



MIT
MANAGEMENT
BUSINESS ANALYTICS



Rachit Jain



Chloe Wu

Candidates of Master of Business Analytics, MIT

CAPSTONE PROJECT

Document Classification Capability

Revvng 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

Agenda

Introduction

Overview

Challenges

Motivation

The Process

Solution

Methodology
(Re-labelling, Modelling)

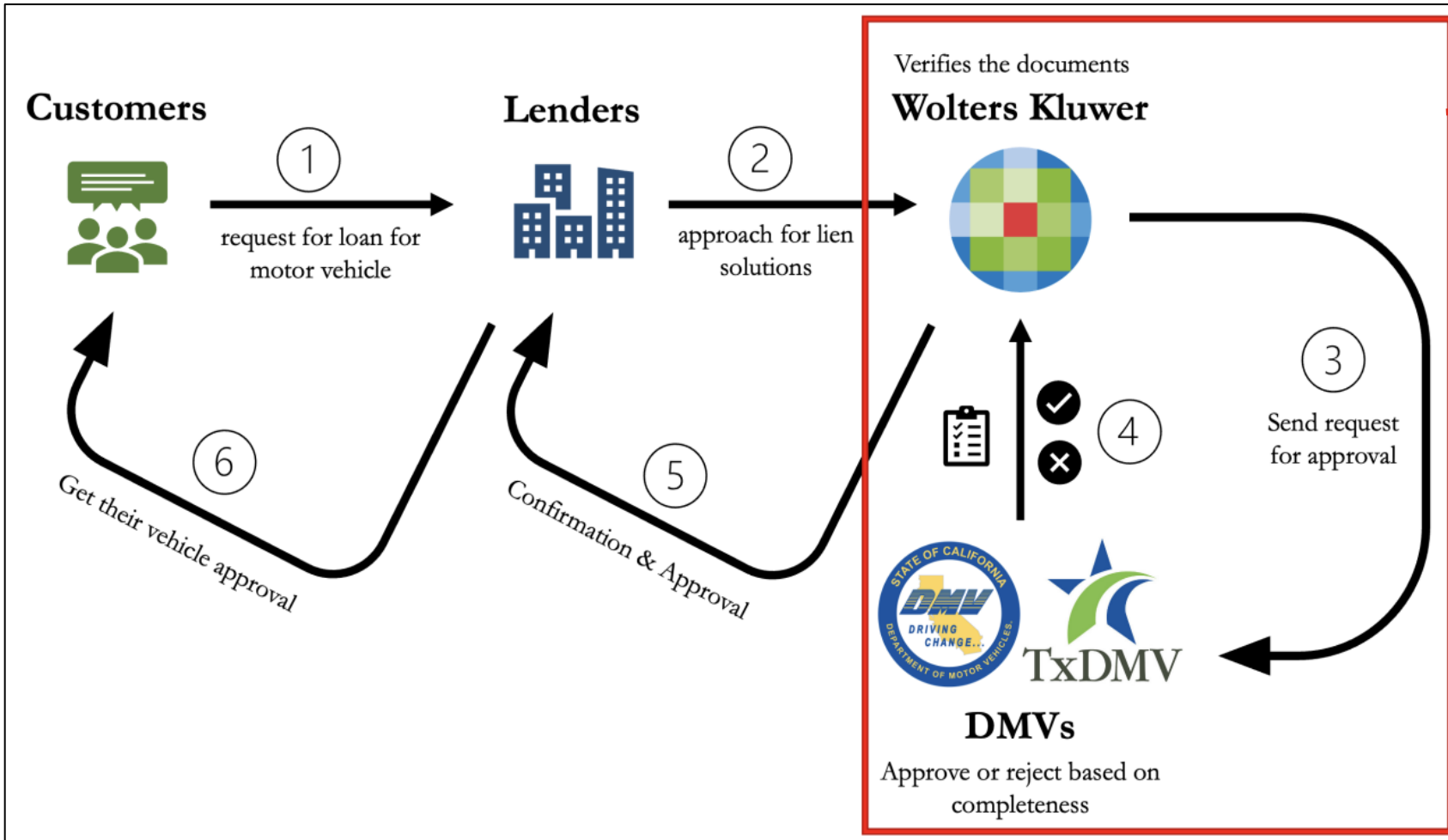
Results

Demo

Deliverables

Impact & Business Value

Overview | Motor Vehicle registration process is error-prone



Challenges At Scale

Huge Volume

50k+ pages¹
per day

Multiple rejections

10%
rejection rate

High processing time

10 mins per
request (~20 pages)

