

Document Layout Analysis and DocVQA – Visual Question Answering for Medical Forms

1. Background & Business Purpose

The prototype was developed to address the challenge of extracting data from images of medical forms/records. The real-world business value of the developed prototype is to assist in the healthcare field by simplifying the process of extracting information from medical documents, saving time and reducing the burden of paperwork-related tasks on healthcare providers.

To achieve this aim, the project goal was to develop a tool that would be able to extract document information and answer user questions based on the data extract from the image file provided. The user input requirements for the tool were (1) an image of a medical form to be queried, and (2) the question to be answered by the model based on the uploaded form.

2. Workflow (Solution/Approach)

The following flow diagram (Figure 1, below) demonstrates how this solution was designed and implemented. A more detailed breakdown of each step in the process is then described below. In summary, the model performs image processing and then detects the layout of the form image. OCR is used to extract text data, which is then converted into a structured form for the Q&A model. Gradio is used for the front end to allow user input (Image upload and Question) and output (answer and form layout with highlighted boxes).

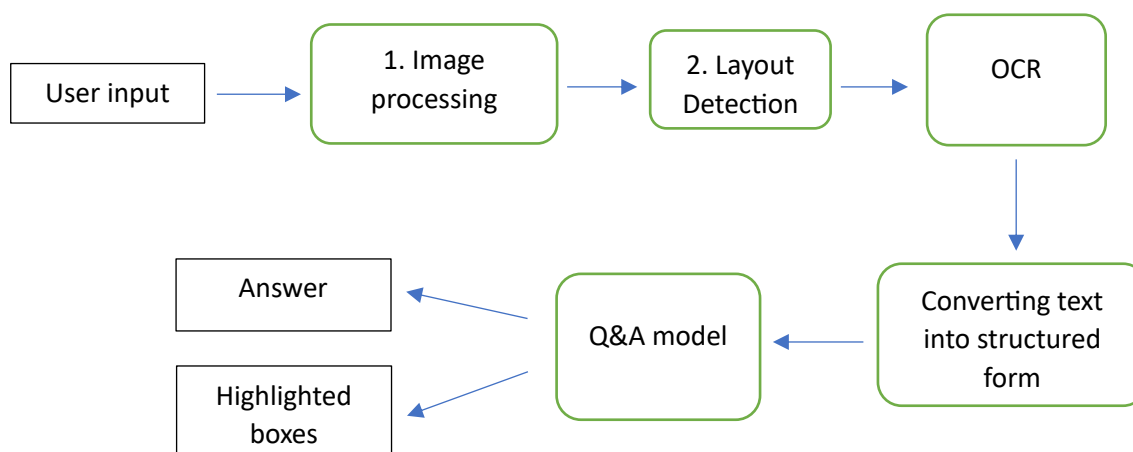


Figure 1. Document Layout Analysis and DocVQA – Visual Question Answering for Medical Forms

1. Image preprocessing

Image processing was performed using the following steps:

- Converted image to gray scale to improve simplification, computational load and focus on structure
- Image enhancement by applying median & min filter to remove noise from the image
- Enhanced image sharpness to make text boundaries more distinct and improve feature clarity for both object detection and OCR engines
- Image binarization with adaptive thresholding to increase the contrast between foreground text and background and make it easier to segment region of interest

2. Layout component detection

Multiple approaches were implemented to detect and extract the structural layout of medical document images, with the goal of identifying the most effective method.

- One approach utilized the **LayoutParser** Python library, which analyzes document structure using pretrained models.
- Another method involved applying a **pretrained object detection model** to identify layout elements such as text blocks and form fields.

After evaluation, the **object detection-based approach** demonstrated superior performance in accurately identifying and segmenting the layout components, making it the preferred method for this pipeline.

2.1 LayoutParser:


An attempt was made to use LayoutParser to extract the structural layout of medical forms. However, the model did not generalize well to the specific formatting and complexity of these documents, resulting in poor detection accuracy.


The image below (Figure 2) illustrates the output produced by LayoutParser on a sample medical form.

