



1

Data understanding

2

Stamp Data

3

LayoutLM data - Preparing Data (OCR & Transformation)

4

Preprocessing results example

5

Overlap and merge steps

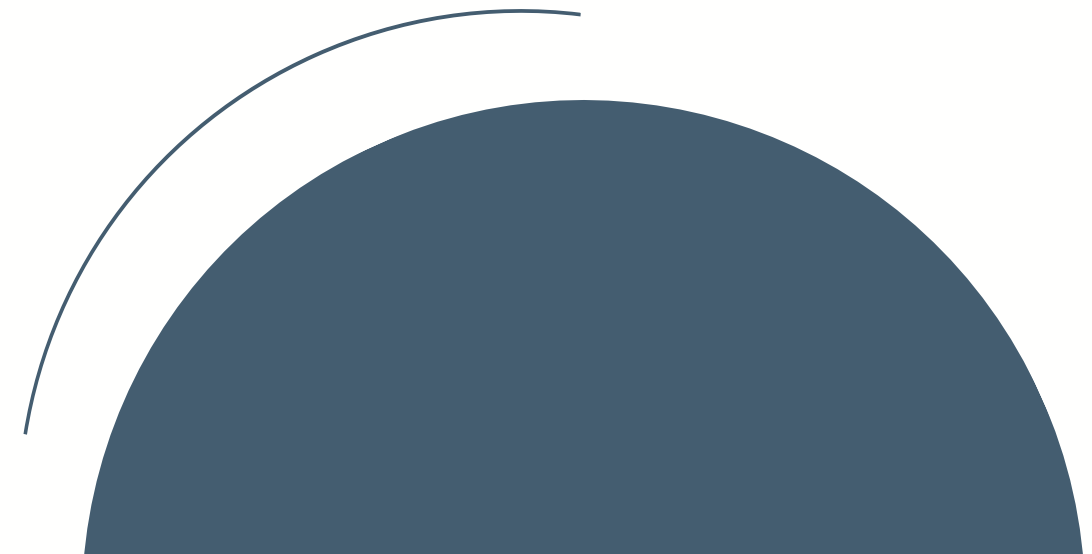
6

7

Next steps

8

9





Advenssement

...



2

3

4

5

6

7

Data understanding

Missing Files

```

non existing files

image_files = set(os.listdir(image_folder_path))

# the expected filenames (adding the suffix '_1.png')
expected_filenames = set(data['filename'].apply(lambda x: f"{x}_1.png"))

csv_not_in_folder = expected_filenames - image_files
folder_not_in_csv = image_files - expected_filenames

print("Files listed in CSV but not found in the folder:\n", csv_not_in_folder)
print('\n')
print("Files in the folder but not listed in CSV:\n", folder_not_in_csv)

Files listed in CSV but not found in the folder:
{'IEEE fAC NÃ84.pdf_1.png', 'ilovepdf_merged-2(1).pdf_1.png', 'ilovepdf_merged-4(1).pdf_1.png', '2022-09-14--2022-10-12_RÃsumÃ_Facture.pdf_1.png',
_merged-3(1).pdf_1.png'}

```

Empty Fields

```
Empty fields in each column:
filename           0
title_bbx          26
date_bbx           9
stamp_bbx          52
signature_bbx      65
ieee_bbx           26
total_bbx          12
totalValue_bbx     13
title              27
handwritten         0
```

As the reaction of our models to nan fields in json files, Two repositories were created. These initial files are designed to stamp/signature and Donut model

The annotations are now aligned

		Livré à: cte olympique manzaht tunis Transporteur: .	
STE MYTEK INFORMATIQUE 58 RUE DE L'INDUSTRIE CHAROUA 1 2035 TUNIS <small>Tel : +216 96 010 910 Fax : +216 71 804 908 email : Site web : www.mytek.tn</small>		ASSOCIATION COGATUNISIA SECTION cte olympique manzaht tunis TUNIS 1002 Tel: 53535301 MF/CIN: 14400882 RC Mode de règlement: CAISSE	

Bon de Livraison
Facture
FAC-23M01LIV-263726

Date	Número pièce	Client	Votre référence	Chargé(e)
Date: 2023	FAC-23M01LIV-263726	C126439	Facture Caisse	SRA INTEGRATION