3 ¹80 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+ pages¹

per day

Huge Volume

10%

rejection rate

Multiple rejections

10 mins per

request (~20 pages)

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

CT Lien Solutions
a Wolters Kluwer Business

ASHLEY HOOBKAMP
Albany Team 3
187 Wolf Road,
Suite 101
Albany, NY 12205
8003423676
liensolutions.dmvteam@wolterskluwer.com

?

TASHA MONROE
Autopay Direct, Inc
8055 East Tufts
Suite 1100
Denver, CO

Order#: [REDACTED]
Customer: [REDACTED]
Date: [REDACTED]
AP Acct#: AP22063641962
AP Client: [REDACTED]

VISION

VIN#: 1FM5K7D8XHGD53336
BorrowerName: COURTNEY NALLY
Jurisdiction: Texas-Travis
Transaction Type: Lien Add - Add/Remove Spouse

Numbers of days on-hold: 57
Comments:
Tracking 1ZX17318NT76100506

Actual Fees: WKLS Fee: \$59.74
Total: \$59.74

This report contains information compiled from sources which Lien Solutions considers reliable, but does not control. Information provided is non-certified unless otherwise indicated. Lien Solutions in no way undertakes or assumes any part of the customer's business, legal or similar risks, and does not guarantee the accuracy, completion or timeliness of the information provided and shall not be liable for any losses or injuries whatever resulting from any contingency beyond its control, or from negligence regardless of the cause. The categorization of filings is provided for the convenience of the customer and is not to be construed as a legal opinion concerning the status of the filings.

(2) Search Request Form

CT Lien Solutions
a Wolters Kluwer Business

Wolters Kluwer's Lien Solutions
187 Wolf RD STE 101
Albany NY 12205
LienSolutions.DMVTeam@wolterskluwer.com
800-833-5778 option 2 then option 3

?

To: TRAVIS COUNTY TAX OFFICE
MOTOR VEHICLE DEPT.
2433 Ridgepoint Dr.

Order#: 88547366
Date: 2/2/2023

BorrowerName: CANDICE RENEE CALVERT
Jurisdiction: Texas-Travis
Transaction Type: Lien Add

Comments:
Please process Lien Add.
For any questions, please contact at 800-833-5778.
Please include titles/receipts in enclosed UPS envelope.
Thank you.

Disclaimer:
The following obligation applies only to non-governmental agencies or entities:
By performing the services requested hereunder, you hereby agree to be subject to, and to comply with, CT's Correspondent Terms and Conditions located at <https://www.wolterskluwer.com/en/solutions/ct-corporation/resources/terms-and-conditions-correspondent>

(3) Identification Card

Texas

?

20470190

TEXT

MISSOURI CITY, TX 77489-5200

5'06" 15.00" F 18.00" BRO

18220170138157767052

?

Pending issuance and delivery of a policy pursuant to the application of the insured and to all the terms and conditions of the policy issued by the company.

State Farm
Does here

HEURISTIC

with loss payable to: CONSUMERS CREDIT UNION
PO BOX 2233
SIOUX CITY IA 51104-0233