LABORATORY REPORT HUDSON

Patient Name:	Art Termaine	Age:	99
DOB:	1925-01-26	Gender:	male
Address:	507 Douglas Overpass Unit 38	City:	Stoughton
State:	MA	Country:	US
PostalCode:	01000	MaritalStatus:	Married
MKN:	f5c2a38a-914d-b5ab-10f3-93e427a8be02	Report Date:	Nov 16, 2014, 16:44:43

Test	Result	Units
Glucose [Mass/volume] in Serum or Plasma	77.19	mg/dL
Urea nitrogen [Mass/volume] in Serum or Plasma	16.09	mg/dL
Creatinine [Mass/volume] in Serum or Plasma	1.9304	mg/dL
Calcium [Mass/volume] in Serum or Plasma	9.59	mg/dL
Sodium [Moles/volume] in Serum or Plasma	142.79	mmol/L
Potassium [Moles/volume] in Serum or Plasma	3.88	mmol/L
Chloride [Moles/volume] in Serum or Plasma	104.3	mmol/L
Carbon dioxide, total [Moles/volume] in Serum or Plasma	22.75	mmol/L
Glomerular filtration rate/1.73 sq M predicted [Volume Rate/Area] in Serum or Plasma by Creatinine-based formula (MDRD)	81.51	mL/min/(1.73_m2)
Glucose [Mass/volume] in Urine by Test strip	0.95147	mg/dL
Bilirubin total [Mass/volume] in Urine by Test strip	0.61404	mg/dL
Ketones [Mass/volume] in Urine by Test strip	18.312	mg/dL
Specific gravity of Urine by Test strip	1.0264	(nominal)
pH of Urine by Test strip	6.1628	pH
Protein [Mass/volume] in Urine by Test strip	197.9	mg/dL


Dr. Victoria Delvalle



Looking Centre: CURAHEALTH STOUGHTON LLC

855 362-3646 louis38@example.org <https://moris.com>

All data is subject to clinical interpretation by qualified medical professionals and this report is not subject to use for any medical-legal purpose.

Figure 2. Output of LayoutParser

2.2 Object detection model:

To address the limitations of LayoutParser and better adapt to the specific characteristics of the dataset, a **pretrained object detection model**, **YOLOv10** (doclayout_yolo_docstructbench_imgs1024.pt), was employed for layout extraction. This model demonstrated significantly improved performance in identifying the structural components of the medical forms.

This approach was capable of detecting high-level layout regions, including headers, text blocks, tables, and form sections. However, it was not specifically trained to detect fine-grained, low-level elements

such as individual form fields, lines of text, or checkboxes. In this application, such low-level detection was not essential, as the objective was to extract semantically meaningful text from broader structural regions for use in downstream OCR and question-answering tasks.

Figure 3 (below) illustrates the output produced by object detection model on a sample medical form.

