OPERATIONS RESEARCH CENTER

MIT MANAGEMENT BUSINESS ANALYTICS

CAPSTONE PROJECT

Document Classification Capability

Revving up manual paperwork with Computer Vision & NLP

MIT x Wolters Kluwer

Faculty Advisor: Dr. Ilya Jackson

WK Advisors: Pooja Srivastava & Varun Dixit

18th August 2023

Rachit Jain Chloe Wu

Candidates of Master of Business Analytics, MIT

LET'S SOAR TOGETHER 2022 New Heights New Frontiers

Agenda

Introduction

Overview

Challenges

Motivation

The Process

Solution

Methodology
(Re-labelling, Modelling)

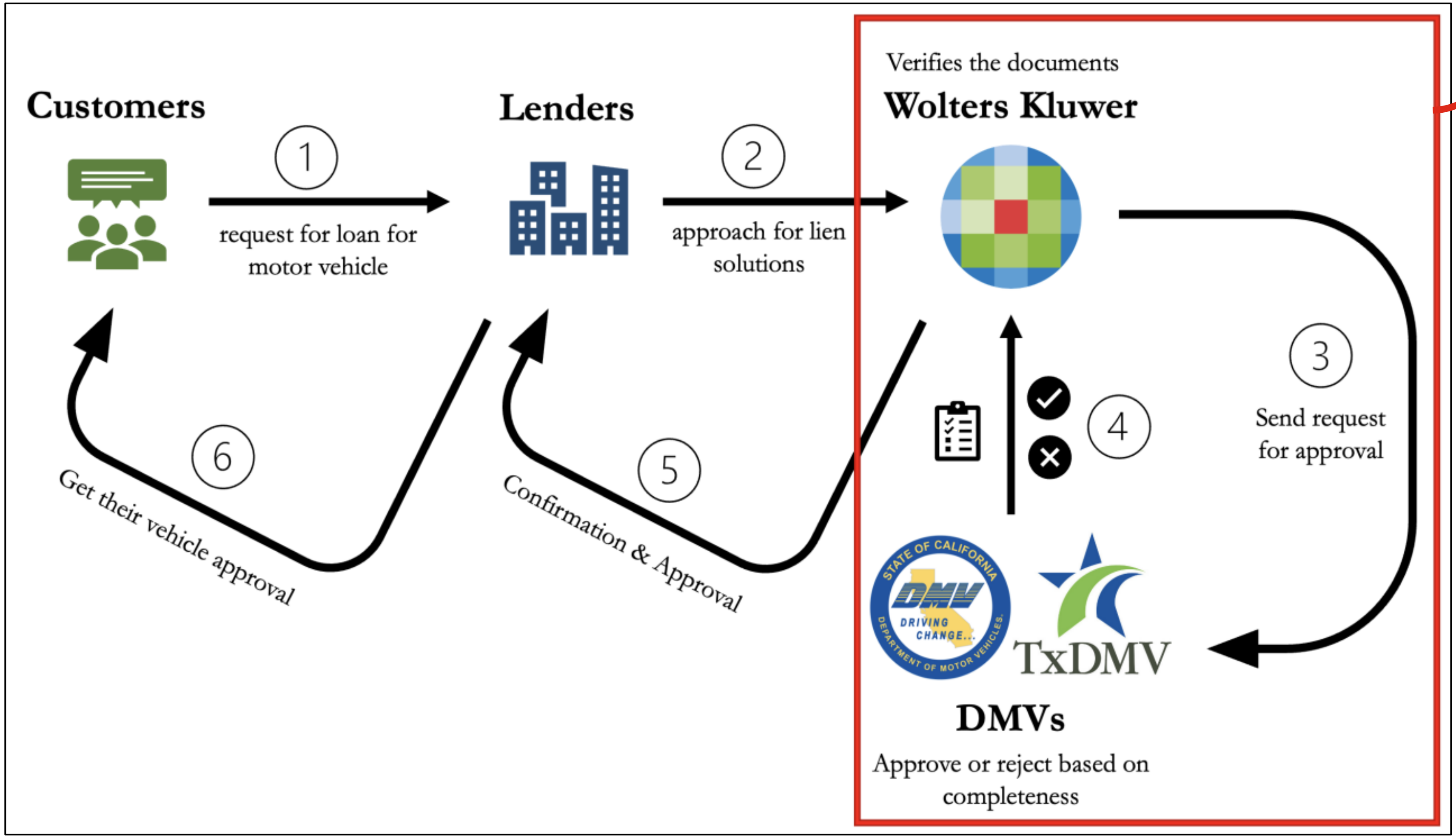
Results

Demo

Deliverables

Impact & Business Value

Overview | Motor Vehicle registration process is error-prone



Challenges At Scale

Huge Volume

50k+ pages[1] per day

Multiple rejections

10% rejection rate

High processing time

10 mins per request (~20 pages)

[1] 80 requests per day only for Lien-Add document type (1 on 8) in Texas state (1 on 50) with 20 pages per request