Policy Number: 483 8561-F06-53A

(4) Binder of Insurance



Our Solution



== Document Classification Capability

== Capstone of the Year? 🙄

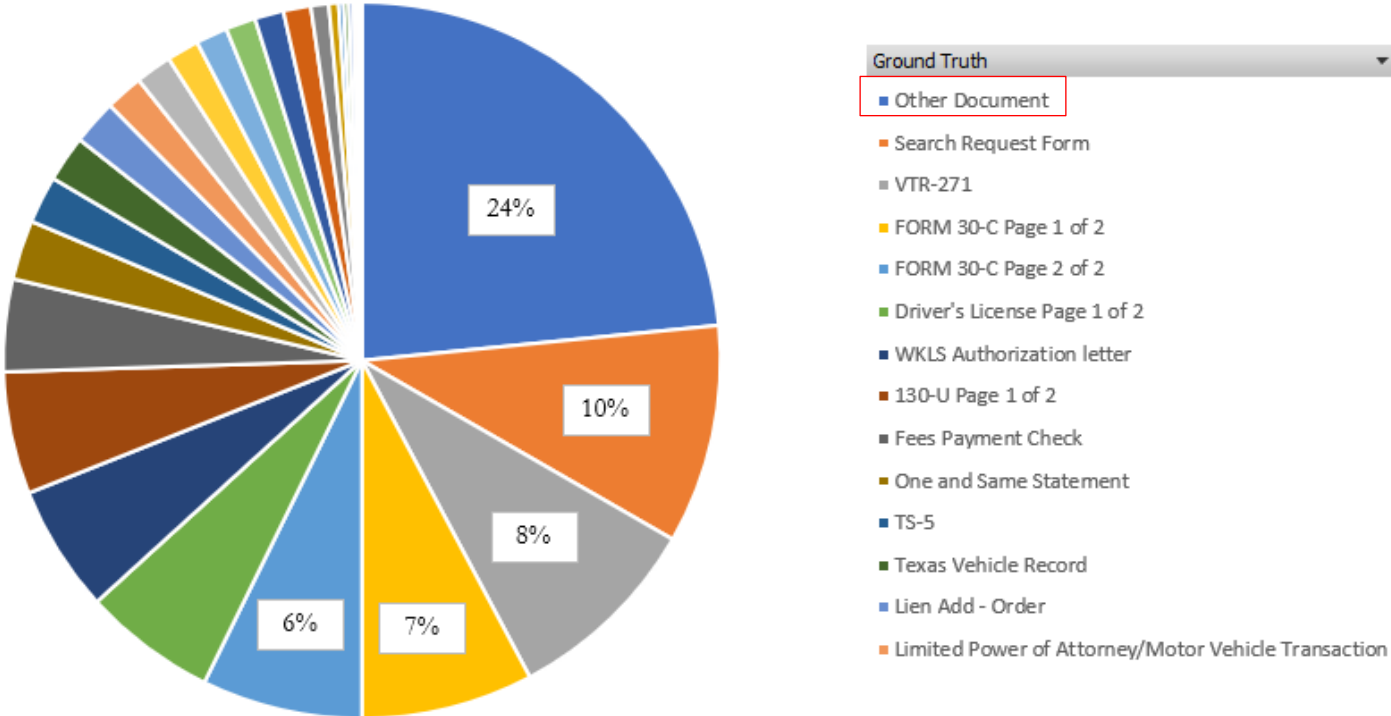
Imbalanced dataset across 120+ categories; 12k scanned pages

