OPERATIONS RESEARCH CENTER

MIT MANAGEMENT BUSINESS ANALYTICS

CAPSTONE PROJECT

Document Classification Capability

Categorization of scanned documents

MIT x Wolters Kluwer

Rachit Jain Chloe Wu

Candidates of Master of Business Analytics, MIT

Faculty Advisor: Dr. Ilya Jackson

18th August 2023

LET'S SOAR TOGETHER 2022 New Heights New Frontiers

Agenda

Introduction

Overview

Challenges

Motivation

The Process

Solution

Methodology
(Re-labelling, Modelling)

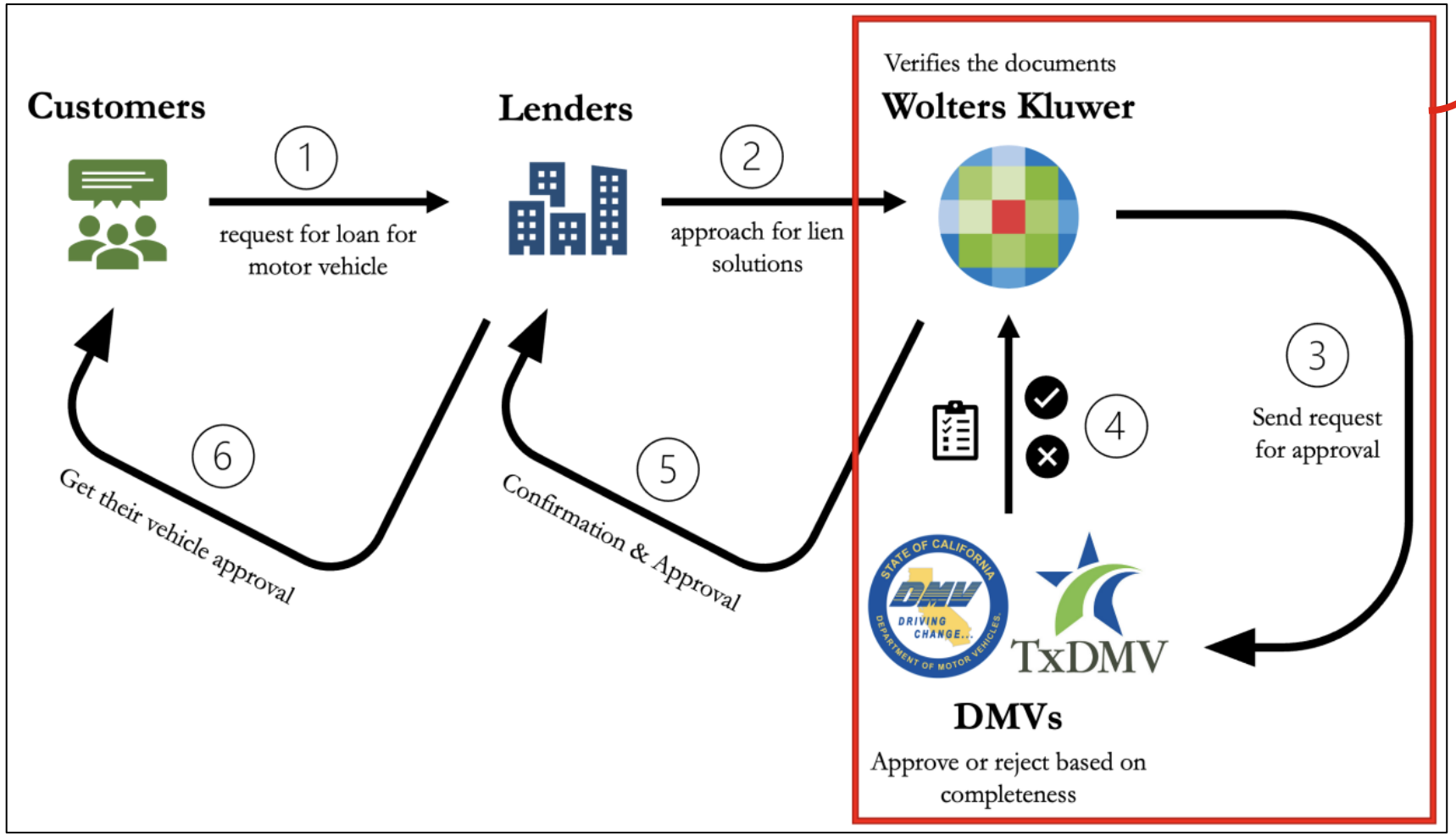
Results

Demo

Deliverables

Business Value & Impact

Overview | Motor Vehicle registration process is error-prone



Challenges At Scale

Huge Volume

100k+ pages* come in per day

Multiple rejections

10% rejections

High processing time

10 mins per request (~20 pages)

*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 

100k+ pages* come in per day

Huge Volume

10% rejections

Multiple rejections

10 mins per request (~20 pages)

High processing time

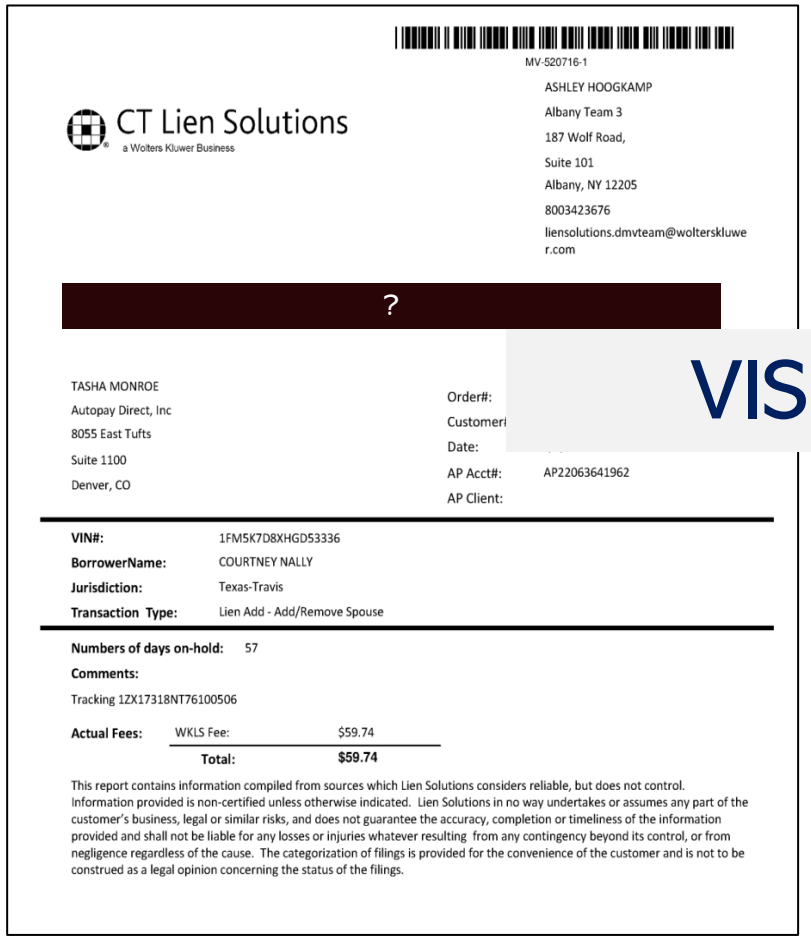
Final Goal

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



MV-520716-1

CT Lien Solutions
a Wolters Kluwer Business

ASHLEY HOOGKAMP
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#:
Customer
Date:
AP Acct#: AP22063641962
AP Client:

