

Samuel Leal Rodríguez

# OCR-Based Pipeline for Table Extraction from Spine Dual X-Ray Absorptiometry Reports

MASTER'S THESIS

Directors:

Dr. Oliver Díaz Montesdeoca

Dr. José Andrés Román Ivorra

## Master's Degree in Health Data Science



Valencia, 2025

# Table of Contents

|                                                                 |    |
|-----------------------------------------------------------------|----|
| 1. Background .....                                             | 1  |
| 1.1. Osteoporosis: A silent disease .....                       | 1  |
| 1.2. Imaging principles of DXA .....                            | 2  |
| 1.3. DXA accessibility and screening alternatives .....         | 4  |
| 1.4. The problem of Health Data Integration .....               | 6  |
| 1.5. Contribution of the current work .....                     | 7  |
| 2. Objectives .....                                             | 8  |
| 3. State-of-the-art .....                                       | 9  |
| 3.1. Opportunistic screening of osteoporosis .....              | 9  |
| 3.2. Methods for extracting tables from images .....            | 12 |
| 4. Design and development .....                                 | 15 |
| 4.1. Step 1. Inspection of images: Identification of ROIs ..... | 16 |
| 4.2. Step 2. Annotating the ROIs .....                          | 20 |
| 4.3. Step 3. Cropping the tables .....                          | 21 |
| 4.4. Step 4. Refining the table boundaries. ....                | 21 |
| 4.5. Step 5. Image to Text conversion with OCR. ....            | 23 |
| 4.6. Step 6. Parsing OCR output & Data frame creation. ....     | 25 |
| 5. Experiments and results .....                                | 27 |
| 5.1. Experiments .....                                          | 27 |
| 5.1.1. Cropping with OpenCV Template matching .....             | 27 |
| 5.1.2. Table layout detection with img2Table .....              | 28 |
| 5.2. Results .....                                              | 29 |
| 5.2.1. Performance on the patient_data table .....              | 29 |
| 5.2.2. Performance on the exam_data table .....                 | 32 |
| 5.2.3. Performance on the dmo_data table .....                  | 33 |
| 5.2.4. Resulting unified data frame .....                       | 38 |
| 6. Discussion .....                                             | 39 |
| 7. Conclusions and future work .....                            | 43 |
| 8. References .....                                             | 45 |
| 9. Appendix .....                                               | 48 |

# List of Tables

|                                                                                                  |    |
|--------------------------------------------------------------------------------------------------|----|
| Table 1. WHO definition of osteoporosis.....                                                     | 3  |
| Table 2. Summary of key studies on DL-Based BMD estimation from plain radiographs. ....          | 11 |
| Table 3. Number of images according to acquisition year .....                                    | 16 |
| Table 4. Main OCR engines available in Python.....                                               | 24 |
| Table 5. Performance of the pipeline on the patient_data tables in the training dataset.....     | 29 |
| Table 6. Performance of the pipeline on the patient_data tables in the test dataset.....         | 30 |
| Table 7. Performance of the pipeline on the exam_data tables in the training dataset.....        | 32 |
| Table 8. Performance of the pipeline on the exam_data tables in the test dataset.....            | 33 |
| Table 9. Summary performance of the pipeline for the 'dmo_table' containing the BMD values. .... | 34 |
| Table 10. Performance metrics for 'dmo_table' in the test dataset.....                           | 38 |
| Table 11. Frameworks for document layout and table recognition. ....                             | 42 |
| Table S1. OCR performance for PyTesseract and EasyOCR on a sample 'dmo_table'.....               | 48 |
| Table S2. View of the resulting unified data frame.....                                          | 49 |

# List of figures

|                                                                                       |    |
|---------------------------------------------------------------------------------------|----|
| Figure 1. SCOPE 2021 Summary Report for Osteoporosis in Europe. ....                  | 1  |
| Figure 2. Example image of a DXA Scan. ....                                           | 3  |
| Figure 3. Summary of the pipeline. ....                                               | 15 |
| Figure 4. Example of one raw DXA report. ....                                         | 17 |
| Figure 5. Resulting tables after applying coordinate cropping. ....                   | 22 |
| Figure 6. Refined tables after cropping.. ....                                        | 23 |
| Figure 7. Results of obtaining the tables using template matching. ....               | 28 |
| Figure 8. Table Layout detection using img2table. ....                                | 29 |
| Figure 9. The original sample 'dmo_table' after refinement. ....                      | 36 |
| Figure 10. Native Surya-OCR image recognition. ....                                   | 36 |
| Figure 11. Sample 'dmo_table' after morphological kernel and background removal. .... | 36 |
| Figure 12. Surya OCR character recognition after background removal. ....             | 37 |

