Placement Assignment,
Indian Institute Of Technology Madras

Invoice Data Extraction

Goutham B (*BE20B010*)

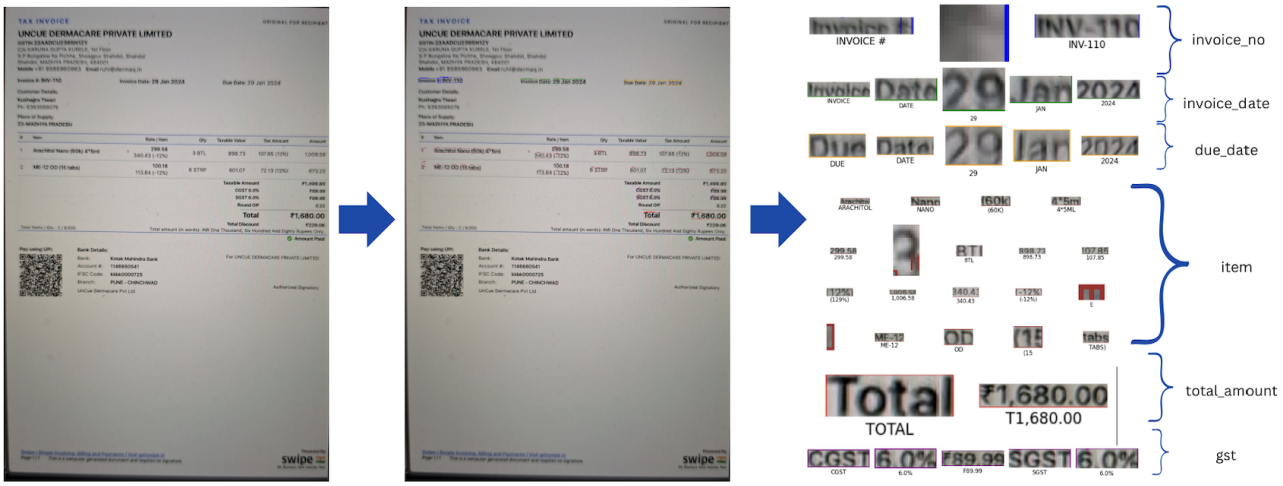
October 20, 2024

Contents

1 Overview of My Work

I have split my work into 3 sections:

- Data Labeling and Processing using `label-studio` and `pytesseract`
- Text Detection Module using `microsoft layoutlmv3`
- Text Recognition Module using `microsoft TrOCR`



Test Image (type: Image, taken from mobile camera) is unseen and this type is not part of training data

Text Detection using LayoutLMv3

Text Recognition using TrOCR

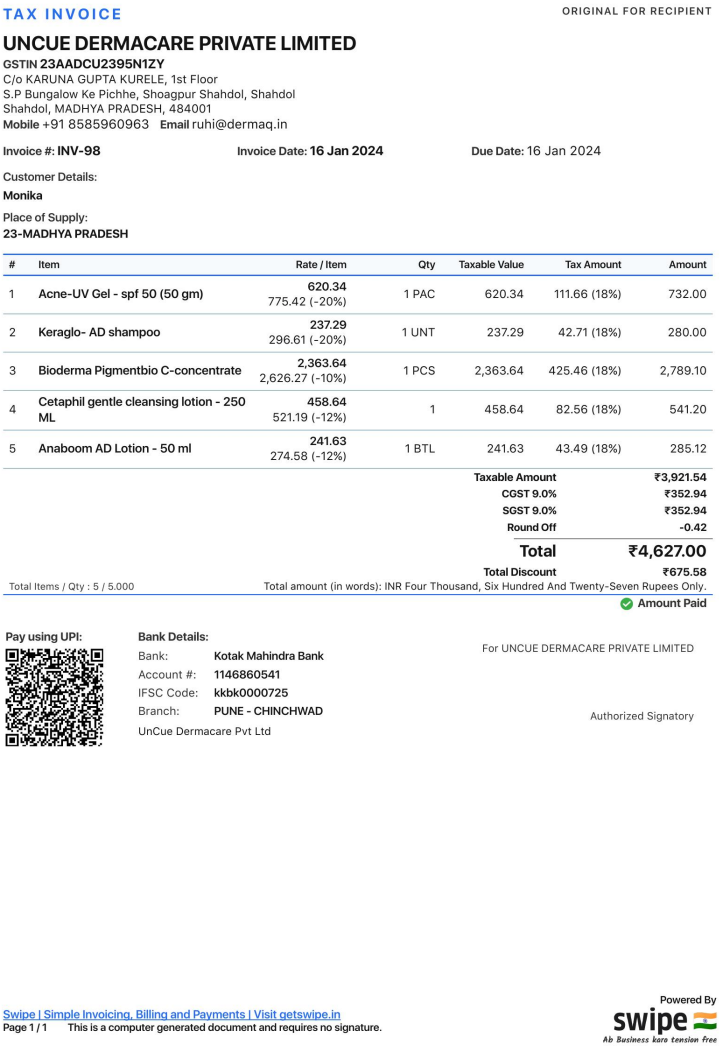
Figure 1: Data Extraction Pipeline

2 Data Labeling and Processing

2.1 Data

There are three types of data in my dataset: *'Regular'*, *'Scanned'* and *'Image'*. The *'Regular'* images are images that are direct conversion of the pdf data files (shared to us) into JPG format. The *'Scanned'* images are images that are scanned using the mobile application DOC-SCANNER. The *'Image'* images are images that are taken using mobile phone camera. Each type contains 5 images making a total of 15 image used for labeling task.