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

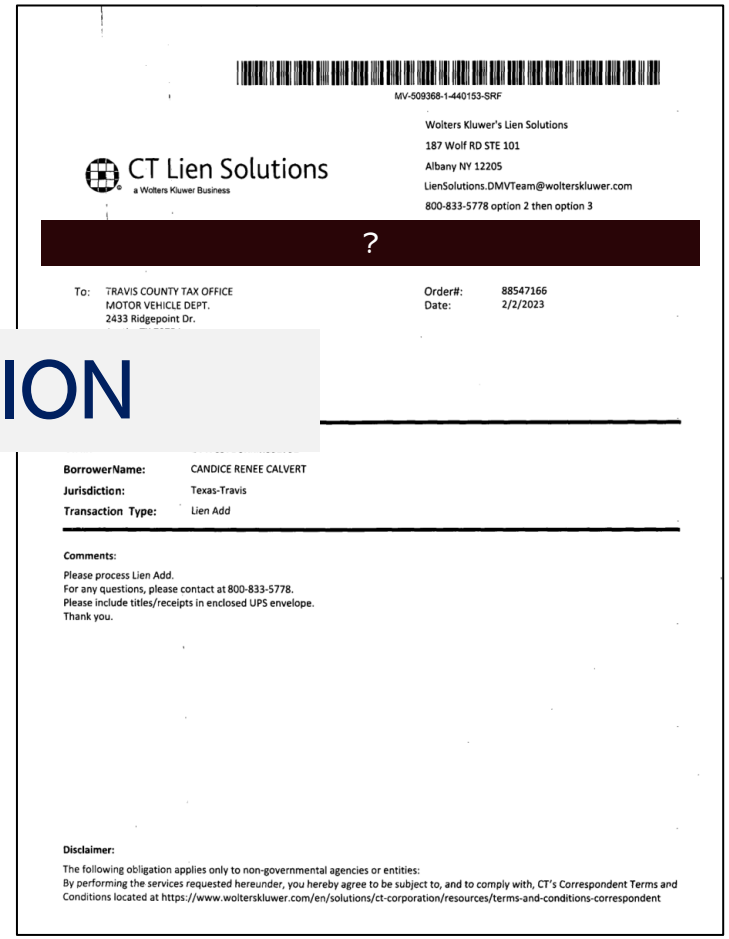
Numbers of days on-hold: 57
Comments:
Tracking 1ZX17318NT76100506

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.

VISION

(2) Search Request Form



MV-509368-1-440153-SRF

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.

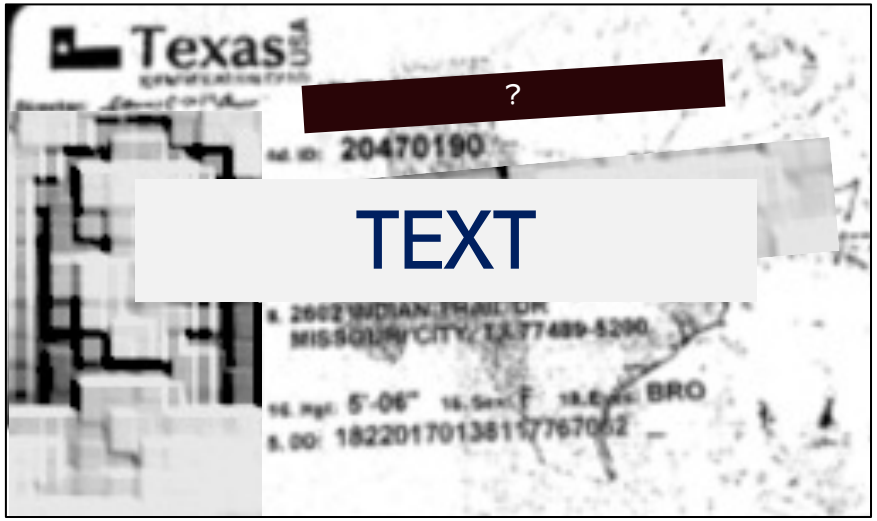
Order#: 88547166
Date: 2/2/2023

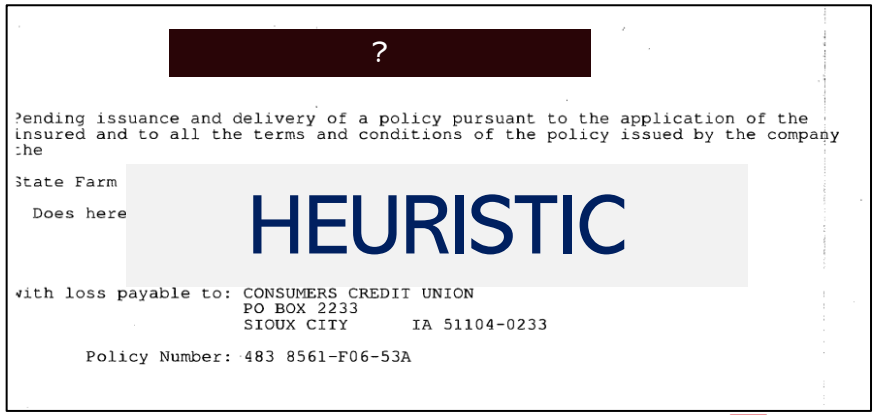
BorrowerName: CANDICE RENEE CALVERT
Jurisdiction: Texas-Travis
Transaction Type: Lien Add

