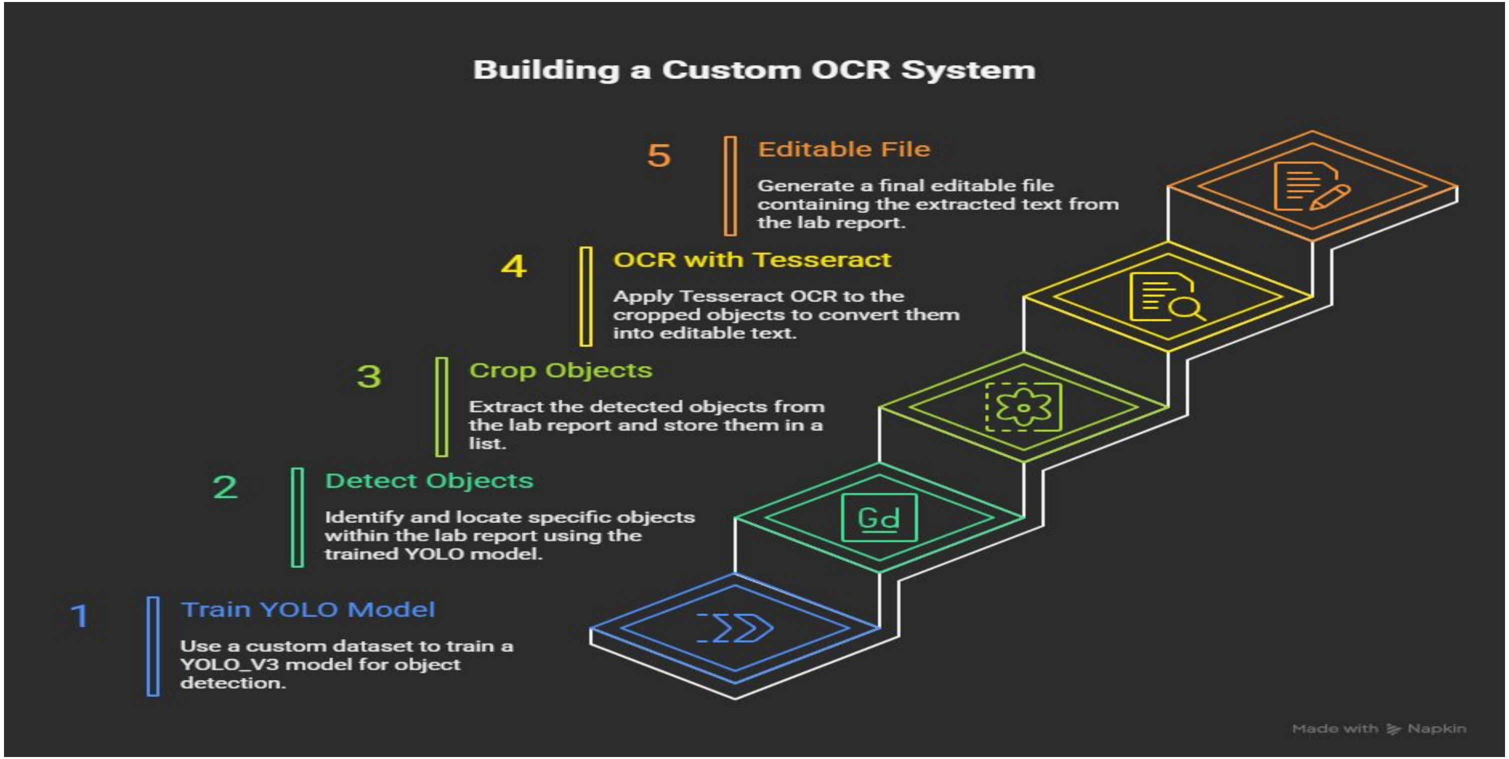
Custom-Object Character Recognition(OCR) on AWS (Google Drive/ Cloud Storage)

- Next Hikes IT Solutions
- Project_10:OCR on Google drive
- Prepared by: Kalavathi Alegapalli
- Date:10/11/2025



Building a Custom OCR System

5 **Editable File**

Generate a final editable file containing the extracted text from the lab report.

4 **OCR with Tesseract**

Apply Tesseract OCR to the cropped objects to convert them into editable text.

3 **Crop Objects**

Extract the detected objects from the lab report and store them in a list.