Référence	Code à Barre	Désignation	Qsm	PU TTC	R.	Montant	TVA
BU-TV-TCL-4056500		BUNDEL TCL TELEVISEUR 40" FULL HD SMART ANDROID S6500 GAR 2 ANS +WAVES ABOONEMENT 1AN	1	769.000	0	769.000	1
C-IP TV-1AN	121209043882	CARTE ABOONEMENT IPTV 1 AN GLOBAL TV PRO	1	29.900	0	0.000	1
ABNT-WAVES-1A N	6190400520230	WAVES ABOONEMENT 1AN	1	0.000	0	0.000	1
TV-TCL-4056500	6192901207832	TCL TELEVISEUR 40" FULL HD SMART ANDROID S6500 GAR 2 ANS S N : Z306ELL196355AG0005	1	949.000	0	0.000	1
V11H73040	8715946680569	EPSON VIDEOPROJECTEUR EB-V06 WXGA 2 ANS GARANTIE S N : X89E3200061	1	1.659.000	0	1.659.000	7
HDL-117B-2	121254876456	HOME DESIGN SUPPORT MURALE MOBILE DE 14" AU 55"-80KG	1	39.000	0	39.000	1

Transport: 0,000		Installation: 0,000	Frais: 0,000	Remises: 0,000	Garantie: 0,000		
TOTAL HT:	Base TVA	Taux	Montant TVA				
2 230,459	1 559,467	7	109,533				
	678,992	19	129,008				

Arrêtée la présente pièce à la somme de: Deux Mille Quatre Cent Soixante-Huit Dinars.

IMPORTANT: Le Client reconnaît avoir reçu les marchandises et que la quantité de marchandise livrée est conforme à celle indiquée sur le bon de livraison. La marchandise reste la propriété de l'entreprise jusqu'à paiement intégral. Les frais seront considérés comme exclusifs des livraisons en cas de pourcentage de retour. Toute réclamation devra être faite au plus tard dans les 15 jours suivant la date de livraison.

Timbre fiscal: 1,000

Total avec Timbre: 2 230,459

TVA :237,541

MONTANT TOTAL

RIB: 04115080004598725258
MF: 1194759K/B/E/002
RC: B0124512011

Création: 29/09/2023 16:00:26
Créé par: LTFI/MARNA
Date d'émission:



1

2

3

4

5

6

7

Stamp Data

```
1 {"filename": "files (65).pdf_1.png",
2  "title_bbx": NaN,
3  "date_bbx": "[0.8035055026194582, 0.22101015654140369, 0.10984140250389518, 0.019127320758753452]",
4  "stamp_bbx": NaN,
5  "signature_bbx": NaN,
6  "ieee_bbx": NaN,
7  "total_bbx": "[0.5312225544596436, 0.5808427928096865, 0.058788357100898025, 0.01852960124816716]",
8  "totalValue_bbx": "[0.8398615030090394, 0.5838314359656938, 0.08972960389250173, 0.019127320758753452]"}

```

```
1 {"filename": "files (65).pdf_1.png",
2  "date_bbx": "[0.8035055026194582, 0.22101015654140369, 0.10984140250389518, 0.019127320758753452]",
3  "total_bbx": "[0.5312225544596436, 0.5808427928096865, 0.058788357100898025, 0.01852960124816716]",
4  "totalValue_bbx": "[0.8398615030090394, 0.5838314359656938, 0.08972960389250173, 0.019127320758753452]"}

```


LayoutLM data - Preparing Data (OCR & Transformation)

- 1
- 2
- 3
- 4
- 5
- 6
- 7



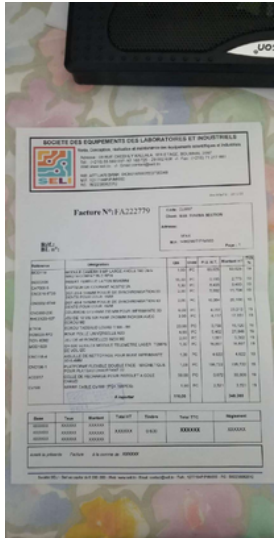
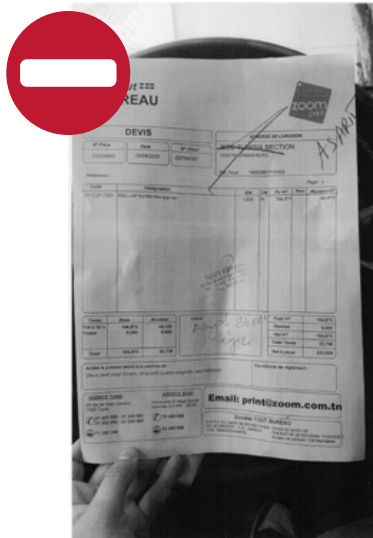
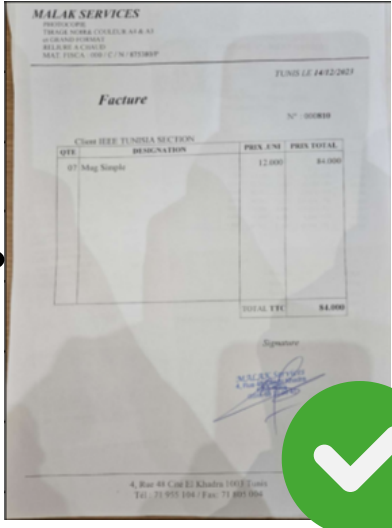
Images:
bad quality, noise,... =>
Preprocessing needed