Challenges | Manual processing is a bottleneck

Challenges At Scale

50k+ pages[1] per day	10% rejection rate	10 mins per request (~20 pages)
Huge Volume	Multiple rejections	High processing time

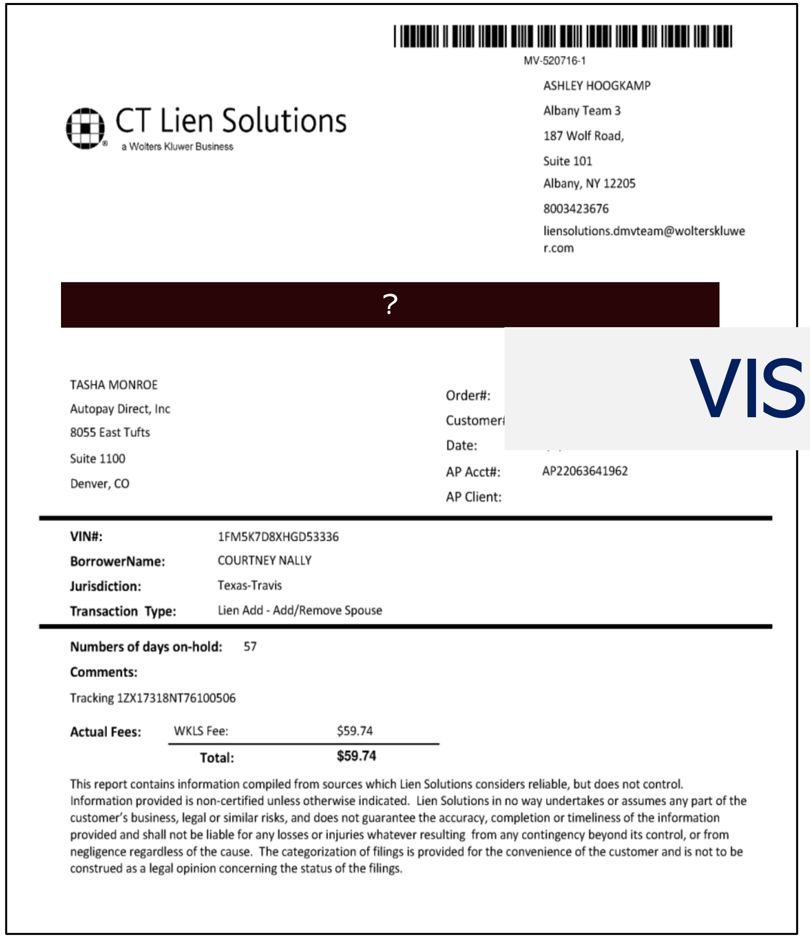
Final Goal

Build an automated, generalized document classification capability to make historically manual logistics paperwork easier to execute and more accurate

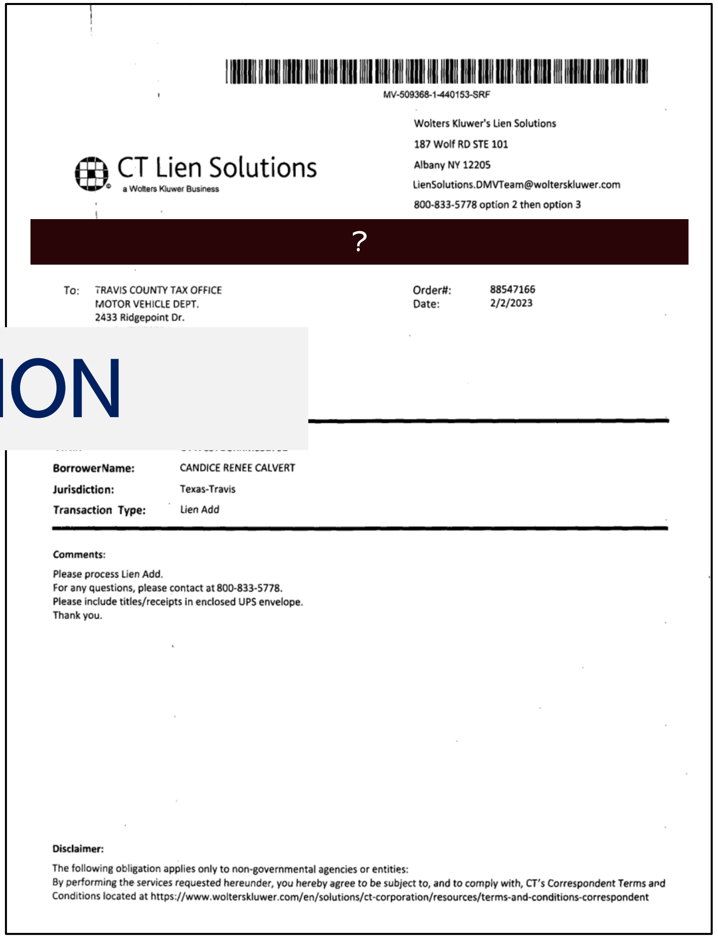
Motivation | Need for more than rule-based systems