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





(4) Binder of Insurance

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

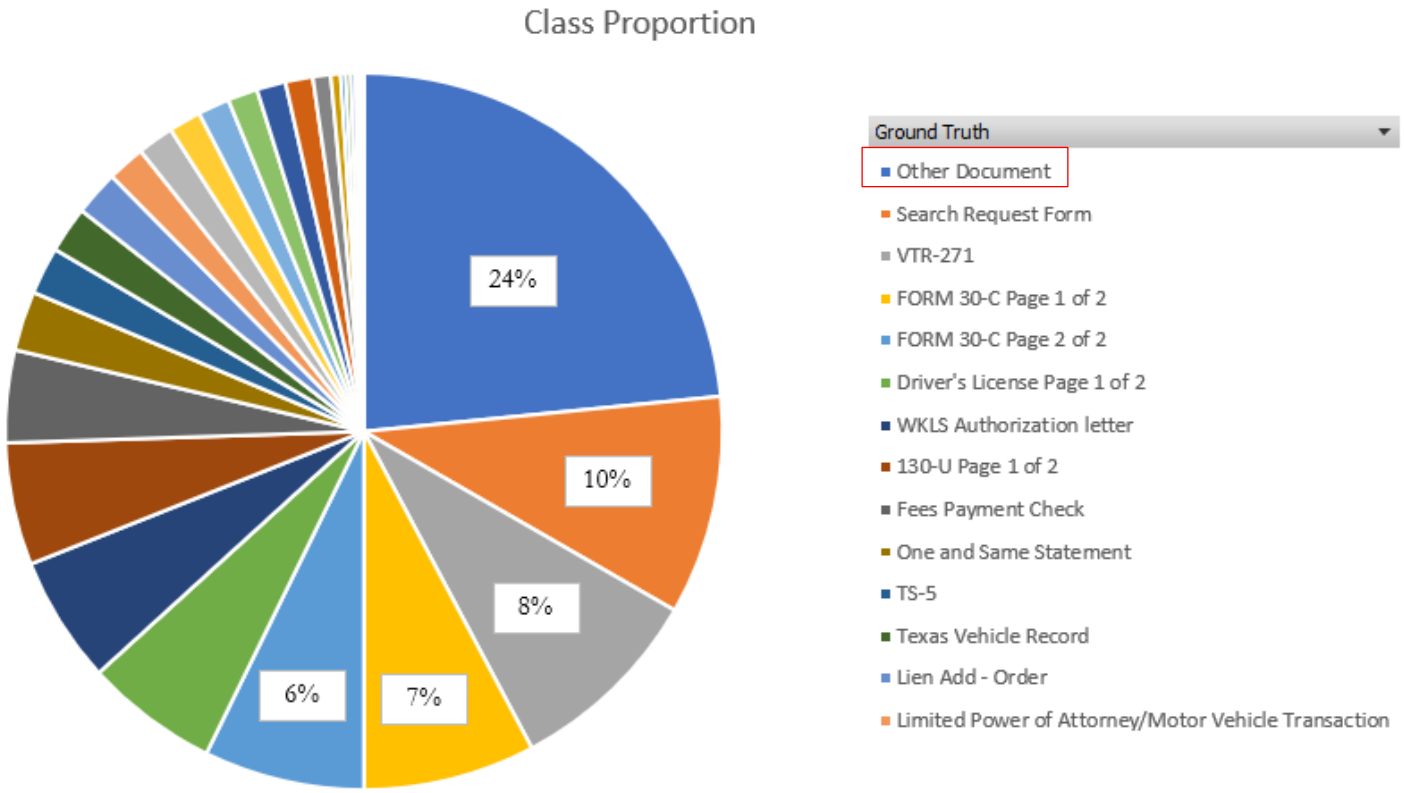
Training Data

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

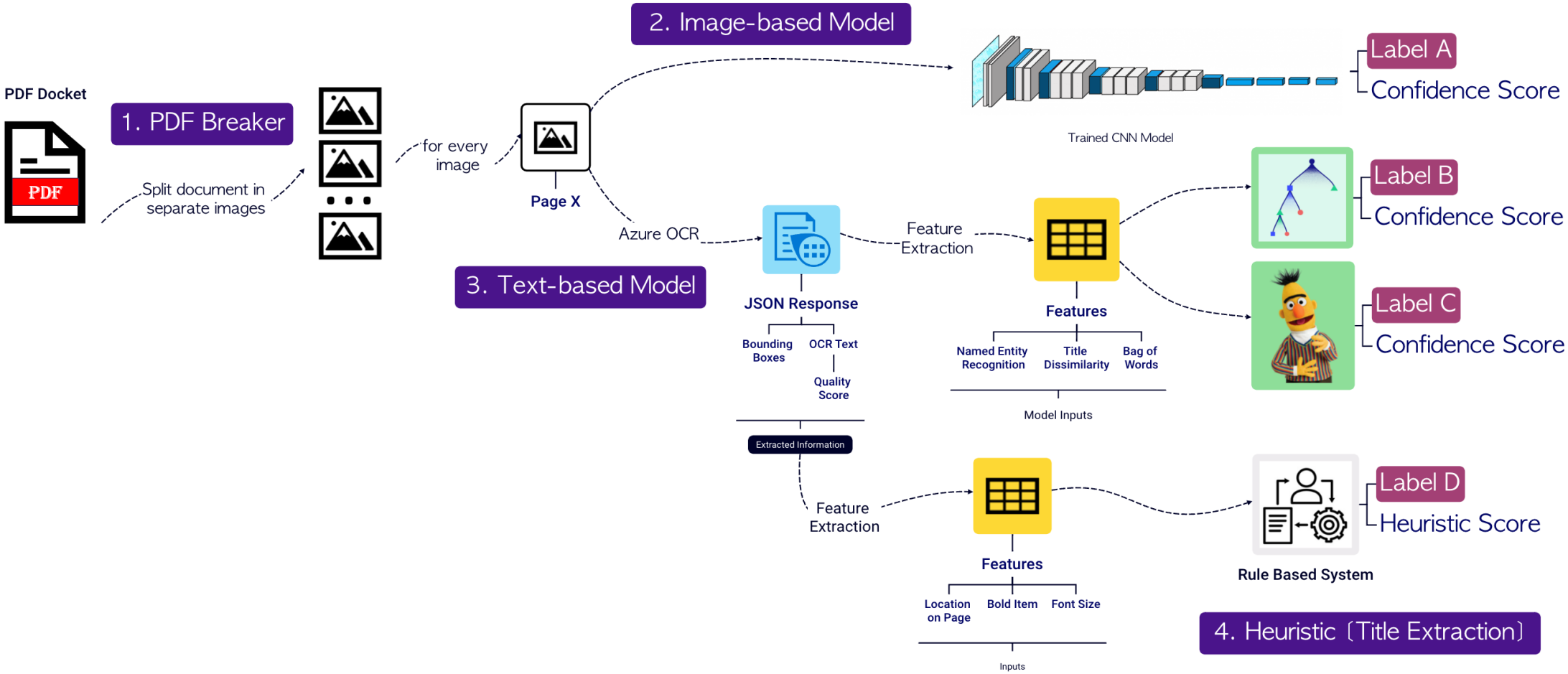


Fig: End-to-End Workflow

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

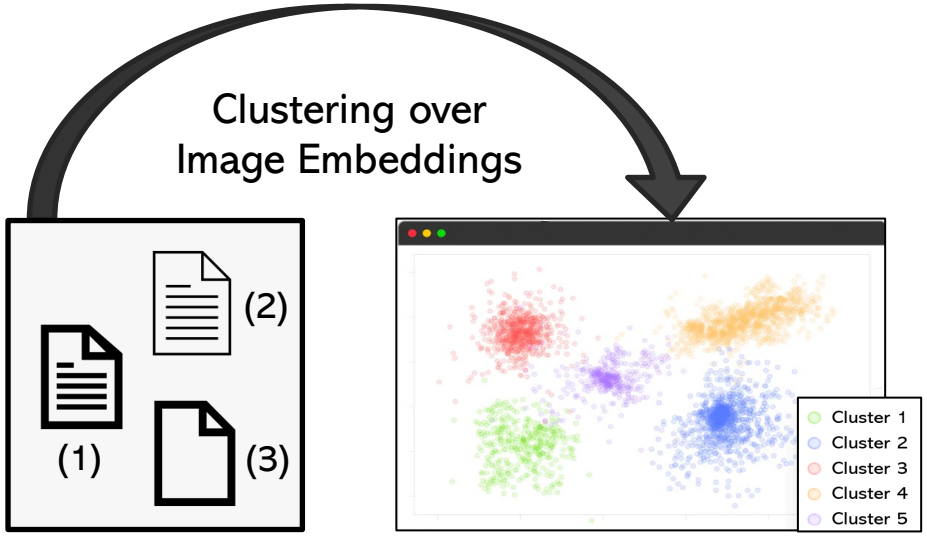


Fig: Embeddings for each image clustered based on similarity

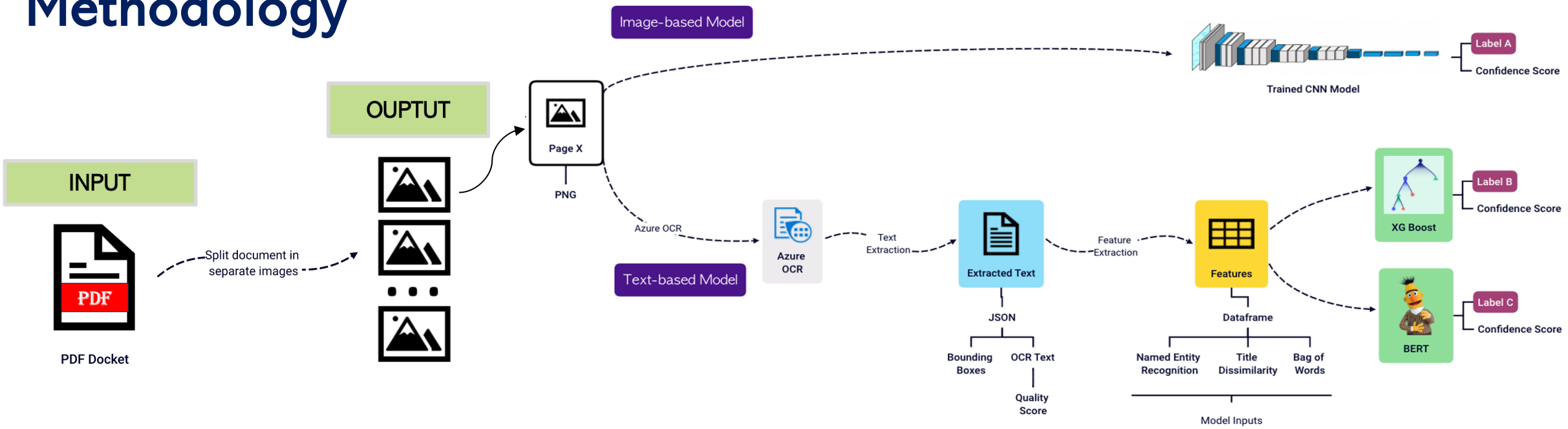
Results

88%
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