# List of Abbreviations

|                 |                                            |
|-----------------|--------------------------------------------|
| <b>AI</b>       | Artificial Intelligence                    |
| <b>AM</b>       | Age-matched                                |
| <b>API</b>      | Application Programming Interface          |
| <b>AUC</b>      | Area Under the Curve                       |
| <b>BMC</b>      | Bone Mineral Content                       |
| <b>BMD</b>      | Bone Mineral Density                       |
| <b>CEIm</b>     | Clinical Research Ethics Committee         |
| <b>CNN</b>      | Convolutional Neural Network               |
| <b>CRNN</b>     | Convolutional Recurrent Neural Network     |
| <b>CSV</b>      | Comma Separated Value                      |
| <b>CT</b>       | Computed Tomography                        |
| <b>DL</b>       | Deep Learning                              |
| <b>DXA/DEXA</b> | Dual X-Ray Absorptiometry                  |
| <b>EHR</b>      | Electronic Health Record                   |
| <b>FCN</b>      | Fully Convolutional Network                |
| <b>FRAX</b>     | Fracture Risk Assessment Tool              |
| <b>GPU</b>      | Graphics Processing Unit                   |
| <b>HU</b>       | Hounsfield Units                           |
| <b>JSON</b>     | JavaScript Object Notation                 |
| <b>LIS</b>      | Laboratory Information System              |
| <b>LSTM</b>     | Long Short-Term Memory                     |
| <b>ML</b>       | Machine Learning                           |
| <b>NN</b>       | Neural Network                             |
| <b>OCR</b>      | Optical Character Recognition              |
| <b>PACS</b>     | Picture Archiving and Communication System |
| <b>PSM</b>      | Page Segmentation Mode                     |
| <b>RM</b>       | Maximum Reference (Referencia Máxima)      |
| <b>ROI</b>      | Region of Interest                         |
| <b>RGB</b>      | Red Green Blue                             |
| <b>TCO</b>      | Total Cost of Ownership                    |
| <b>TD</b>       | Table Detection                            |
| <b>TR</b>       | Table Recognition                          |
| <b>TSR</b>      | Table Structure Recognition                |
| <b>UDCA</b>     | Clinical Documentation and Admissions Unit |
| <b>VRAM</b>     | Video Random Access Memory                 |
| <b>WHO</b>      | World Health Organization                  |
| <b>XML</b>      | eXtensible Markup Language                 |

# 1. Background

## 1.1. Osteoporosis: A silent disease

The term osteoporosis (from Greek *osteo* or ‘bone’ and *poros* or ‘pore’) describes a disease where bone density and mass are decreased, making bones weaker and more susceptible to fractures. In Europe, **more than 32 million individuals** aged 50 or more are estimated to have osteoporosis, equivalent to 5.6% of the total European population over 50, or approximately 25.5 million women (22.1% of women aged +50) and 6.5 million men (6.6% of men aged +50) (1–3).

Individuals with osteoporosis in the EU27+2



Prevalence of osteoporosis across the EU27+2 in 2019<sup>3</sup>



*Figure 1. SCOPE 2021 Summary Report for Osteoporosis in Europe.*

Because the main risk factor for developing osteoporosis is age, it does not only cause fractures, but it also causes people to become bedridden with **secondary complications** that may be life threatening in the elderly. Since osteoporosis also causes back pain and loss of height, prevention of the disease and its associated fractures is essential for maintaining health, quality of life, and independence among the elderly (4,5).

**Osteoporosis is a silent disease**, meaning that it causes no symptoms to the patients, and very often the diagnosis is made late in the course of the disease (e.g.: when bone fractures are already present). Until recently, osteoporosis was an under-recognized disease and considered an inevitable consequence of ageing. Careful attention has to be paid to people having several risk factors

for its development, such as female gender, menopause, prior gastric surgery, unhealthy lifestyle habits, chronic diseases or family history of fractures (6).

Fortunately, improvements in diagnostic technology and assessment facilities over the past decades now mean that it is possible to detect the disease before fractures occur. **The cornerstone of diagnosis nowadays is the measurement of bone mineral density by Dual X-Ray Absorptiometry**, commonly known as a DXA or DEXA scan, a highly precise and low-radiation imaging technique that measures bone mineral density (BMD). Currently, DXA is now considered the gold standard for measuring BMD and assessing fracture risk (7).