The different types of images are as follows:



TAX INVOICE

ORIGINAL FOR RECIPIENT

UNCUE DERMACARE PRIVATE LIMITED

GSTIN 23AADCU2395N1ZY
C/o KARUNA GUPTA KURELE, 1st Floor
S.P Bungalow Ke Pichhe, Shoagpur Shahdol, Shahdol
Shahdol, MADHYA PRADESH, 484001
Mobile +91 8585960963 **Email** ruhi@dermaq.in

Invoice #: INV-98 **Invoice Date: 16 Jan 2024** **Due Date:** 16 Jan 2024

Customer Details:
Monika

Place of Supply:
23-MADHYA PRADESH

#	Item	Rate / Item	Qty	Taxable Value	Tax Amount	Amount
1	Acne-UV Gel - spf 50 (50 gm)	620.34 775.42 (-20%)	1 PAC	620.34	111.66 (18%)	732.00
2	Keraglo- AD shampoo	237.29 296.61 (-20%)	1 UNT	237.29	42.71 (18%)	280.00
3	Bioderma Pigmentbio C-concentrate	2,363.64 2,626.27 (-10%)	1 PCS	2,363.64	425.46 (18%)	2,789.10
4	Cetaphil gentle cleansing lotion - 250 ML	458.64 521.19 (-12%)	1	458.64	82.56 (18%)	541.20
5	Anaboom AD Lotion - 50 ml	241.63 274.58 (-12%)	1 BTL	241.63	43.49 (18%)	285.12

Taxable Amount	**₹3,921.54**
CGST 9.0%	**₹352.94**
SGST 9.0%	**₹352.94**
Round Off	**-0.42**
Total	**₹4,627.00**
Total Discount	**₹675.58**

Total Items / Qty : 5 / 5.000

Total amount (in words): INR Four Thousand, Six Hundred And Twenty-Seven Rupees Only.

Amount Paid

Pay using UPI:

Bank Details:
Bank: **Kotak Mahindra Bank**
Account #: **1146860541**
IFSC Code: **kkbk0000725**
Branch: **PUNE - CHINCHWAD**
UnCue Dermacare Pvt Ltd

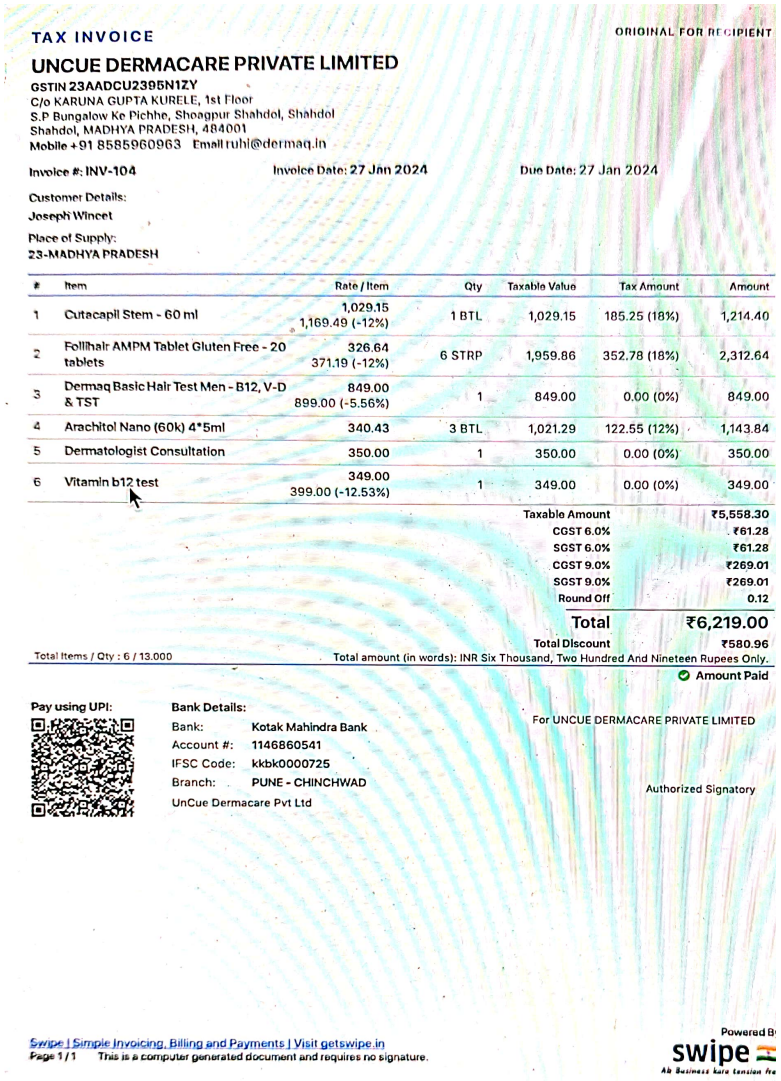
For UNCUE DERMACARE PRIVATE LIMITED

Authorized Signatory

Swipe | Simple Invoicing, Billing and Payments | Visit getswipe.in
Page 1 / 1 This is a computer generated document and requires no signature.