2 **Detect Objects**

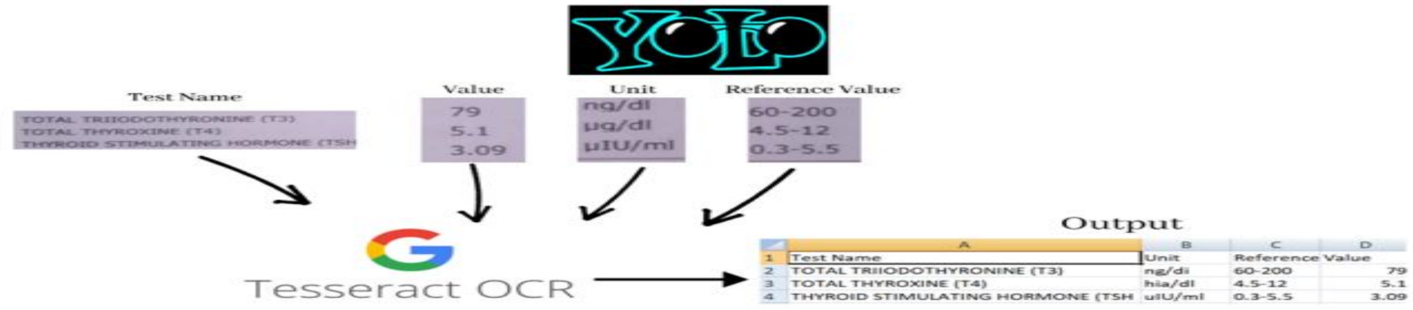
Identify and locate specific objects within the lab report using the trained YOLO model.

1 **Train YOLO Model**

Use a custom dataset to train a YOLO_V3 model for object detection.

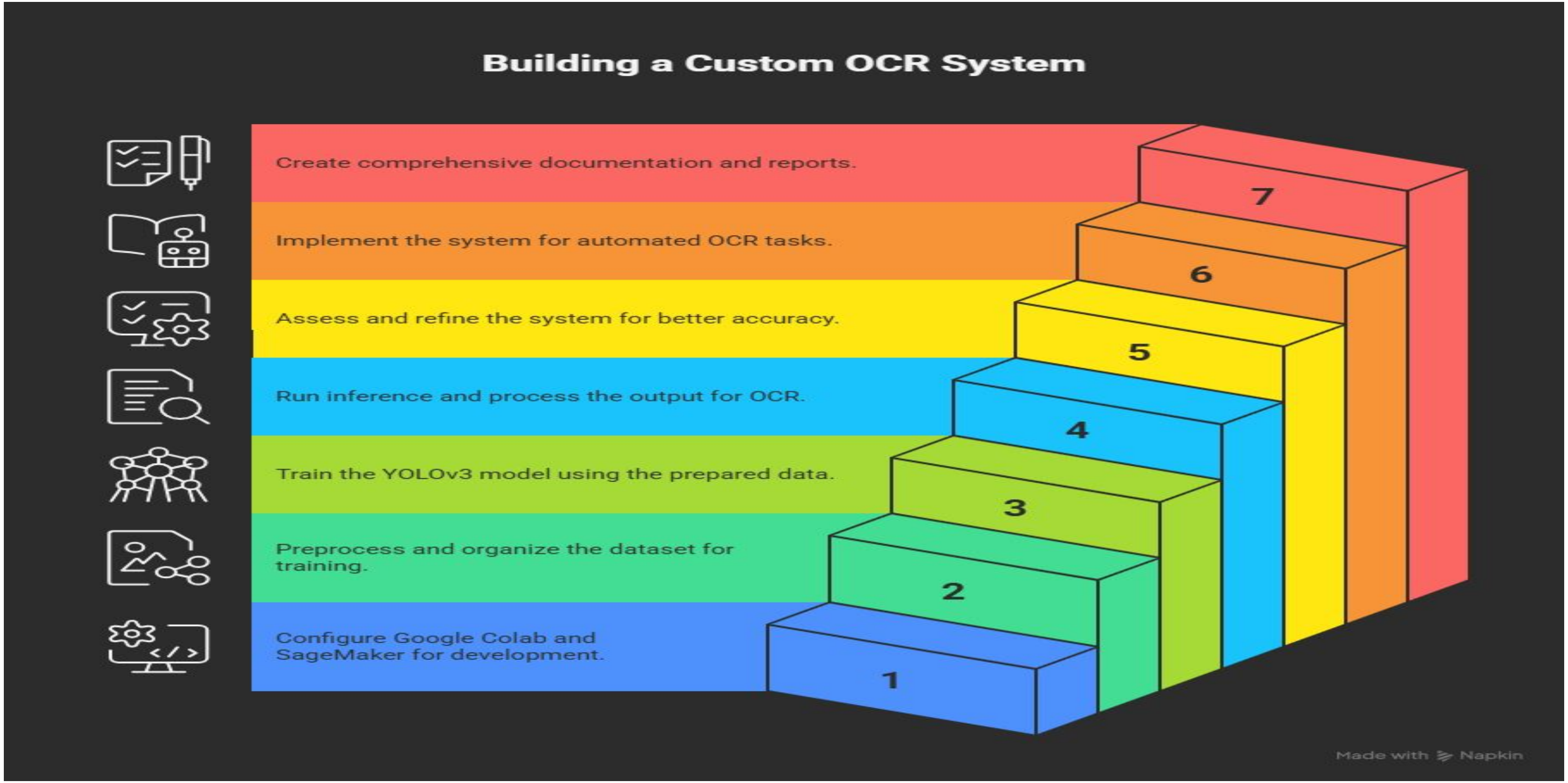
📄 Structured Medical Data Extraction Using OCR