## 1.2. Imaging principles of DXA

At its core, a DXA scan utilises two X-ray beams with different energy levels. These beams are directed towards the bones, typically at the hip and lumbar spine, as these are common sites for osteoporotic fractures. The fundamental physics lies in the differential X-ray absorption behaviour of these X-ray photons by bone and soft tissue.

The bone, being much denser than the surrounding soft tissues, absorbs more of the X-ray energy than soft tissue. By measuring the amount of each X-ray beam that passes through the body, the DXA machine's software can calculate a precise value for bone mineral content. This is then used to determine the BMD, a key indicator of bone strength.

The clinical procedure itself is straightforward. The patient lies on a padded table while a scanning arm passes over the targeted area. The radiation dose from a DXA scan is minimal, significantly lower than a standard chest X-ray, making it a very safe diagnostic tool.



**Figure 2. Example image of a DXA Scan.** The X-ray detector system is on the arm above the table. The X-ray source lies below.

The primary importance of DXA in the context of osteoporosis is its ability to provide a quantitative and accurate assessment of bone density. The results of a DXA scan are typically reported as a "T-score" and a "Z-score". Diagnostic thresholds offered by the World Health Organization have been widely accepted (Table 1) (8), and these are:

| WHO classification         | T-Score                   |
|----------------------------|---------------------------|
| Normal                     | $\geq -1.0$               |
| Low Bone Mass (Osteopenia) | $< -1.0$ to $> -2.5$      |
| Osteoporosis               | $\leq -2.5$               |
| Severe Osteoporosis        | $\leq -2.5$ with fracture |

**Table 1. WHO definition of osteoporosis.**

- **T-score:** This is the most critical value for diagnosing osteoporosis in postmenopausal women and men over 50. It compares the measured BMD to that of a healthy young adult of the same sex at their peak bone mass. A T-score of  $-1.0$  or above is considered normal. A T-score between  $-1.0$  and  $-2.5$  indicates osteopenia, which is a state of lower-than-normal bone

density that may progress to osteoporosis. A T-score of -2.5 or below is the diagnostic threshold for osteoporosis.

- **Z-score:** This score compares measured BMD to that of an average person of the same age, sex, and ethnicity. A significantly low Z-score may suggest that a factor other than age is contributing to bone loss.

The early detection afforded by DXA is paramount, as it enables timely intervention through lifestyle modifications, dietary changes, and medication to help preserve bone mass and reduce the likelihood of debilitating fractures.

### 1.3. DXA accessibility and screening alternatives

The International Osteoporosis Foundation has reported that in several regions, such as Eastern Europe, Central Asia, the Middle East, Africa, and Latin America, **access to DXA is insufficient** (9). Even in well-equipped countries like the US, DXA remains under-utilised. This highlights the need for alternative strategies for the early detection of osteoporosis and fracture risk stratification.

Despite its prevalence, awareness of the risk factors for osteoporosis among the general population can be low. Many people may attribute a loss of height or a stooped posture to normal ageing, rather than potential spinal fractures caused by osteoporosis. This can lead to a failure to seek medical advice. Furthermore, in a busy primary care setting, **assessing osteoporosis risk may not always be a top priority**, leading to missed opportunities for referral.

Even when a referral for a DXA scan is made, **patients can face lengthy waiting lists**. It is our personal experience in our institution that waiting time for a DXA scan since ordered can be as long as 6-12 months. This translates into a delay in the early identification of the disease and narrows the time window for optimal preventive strategies to be established.

**Opportunistic screening** using medical imaging has emerged as a viable strategy to identify at-risk populations without additional costs or extra radiation exposure. Various studies have explored the use of computed

tomography (CT) to estimate BMD, classify osteoporosis, and predict fracture risk (10,11). However, its implementation in clinical settings faces limitations related to radiation dose and accessibility. In contrast, **conventional radiography offers a more accessible and lower-cost alternative**, with sufficient spatial resolution to visualise bone texture, which correlates with bone density and the detection of osteoporotic fractures (12).

Lower back pain has the highest global prevalence among musculoskeletal conditions and is the leading cause of disability worldwide. It is estimated that low back pain affects more than 619 million people globally, with a projection of 843 million prevalent cases by 2050 (13). People of any age can suffer from low back pain, including children and adolescents. The highest number of cases occurs between the ages of 50 and 55, and women suffer from it more frequently than men. The prevalence of lower back pain and the resulting disability are highest among people aged 80 to 85. Recurrent episodes of lower back pain become more common with age.