Powered By swipe
Ab Business karo tension free

Figure 2: Regular (Direct PDF to JPG File)



TAX INVOICE

ORIGINAL FOR RECIPIENT

UNCUE DERMACARE PRIVATE LIMITED

GSTIN 23AADCU2305N1ZY
C/o KARUNA GUPTA KURELE, 1st Floor
S.P Bungalow Ke Pichhe, Sheagpur Shahdol, Shahdol
Shahdol, MADHYA PRADESH, 484001
Mobile +91 8585960963 Email ruhi@dermaq.in

Invoice #: INV-104 Invoice Date: 27 Jan 2024 Due Date: 27 Jan 2024

Customer Details:
Joseph Wincet

Place of Supply:
23-MADHYA PRADESH

#	Item	Rate / Item	Qty	Taxable Value	Tax Amount	Amount
1	Cutacapil Stem - 60 ml	1,029.15 1,169.49 (-12%)	1 BTL	1,029.15	185.25 (18%)	1,214.40
2	Follihair AMPM Tablet Gluten Free - 20 tablets	326.64 371.19 (-12%)	6 STRP	1,959.86	352.78 (18%)	2,312.64
3	Dermaq Basic Hair Test Men - B12, V-D & TST	849.00 899.00 (-5.56%)	1	849.00	0.00 (0%)	849.00
4	Arachitol Nano (60k) 4*5ml	340.43	3 BTL	1,021.29	122.55 (12%)	1,143.84
5	Dermatologist Consultation	350.00	1	350.00	0.00 (0%)	350.00
6	Vitamin b12 test	349.00 399.00 (-12.53%)	1	349.00	0.00 (0%)	349.00

Taxable Amount	₹5,558.30
CGST 6.0%	₹61.28
SGST 6.0%	₹61.28
CGST 9.0%	₹269.01
SGST 9.0%	₹269.01
Round Off	0.12
Total	**₹6,219.00**
Total Discount	₹580.96

Total Items / Qty : 6 / 13.000

Total amount (in words): INR Six Thousand, Two Hundred And Nineteen Rupees Only.

Amount Paid

Pay using UPI:

Bank Details:
Bank: Kotak Mahindra Bank
Account #: 1146860541
IFSC Code: kkbk0000725
Branch: PUNE - CHINCHWAD
UnCue Dermacare Pvt Ltd

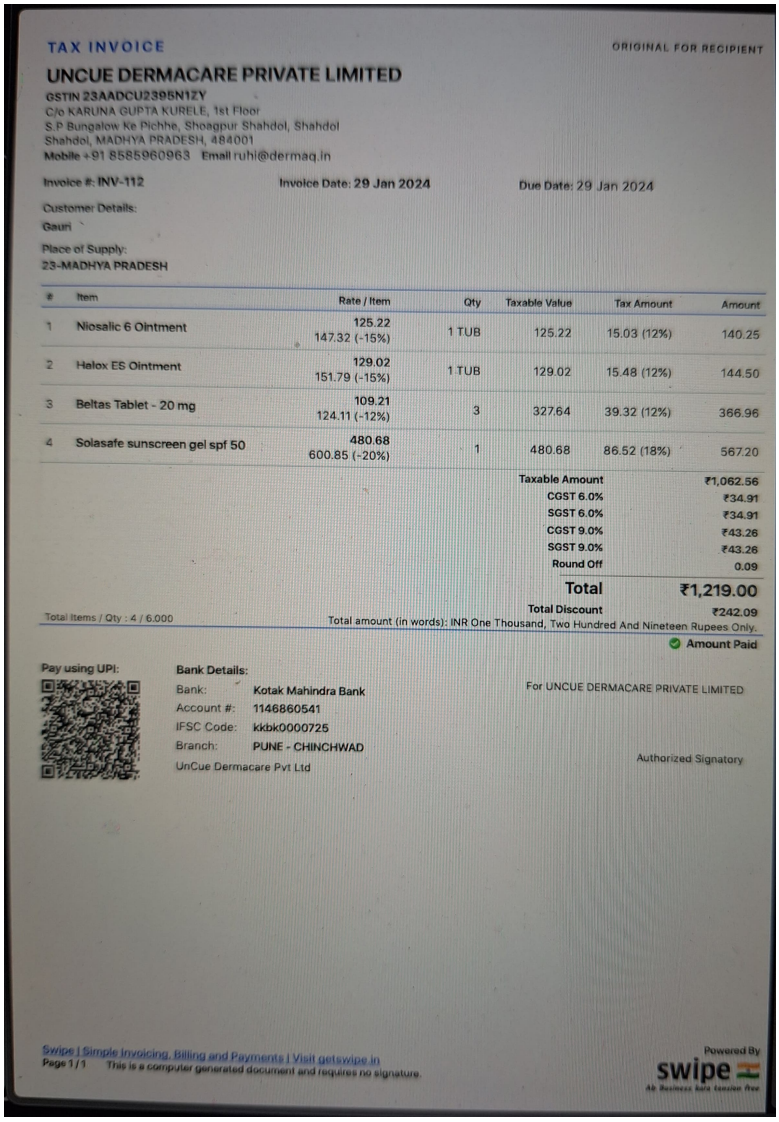
For UNCUE DERMACARE PRIVATE LIMITED

Authorized Signatory

Swipe | Simple Invoicing, Billing and Payments | Visit getswipe.in
Page 1 / 1 This is a computer generated document and requires no signature.

