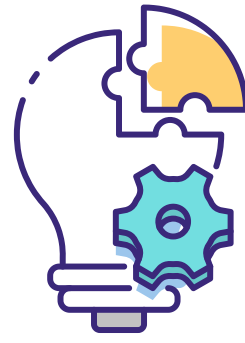


Instructor: Dr. Masoud Sadjadi
Project Manager: Karthikeyan Saravanan
College of Engineering & Computing — Florida International University
Students : Carlos Hernandez, Luis Jimenez, Ernesto Blanco, Christoph McPhee

Problem Statement



Loan applicants applying for loans submit various financial documents such as bank statements, pay stubs, IDs, and tax forms. Loan processors must manually verify all details, detect fraud, check for inconsistencies, and cross-reference values across multiple documents. This manual process is slow, prone to human error, and makes fraud easier to miss.

Foreign or non-standard bank statements add even more difficulty.

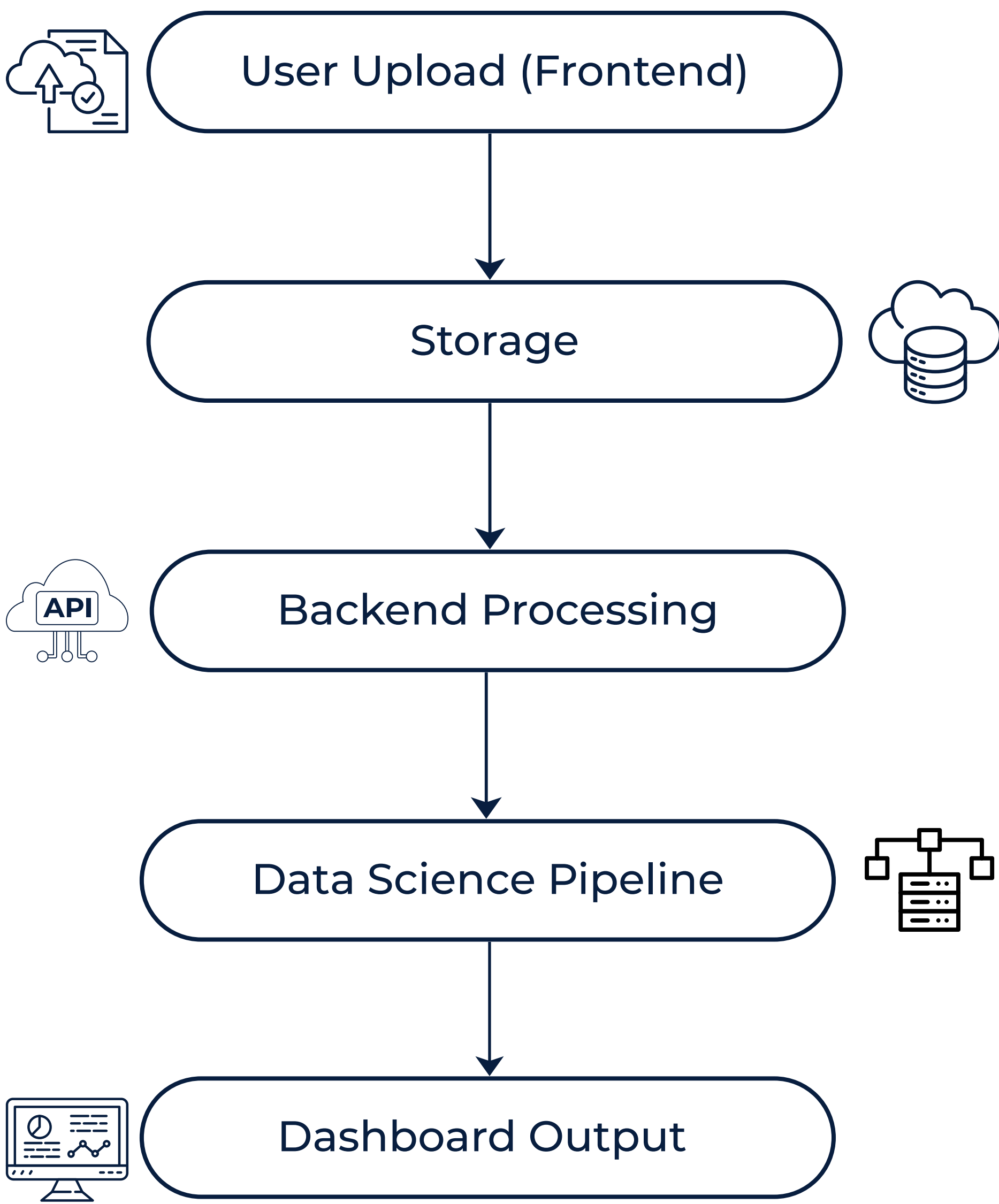
Our Goal: Build an automated AI-powered system to extract, validate, and cross-check applicant information with high accuracy.