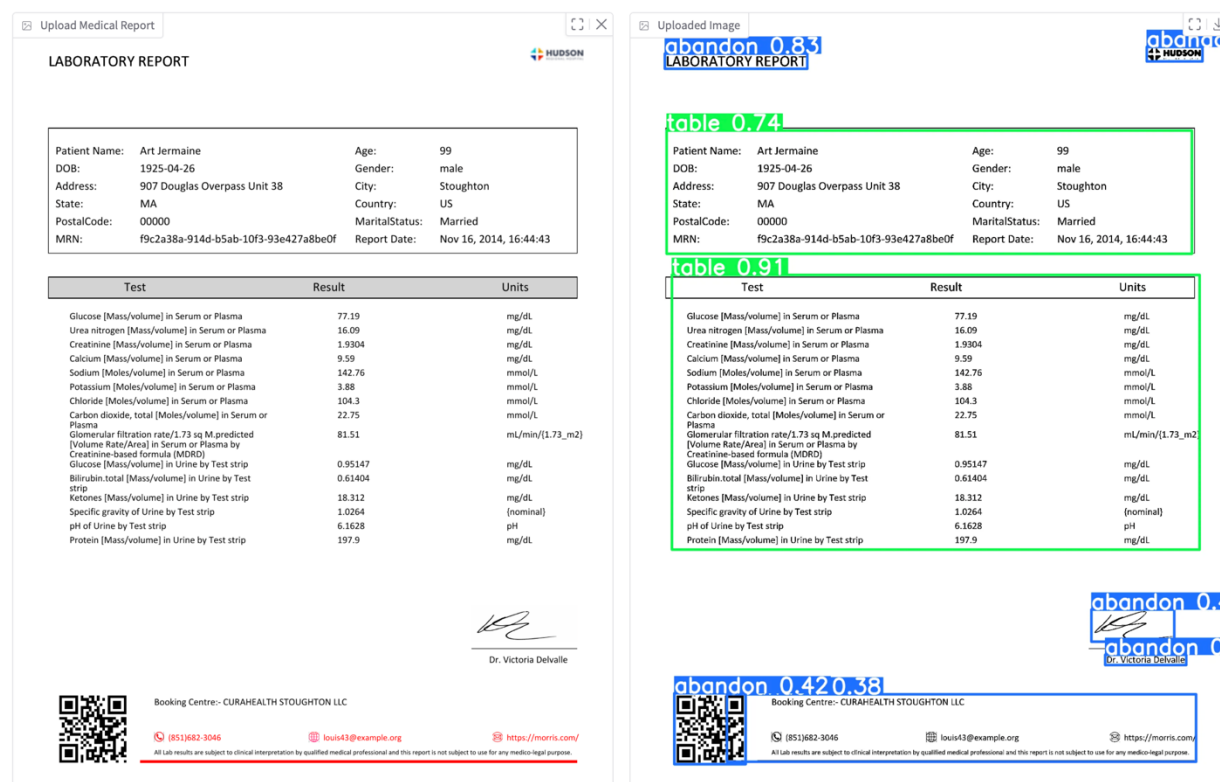


Figure 3. Output produced by Yolo object detection model

However, the detection output included several nested bounding boxes, which could potentially lead to duplicate text extraction during the OCR phase. To mitigate this, a post-processing step was implemented to eliminate nested boxes, ensuring cleaner and more accurate downstream text extraction.

4. OCR and NLP for field extraction

The following models were compared in order to evaluate which method of OCR gave the best performance. In the final model, Tesseract (4.2, below) was used because of its superior performance.

4.1 PaddleOCR

Initial text extraction was performed using PaddleOCR. However, the results were inconsistent, as the extracted text did not follow a uniform reading direction (e.g., left-to-right or top-to-bottom). This lack of directional consistency made it challenging to structure the extracted content into a coherent and reliable format.

4.2 Tesseract

Text extracted using Tesseract OCR followed a consistent reading direction, which made it easier to identify key-value pairs and separate them accurately. This consistency greatly facilitated the transformation of raw OCR output into a structured format.

5. Transforming unstructured text to structured format

To convert the unstructured OCR output from Tesseract into a more organized form, regular expressions (Regex) were used to extract specific fields and format them as dictionaries. An example of the structured output generated from Tesseract OCR results is included in Figure 4.

```

Patient information: {'patient_name': 'Art Jermaine', 'age': '99', 'dob': '1925-04-26', 'gender': 'male', 'address': '907 Douglas Overpass Unit 38', 'city': 'Stoughton', 'state': 'MA', 'country': 'US', 'postal_code': '00000', 'marital_status': 'Married', 'mrn': 'f9c2a38a-914d-b5ab-10f3-93e427a8be0f', 'report_date': 'Nov 16, 2014, 16:44:43'}

Patient test results: {'Test Result Units Glucose [Mass/volume] in Serum or Plasma': '77.19 mg/dL', 'Urea nitrogen [Mass/volume] in Serum or Plasma': '16.09 mg/dL', 'Creatinine [Mass/volume] in Serum or Plasma': '1.9304 mg/dL', 'Calcium [Mass/volume] in Serum or Plasma': '9.59 mg/dL', 'Sodium [Moles/volume] in Serum or Plasma': '142.76 mmol/L', 'Potassium [Moles/volume] in Serum or Plasma': '3.88 mmol/L', 'Chloride [Moles/volume] in Serum or Plasma': '104.3 mmol/L', 'Carbon dioxide, total [Moles/volume] in Serum or Plasma': '22.75 mmol/L', 'Glomerular filtration rate/1.73 sq M.predicted': '81.51 mL/min/{1.73_m2}', '[Volume Rate/Area] in Serum or Plasma by Creatinine-based formula (MDRD) Glucose [Mass/volume] in Urine by Test strip': '0.95147 mg/dL', 'Bilirubin.total [Mass/volume] in Urine by Test': '0.61404 mg/dL', 'Ketones [Mass/volume] in Urine by Test strip': '18.312 mg/dL', 'Specific gravity of Urine by Test strip': '1.0264 {nominal}', 'pH of Urine by Test strip': '6.1628 pH', 'Protein [Mass/volume] in Urine by Test strip': '197.9 mg/dL'}

Clinic Information: {'Doctor name': 'Dr. Victoria Delvalle', 'Booking Centre': 'CURAHEALTH STOUGHTON LLC', 'Phone Number': '(851)682-3046', 'Email': 'louis43@example.org', 'URL': 'https://morris.com/'}

