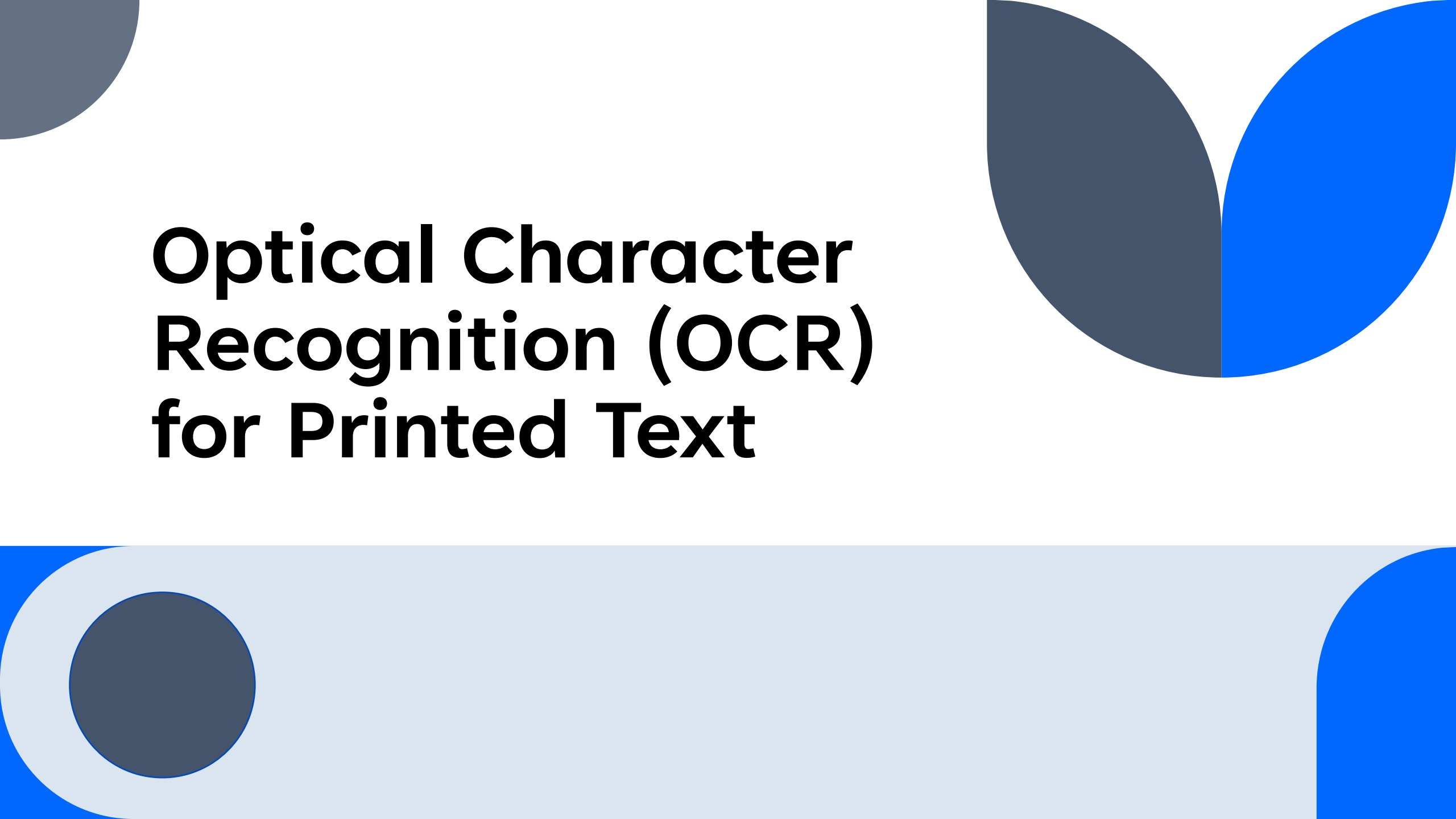# Optical Character Recognition (OCR) for Printed Text

Project Overview

- **Optical Character Recognition (OCR) for Printed Text**

- **Goal: Extract structured data from images containing tables using OCR.**

- **Tools: OpenCV and PaddleOCR.**

Approach Summary

Stage 1: Table Detection

Stage 2: Line Removal for Clearer OCR
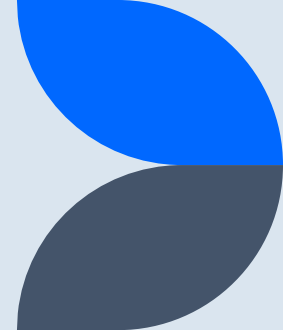
Stage 3: Cell Detection and Text Extraction