Annotations:
Missing Text Values and
unsuitable for Layoutlmv3

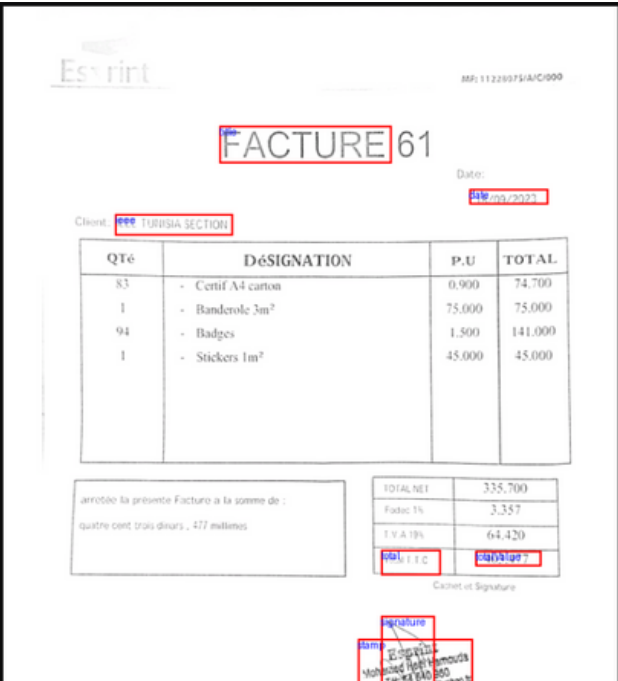
- Gray Scaling
- Fix Illumination (CLAHE)
- Denoising (NLM)
- Skew angle determination
- rotation
- Resizing
- DPI Scaling
- Binarization (Adaptive Thresholding, OITSU ...)
- ~~Dilation~~, ~~Erosion~~
- Wrapping

- Change annotation shape => Json file
 - OCR Application
 - Overlapping Test with our annotation
- if yes => label<- annotation.label
else label = other