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

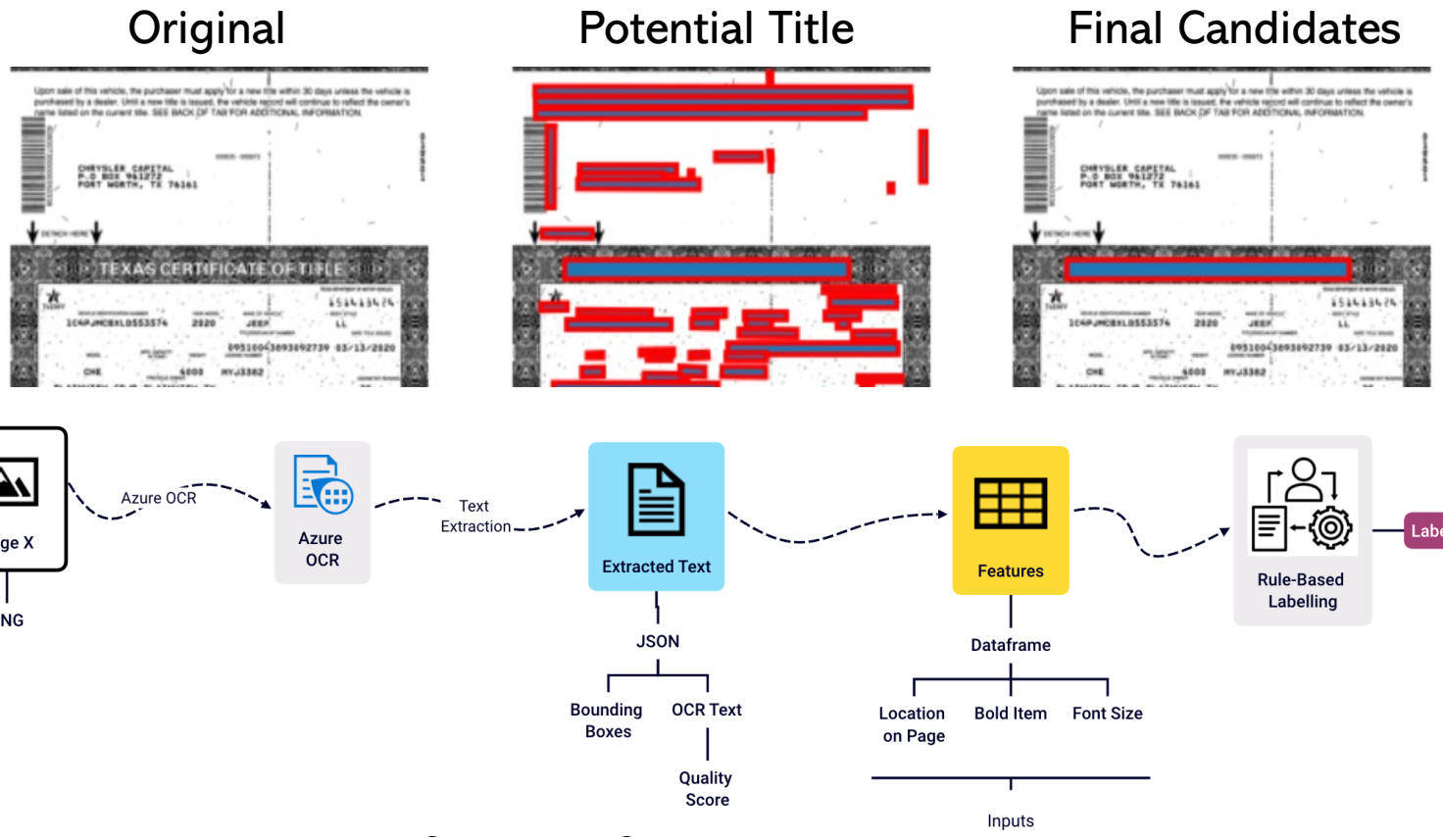
Each page is extracted as an image and saved in folders

2 Vision-based Model

CNN model fine-tuned on 12k scanned documents

3 Text-based Model

BERT + XGBoost running on text extracted from AzureOCR



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

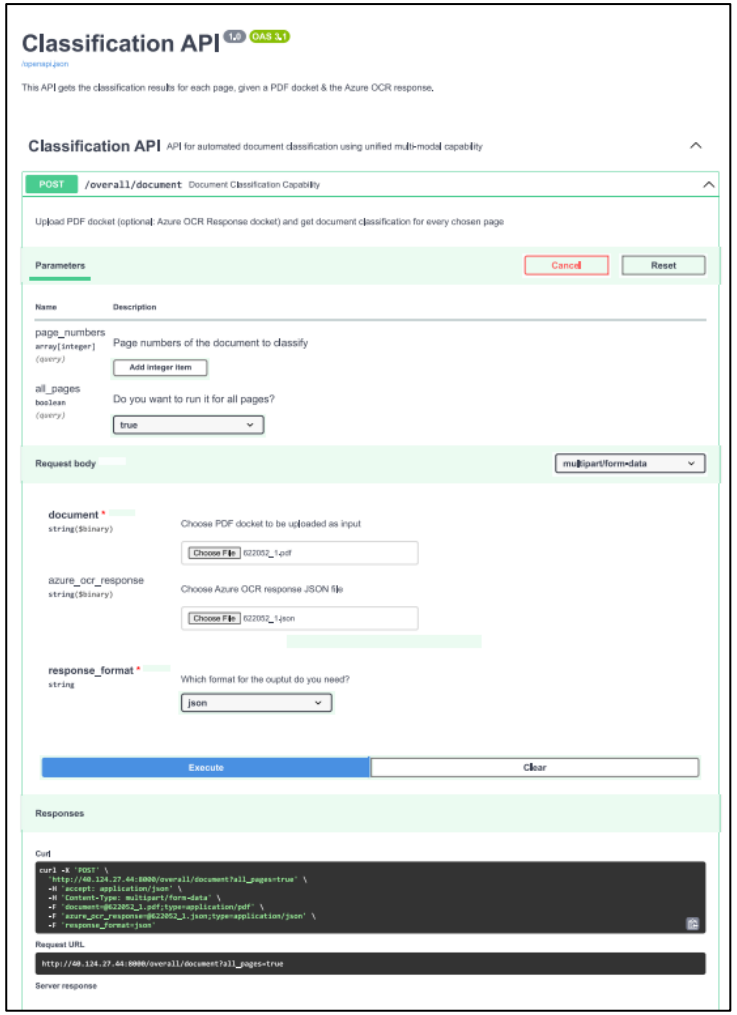
Model Supremacy from Vision and Text with Confidence Score

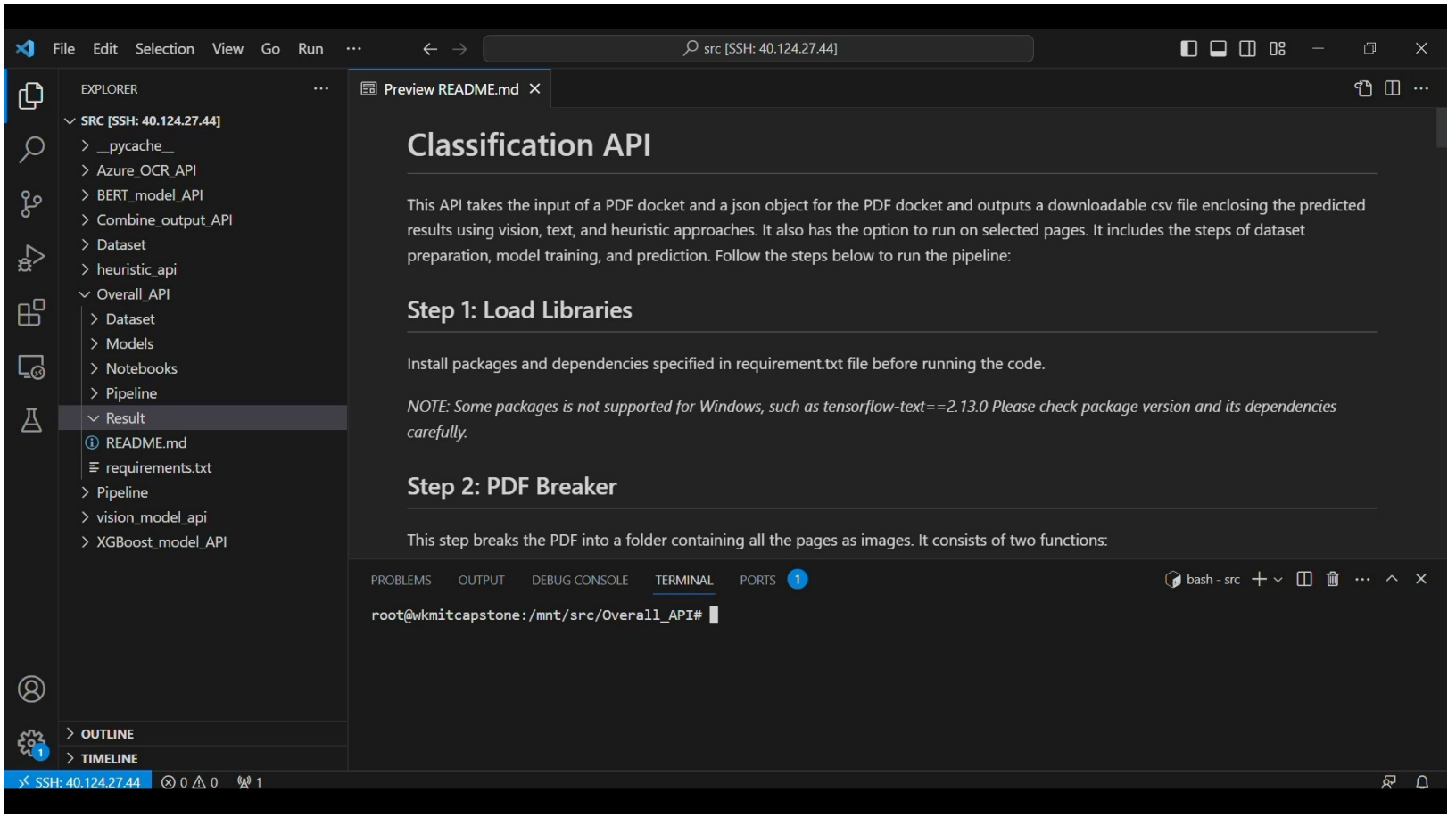
The final model is chosen based on the F1 score & rules based on heuristic title