Table Detection

Group Number	1305693	Overall Benefit Limit	1,000,000
Policy Inception Date	Feb 17, 2022	Inpatient/Outpatient Limit	1,000,000
Policy Expiry Date	Feb 16, 2023	Dental Limit	3,000
Class	B	Optical Limit	1,000
Deductible	20 % up to SR 200.0	Maternity Limit	15,000

Monthly Claims	Number of Lives Insured	Number of Paid Claims	Amount of Paid Claims	Amount of Paid Claims with VAT
Policy Year - 2 years prior: Number of lives at start 0				
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
		0	0	0
Prior Policy Year: Number of lives at start 13				
202102	13	15	4,749.6	4,800.9
202103	13	23	9,598.35	9,772.15
202104	13	17	7,226.49	7,435.04
202105	13	21	2,449.17	2,573.31
202106	13	14	4,401	4,487.1
202107	13	17	8,572.48	8,640.11
202108	13	17	1,843.14	1,880.25
202109	13	8	4,749.6	4,800.9
202110	13	19	9,598.35	9,772.15
202111	13	20	7,226.49	7,435.04
202112	13	5	2,449.17	2,573.31
202201	13	18	4,401	4,487.1
202202	13	16	8,572.48	8,640.11
		210	75,837.32	87,212.92
Last Policy Year: Number of lives at start 27				
202202	27	24	4,749.6	4,800.9
202203	27	30	9,598.35	9,772.15
202204	27	21	7,226.49	7,435.04
202205	27	17	2,449.17	2,573.31
202206	27	19	4,401	4,487.1
202207	27	15	8,572.48	8,640.11
202208	27	28	9,598.35	9,772.15
202209	27	25	7,226.49	7,435.04
Overall - Total		179	53,821.93	61,895.22

Overall Benefits				
Last Policy Year				
Benefit_Sama	Number of Paid Claims	Amount of Paid Claims	Amt of Claims (VAT)	Notes
1. Outpatient	44	37,035.25	42590.53	Paid 20 % up to 100.0
2. OP Lab & Diagnostics	17	23,346.47	26848.44	Paid 15 % up to 125.0
3. OP Consultation	10	2,040.89	2347.02	
4. OP Pharmacy	17	11,867.51	13647.64	
5. Dental	1	385	442.75	
6. Optical	1	334	384.1	
7. Maternity	1	875	1006.25	Cesarean is covered
Overall	47	75884.12	87266.74	

Remove Line
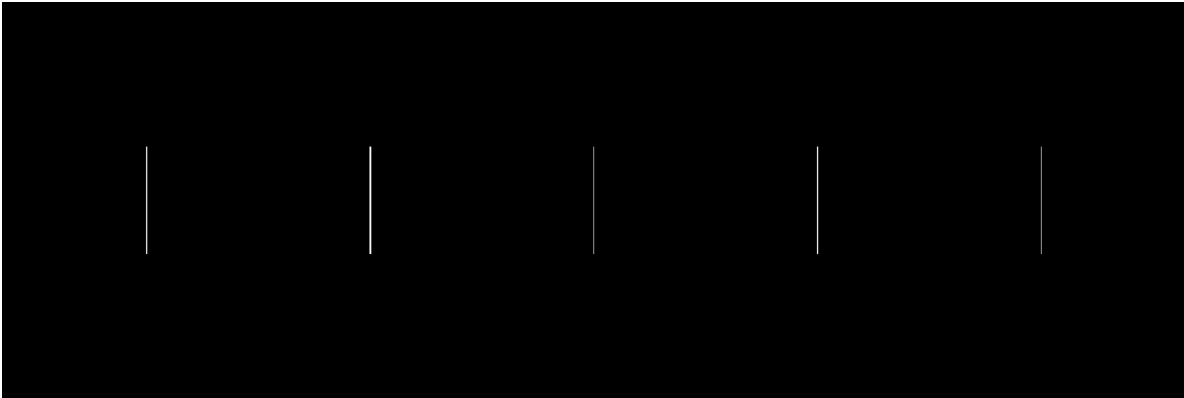
Combined Line and Subtract

Group Number	1305693	Overall Benefit Limit	1,000,000
Policy Inception Date	Feb 17, 2022	Inpatient/Outpatient Limit	1,000,000
Policy Expiry Date	Feb 16, 2023	Dental Limit	3,000
Class	B	Optical Limit	1,000
Deductible	20 % up to SR 200.0	Maternity Limit	15,000