- **Efficient Digitization**: Tesseract OCR accurately extracts test names, values, units, and reference ranges from medical reports, enabling automated data entry.
- **Structured Output**: Extracted information is organized into spreadsheet columns for streamlined analysis and record-keeping.
- **Clinical Utility**: This process supports faster diagnostics, trend tracking, and integration into electronic health systems.



	A	B	C	D
1	Test Name	Unit	Reference Value	
2	TOTAL TRIIODOTHYRONINE (T3)	ng/di	60-200	79
3	TOTAL THYROXINE (T4)	hia/dl	4.5-12	5.1
4	THYROID STIMULATING HORMONE (TSH	uIU/ml	0.3-5.5	3.09

Building a Custom OCR System



Building a Custom OCR System

Create comprehensive documentation and reports.

Implement the system for automated OCR tasks.

Assess and refine the system for better accuracy.

Run inference and process the output for OCR.

Train the YOLOv3 model using the prepared data.

Preprocess and organize the dataset for training.

Configure Google Colab and SageMaker for development.

Workflow Overview

Step 1: Label images using LabelImg (Test Name, Value, Units, Reference Range).

Step 2: Split the dataset into training and testing sets.

Step 3: Train YOLO_V5 model in Google Colab for object detection.

Step 4: Use YOLO_V5 to detect key regions in lab reports.

Step 5: Extract text using Tesseract OCR.

Step 6: Build a Streamlit app for user interaction and testing.

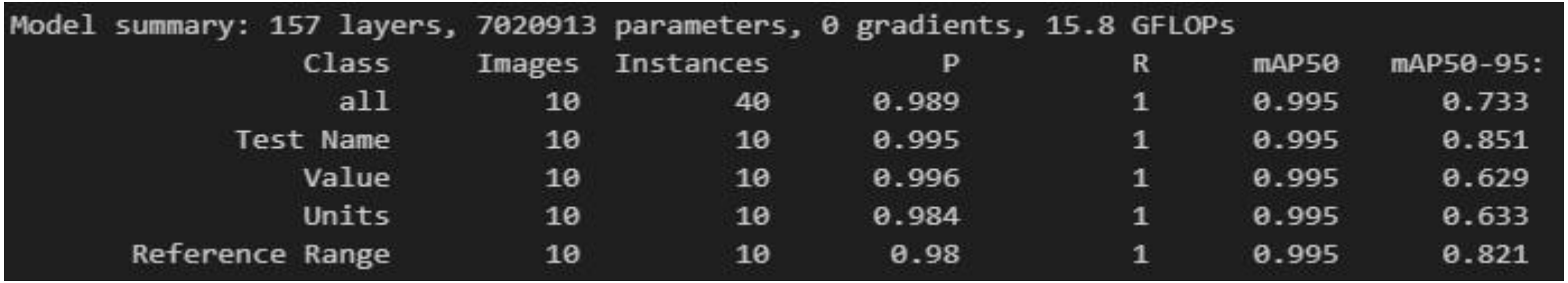
STEP 1 & 2 – IMAGE LABELING & DATA PREPARATION