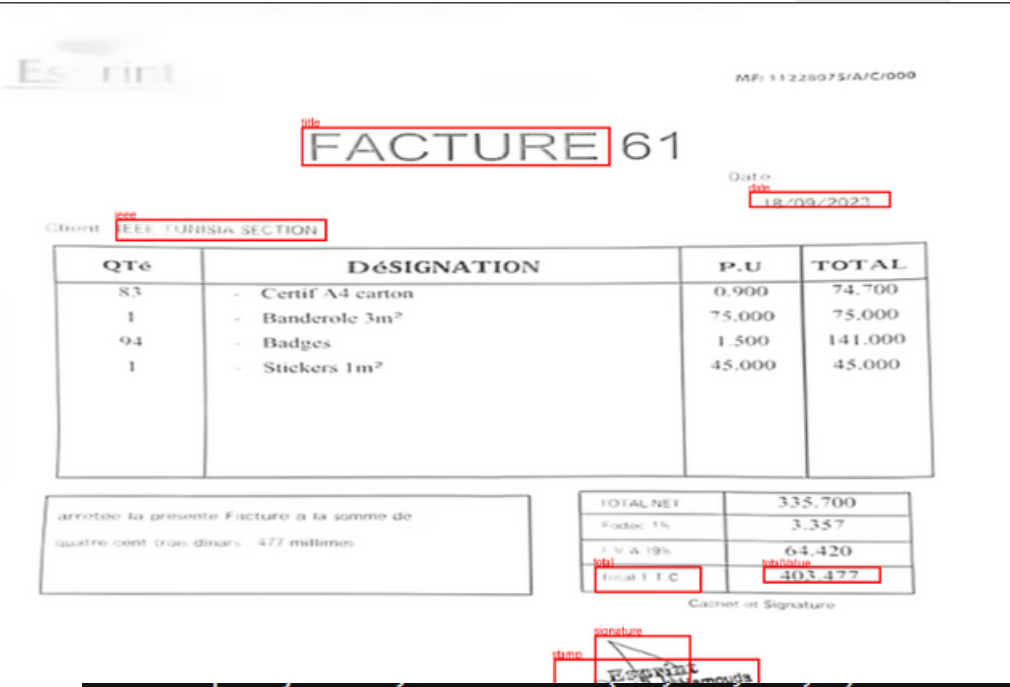
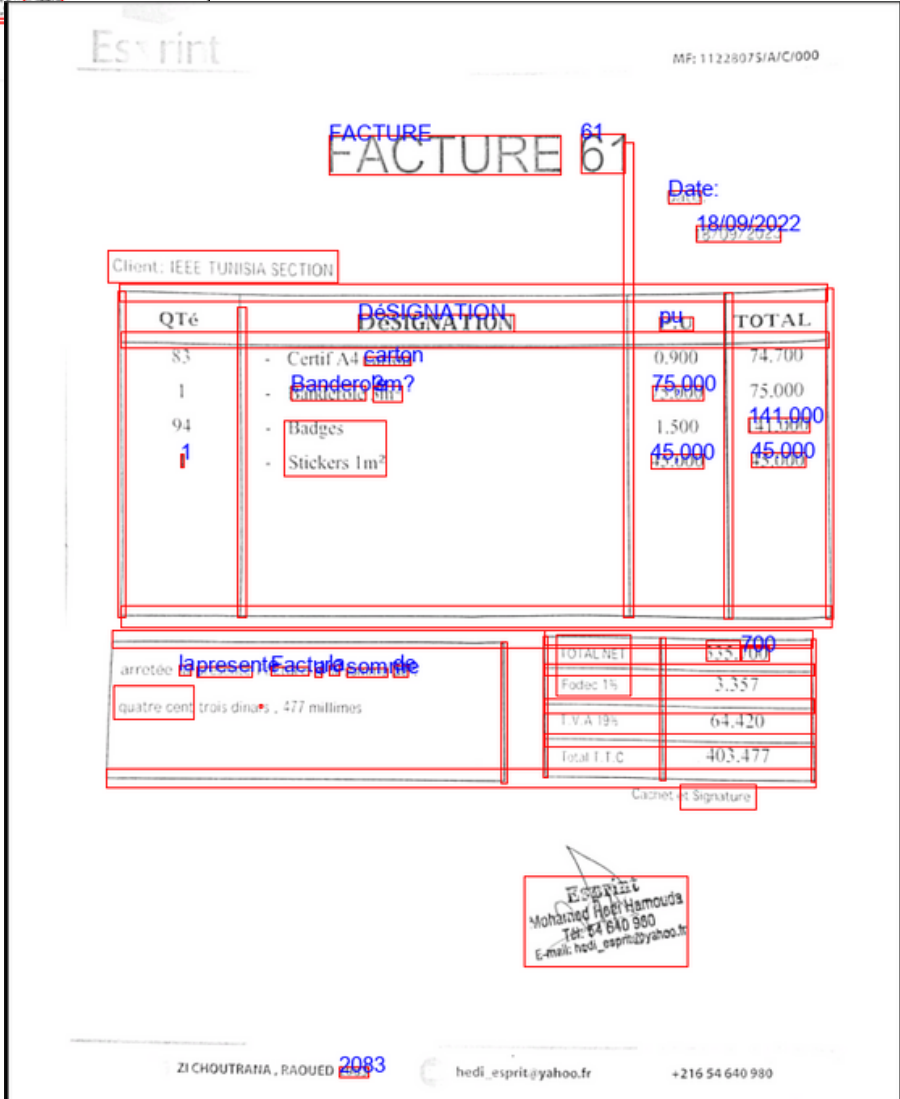
- if an annotation.label doesn't exist => merge the "item"
- Merge repeated labels