Training Data

12k+
Data Points

10k+
from top 31
classes

Class Proportion



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

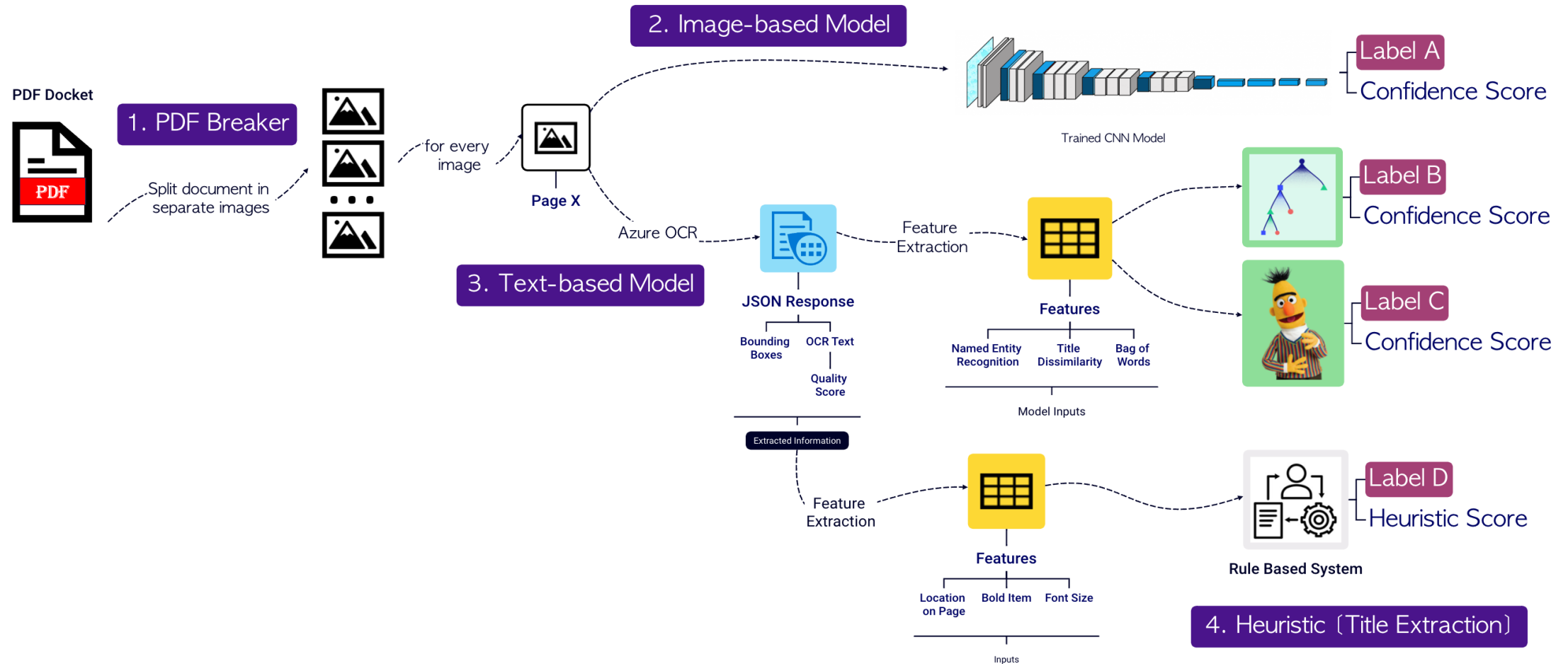


Fig: End-to-End Workflow

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

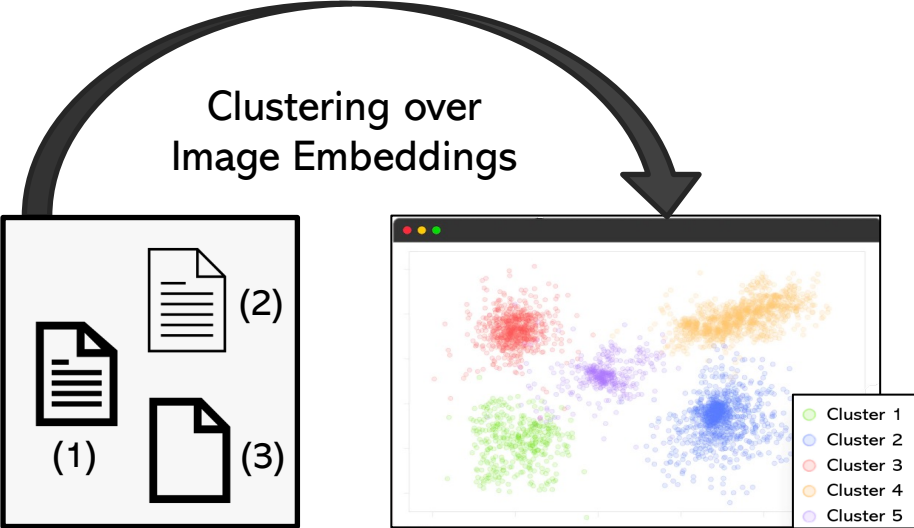


Fig: Embeddings for each image clustered based on similarity

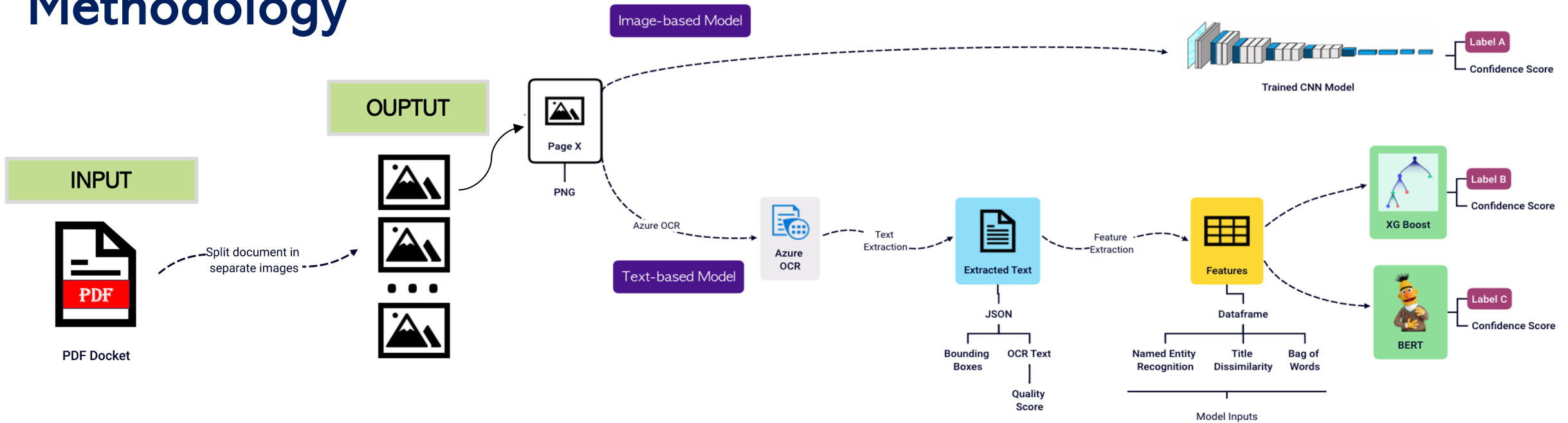
Results

88%
saving in time for manual labelling

10.6%
Mis-classified labels identified

8%
F1 Score jump!