Similar formats, similar text, but different titles

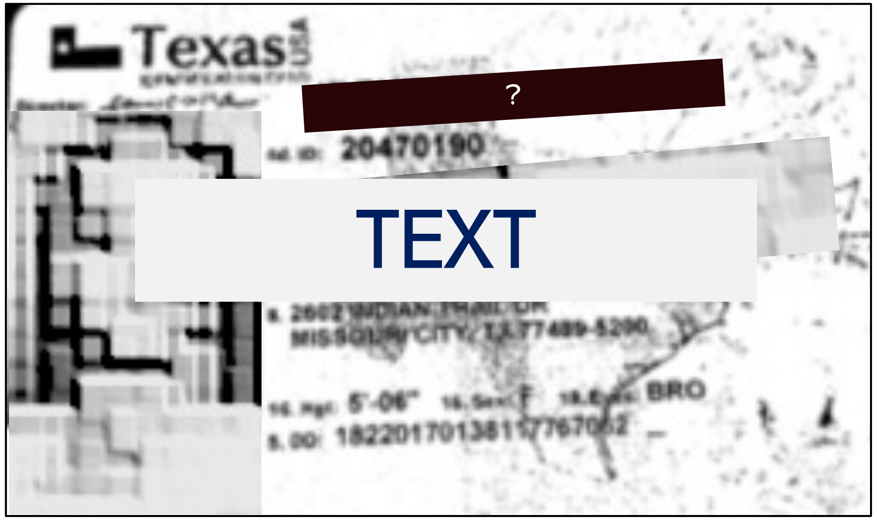
(1) Return Unprocessed

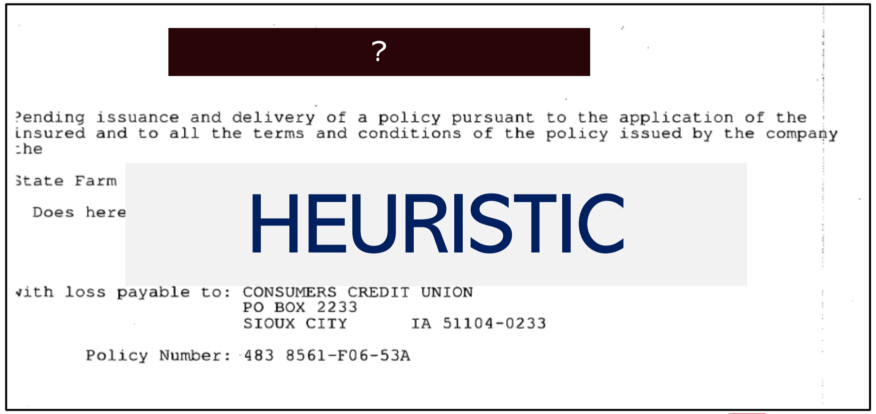


(2) Search Request Form



(3) Identification Card





(4) Binder of Insurance

LET'S SOAR TOGETHER 2022 New Heights New Frontiers

 Highlighted texts are results for the associated pages using our capability

Our Solution

CV + NLP + HEURISTIC

Computer Vision | Natural Language Processing | Title Extraction

= **Document Classification Capability**

= **Capstone of the Year?** 🫣

Imbalanced dataset across 120+ categories; 12k scanned pages

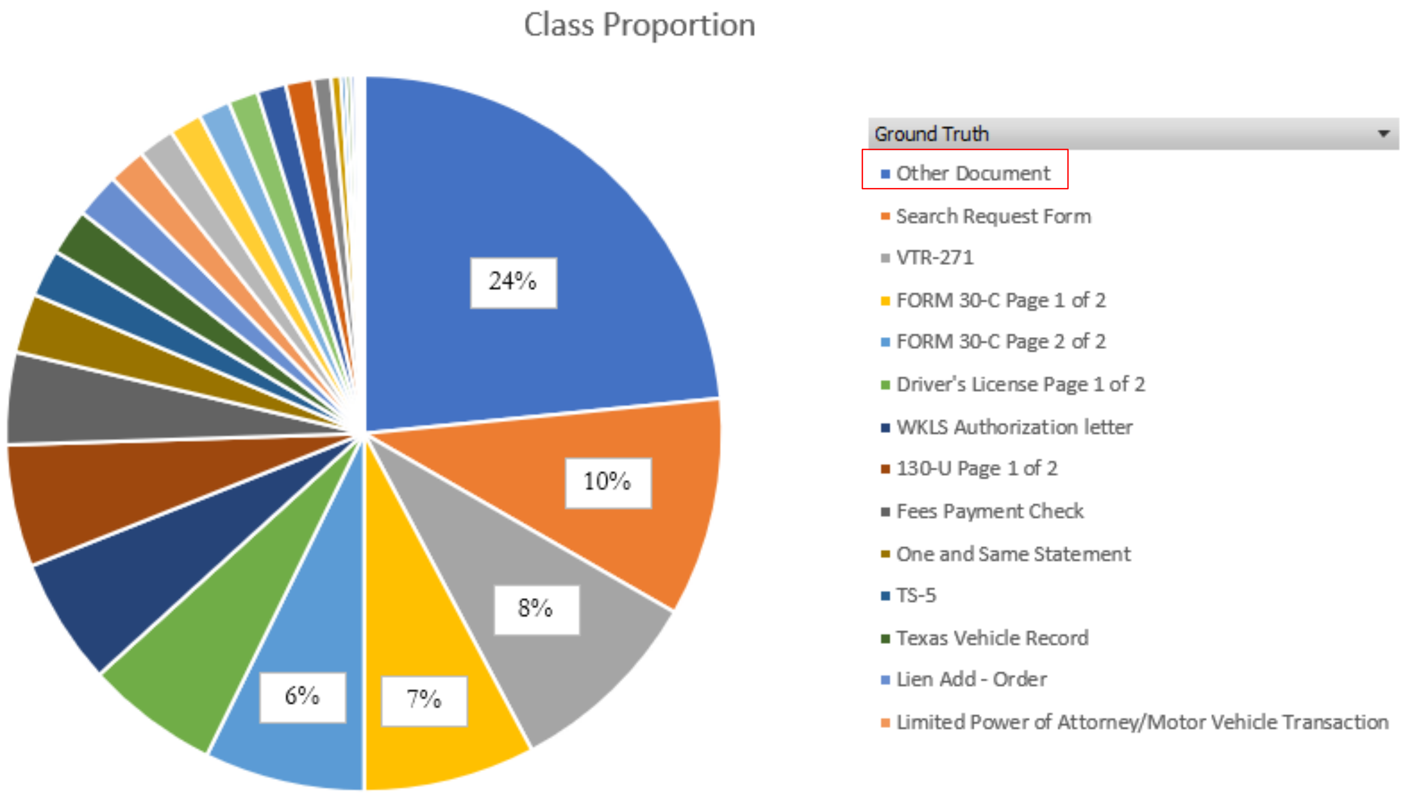
Training Data

12k+

Data Points

10k+

from top 31 classes



Represented Counties:
Bexar, Brazoria, Dallas, El Paso, Fort Bend, Harris, Hidalgo, Lubbock, Travis, Van Zandt

Solution | End-to-End Workflow takes PDF input and gives multiple labels for each page, along with confidence scores