- **Tool Used:** LabelImg
  → Labeled lab report images in data_images folder with bounding boxes for **Test Name, Value, Units, Reference Range** to train YOLO_V5.
- **XML Parsing & Structuring:**
  → Imported libraries for file handling and XML parsing
  → Extracted image paths and parsed XML for filenames, dimensions, and bounding boxes
  → Combined into a structured **pandas DataFrame**
- **Coordinate Normalization:**
  → Computed normalized center, width, and height for YOLO format
- **Dataset Splitting & Label Encoding:**
  → Split into **80% training / 20% testing** by unique filenames
  → Encoded object labels (Test Name, Value, Units, Reference Range)
- **Directory Setup:**
  → Created train/test folders
  → Moved images and saved labels in .txt format

3. Model Training:

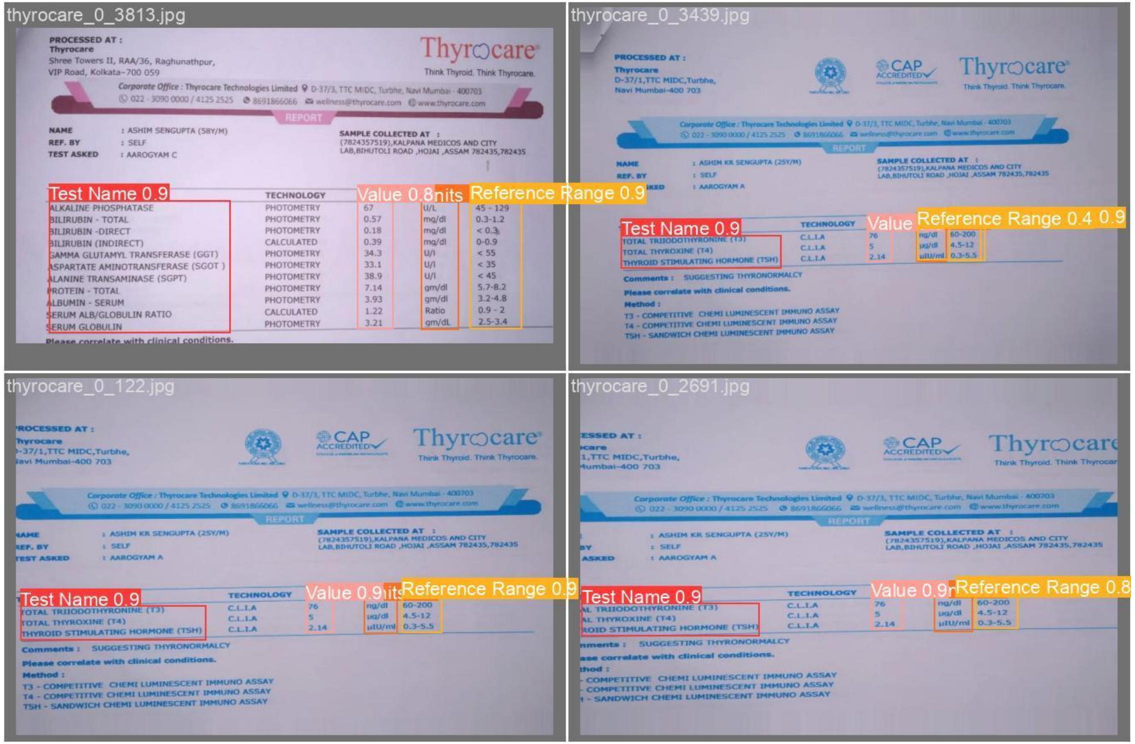
Model trained in Google Colab using YOLOv5 with pretrained weights, 640px images, batch size 12, over 200 epochs.