Project Objectives

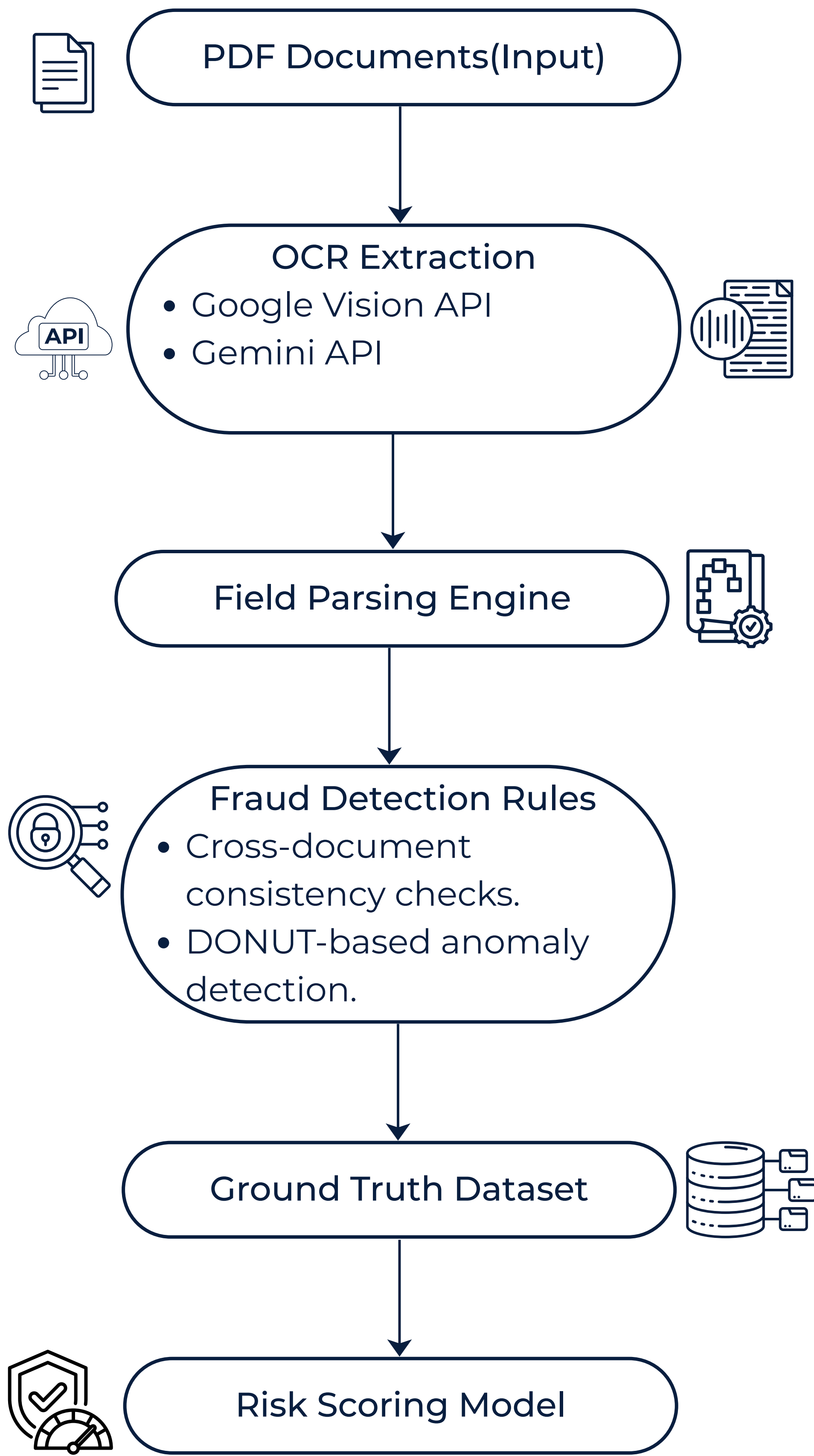


- Automate the extraction of key fields from financial documents using OCR and AI.
- Detect fraudulent or altered documents using machine learning and rule-based methods.
- Cross-validate applicant data across multiple document types (bank statements, pay stubs, IDs, tax forms).
- Provide a secure web dashboard for document upload, review, and fraud analysis.
- Improve processing speed, reduce manual workload, and increase accuracy.