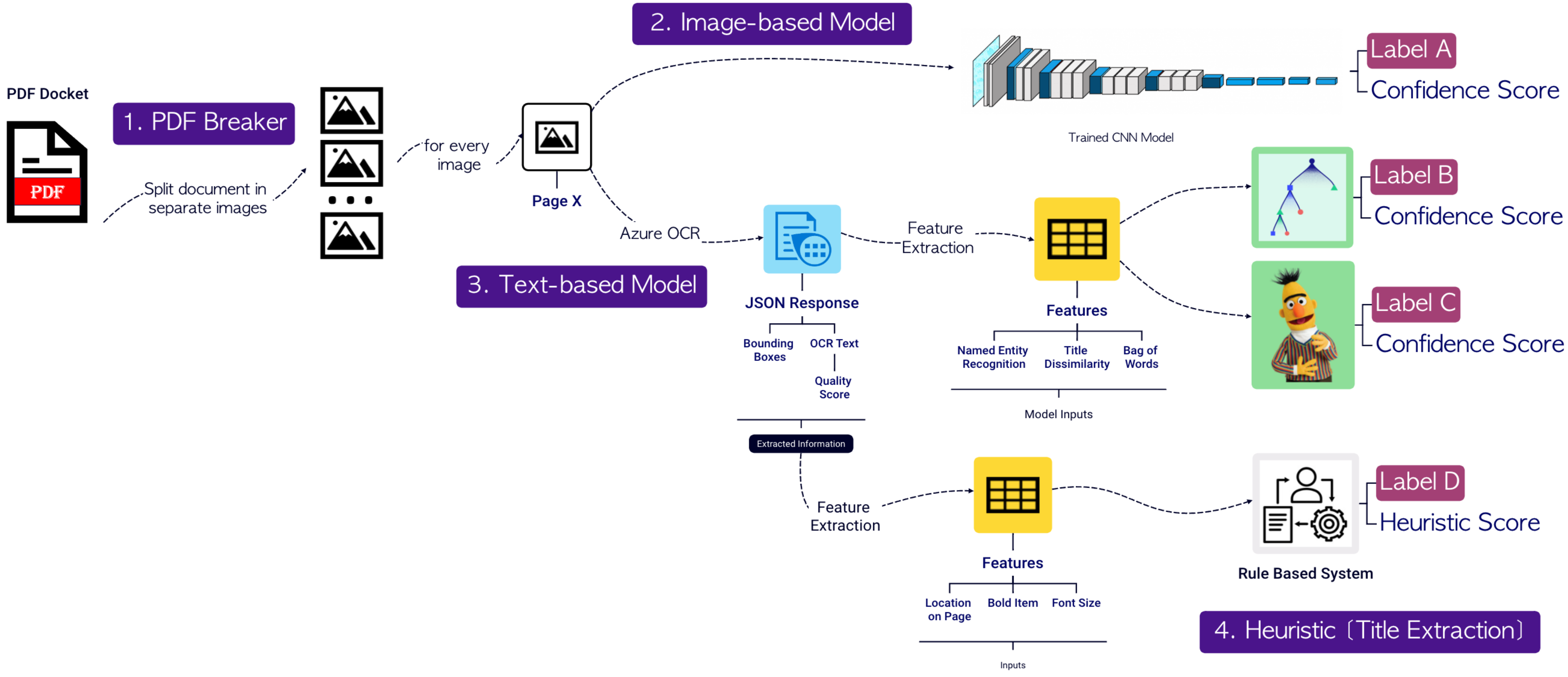


Fig: End-to-End Workflow

Step 0: Need for relabeling - Mystery behind ground truth labels

Status Quo

Current 'Ground Truth' Label = Output of Champion model

Wrong Labels → Poor Models

Solution?

'Smarter' Manual Labelling