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

Deliverables | Scalable model pipeline deployed over WK Cloud

Result Summary

0.86
F1 Score

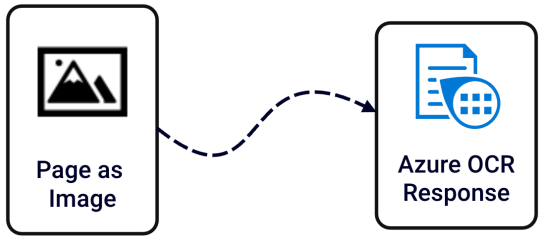
Over 31 document types over 2.1k highly noisy test dataset

Challenger > Champion

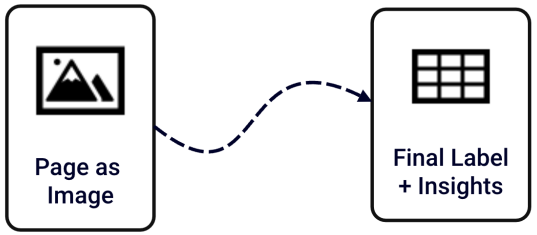
Our implementation beats the status quo with 5% higher F1 score

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

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-INTEGERATE capability

Scalable

and reproducible

OPERATIONS RESEARCH CENTER

MIT MANAGEMENT BUSINESS ANALYTICS

Thank You!

Rachit Jain Chloe Wu
Candidates of Master of Business Analytics, MIT

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

~ T.S. Eliot

LET'S SOAR TOGETHER 2022 New Heights New Frontiers

Appendix

The mystery behind ground truth labels: need for relabeling

Status Quo

Output of Champion model = 'Ground' Truth Label

Challenge

Wrong Labels → Poor Models
The models learn to predict incorrectly

Realization

Data-Centric approach 'Confident Learning'

Solution?

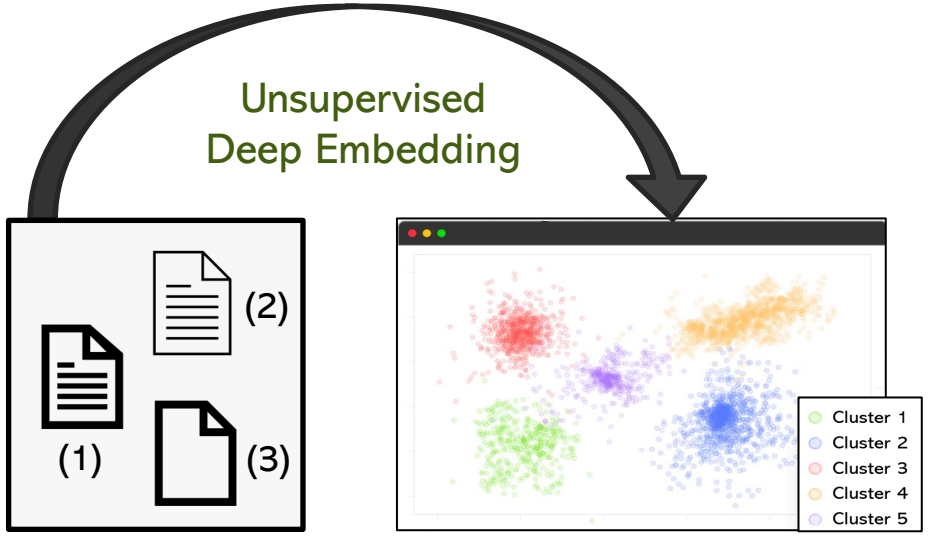
Manually label all 12k training images, but smartly!!

Category	%age Mislabelled	Count of Documents
130-U Page 1 of 2	4.90%	572
Check Endorsement	11.68%	137
Driver's License Page 1 of 2	15.90%	547
Fees Payment Check	1.68%	417
FORM 30-C Page 1 of 2	1.15%	786
FORM 30-C Page 2 of 2	2.24%	758
Lien Add - Order	13.85%	231
Limited Power of Attorney/Motor Vehicle Transaction	1.68%	179
MV-50	5.93%	135
Odometer Disclosure Statement	56.92%	130
One and Same Statement	29.41%	374
Other Document	16.97%	2841
Search Request Form	0.19%	1045
Texas Liability Insurance Card	59.57%	235
Texas Vehicle Record	8.71%	241
TS-5	0.00%	226
TS-8	18.93%	169
VTR-271	2.00%	948
WKLS Authorization letter	5.60%	607
Grand Total	**10.60%**	**10578**

The mystery behind ground truth labels: need for relabeling

Our Solution

Smarter Labelling [Unsupervised Clustering on Deep Embeddings]



Intuition

'Can you spot the (major visual) difference?' is easier while scrolling through 100 similarly formatted images!

Solution Process

- Get image embeddings from trained vision model
- Cluster based on embedding similarity
- Assign the most prominent category
- Merge clusters on common categories
- Create sub-clusters within the clusters
- Smarter manual labelling

Results

84hrs → 10hrs

Assuming 30s to classify a page

10.6%

Mis-classified labels corrected

~8%

F1 Score jump!

PDF Breaker extracts every page from the docket as an image

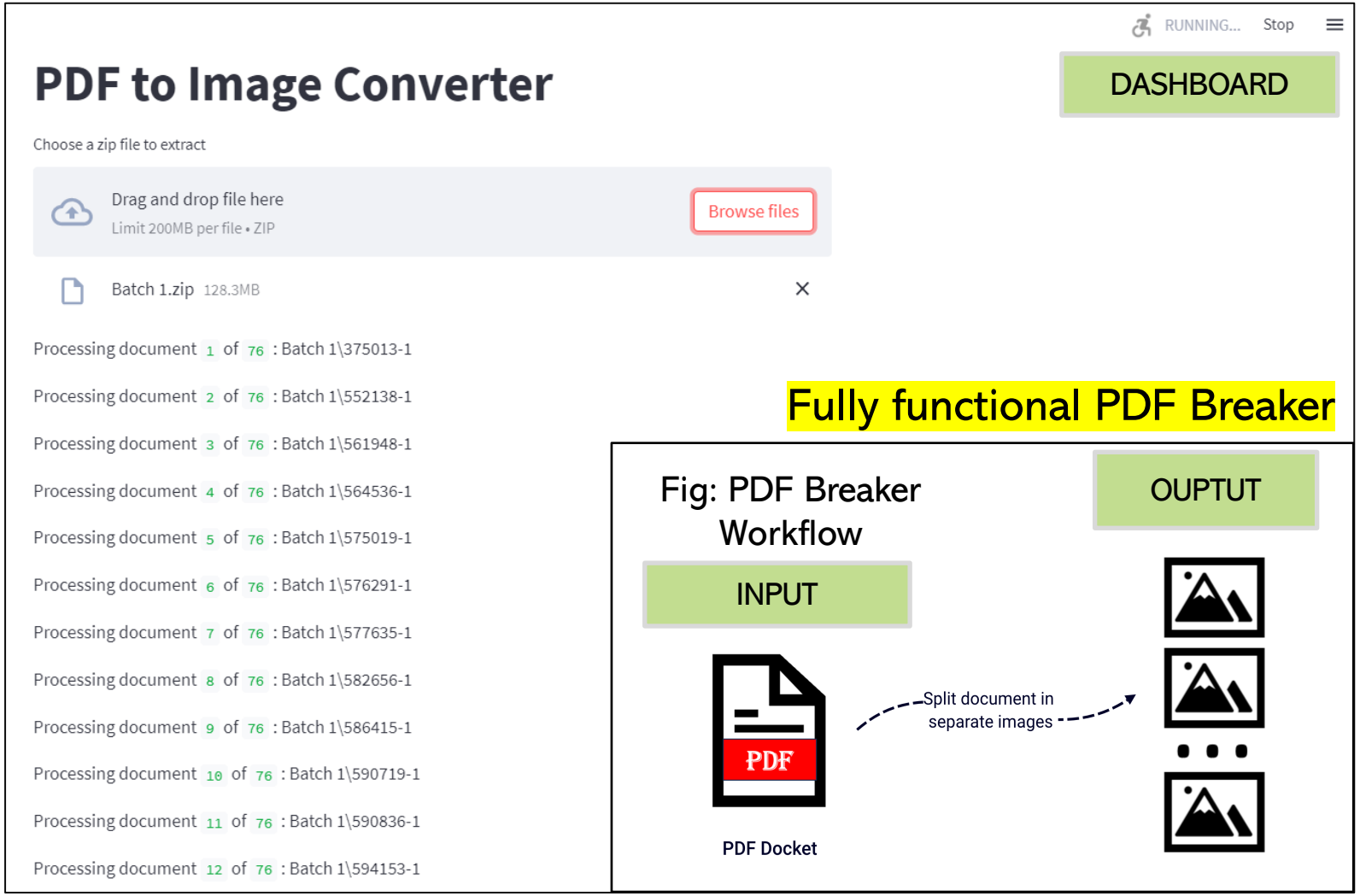
Our Implementation

PDF Breaker on a user-friendly Streamlit dashboard

Input: ZIP folder with any number of PDFs within county folders
Output: ZIP folder with all images stored in respective folders