- Develop a risk scoring model to classify applicants (Low/Medium/High) based on extracted and validated financial information.
- Provide an executive summary of overall applicant risk by analyzing fraudulent documents, identifying discrepancies, and cross-validating key financial information.

System Architecture



Data Science Pipeline



Final Dashboard Output

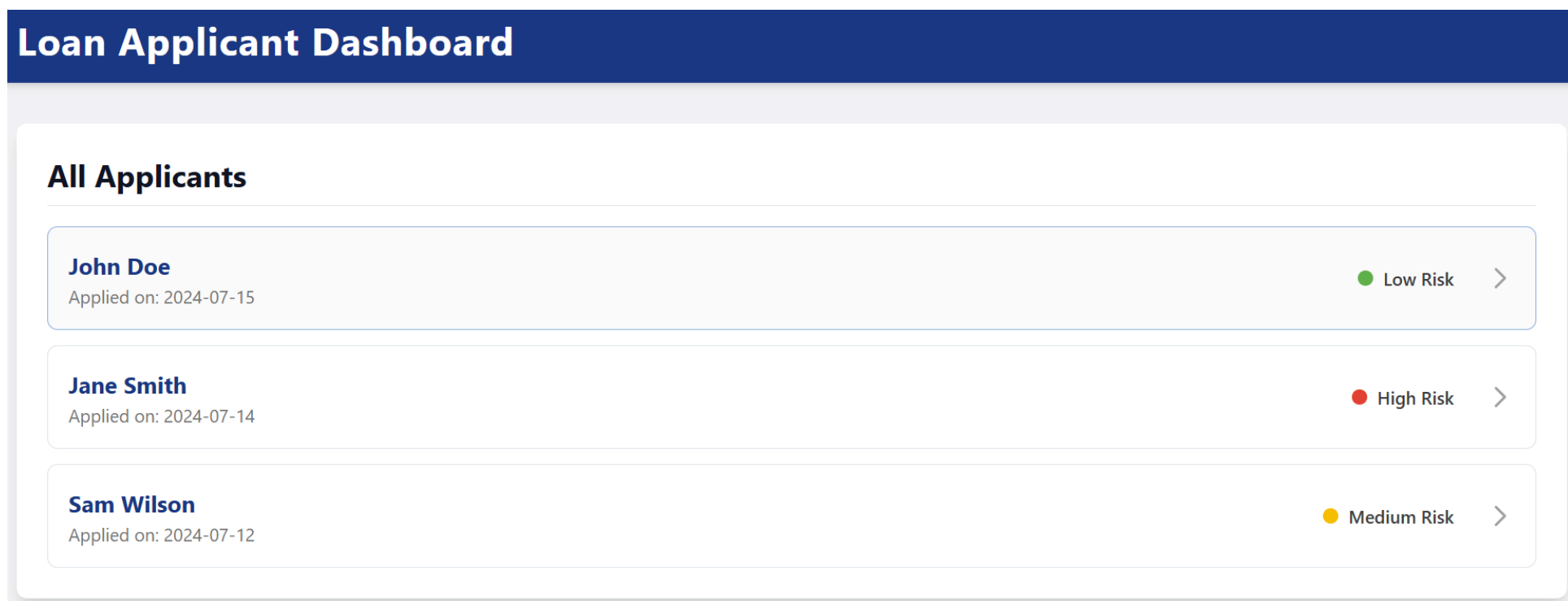


Figure 1. Applicant Dashboard showing risk levels for each applicant.