[Unsupervised Clustering on Deep Embeddings]

- Cluster similar image embeddings from trained vision model
- Merge clusters on common categories & create sub-clusters
- Smarter Manual Label

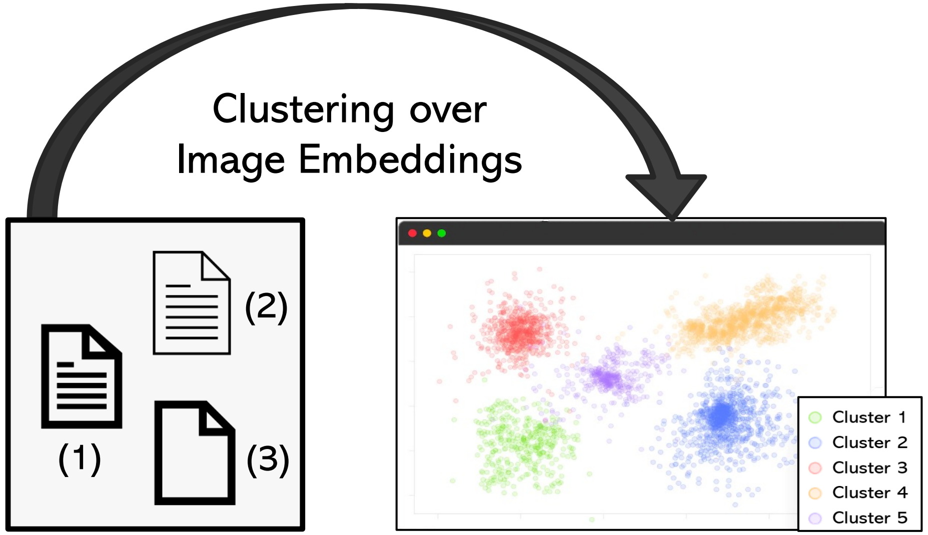
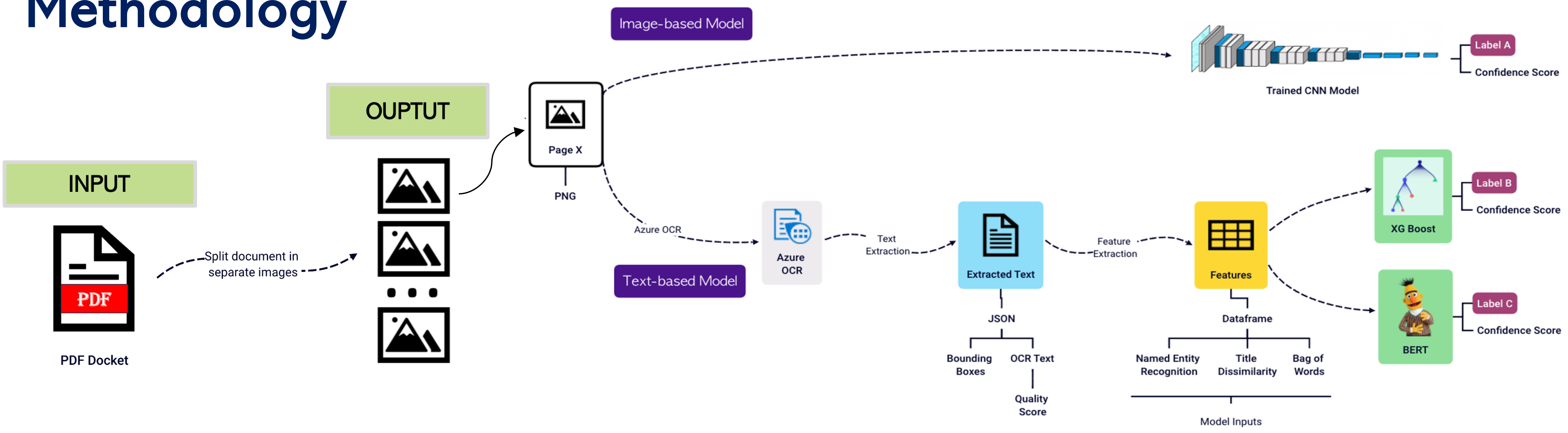