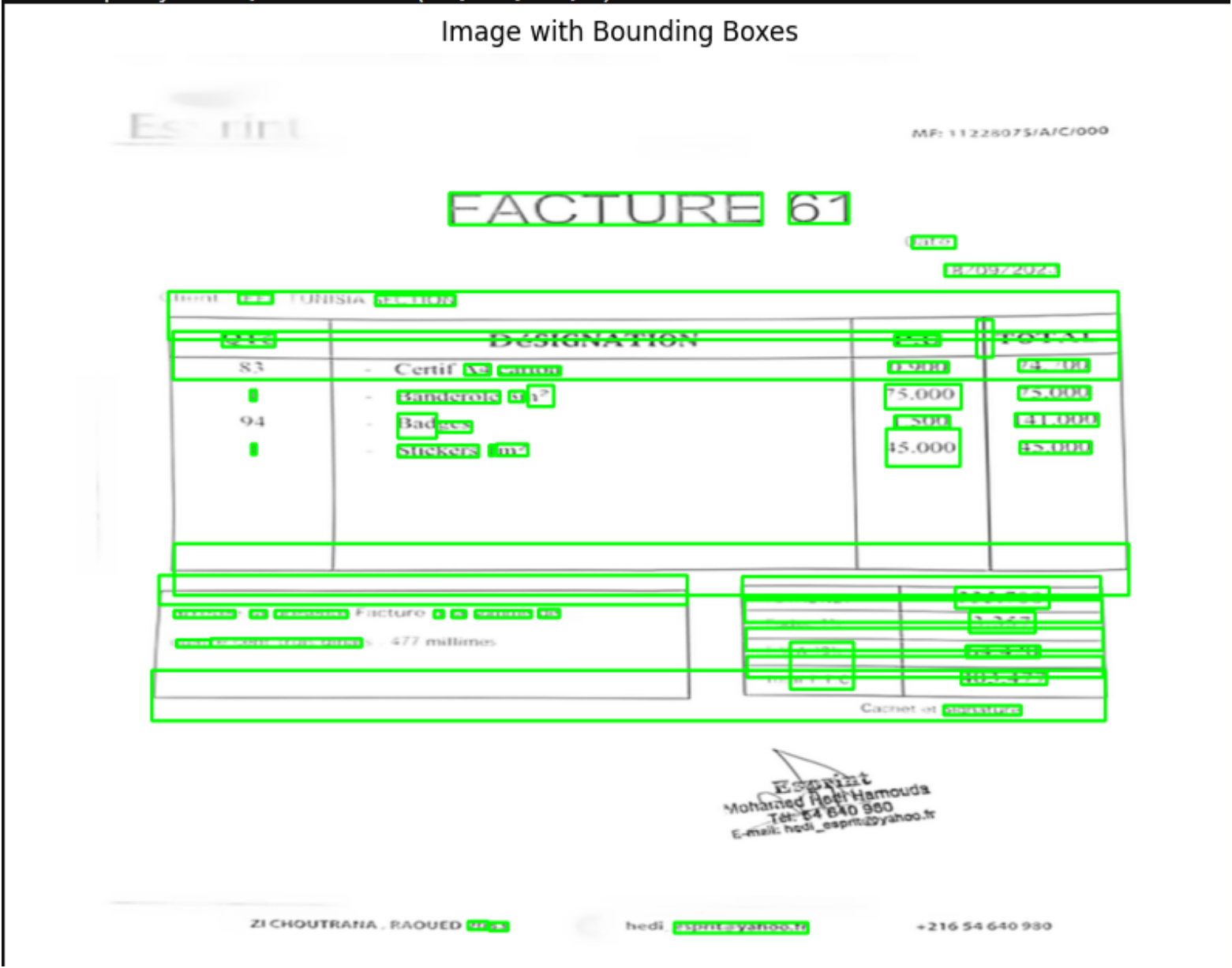
- 1
- 2
- 3
- 4
- 5
- 6
- 7



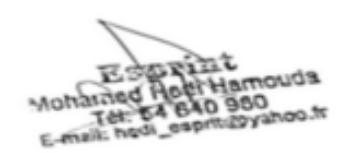
Original image



Preprocessed image



7



Esprit

MF: 11228075/A/C/000

FACTURE 61

Client: TUNISIA SECTION

180002022

QTY	DESIGNATION	PU	TOTAL
83	- Certif	86,64 DM	7.4200 DM
1	- Contrats	75000,00	75000 DM
94	- Bdes	8.600 DM	8.44,00 DM
1	- Stickers	45.000	45.000 DM