Powered B
swipe
Ab Business kara tension fre

Figure 3: Scanned (Using DocScanner Application)



TAX INVOICE

ORIGINAL FOR RECIPIENT

UNCUE DERMACARE PRIVATE LIMITED

GSTIN 23AADCU2395N1ZY
C/o KARUNA GUPTA KURELE, 1st Floor
S.P Bungalow Ke Pichhe, Shoagpur Shahdol, Shahdol
Shahdol, MADHYA PRADESH, 484001
Mobile +91 8585960963 Email ruhi@dermaq.in

Invoice #: INV-112 Invoice Date: 29 Jan 2024 Due Date: 29 Jan 2024

Customer Details:
Gauri

Place of Supply:
23-MADHYA PRADESH

#	Item	Rate / Item	Qty	Taxable Value	Tax Amount	Amount
1	Niosalic 6 Ointment	125.22 147.32 (-15%)	1 TUB	125.22	15.03 (12%)	140.25
2	Halox ES Ointment	129.02 151.79 (-15%)	1 TUB	129.02	15.48 (12%)	144.50
3	Beltas Tablet - 20 mg	109.21 124.11 (-12%)	3	327.64	39.32 (12%)	366.96
4	Solasafe sunscreen gel spf 50	480.68 600.85 (-20%)	1	480.68	86.52 (18%)	567.20

Taxable Amount	₹1,062.56
CGST 6.0%	₹34.91
SGST 6.0%	₹34.91
CGST 9.0%	₹43.26
SGST 9.0%	₹43.26
Round Off	0.09
Total	**₹1,219.00**
Total Discount	₹242.09

Total Items / Qty : 4 / 6.000

Total amount (in words): INR One Thousand, Two Hundred And Nineteen Rupees Only.

Amount Paid

Pay using UPI:

Bank Details:
Bank: Kotak Mahindra Bank
Account #: 1146860541
IFSC Code: kkbk0000725
Branch: PUNE - CHINCHWAD
UnCue Dermacare Pvt Ltd

For UNCUE DERMACARE PRIVATE LIMITED

Authorized Signatory

Swipe | Simple Invoicing, Billing and Payments | Visit getswipe.in
Page 1 / 1 This is a computer generated document and requires no signature.

Powered By swipe

Figure 4: Image (Using Mobile Phone Camera)

2.2 Labeling with Label-Studio

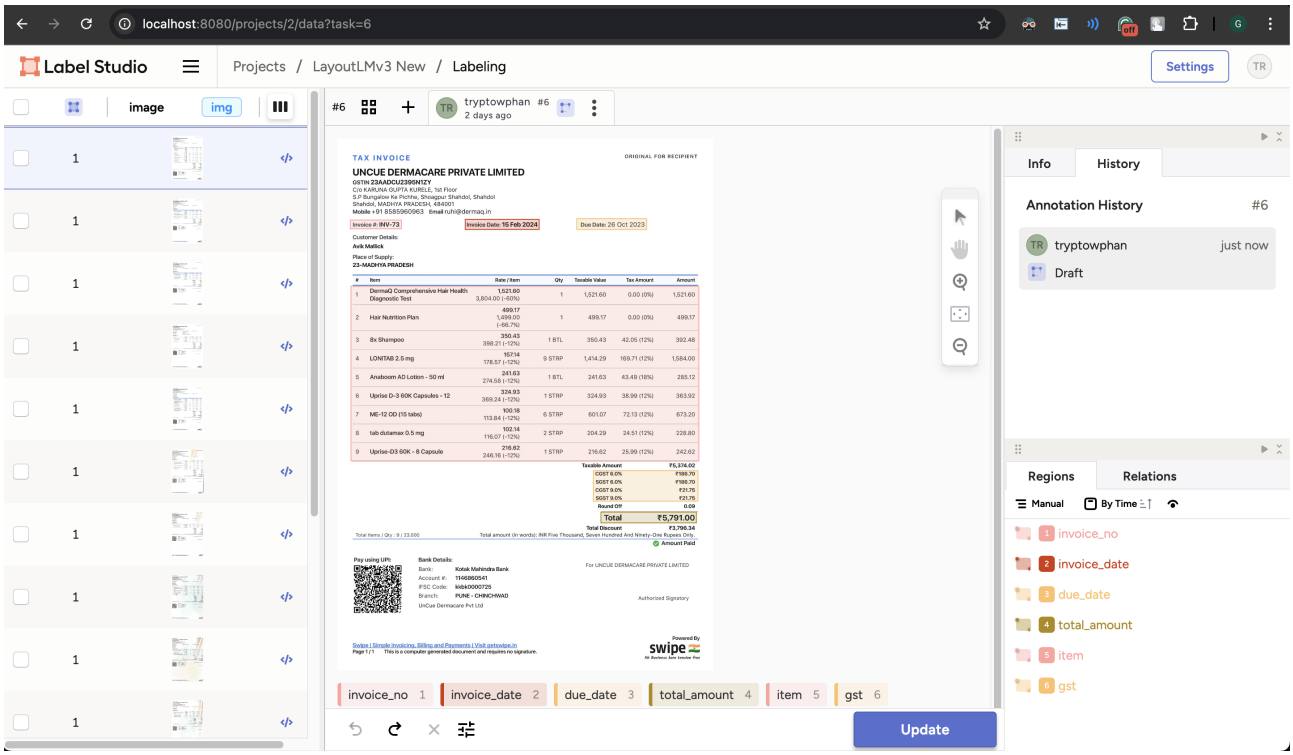


Figure 5: Label Studio

Prior to training, we have to label the data. I have used the label-studio python package for the same. The labels have been clearly mentioned in 5.