Fig: Embeddings for each image clustered based on similarity

Results

88%	10.6%	8%
saving in time for manual labelling	Mis-classified labels identified	F1 Score jump!

Methodology



1 PDF Breaker API

PDF Breaker on a Streamlit dashboard

2 Vision-based Model

CNN model fine-tuned on 12k scanned documents

3 Text-based Model

BERT + XGBoost running on text extracted from AzureOCR

Each page is extracted as an image and saved in folders

Highlights: Image padding, Image processing, Hyper-parameter tuning, Data Augmentation, Feature Engineering…

Methodology

① PDF Breaker API

PDF Breaker on a
Streamlit dashboard

Each page is extracted as an
image and saved in folders

② Vision-based Model

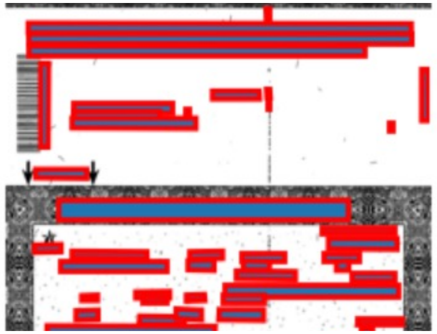
CNN model fine-tuned
on 12k scanned documents

③ Text-based Model

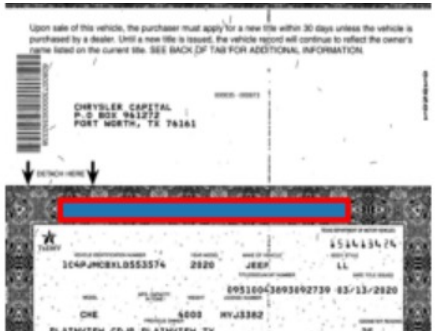
BERT + XGBoost running on
text extracted from AzureOCR

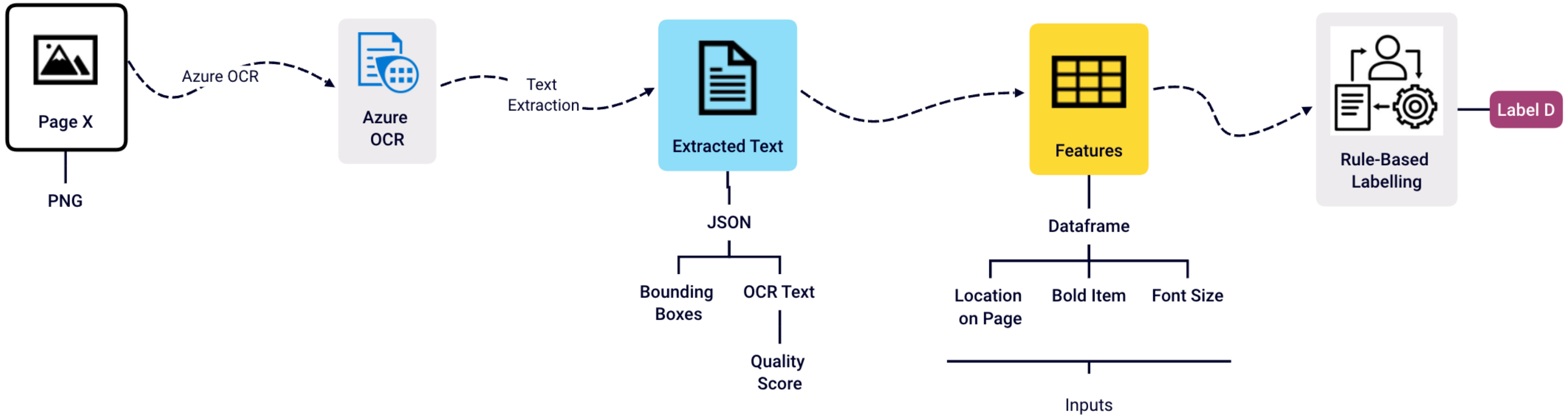
Original

Potential Title



Final Candidates





④ Heuristic-based Model

Potential title extraction
from any document type

⑤ Ensembling

Selecting 'supreme' model;
if clash, choose heuristic

The labels need to be ensembled into one single prediction

Predicted Category

Category	Vision Supremacy	Text Supremacy
Driving License	1	0
Lien Add – Order	0	1
…	…	…
Other Document	0	0