When a patient seeks medical attention for low back pain, the first imaging technique that is indicated due to its high accessibility and low cost is a plain **X-ray of the lumbar spine**, most often with anteroposterior and lateral views. Because osteoporosis and low back pain share common risk factors, they tend to occur in the same population rich in postmenopausal women. Thus, **the likelihood of having a lumbar spine X-ray available for a late middle-aged person in most modern healthcare settings is higher than any other imaging modalities**. Using already existent information in electronic health records (EHRs) for the screening of disease is a rapidly growing area of research. Recent studies are leveraging advanced techniques, particularly machine learning (ML) and artificial intelligence (AI), to analyse the complex information contained within EHRs to predict and screen for a wide range of conditions, mostly cardiovascular disease, chronic kidney disease or even dementia (14–16).

Advances in AI and deep learning (DL) have enabled the development of automated tools for the analysis of medical images, with promising results in

fracture detection, BMD assessment, and fracture risk estimation from plain radiographs (17). In a recent study, a DL-based algorithm achieved correlations of 0.92 and 0.90 between estimated and DXA-measured BMD in the hip and lumbar spine, respectively. These results suggest that DL algorithms can be effective tools for expanding osteoporosis screening and reducing the burden associated with fragility fractures.

## 1.4. The problem of Health Data Integration

Healthcare systems today generate vast amounts of data across multiple platforms and formats, creating a complex digital ecosystem. Modern medical institutions rely on various specialised systems including Picture Archiving and Communication Systems (PACS) for medical imaging, EHRs for patient documentation, Laboratory Information Systems (LIS) for test results, and numerous monitoring devices that capture real-time physiological data. Each system serves a specific purpose but often operates in isolation, creating what is commonly referred to as data silos.

The **fragmentation of healthcare data** presents significant challenges for clinical decision-making and research advancement. In the specific case involving DXA scans, critical clinical information such as patient demographics, imaging system information and bone mineral measurements is **embedded within JPEG images** stored in the PACS. Healthcare professionals must manually extract and transcribe this information, a process that is **time-consuming, prone to human error, and fundamentally unscalable** when dealing with large patient populations or retrospective studies.

The transformation of unstructured medical images into structured, machine-readable data opens numerous possibilities for advanced clinical analytics. An automated system that can extract structured information from medical images not only reduces administrative burden but also improves data consistency and reduces the potential for transcription errors.

This project seeks to develop a DL-based algorithm capable of detecting osteoporosis and estimating BMD using plain radiographs, with the aim of **reducing waiting times for DXA and the costs associated with diagnosis**. This strategy has the potential to **improve the early detection of osteoporosis**, allowing for early intervention and better management of fracture risk in the general population.

## 1.5. Contribution of the current work

The current work presents an Optical Character Recognition (OCR)-based pipeline unlocks the potential of the unexploited information contained in DXA reports by converting raw images into editable, integrable and workable text. **The pipeline enables to:**

- **Extract information from large image datasets** in a fast, accurate and unassisted way. **Batch-processing DXA images** can generate big datasets that can be further used to train ML or DL models for osteoporosis and fracture prediction.
- **Integrate complete and error-free DXA information into EHRs**, where plain text information is recorded.
- **Save time to healthcare professionals** by eliminating the manual copy and paste tactic employed until now.

## 2.Objectives

The **primary objective** of this project is to develop a pipeline that:

- Selects and extracts tables of interest from a DXA report as image format (JPEG),
- Converts each table into text using OCR technology, and
- Merges the data extracted from all tables in a unified data frame containing all the information of interest.

This pipeline is framed within a larger main branch of a research project seeking to estimate bone mineral density and fracture risk from plain X-ray radiographs of the lumbar spine. Workable data in a tabular readable format is a prerequisite to develop more complex models.

The **secondary objectives** of our project are:

- Enabling healthcare professionals to reduce data entry errors in EHRs and improve efficiency by reducing hand typing DXA results.
- Facilitating Health Data Integration of unstructured image data for clinical use or further research.

## 3.State-of-the-art

### 3.1. Opportunistic screening of osteoporosis

The use of routine scans for opportunistic osteoporosis screening was one of the earliest and most successful applications of this concept. The methodology is based on the principle that the X-ray attenuation values measured by a CT scanner, expressed in Hounsfield Units (HU), are directly correlated with tissue density. By measuring the mean HU value within the trabecular (spongy) bone of a vertebral body—typically the first lumbar vertebra (L1), as it is commonly included in both chest and abdominal CT scans—a reliable estimate of bone density can be obtained (18). The clinical and economic impact of this technology has been substantial. A recent large-scale study of the Medicare population demonstrated that opportunistic CT screening could more than double the overall screening rate (an increase of 113%) without any additional imaging (19). The study projected that the early identification and subsequent treatment of at-risk patients could result in annual healthcare cost savings of nearly \$2.5 billion in the U.S. by preventing costly fragility fractures.