The 5 different labels used are: ***invoice_no***, ***invoice_date***, ***due_date***, ***total_amount***, ***item***, ***gst***.

2.3 Processing

The `initial_processing.ipynb` notebook processes image data annotated in Label Studio for Optical Character Recognition (OCR) tasks using Tesseract. It begins by loading JSON data exported from Label Studio, which contains image annotations, and initializing the necessary libraries for image processing and geometric calculations.

The function `calculate_iou` computes the Intersection over Union (IoU) between bounding boxes to assess overlap, while `hocr_to_dataframe` parses the generated hOCR files, extracting recognized words, their coordinates, and confidence scores into a DataFrame. The script iterates through the labeled data, converting bounding box coordinates from percentages to absolute pixel values, and generates annotations for each image. Using Tesseract, it performs OCR on the images and compares the extracted words with the labeled bounding boxes using IoU, filtering out those with less than 80% overlap. The resulting tokens, bounding boxes, and corresponding labels are compiled into a final list, which is then split into training and testing datasets.

The training dataset contains five *Regular* (.jpg) files and four *Scanned* (.jpg) files, while the testing dataset includes one *Scanned* (.jpg) file and five *Image* (.jpeg) type files.

3 Text Detection Module

I fine tuned the pre-trained `microsoft layoutlmv3` on the training dataset (9 files: 5 *Regular* + 4 *Scanned*) for the text detection module. The tokens created using Tesseract is fed into LayoutLMv3 for fine tuning.

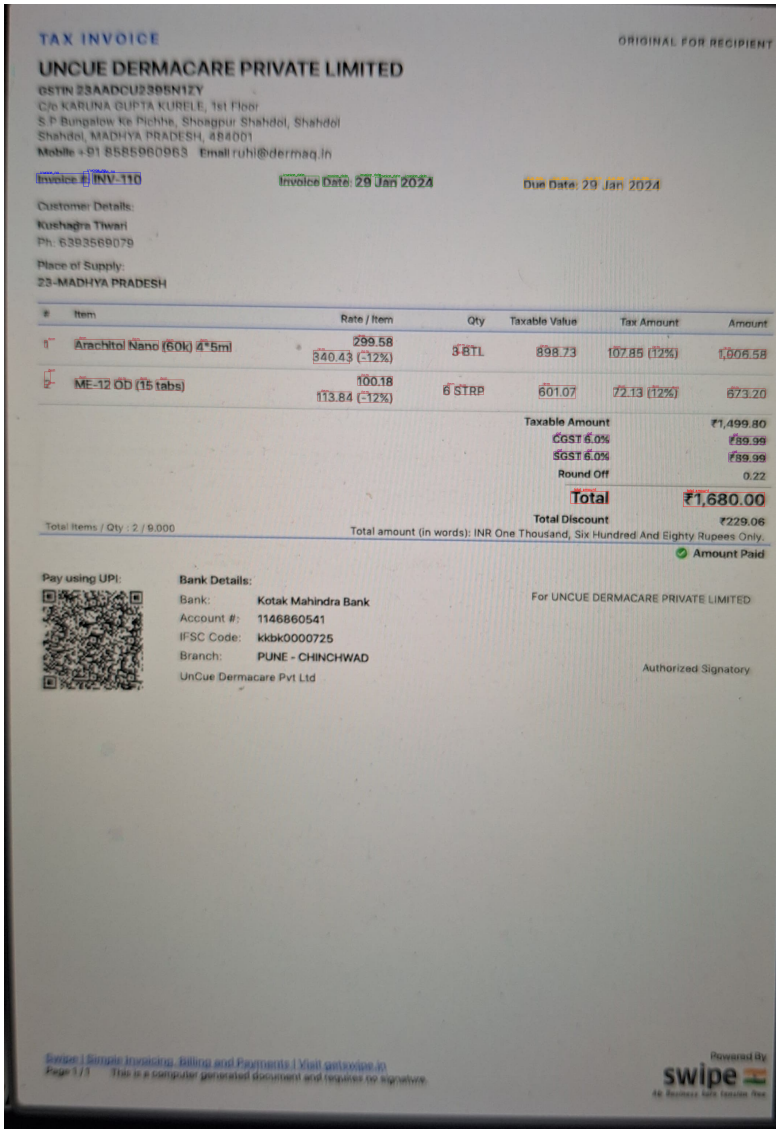


Figure 6: LayoutLMv3 is successfully able to detect bounding boxes corresponding to each label with **99.46% Accuracy** on evaluation dataset.