Achieved high precision, recall, and mAP@0.50–0.95, indicating strong, generalizable object detection performance.



```
Model summary: 157 layers, 7020913 parameters, 0 gradients, 15.8 GFLOPs
                 Class     Images  Instances          P          R      mAP50   mAP50-95:
                   all         10         40      0.989          1      0.995      0.733
             Test Name         10         10      0.995          1      0.995      0.851
                 Value         10         10      0.996          1      0.995      0.629
                 Units         10         10      0.984          1      0.995      0.633
       Reference Range         10         10       0.98          1      0.995      0.821
```

Step4:Object Detection



PROCESSED AT :
Thyrocare
Shree Towers II, RAA/36, Raghunathpur,
VIP Road, Kolkata-700 059

Thyrocare
Think Thyroid. Think Thyrocare.

Corporate Office : Thyrocare Technologies Limited D-37/3, TTC MIDC, Turbhe, Navi Mumbai - 400703
022 - 3090 0000 / 4125 2525 8691866066 wellness@thyrocare.com www.thyrocare.com

REPORT

NAME : ASHIM SENGUPTA (58Y/M)
REF. BY : SELF
TEST ASKED : AAROGYAM C

SAMPLE COLLECTED AT :
(7824357519),KALPANA MEDICOS AND CITY
LAB,BIHUTOLI ROAD ,HOJAI ,ASSAM 782435,782435

	TECHNOLOGY			
ALKALINE PHOSPHATASE	PHOTOMETRY	67	U/L	45 - 129
BILIRUBIN - TOTAL	PHOTOMETRY	0.57	mg/dl	0.3-1.2
BILIRUBIN -DIRECT	PHOTOMETRY	0.18	mg/dl	< 0.3
BILIRUBIN (INDIRECT)	CALCULATED	0.39	mg/dl	0-0.9
GAMMA GLUTAMYL TRANSFERASE (GGT)	PHOTOMETRY	34.3	U/l	< 55
ASPARTATE AMINOTRANSFERASE (SGOT)	PHOTOMETRY	33.1	U/l	< 35
ALANINE TRANSAMINASE (SGPT)	PHOTOMETRY	38.9	U/l	< 45
PROTEIN - TOTAL	PHOTOMETRY	7.14	gm/dl	5.7-8.2
ALBUMIN - SERUM	PHOTOMETRY	3.93	gm/dl	3.2-4.8
SERUM ALB/GLOBULIN RATIO	CALCULATED	1.22	Ratio	0.9 - 2
SERUM GLOBULIN	PHOTOMETRY	3.21	gm/dL	2.5-3.4

Please correlate with clinical conditions.

PROCESSED AT :
Thyrocare
D-37/1,TTC MIDC,Turbhe,
Navi Mumbai-400 703

CAP ACCREDITED

Thyrocare
Think Thyroid. Think Thyrocare.

Corporate Office : Thyrocare Technologies Limited D-37/3, TTC MIDC, Turbhe, Navi Mumbai - 400703
022 - 3090 0000 / 4125 2525 8691866066 wellness@thyrocare.com www.thyrocare.com

REPORT

NAME : ASHIM KR SENGUPTA (25Y/M)
REF. BY : SELF
TEST ASKED : AAROGYAM A

SAMPLE COLLECTED AT :
(7824357519),KALPANA MEDICOS AND CITY
LAB,BIHUTOLI ROAD ,HOJAI ,ASSAM 782435,782435

	TECHNOLOGY			
TOTAL TRIIODOTHYRONINE (T3)	C.L.I.A	76	ng/dl	60-200
TOTAL THYROXINE (T4)	C.L.I.A	5	µg/dl	4.5-12
THYROID STIMULATING HORMONE (TSH)	C.L.I.A	2.14	µIU/ml	0.3-5.5

Comments : SUGGESTING THYRONORMALCY
Please correlate with clinical conditions.
Method :
T3 - COMPETITIVE CHEMI LUMINESCENT IMMUNO ASSAY
T4 - COMPETITIVE CHEMI LUMINESCENT IMMUNO ASSAY
TSH - SANDWICH CHEMI LUMINESCENT IMMUNO ASSAY

ROCESSED AT :
hyrocare
-37/1,TTC MIDC,Turbhe,
avi Mumbai-400 703

CAP ACCREDITED

Thyrocare
Think Thyroid. Think Thyrocare.

