Document Classification Capability

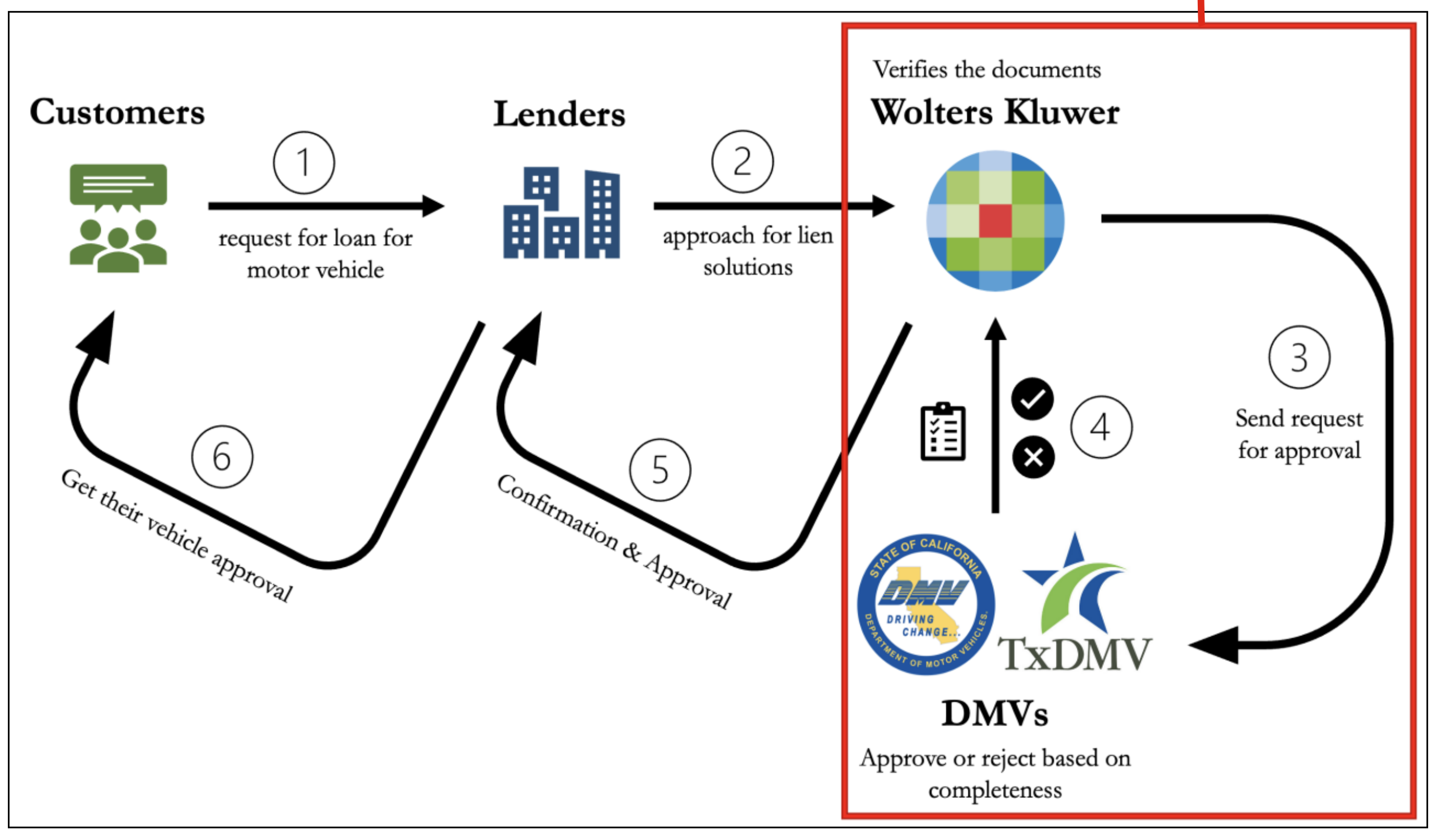
Revving up manual paperwork with Computer Vision & NLP

Faculty Advisor: Dr. Ilya Jackson
Wolters Kluwer Supervisors: Pooja Srivastava, Varun Dixit