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

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

3 Text-based Model

BERT + XGBoost running on text extracted from AzureOCR

Methodology

1 PDF Breaker API

PDF Breaker on a Streamlit dashboard

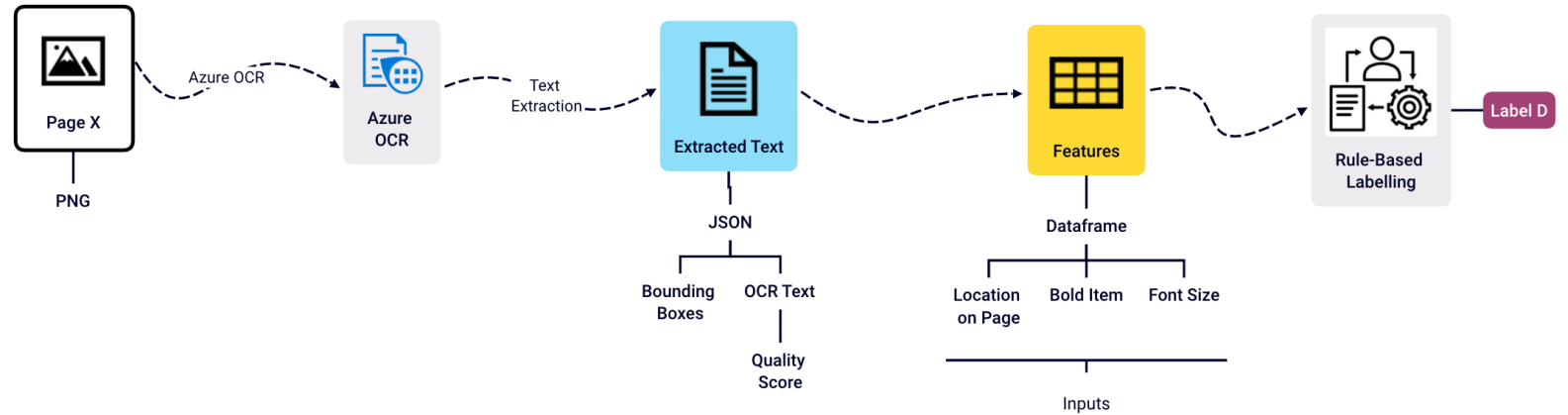
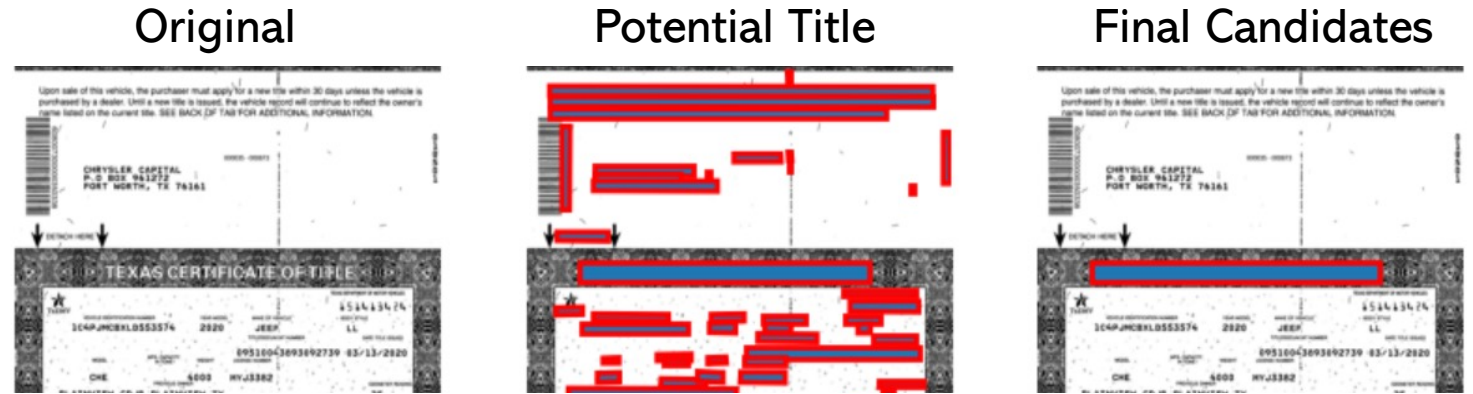
Each page is extracted as an image and saved in folders