```

Figure 4. structured output from OCR detection

6. Question & Answering model

Utilized a HuggingFace question-answering model (model=deepset/roberta-base-squad2) to extract insights from transformed form records.

7. Utilizing Gradio for Interface

The interface allows users to upload medical report images, view layout detection overlays, and ask questions to extract patient details, test results, and booking information using a QA pipeline. Below is a screenshot of the Gradio interface integrated with the NLP and layout analysis pipeline.

The screenshot displays the Gradio interface for a medical report analysis pipeline. It features a table of test results, a Q&A section, and a submission button.

Test	Result	Units
Glucose [Mass/volume] in Serum or Plasma	77.19	mg/dL
Urea nitrogen [Mass/volume] in Serum or Plasma	16.09	mg/dL
Creatinine [Mass/volume] in Serum or Plasma	1.9304	mg/dL
Calcium [Mass/volume] in Serum or Plasma	9.59	mg/dL
Sodium [Moles/volume] in Serum or Plasma	142.76	mmol/L
Potassium [Moles/volume] in Serum or Plasma	3.88	mmol/L
Chloride [Moles/volume] in Serum or Plasma	104.3	mmol/L
Carbon dioxide, total [Moles/volume] in Serum or Plasma	22.75	mmol/L
Glomerular filtration rate/1.73 sq M.predicted [Volume Rate/Area] in Serum or Plasma by Creatinine-based formula (MDRD)	81.51	mL/min/{1.73_m2}
Glucose [Mass/volume] in Urine by Test strip	0.95147	mg/dL
Bilirubin.total [Mass/volume] in Urine by Test strip	0.61404	mg/dL
Ketones [Mass/volume] in Urine by Test strip	18.312	mg/dL
Specific gravity of Urine by Test strip	1.0264	{nominal}
pH of Urine by Test strip	6.1628	pH
Protein [Mass/volume] in Urine by Test strip	197.9	mg/dL

Below the table, there is a section for asking questions. The question "What is the glucose?" is entered, and the answer "77.19 mg/dL" is displayed. A "Flag" button is also present.

The interface includes a QR code, a booking center name (CURAHEALTH STOUGHTON LLC), a phone number ((851)682-3046), an email (louis43@example.org), and a URL (https://morris.com/). It also features a signature of Dr. Victoria Delvalle and a disclaimer: "All Lab results are subject to clinical interpretation by qualified medical professional and this report is not subject to use for any medical legal purposes."

Figure 5. Screenshot of the Gradio interface integrated Document Layout Analysis and DocVQ

8. Limitations / design trade-offs

What does the current model do?

- It segments the layout and successfully extracts and transforms unstructured text to meaningful form.
- It supports questions related to booking information, patient details, and test results—returning precise answers from structured context.
- Currently supports processing one medical report image at a time.

What doesn't the model do?

- Cannot answer multiple questions – for example: what is the body height & heart rate? The current model only gives output for the first field body height.
- Cannot structure the dense layout structure of medical form like Ontario health referral forms where it has multiple different structures
- Cannot handle multiple pages report

9. Future directions

- The current model can be further specialized by training an object detection model to identify structures in more complex medical forms
- Implement a tracking feature to flag changes over time for improved longitudinal analysis of real-world patient data
- Add support for multi-file handling to enable batch processing of documents
- Introduce a summarization feature to highlight key findings, allowing for rapid evaluation of lengthy forms or large volumes of data