Binary Model Strength Look-Up Table (based on F1 Score)

Fig: Logic workflow to combine the labels

Our Implementation

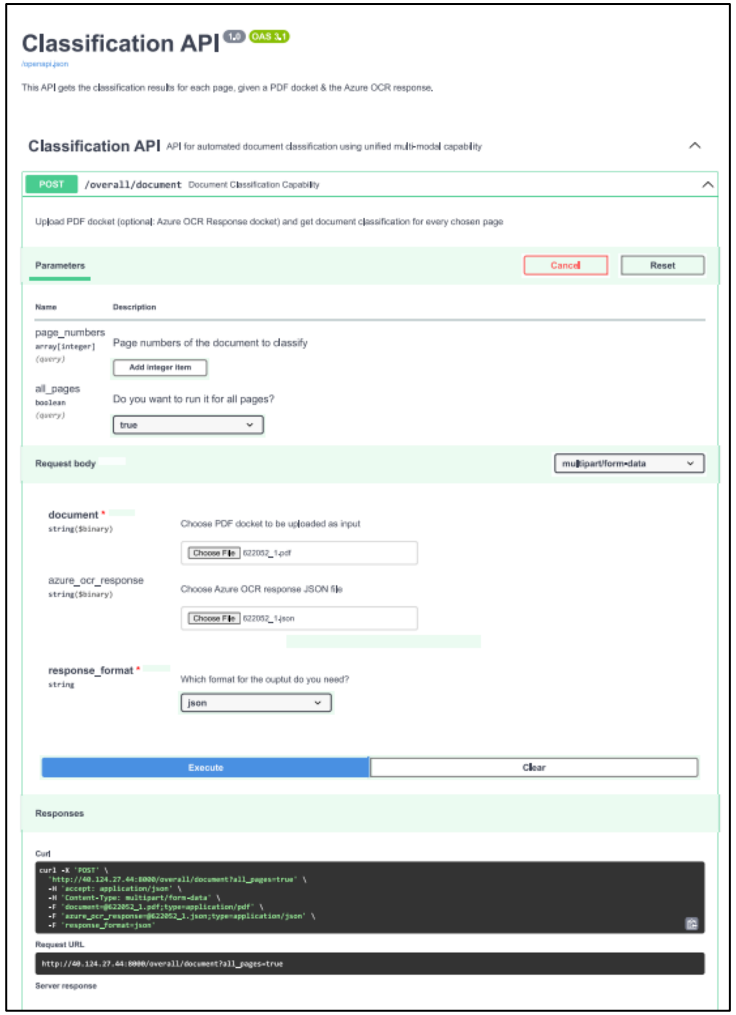
Model Supremacy for Vision and Text based models

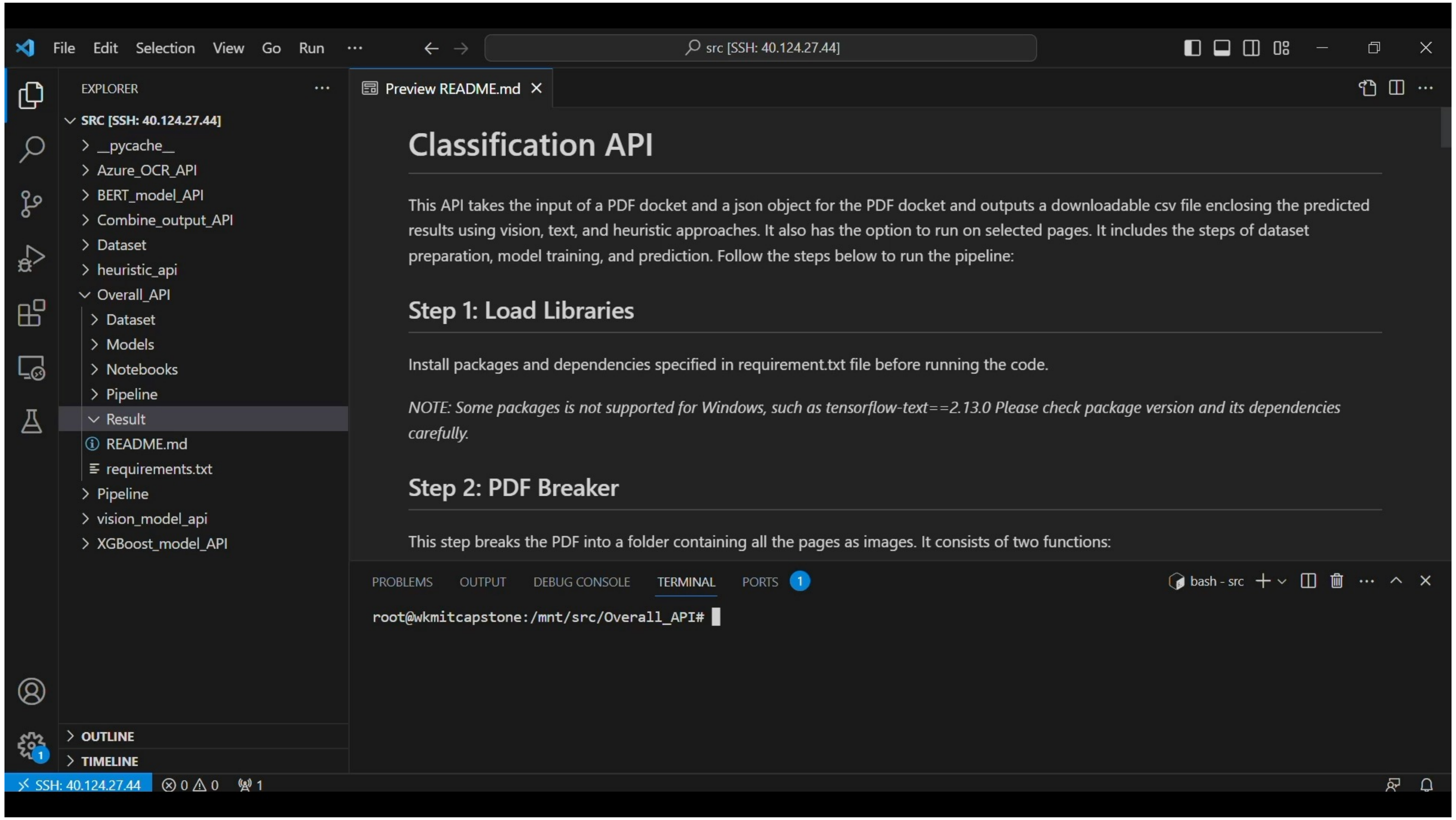
Training F1 score used to assign 'supreme' model for each category