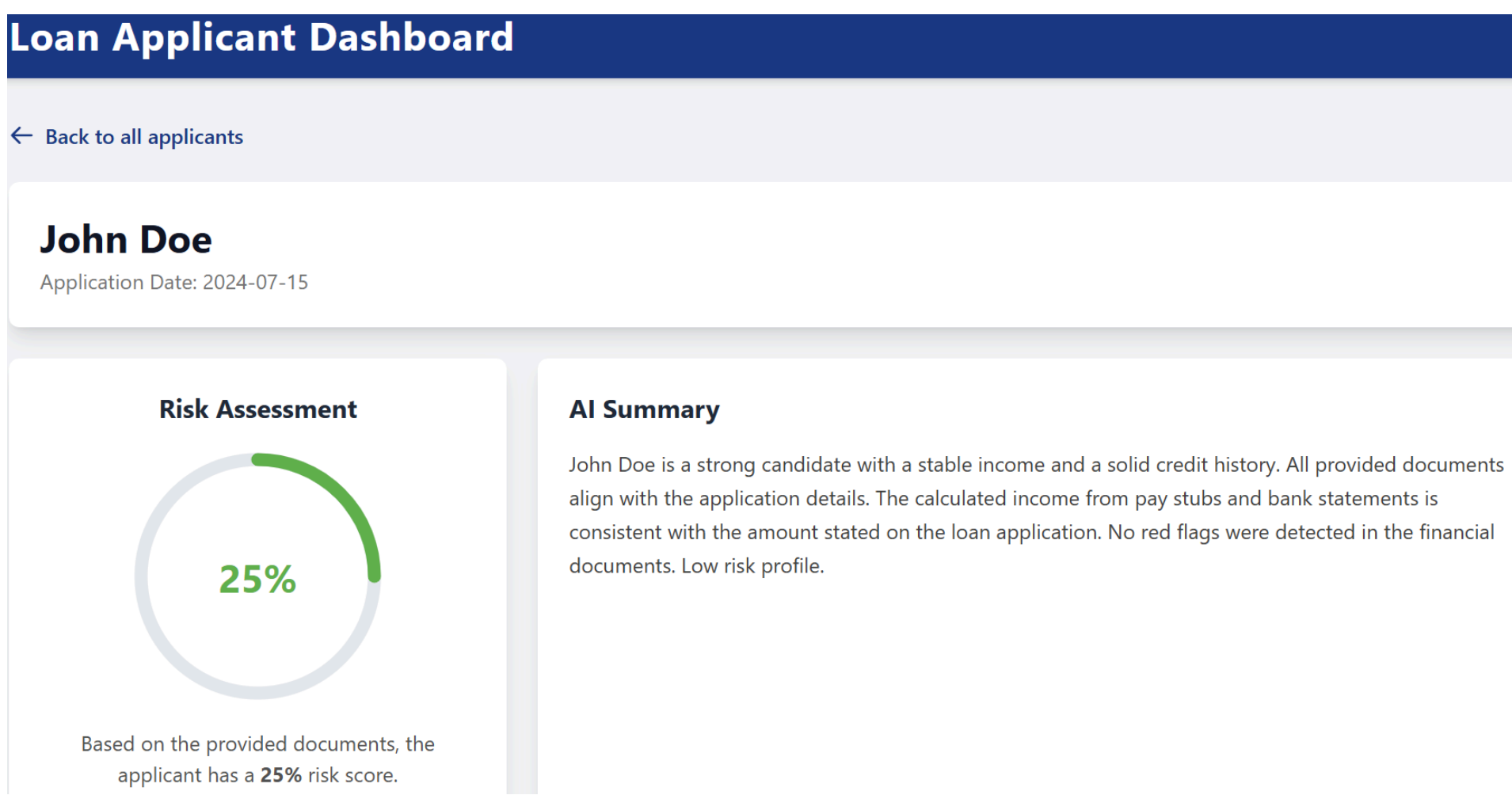


Figure 2. Applicant Detail View with automated risk score and AI-generated summary.

Summary:

The Data Science pipeline uses advanced OCR models (Google Vision + Gemini API) to extract text from financial documents. It also integrates the DONUT document understanding model to enhance OCR accuracy and enable anomaly detection for fraud identification. Extracted text is processed using rule-based and heuristic parsers to derive key fields, and a fraud detection engine performs cross-document consistency checks. A manually curated ground truth dataset is used to validate and improve system accuracy.