Rachit Jain Chloe Wu
MIT Masters of Business Analytics (MBAn)

Workflow

Motor Vehicle registration process is error-prone


Customers
1
request for loan for motor vehicle
Lenders
2
approach for lien solutions
Verifies the documents
Wolters Kluwer
3
Send request for approval
4
TxDMV
DMVs
Approve or reject based on completeness
5
Confirmation & Approval
6
Get their vehicle approval

Challenges

Manual processing
→ bottleneck

Huge Volume

50k+ pages*
per day

Multiple rejections

10%
rejection rate

High processing time

10 mins per request (~20 pages)

Goal

Need for more than just rule-based systems

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

Solution

Multi-modal approach for document types inside & outside the training set

VISION + TEXT + HEURISTIC

= CV + NLP + HEURISTIC

Computer Vision | Natural Language Processing | Title Extraction

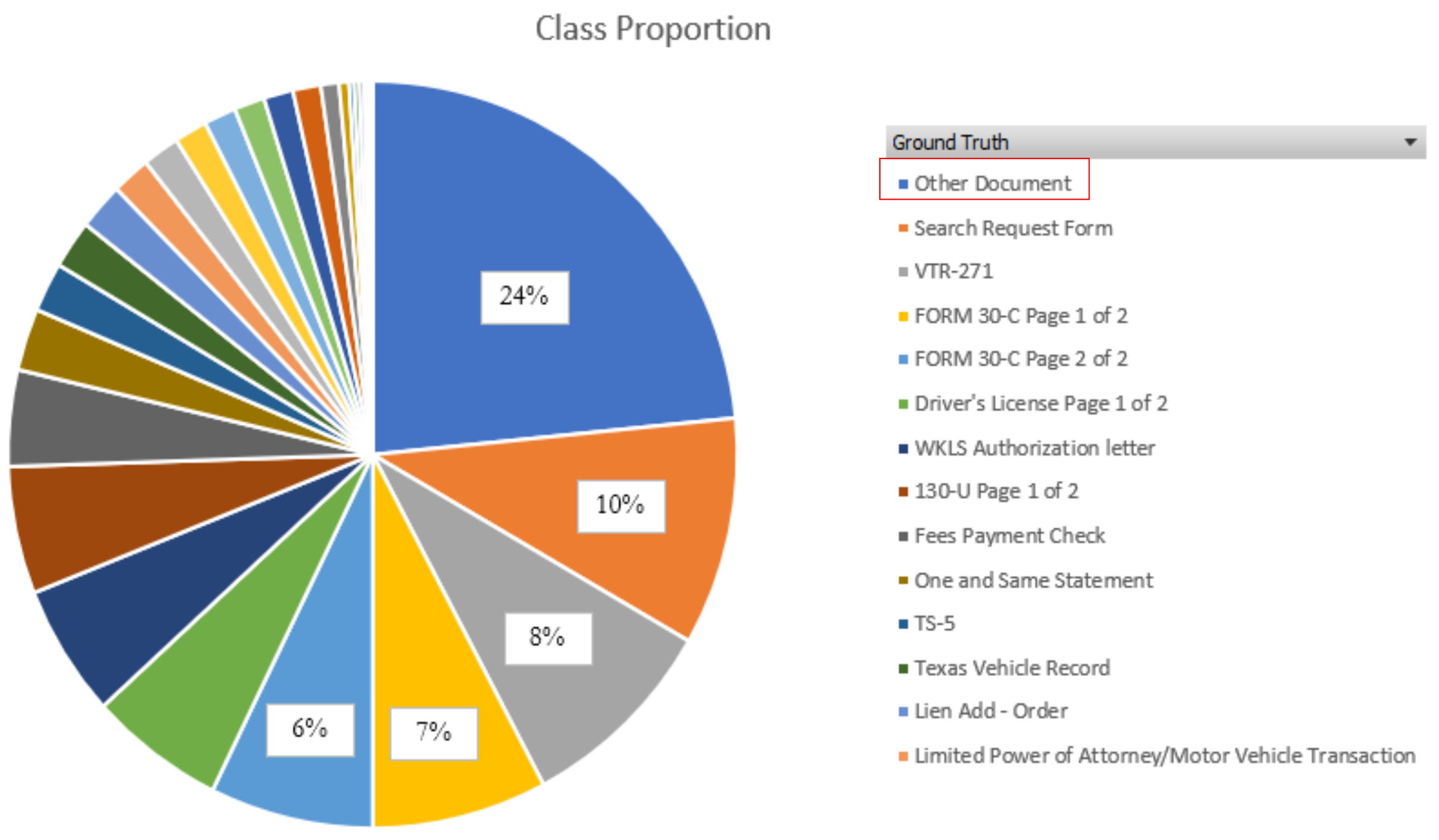
Data Exploration

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

Data Processing

Mystery behind ground truth labels - Need for relabeling

Status Quo

Current 'Ground Truth' Label = Output of Champion model

Wrong Labels → Poor Models

Solution?

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

- Image embeddings from trained vision model
- Merge clusters on common categories & create sub-clusters
- Create sub-clusters within the clusters
- Manually label documents, but smartly!