2 Vision-based Model

CNN model fine-tuned on 12k scanned documents

3 Text-based Model

BERT + XGBoost running on text extracted from AzureOCR



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

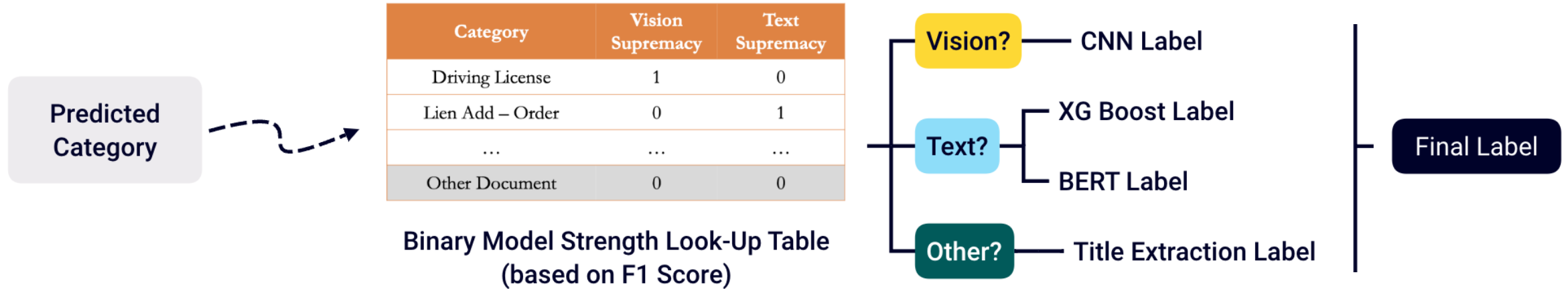


Fig: Logic workflow to combine the labels

Our Implementation

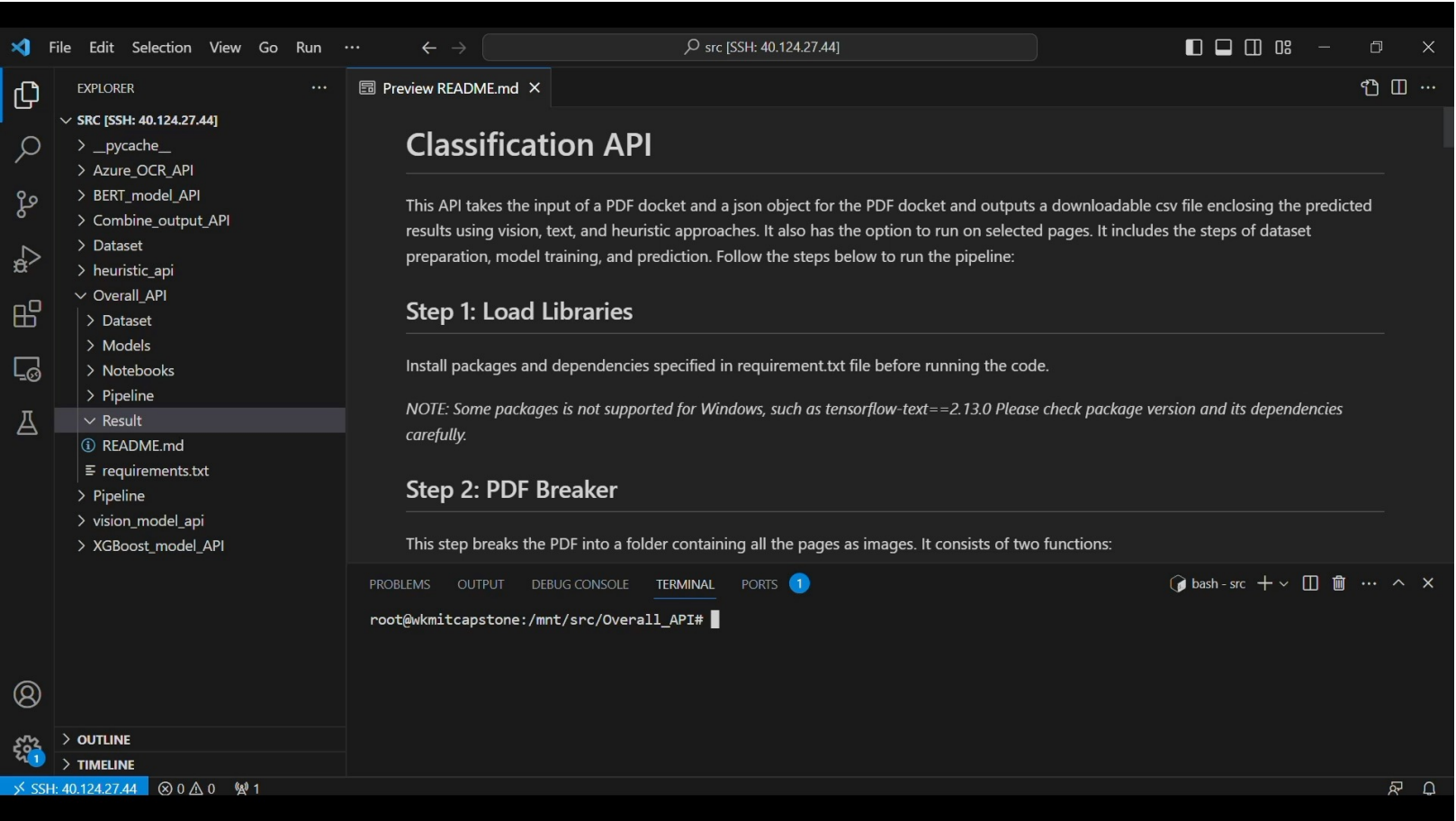
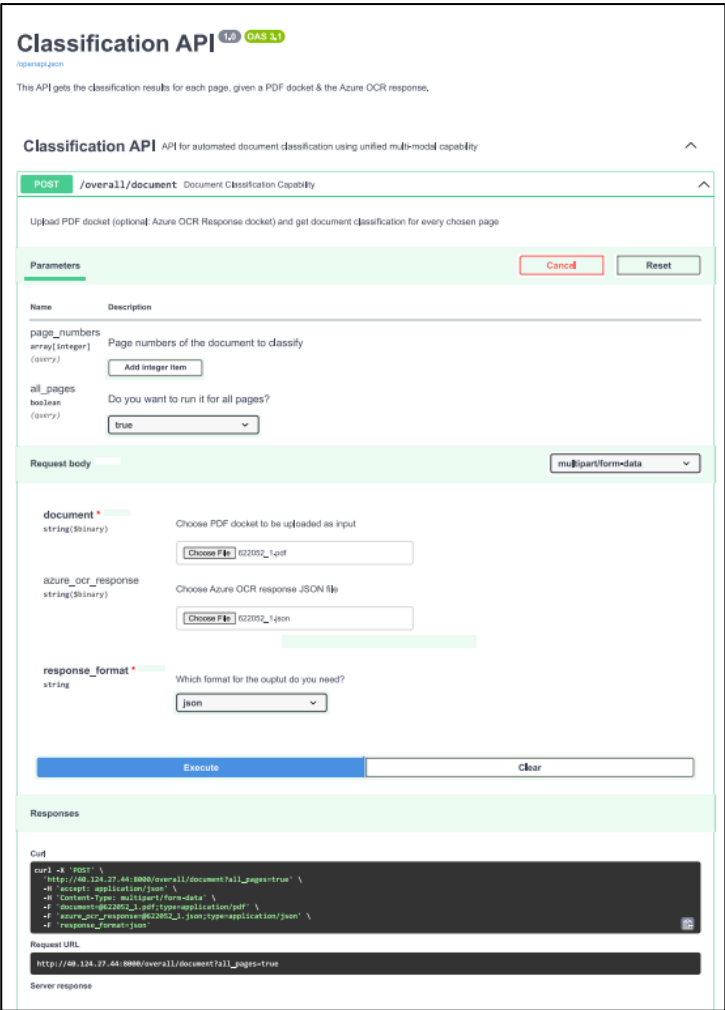
Model Supremacy for Vision and Text based models