Process:

- Every docket contains 'n' number of page
- Each page is extracted as an image and saved in respective folder
- The models use these images as inputs



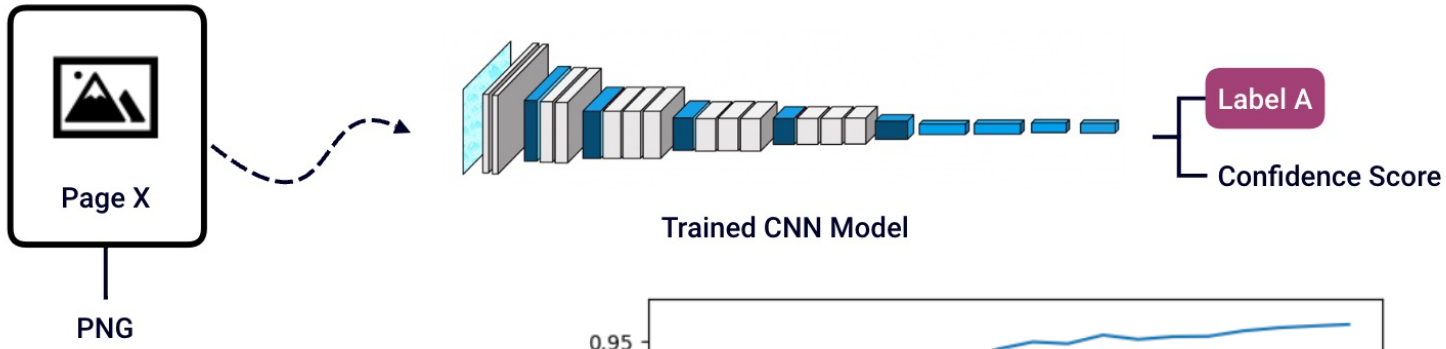
Vision-based model harnesses the power of visual formatting

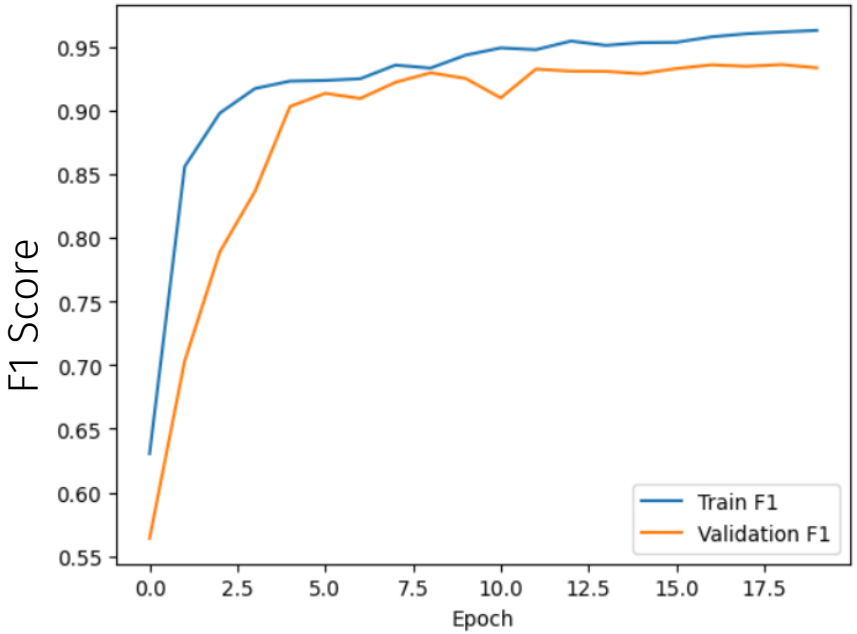
Our Implementation

ResNet50 Model fine-tuned on 12k scanned documents

Input: Image (document)
Output: Category label and confidence score

Fig: Image-based model workflow





Highlights:

- Image padding while maintaining resolution
- Fine-Tuned instead of simply using pre-trained model
- Image processing (adaptive document image binarization, contrast reduction, noise removal…)
- Hyper-parameter tuning (contributed to ~2%+)
- Data Augmentation for training + validation (further 3%+)
- Pipeline ready!

Text-based model harnesses the power of document content

Our Implementation

BERT + XGBoost running on text extracted from AzureOCR

Input: Image (document)
Output: Category label and confidence score (for both models)

Highlights for BERT:
- Text Preprocessing to remove noise
- Transfer Learning with model fine-tuning
- Hyper-parameter Tuning

Highlights for XGBoost:
- NLP Feature Engineering (Bag of Words, Named Entity Recognition, Title Dissimilarity)
- Hyper-parameter Tuning
- Pipeline ready

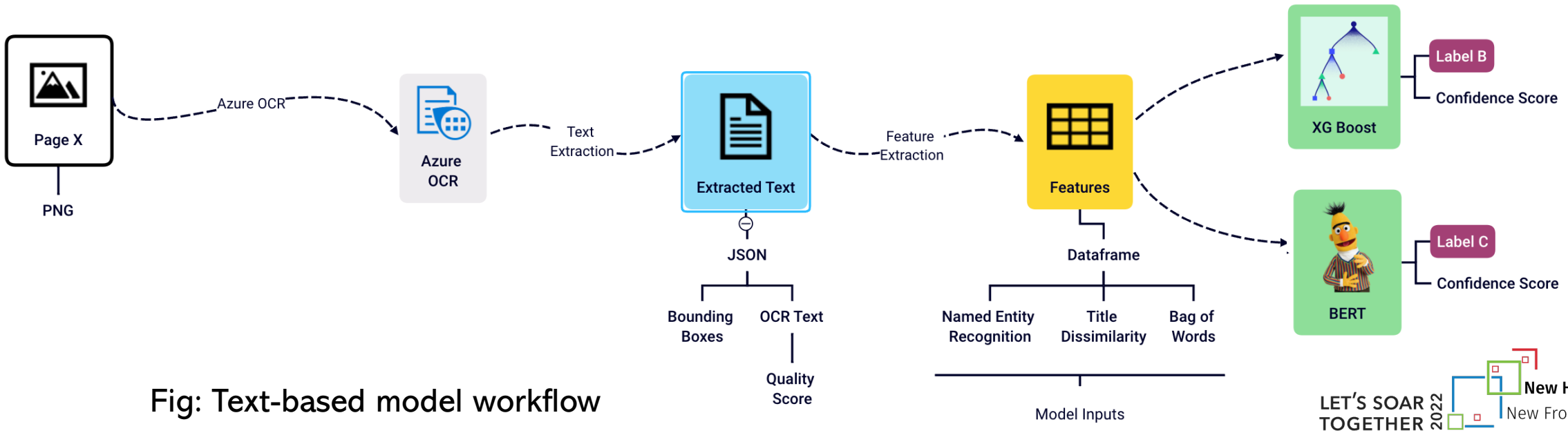


Fig: Text-based model workflow

Heuristic-based used in case of uncertain documents (Capability Design)

Our Implementation

Title Extraction by heuristically narrowing on 'potential' title of the document

Input: Image (document)
Output: Category label and heuristic title score

Intuition: A heuristic title is informative than predicting 'Other Document'

Highlights:

- Used all bounding boxes from Azure OCR as potential titles; narrowed search space
- Engineered features like location on page, bold, font size, lone item, etc.