Step	Training Loss	Validation Loss	Precision	Recall	F1	Accuracy
100	No log	0.059284	0.891892	0.916667	0.904110	0.978202
200	No log	0.061095	0.971429	0.944444	0.957746	0.980926
300	No log	0.064241	0.971429	0.944444	0.957746	0.980926
400	No log	0.072896	0.891892	0.916667	0.904110	0.978202
500	0.045400	0.076554	0.891892	0.916667	0.904110	0.978202
600	0.045400	0.076316	0.891892	0.916667	0.904110	0.978202
700	0.045400	0.079028	0.891892	0.916667	0.904110	0.978202
800	0.045400	0.079596	0.891892	0.916667	0.904110	0.978202
900	0.045400	0.079808	0.891892	0.916667	0.904110	0.978202
1000	0.000400	0.080063	0.891892	0.916667	0.904110	0.978202

Figure 7: The training loss and other accuracy parameters after every 100 steps.

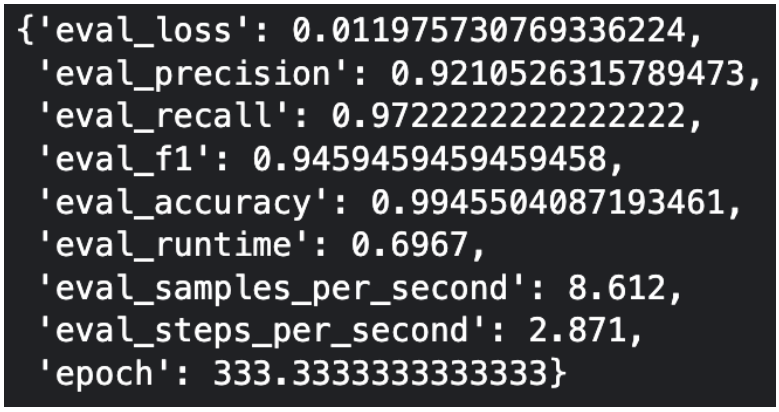


Figure 8: The evaluation loss and other accuracy parameters.

4 Text Recognition Module

I used `microsoft TrOCR` for converting the labelled images produced from LayoutLMv3 into text. Each of the bounding box images for each label type is fed into TrOCR and a resulting dictionary is formed which contains the images and corresponding text that is extracted after running through TrOCR. An example of the output is:

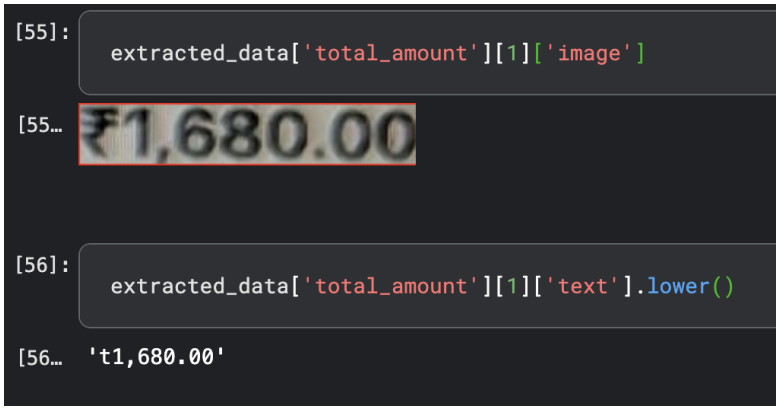


Figure 9: TrOCR applied on the image bounded by ***total_amount*** label.

5 Accuracy Check Logic Implementation & Trust Determination

Due to time constraints, I could not implement accuracy logic in code. But here I extensively analyse loopholes and try to address with potential solutions to tackle them. The proposed solution can easily be implemented in code.

Accuracy check logic should be implemented on three different sections:

- Text extracted using Tesseract: I have used Tesseract for extracting the ground truth from the images. We could however use TrOCR also for this task. However, **this is trade-off between Time and Accuracy** as TrOCR takes a significantly longer time to extract the entire textual data from the image. During production, we could however use TrOCR with good GPU for better ground truth tokens.

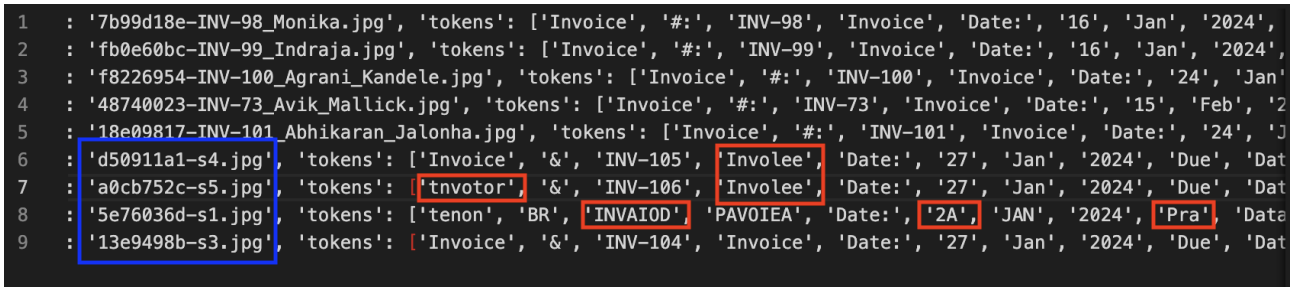


Figure 10: The red blocks shows the incorrectly converted text by Tesseract. The blue block shows that this is a problem for *Scanned* type images, and not for *Regular* type.

