-friendly review interface, it reduces manual data entry effort, improves accuracy, and streamlines the document processing workflow for Banks.