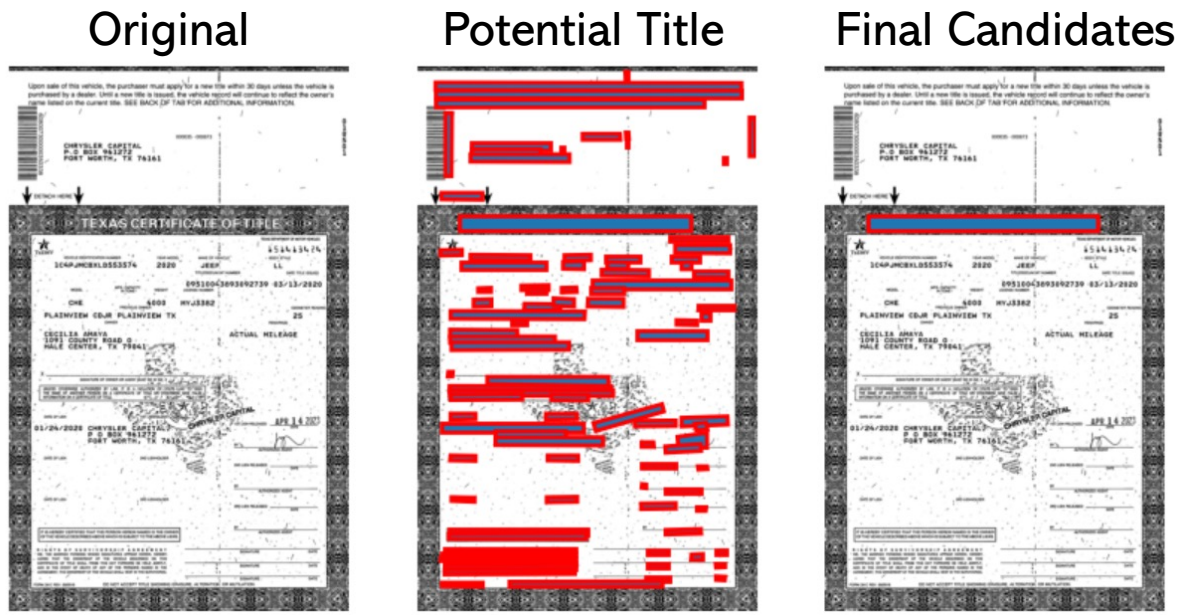


Fig: Heuristic-based model workflow

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)

Ensemble Technique (Expanded)

Case	Vision Label	Text Label	Vision Supremacy	Text Supremacy	Final Model	Final Label	Comments
Case 1: Same output	A	A	1	0	Same	A	
Case 2: Conflicted output	A	B	1	1	NA	NA	Manual Review
Case 3: One model is supreme	A	B	1	0	Vision	A	
	A	B	0	1	Text	B	
Case 4: Both "Other"	Other	Other	0	0	Heuristic	Heuristic Title	Manual Review
Case 5: One "Other" and heuristic title contains another model's output	Other	B	0	0	Text	(Heuristic ∩ B) = B	
	A	Other	0	0	Vision	(Heuristic ∩ A) = A	
Case 6: One "Other" and heuristic does NOT contain another model's output	Other	B	0	0	Heuristic	Heuristic Title	Manual Review
	A	Other	0	0	Heuristic	Heuristic Title	Manual Review
Case 7: Case 7: Both non-"Other" and None supreme	A	B	0	0	Heuristic	Heuristic Title	Manual Review

Result | Final model predictions with added visibility

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

Fig: Result Example

Our Implementation

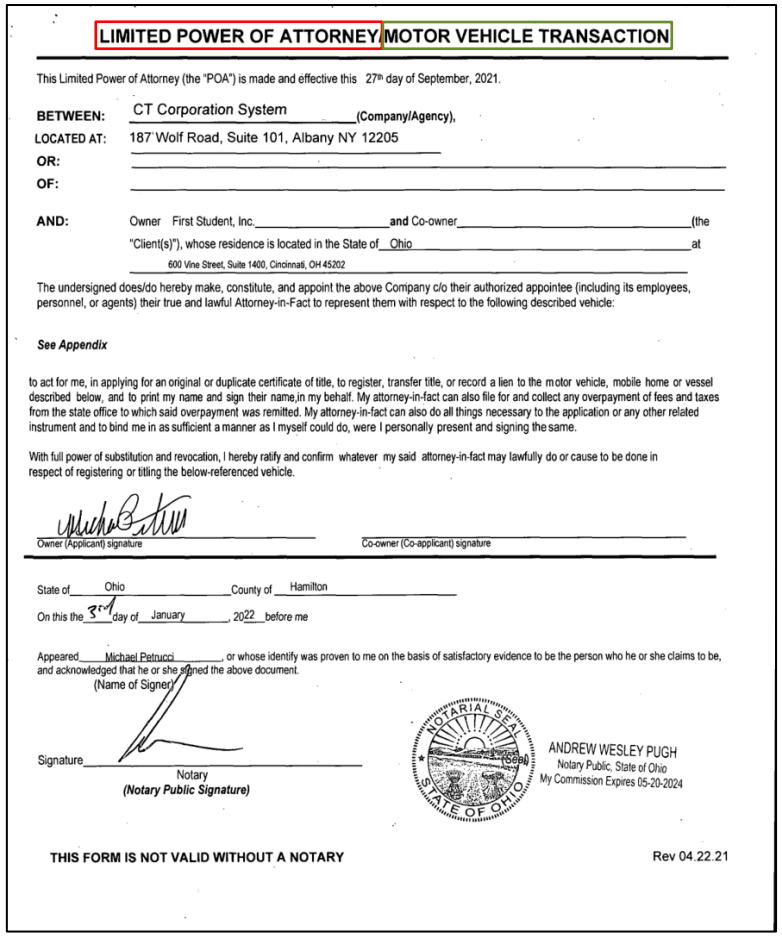
'Classification API' runs the entire pipeline on Virtual Machine

Highlights for Results:

- The results show confidence score for the chosen model
- OCR score give a proxy to the document quality
- Manual review in the 'notes' gives user pin-pointed areas to spend more time labelling

A glimpse of Champion vs Challenger capabilities

(1) Limited Power of Attorney/Motor Vehicle Transaction



LIMITED POWER OF ATTORNEY MOTOR VEHICLE TRANSACTION

This Limited Power of Attorney (the "POA") is made and effective this 27th day of September, 2021.

BETWEEN: CT Corporation System (Company/Agency),
LOCATED AT: 187 Wolf Road, Suite 101, Albany NY 12205
OR:
OF:

AND: Owner First Student, Inc. and Co-owner (the "Client(s)"), whose residence is located in the State of Ohio at 600 Vine Street, Suite 1400, Cincinnati, OH 45202

The undersigned does/do hereby make, constitute, and appoint the above Company c/o their authorized appointee (including its employees, personnel, or agents) their true and lawful Attorney-in-Fact to represent them with respect to the following described vehicle:

See Appendix

to act for me, in applying for an original or duplicate certificate of title, to register, transfer title, or record a lien to the motor vehicle, mobile home or vessel described below, and to print my name and sign their name,in my behalf. My attorney-in-fact can also file for and collect any overpayment of fees and taxes from the state office to which said overpayment was remitted. My attorney-in-fact can also do all things necessary to the application or any other related instrument and to bind me in as sufficient a manner as I myself could do, were I personally present and signing the same.

With full power of substitution and revocation, I hereby ratify and confirm whatever my said attorney-in-fact may lawfully do or cause to be done in respect of registering or titling the below-referenced vehicle.