The Trust Determination here should be focused on each individual label separately

- **For Date fields, flag all elements which are not in proper date formats.**
- **For Invoice number, use regular expression to match given syntax. If syntax doesn't match, flag it.**
- **Put threshold (eg. $\geq$ 90%) using the the in-built text confidence score by Tesseract.**

- LayoutLMv3 Classification: Bounding Box score should be given utmost priority. Since my model works with $\geq$99% accuracy, I'm not discussing the improvement/flagging part here.
- Text Extracted using TrOCR: The text extracted using TrOCR have a
  - **100% accuracy on invoice numbers**
  - **100% accuracy on invoice date**
  - **92.59% accuracy (74 correct characters out of 81 characters) on due date**
  - **92.31% accuracy (60 correct characters out of 65 characters) on total amount**

  The average accuracy score considering all 4 labels is: **96.23%**. These accuracy scores are calculated by considering all the test images (5 in total). I haven't calculated accuracy scores for gst and item data as the number of characters are really high and we don't have proper ground truth labels.

  The test images, the labels, bounding boxes, and the texts extracted on all the 5 (the first image is a scanned type and the last four are mobile phone camera images, whose type is not seen during training) test images are as follows (grouped by labels):

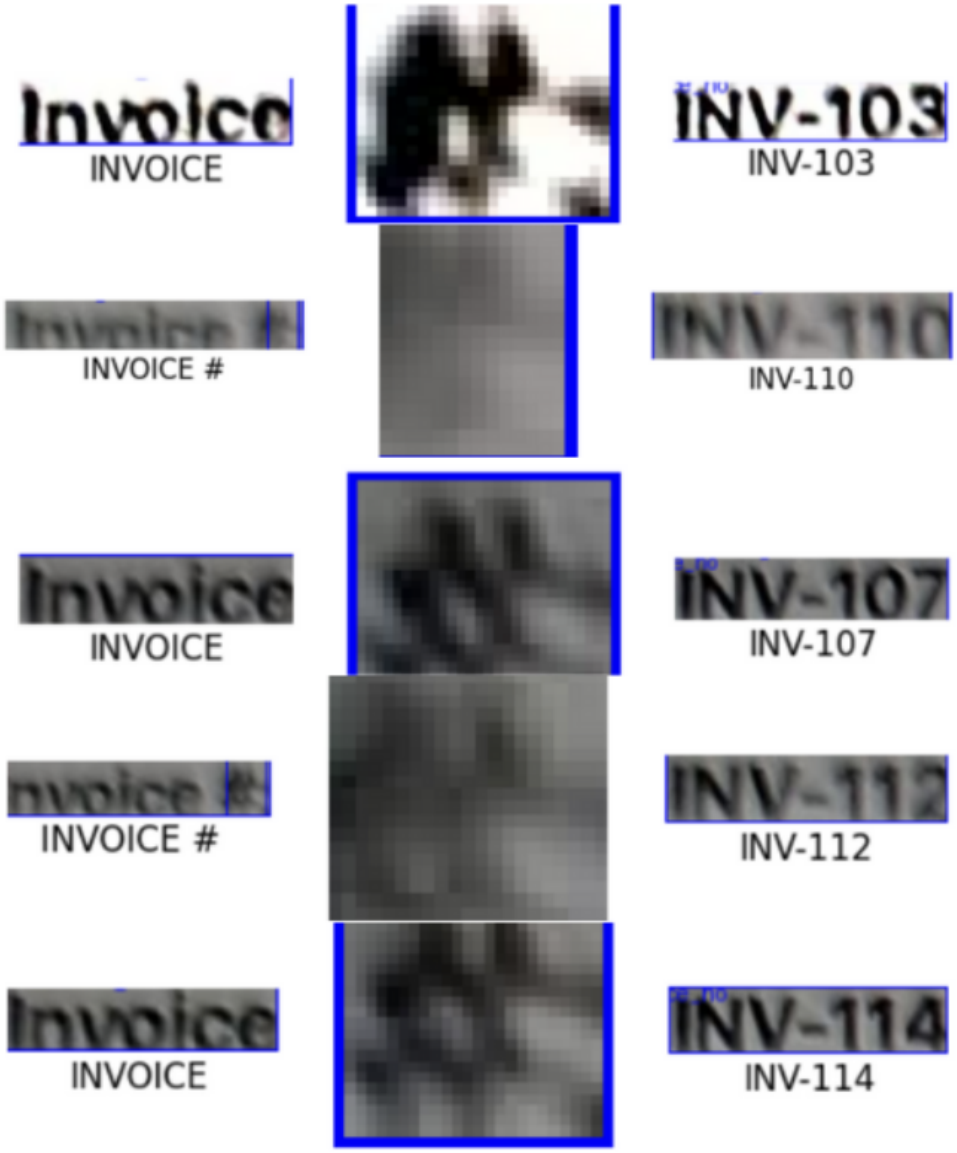


Figure 11: Invoice number

Figure 12: Invoice date

Figure 13: Due date

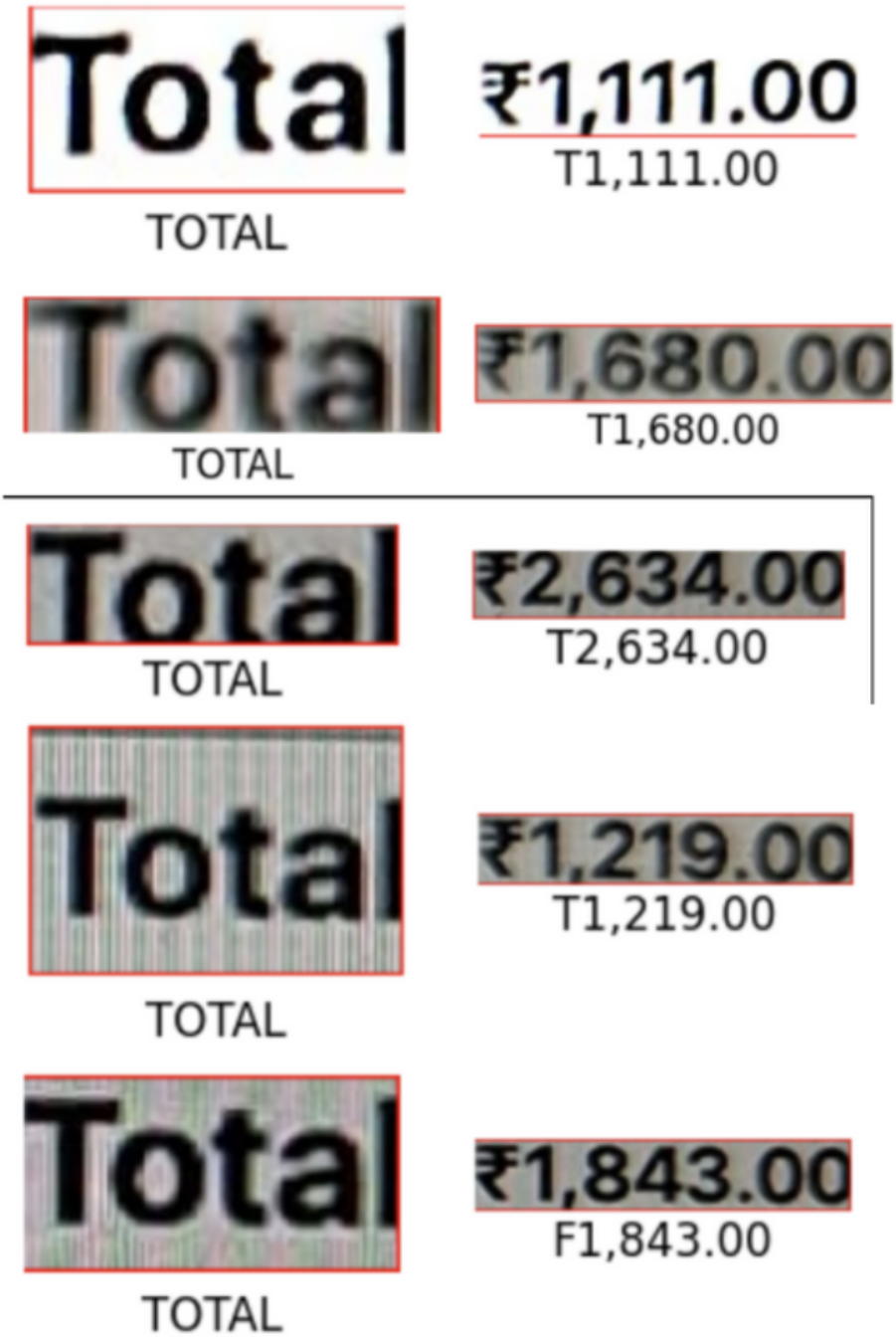


Figure 14: Total amount

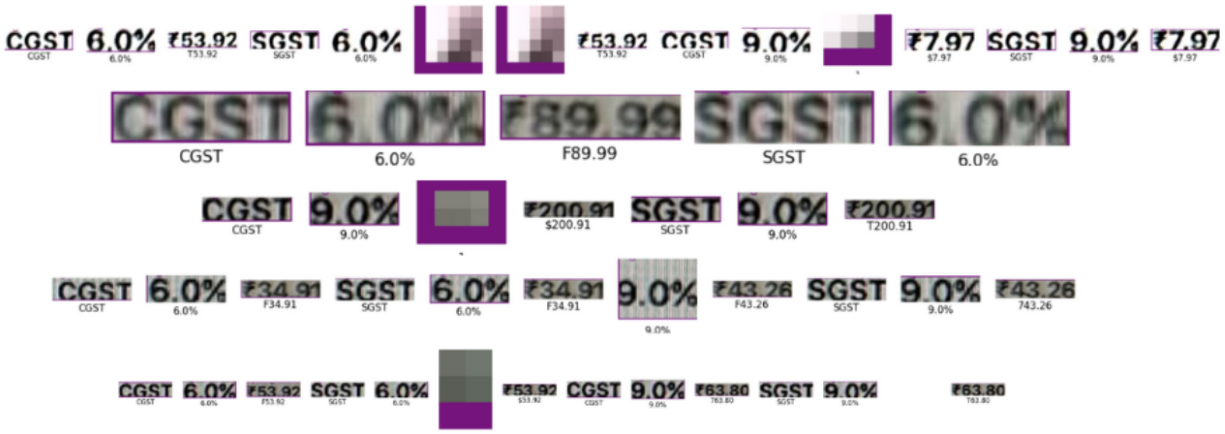


Figure 15: GST

The results shows that although the training was done only on the regular and scanned images, the model is successfully able to extract text even from a different form, i.e the images of the invoices taken using a mobile phone camera with **total accuracy of 95.71% (99.46% for LayoutLMv3 $\times$ 96.23% for TrOCR)**.