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

Ensemble Technique

Case	Same Label	Vision Supremacy?	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

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



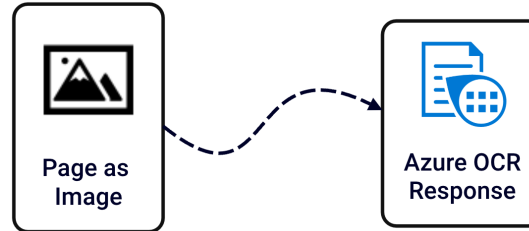
Challenger >

Champion

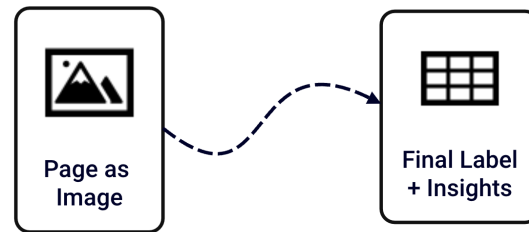
Our implementation beats
the status quo with 5%
higher F1 score

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



Business Skills



LET'S SOAR TOGETHER 2022 New Heights
New Frontiers

Business Value

Shareholders

10X growth
in business

Generalized capability →
Scalable business model

Customers

3X lower
turn-around time

Less headache →
Happy customers

Employees

10X faster
processing

Free from mundane work →
higher productivity

Market
Differentiation

Leveraging AI → journey
to be market-leaders

70% lower
rejection rate

Quick and efficient →
drives customer experience

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-INTEGRATE capability

Scalable

and reproducible



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