Facture

la somme de

le cent trois dinars . 477 millimes

Signature

Esprit

Mohamed Hedi Hamouda

Tel: 54 640 980

E-mail: hedi_esprit@yahoo.fr

Stamp

ZI CHOUTRANA, RAOUED

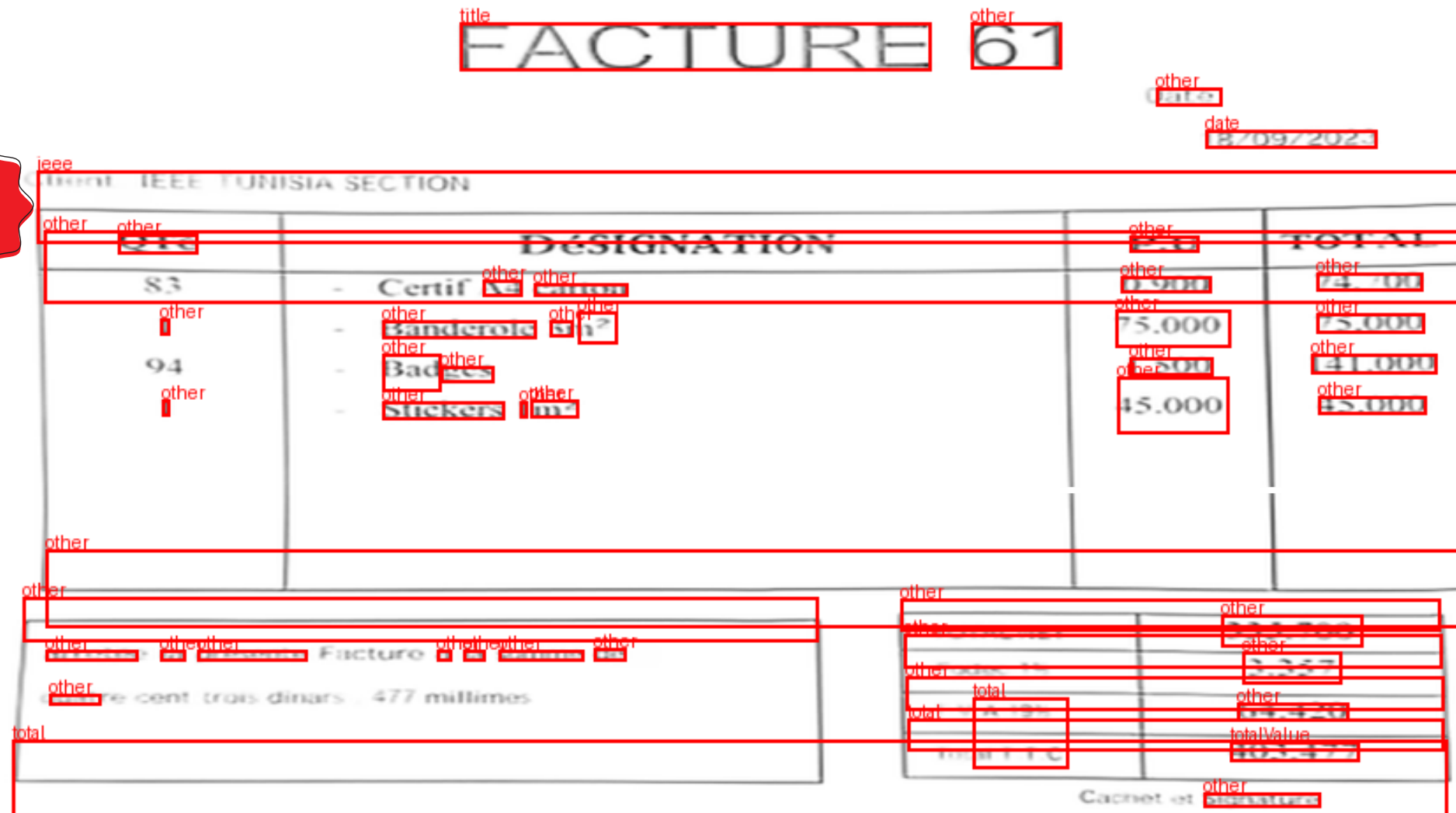
hedi_esprit@yahoo.fr

+216 54 640 980



MF: 11228075/A/C/000

7



ZI CHOUTRANA, RAOUED  hedi_esprit@yahoo.fr +216 54 640 980

1

2

3

4

5

6

7

Next Steps

Model Fine Tuning

Test after Data Augmentation & Better OCR results