Corporate Office : Thyrocare Technologies Limited D-37/3, TTC MIDC, Turbhe, Navi Mumbai - 400703
022 - 3090 0000 / 4125 2525 8691866066 wellness@thyrocare.com www.thyrocare.com

REPORT

AME : ASHIM KR SENGUPTA (25Y/M)
EF. BY : SELF
EST ASKED : AAROGYAM A

SAMPLE COLLECTED AT :
(7824357519),KALPANA MEDICOS AND CITY
LAB,BIHUTOLI ROAD ,HOJAI ,ASSAM 782435,782435

	TECHNOLOGY			
TOTAL TRIIODOTHYRONINE (T3)	C.L.I.A	76	ng/dl	60-200
TOTAL THYROXINE (T4)	C.L.I.A	5	µg/dl	4.5-12
THYROID STIMULATING HORMONE (TSH)	C.L.I.A	2.14	µIU/ml	0.3-5.5

Comments : SUGGESTING THYRONORMALCY
Please correlate with clinical conditions.
Method :
T3 - COMPETITIVE CHEMI LUMINESCENT IMMUNO ASSAY
T4 - COMPETITIVE CHEMI LUMINESCENT IMMUNO ASSAY
TSH - SANDWICH CHEMI LUMINESCENT IMMUNO ASSAY

ESSED AT :
care
1,TTC MIDC,Turbhe,
Mumbai-400 703

CAP ACCREDITED

Thyrocare
Think Thyroid. Think Thyrocar

Corporate Office : Thyrocare Technologies Limited D-37/3, TTC MIDC, Turbhe, Navi Mumbai - 400703
022 - 3090 0000 / 4125 2525 8691866066 wellness@thyrocare.com www.thyrocare.com

REPORT

: ASHIM KR SENGUPTA (25Y/M)
BY : SELF
ASKED : AAROGYAM A

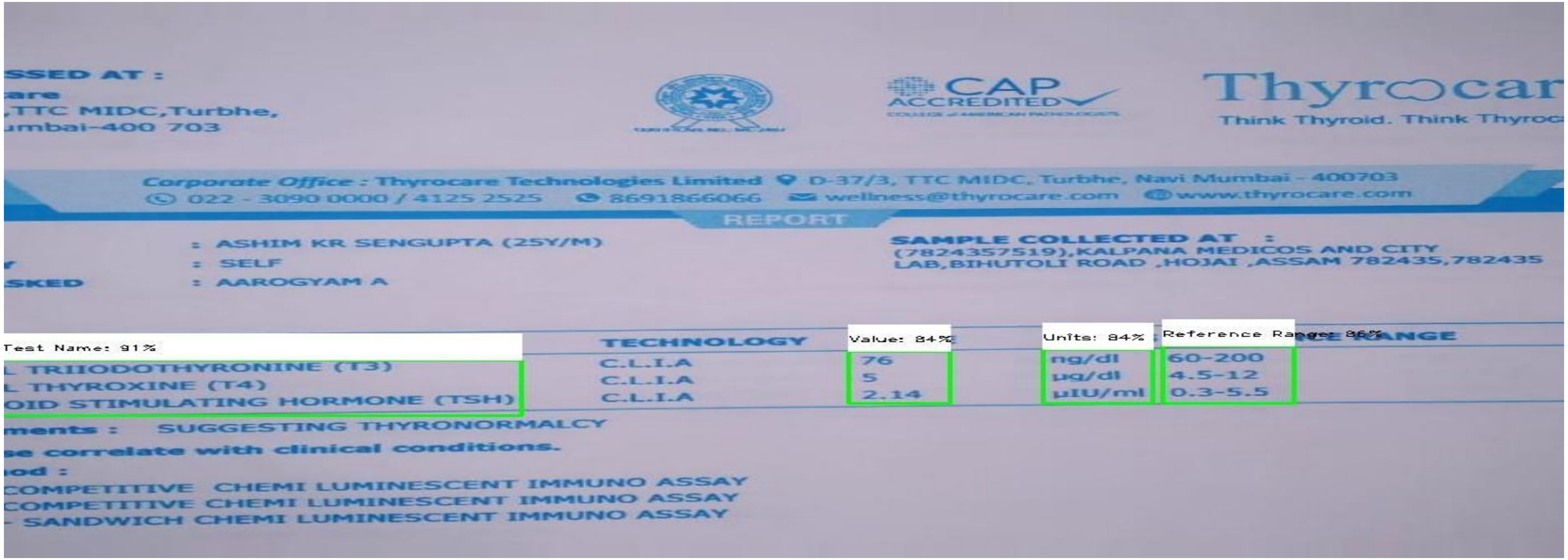
SAMPLE COLLECTED AT :
(7824357519),KALPANA MEDICOS AND CITY
LAB,BIHUTOLI ROAD ,HOJAI ,ASSAM 782435,782435