Owner (Applicant) signature — Co-owner (Co-applicant) signature

State of Ohio County of Hamilton

On this the 31st day of January, 2022 before me

Appeared Michael Petrucci, or whose identify was proven to me on the basis of satisfactory evidence to be the person who he or she claims to be, and acknowledged that he or she signed the above document.
(Name of Signer)

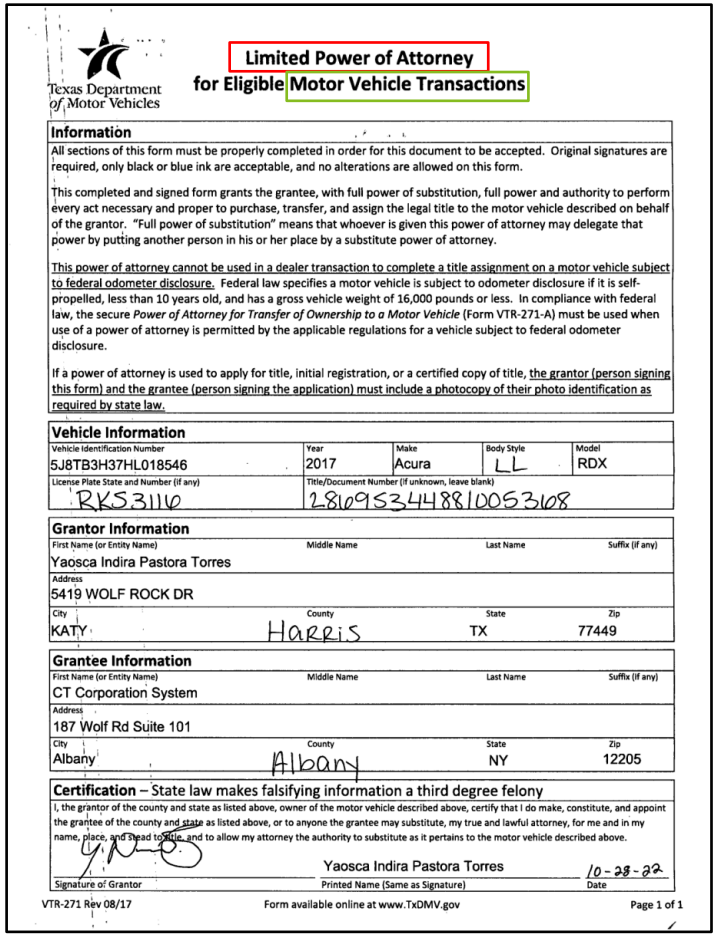
Signature
Notary
(Notary Public Signature)

NOTARIAL SEAL STATE OF OHIO
ANDREW WESLEY PUGH
Notary Public, State of Ohio
My Commission Expires 05-20-2024

THIS FORM IS NOT VALID WITHOUT A NOTARY

Rev 04.22.21

(2) VTR-271



Texas Department of Motor Vehicles

Limited Power of Attorney for Eligible Motor Vehicle Transactions

Information

All sections of this form must be properly completed in order for this document to be accepted. Original signatures are required, only black or blue ink are acceptable, and no alterations are allowed on this form.

This completed and signed form grants the grantee, with full power of substitution, full power and authority to perform every act necessary and proper to purchase, transfer, and assign the legal title to the motor vehicle described on behalf of the grantor. "Full power of substitution" means that whoever is given this power of attorney may delegate that power by putting another person in his or her place by a substitute power of attorney.

This power of attorney cannot be used in a dealer transaction to complete a title assignment on a motor vehicle subject to federal odometer disclosure. Federal law specifies a motor vehicle is subject to odometer disclosure if it is self-propelled, less than 10 years old, and has a gross vehicle weight of 16,000 pounds or less. In compliance with federal law, the secure *Power of Attorney for Transfer of Ownership to a Motor Vehicle* (Form VTR-271-A) must be used when use of a power of attorney is permitted by the applicable regulations for a vehicle subject to federal odometer disclosure.

If a power of attorney is used to apply for title, initial registration, or a certified copy of title, the grantor (person signing this form) and the grantee (person signing the application) must include a photocopy of their photo identification as required by state law.

Vehicle Information

Vehicle Identification Number	Year	Make	Body Style	Model
5J8TB3H37HL018546	2017	Acura	LL	RDX

License Plate State and Number (if any): RKS3116 — Title/Document Number (if unknown, leave blank): 28109534488100S3108

Grantor Information

First Name (or Entity Name)	Middle Name	Last Name	Suffix (if any)
Yaosca Indira Pastora Torres			

Address: 5419 WOLF ROCK DR

City	County	State	Zip
KATY	Harris	TX	77449

Grantee Information

First Name (or Entity Name)	Middle Name	Last Name	Suffix (if any)
CT Corporation System			

Address: 187 Wolf Rd Suite 101

City	County	State	Zip
Albany	Albany	NY	12205

Certification – State law makes falsifying information a third degree felony

I, the grantor of the county and state as listed above, owner of the motor vehicle described above, certify that I do make, constitute, and appoint the grantee of the county and state as listed above, or to anyone the grantee may substitute, my true and lawful attorney, for me and in my name, place, and stead to title, and to allow my attorney the authority to substitute as it pertains to the motor vehicle described above.

Signature of Grantor — Yaosca Indira Pastora Torres (Printed Name (Same as Signature)) — 10-28-22 (Date)

VTR-271 Rev 08/17 — Form available online at www.TxDMV.gov — Page 1 of 1

Champion Model

Keyword matching + extensive rule-based system

Both documents have similar keywords
Champion: (1) VTR-271 (2) VTR-271

Challenger Model

Multi-Modal Approach (Vision + Text)

Formats are different. Text is different.
(1)'s content is like "Limited Power…"
(2)'s format is like VTR-271
Challenger: (1) Limited Power… (2) VTR-271

Result | A glimpse of Champion vs Challenger capabilities

	Champion vs. Challenger Model Performance				
Document Type (19)	**Champion Recall**	**Vision Recall**	**BERT Recall**	**XGBoost Recall**	**Better Model**
130-U Page 1 of 2	90.9%	95.0%	100.0%	100.0%	BERT/XGBoost
Check Endorsement	100.0%	100.0%	100.0%	100.0%	All
Driver's License Page 1 of 2	67.3%	95.0%	92.7%	85.5%	Vision
Fees Payment Check	97.3%	94.0%	88.9%	89.2%	Champion
FORM 30-C Page 1 of 2	97.3%	91.0%	96.2%	60.5%	Champion
FORM 30-C Page 2 of 2	100.0%	98.0%	89.2%	100.0%	XGBoost/Champion
Lien Add - Order	95.5%	95.0%	86.4%	93.2%	Champion
Limited Power of Attorney/Motor Vehicle	100.0%	82.0%	100.0%	100.0%	BERT/XGBoost/Champion
MV-50	100.0%	100.0%	100.0%	100.0%	All
Odometer Disclosure Statement	71.4%	75.0%	71.4%	100.0%	XGBoost
One and Same Statement	93.3%	81.0%	73.3%	73.3%	Champion
Other Document	96.5%	93.0%	87.6%	91.7%	Champion
Search Request Form	93.6%	100.0%	98.9%	100.0%	Vision/XGBoost
Texas Liability Insurance Card	0.0%	5.0%	100.0%	66.7%	BERT
Texas Vehicle Record	0.0%	100.0%	100.0%	100.0%	Vision/BERT/XGBoost
TS-5	100.0%	100.0%	33.3%	100.0%	Vision/XGBoost/Champion
TS-8	100.0%	100.0%	100.0%	100.0%	All
VTR-271	98.3%	99.0%	100.0%	100.0%	BERT/XGBoost
WKLS Authorization letter	97.6%	99.0%	100.0%	100.0%	BERT/XGBoost
Accuracy	**87.6%**	**86.3%**	**88.4%**	**88.4%**	**Challenger Wins: 14/19**

High learning + Solid deliverables + Strong impact

Our Deliverables to Wolters Kluwer

(1) OCR API
Input: PDF Docket
Output: Azure OCR response with document quality proxy

(2) CAPABILITY API
Input: PDF Docket (+OCR Response)
Output: Azure OCR response with document quality score

(3) Documentation
- Used all bounding boxes from Azure OCR as potential titles; narrowed search space
- Engineered features like location on page, bold, font size, lone item, etc.

Our Deliverables to MIT*

(1) Capstone Presentation
It shows the entire journey of the Capstone project and to be presented if selected in Top 10 Capstone teams

(2) Capstone Poster
It is to be presented during Capstone Showcase

(3) Capstone Report
It contains every technical detail of the Capstone project

(4) Capstone Video
A 2-3 minute video explaining the project (voted by public globally)

Learnings from Capstone with WK

(1) Industry Knowledge
(2) Business Acumen
(3) Mentorship
(4) Collaboration
(5) Consistent Support
(6) Challenged!
(7) Awesome Team!

*due on 4th August 2023

A big thanks to the Wolters Kluwer family for the constant support!

Abhishek (VP, Data Analytics), Rajiv (VP, Data Science),
Pooja (Data Scientist Manager),
Varun (Lead Business Systems Analyst)

[Core Team]

Lucas (Onboarding), Savannah (Business), Erin (Proof of Concept), Florence (Experimentation), Piyush (Technical), Ameya (Sr. Data Scientist)

[Mentor Match Series]

*names are in order of the images shown