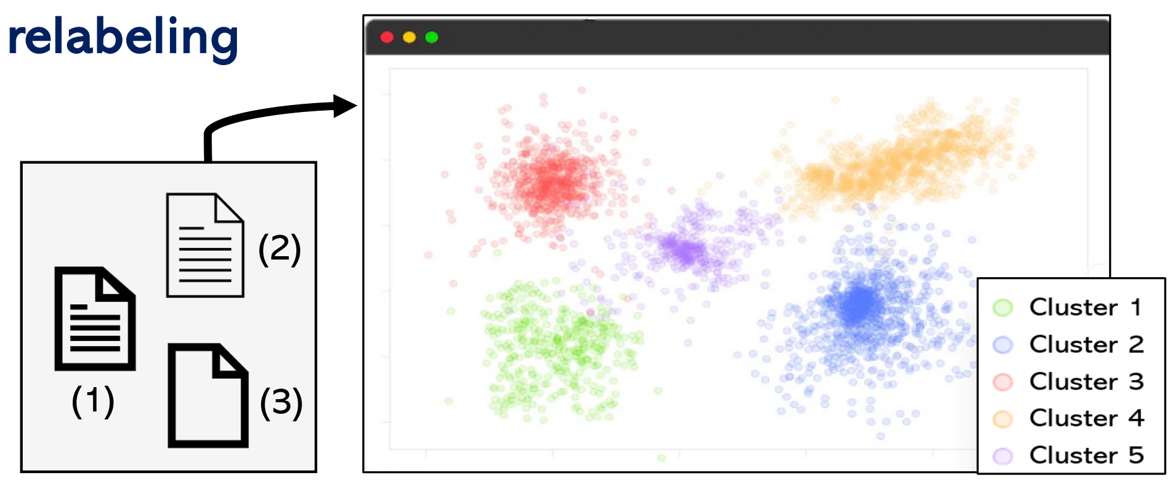


Fig: Clustering based on embedding similarity

Results

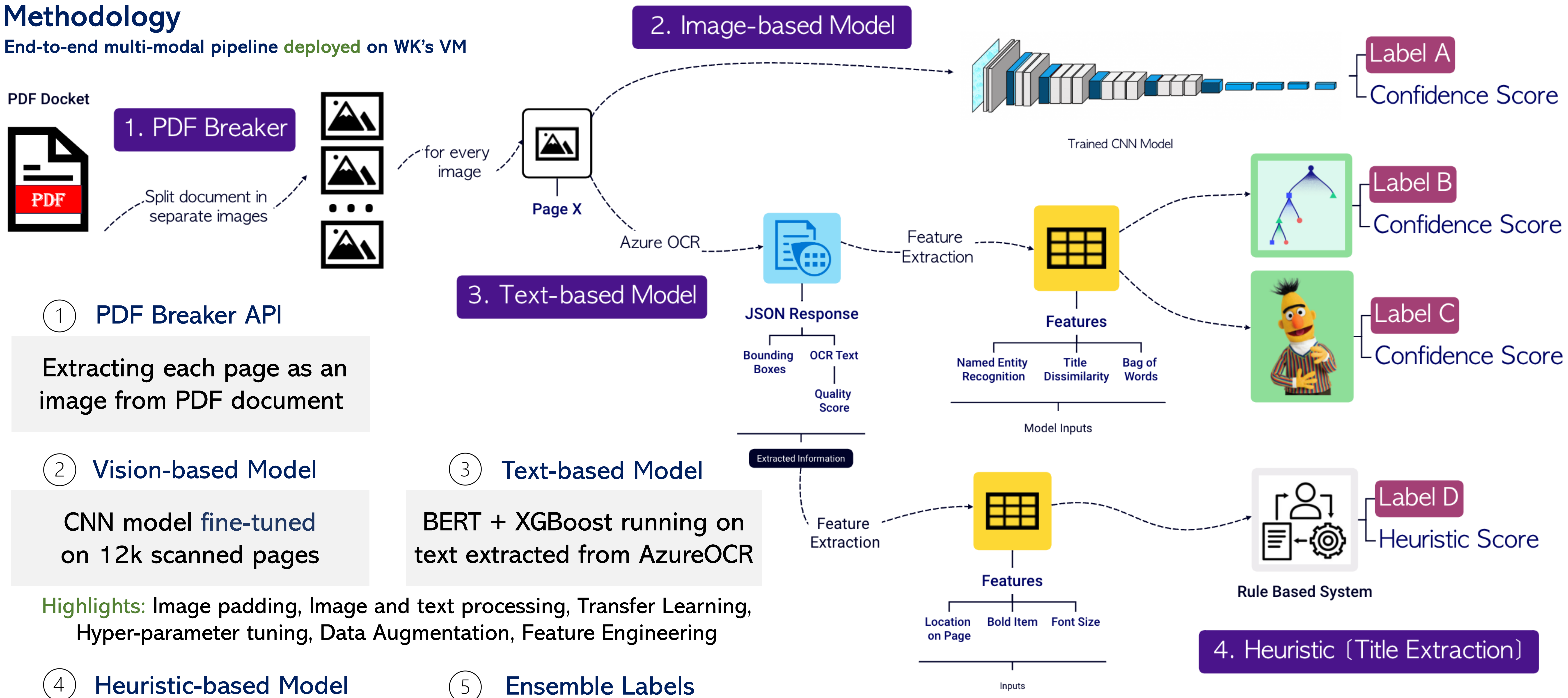
88% saving in time for manual labelling

10.6% Mis-classified labels identified

8% F1 Score jump!

Methodology

End-to-end multi-modal pipeline deployed on WK's VM



1 PDF Breaker API

Extracting each page as an image from PDF document

2 Vision-based Model

CNN model fine-tuned on 12k scanned pages

3 Text-based Model

BERT + XGBoost running on text extracted from AzureOCR

Highlights: Image padding, Image and text processing, Transfer Learning, Hyper-parameter tuning, Data Augmentation, Feature Engineering

4 Heuristic-based Model

Potential title extraction from any document type

Used bounding boxes from Azure OC, narrowed search space and engineered features

5 Ensemble Labels

Selecting 'supreme' model; if clash, choose heuristic

The final model is chosen based on the F1 score during training & rules based on heuristic title

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

Table: Ensemble technique rules

Results + Deliverables

Final model predictions with added visibility

Fully functional Classification API deployed on Wolters Kluwer's virtual machine to be productionized into their current capability, giving added flexibility & scalability!

Docket #	Page #	Final Model	Final Prediction	Prediction Confidence	Document Quality	Notes
616823_1	1	Matched	Lien Add - Order	1.00	0.699	
616823_1	2	Matched	FORM 30-C Page 1 of 2	1.00	0.766	
616823_1	3	Text	MV-50	0.49	0.640	
616823_1	4	Vision	Driver's License Page 1 of 2	0.88	0.292	
616823_1	5	Heuristic	Binder of Insurance	0.68	0.612	Manual Review

Table: Results not only have the final prediction, but chosen model, confidence score & added interpretability

0.86 F1
High Score → Better

1 API
running end-to-end

Deployed
over Wolters Kluwer's VM

Challenger >
Champion (status-quo)

Unrestricted
# of PDFs broken at once

Streamlined
workflow

Business Value

Shareholders

10X growth
with market differentiation

Customers

3X lower
turn-around time

Employees

10X faster
processing leveraging AI