	TECHNOLOGY			
AL TRIIODOTHYRONINE (T3)	C.L.I.A	76	ng/dl	60-200
AL THYROXINE (T4)	C.L.I.A	5	µg/dl	4.5-12
ROID STIMULATING HORMONE (TSH)	C.L.I.A	2.14	µIU/ml	0.3-5.5

nments : SUGGESTING THYRONORMALCY
ase correlate with clinical conditions.
thod :
- COMPETITIVE CHEMI LUMINESCENT IMMUNO ASSAY
- COMPETITIVE CHEMI LUMINESCENT IMMUNO ASSAY
1 - SANDWICH CHEMI LUMINESCENT IMMUNO ASSAY

- The script loads YOLO from an ONNX file, preprocesses and resizes the image to 640x640, runs inference, applies NMS, and draws labeled bounding boxes.
- The final annotated image is saved for visualizing detected objects.

Step5:Text Extraction



Test Name: 91%
Value: 84%
Units: 84%
Reference Range: 86%

SSED AT :
are
,TTC MIDC,Turbhe,
umbai-400 703

Thyrocare
Think Thyroid. Think Thyroc

CAP ACCREDITED
COLLEGE of AMERICAN PATHOLOGISTS

Corporate Office : Thyrocare Technologies Limited D-37/3, TTC MIDC, Turbhe, Navi Mumbai - 400703
022 - 3090 0000 / 4125 2525 8691866066 wellness@thyrocare.com www.thyrocare.com

REPORT

: ASHIM KR SENGUPTA (25Y/M)
y : SELF
SKED : AAROGYAM A

SAMPLE COLLECTED AT :
(7824357519),KALPANA MEDICOS AND CITY
LAB,BIHUTOLI ROAD ,HOJAI ,ASSAM 782435,782435

	TECHNOLOGY			NCE RANGE
L TRIIODOTHYRONINE (T3)	C.L.I.A	76	ng/dl	60-200
L THYROXINE (T4)	C.L.I.A	5	µg/dl	4.5-12
OID STIMULATING HORMONE (TSH)	C.L.I.A	2.14	µIU/ml	0.3-5.5

ments : SUGGESTING THYRONORMALCY
se correlate with clinical conditions.
od :
COMPETITIVE CHEMI LUMINESCENT IMMUNO ASSAY
COMPETITIVE CHEMI LUMINESCENT IMMUNO ASSAY
- SANDWICH CHEMI LUMINESCENT IMMUNO ASSAY

Conclusion

- The pipeline efficiently detects and annotates objects using a YOLO model loaded from an ONNX file.
- Preprocessing ensures input images are resized and padded to meet model requirements (640×640).
- Inference and Non-Maximum Suppression (NMS) refine predictions by filtering low-confidence and overlapping detections.
- Bounding boxes with class labels and confidence scores are drawn for clear visualization.
- The final annotated image is saved, confirming successful end-to-end object detection and rendering.

References:

1. **Ultralytics YOLOv5 Official Documentation**
   Source: Ultralytics. “Train Custom Data with YOLOv5.”
   URL: https://docs.ultralytics.com/yolov5/tutorials/train_custom_data/
   Description: Official guide on preparing datasets for YOLOv5 training, including label formats, folder structure, and common dataset-related errors.

2. **Ultralytics Dataset Format Guide**
   Source: Ultralytics. “Dataset Structure and Label Formats.”
   URL: https://docs.ultralytics.com/datasets/detect/
   Description: Explains the correct directory and label structure required for YOLOv5 and YOLOv8 datasets.

3. **YOLOv5 GitHub Repository**
   Source: Ultralytics GitHub Repository.
   URL: https://github.com/ultralytics/yolov5
   Description: The open-source repository for YOLOv5, including training scripts, dataset requirements, and issue discussions related to label errors.

4. **Stack Overflow Discussion: “YOLOv5 no labels found error”**
   Source: Stack Overflow.
   URL: https://stackoverflow.com/questions/73114650/yolov5-no-labels-found-error
   Description: Community discussion confirming that missing or incorrectly structured label files cause this error, with examples of fixes.

Thank you