=

Group Number	1305693	Overall Benefit Limit	1,000,000
Policy Inception Date	Feb 17, 2022	Inpatient/Outpatient Limit	1,000,000
Policy Expiry Date	Feb 16, 2023	Dental Limit	3,000
Class	B	Optical Limit	1,000
Deductible	20 % up to SR 200.0	Maternity Limit	15,000

Cell Detection & Text Extraction

Group Number	1305693	Overall Benefit Limit	1,000,000
Policy Inception Date	Feb 17, 2022	Inpatient/Outpatient Limit	1,000,000
Policy Expiry Date	Feb 16, 2023	Dental Limit	3,000
Class	B	Optical Limit	1,000
Deductible	20 % up to SR 200.0	Maternity Limit	15,000

Group Number	1305693	Overall Benefit Limit	1,000,000
Policy Inception Date	Feb 17, 2022	Inpatient/Outpatient Limit	1,000,000
Policy Expiry Date	Feb 16, 2023	Dental Limit	3,000
Class	B	Optical Limit	1,000
Deductible	20 % up to SR 200.0	Maternity Limit	15,000

Data Extraction and Output

CSV Output: Each table is exported as a CSV.Data

Manipulation: Organize extracted data into structured data frames:

Claims_data

Benefits_data
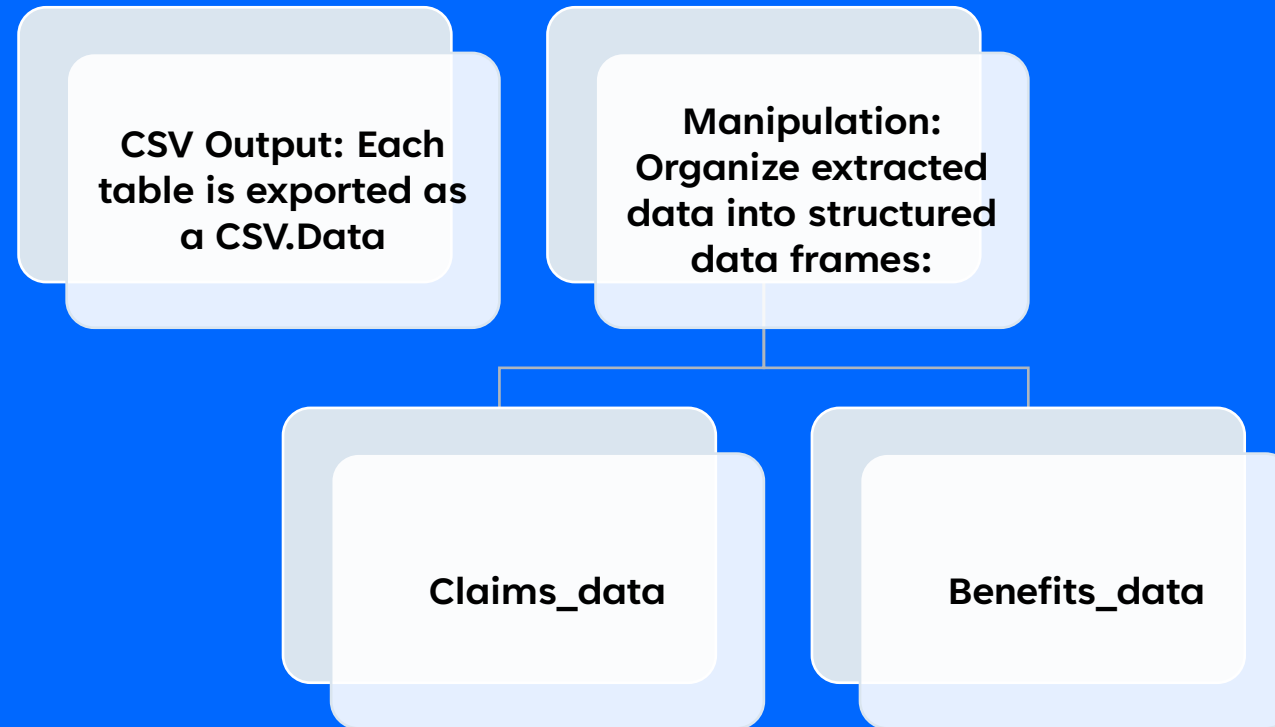
Key Insights

1

Efficient Table Detection: Using contour-based techniques improves accuracy.

2

Line Removal Enhances OCR: Removing grid lines significantly improves text recognition accuracy.

3

Structured Data Extraction: Resulting CSV files allow easy data manipulation and integration.

Thank you

Ayman Attili

009627-8665-6124

engaymanattili@gmail.com