Ensemble Technique

Case	Same Label	Vision Supermacy?	Text Supremacy?	Final Model
Case 1	1	1	0	Same
Case 2	0	1	1	Manual Review
Case 3	0	1	0	Vision
Case 4	0	0	1	Text
Case 5	1	0	0	Heuristic

Result | End-to-End multi-modal architecture deployed





Video: Demo implemented over FastAPI and deployed over WK's Virtual Machine

LET'S SOAR TOGETHER 2022 New Heights New Frontiers

Deliverables | Scalable model pipeline deployed over WK Cloud

Result Summary

0.86

F1 Score

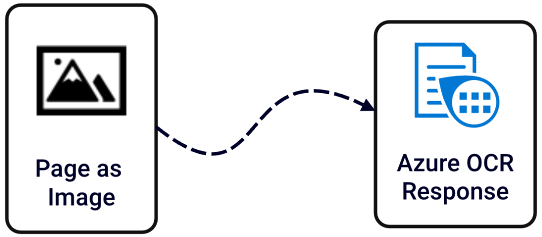
Over 31 document types over 2.1k highly noisy test dataset

Challenger > Champion

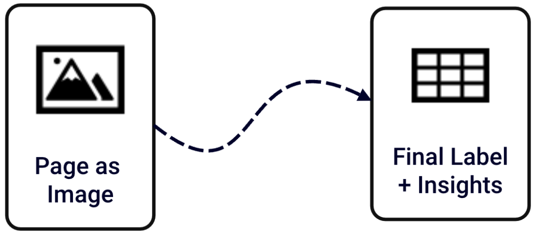
Our implementation beats the status quo with 5% higher F1 score

Our Deliverables to Wolters Kluwer

(1) OCR API



(2) CAPABILITY API



(3) Documentation & Knowledge Transfer

for smooth integration with current system

Learnings

Industry Practices

Dashboarding

Capability Design

Tech Stack

Azure

FastAPI

Business Skills

Business Value

Shareholders

10X growth in business

Generalized capability → Scalable business model

Market Differentiation

Leveraging AI → journey to be market-leaders

Customers

3X lower turn-around time

Less headache → Happy customers

70% lower rejection rate

Quick and efficient → drives customer experience

Employees

10X faster processing

Free from mundane work → higher productivity

Future-Ready by leveraging AI

Machine aided human experts → Upskilling

High learning + Solid deliverables + Strong impact =

More ACCURATE Labelling

0.86 F1
High → Better

AUTOMATED pipeline

1 API
running everything

LOW processing time

10X saving
in processing time

Flexible PDF BREAKER

Unrestricted
# of PDFs broken

FEWER rejections

Challenger >
Champion (status-quo)

PRODUCTIONIZED pipeline

Deployed
over WK's VM

Result INTERPRETABILITY

User Insights
behind predictions

END-TO-END pipeline

Streamlined
workflow

EASY-TO-INTEGEGRATE capability

Scalable
and reproducible

OPERATIONS RESEARCH CENTER

MIT MANAGEMENT BUSINESS ANALYTICS

Rachit Jain Chloe Wu

Candidates of Master of Business Analytics, MIT

Thank You!

"Only those who will risk going too far can possibly find out how far one can go!"

~ T.S. Eliot