While CT-based screening is highly effective, the most ubiquitous, low-cost, and lowest-radiation imaging modality in medicine is the plain radiograph, or X-ray. Consequently, developing methods to estimate BMD from routine X-rays of the hip and spine represents the forefront of research into accessible, population-scale opportunistic screening. The challenge is greater than with CT, as a 2D radiograph is a projection image where the attenuation information from bone is superimposed with that of overlying soft tissues. However, the fine spatial resolution of radiographs allows for the visualization of subtle bone textures and trabecular patterns that are correlated with bone density.

DL, particularly the use of deep Convolutional Neural Networks (CNNs), has proven exceptionally capable of learning these complex, subtle image signatures that are often invisible to the human eye but predictive of BMD

(17). A multitude of studies have successfully applied various CNN architectures to this task, including well-established models like VGG, *ResNet*, *EfficientNet*, and *DenseNet*, often leveraging transfer learning to enhance performance (20).

A landmark study in this domain is that of *Hsieh et al.* which is of direct relevance to this thesis (17). The authors developed a fully automated, three-step DL pipeline:

- Automated ROI extraction with detection of key anatomical landmarks on pelvis or lumbar spine radiographs. These landmarks are then used to precisely and automatically extract the ROIs for the hip and individual lumbar vertebrae (L1-L4).
- Automated quality assessment by dedicated neural networks (NNs) that automatically check the quality of the extracted ROIs, excluding any that contain fractures, surgical implants, or other morphological abnormalities that would confound a BMD measurement.
- BMD prediction using CNNs (using VGG-16 and ResNet-34 backbones) that takes the clean ROIs as input and predicts the corresponding DXA-equivalent BMD value. The model also uses the predicted BMD to calculate a T-score and estimate the 10-year fracture risk using the FRAX® tool (7).

The performance of this system, validated on a large retrospective cohort of over 25,000 patients, was remarkable. The Pearson's correlation coefficient between the AI-predicted BMD and the ground-truth DXA-measured BMD was exceptionally high, reaching 0.92 for the hip and 0.90 for the lumbar spine. The model demonstrated excellent diagnostic accuracy for identifying osteoporosis (defined as a T-score  $\leq -2.5$ ), achieving an area under the curve (AUC) of 0.89 and an overall accuracy of 91.7% for the hip. In a real-world clinical simulation, the tool was able to confidently classify or rule out osteoporosis for nearly 85% of patients, suggesting that its widespread implementation could potentially reduce the number of required DXA scans by a similar margin.

This work is part of a growing body of evidence demonstrating the high potential of this approach. Table 2 summarizes key findings from several influential studies in the field, illustrating the consistency of results across different anatomical sites and model architectures.

| Study (Year)             | Imaging Modality           | DL Architecture(s)            | Correlation with DXA BMD             | AUC for Classification           |
|--------------------------|----------------------------|-------------------------------|--------------------------------------|----------------------------------|
| Hsieh et al. (2021) (17) | Pelvis AP, Lumbar Spine AP | VGG-16, ResNet-34 Ensemble    | Hip: 0.92, Spine: 0.90 (Pearson's r) | Hip: 0.89, Spine: 0.89           |
| Hong et al. (2023) (21)  | Lateral Spine Radiograph   | EfficientNet-B4               | Not Reported                         | 0.85 (Internal), 0.83 (External) |
| Sato et al. (2022) (22)  | Chest X-ray                | ResNet-50                     | Not Reported                         | 0.84                             |
| Jang et al. (2021) (23)  | Hip AP Radiograph          | VGG-16 with Non-local Network | Not Reported                         | 0.87                             |
| Zhang et al. (2023) (24) | Lumbar Spine X-ray         | Not Specified (CNN)           | Not Reported                         | 0.77                             |

*Table 2. Summary of key studies on DL-Based BMD estimation from plain radiographs.*

### 3.2. Methods for extracting tables from images

The automated conversion of visual information within document images into a structured, machine-readable format is a central problem in document analysis, formally known as table recognition (TR). This process is more complex than simple OCR as it requires not only recognizing characters but also understanding their spatial and logical relationships. The TR task is universally deconstructed into two primary, sequential sub-tasks: Table Detection (TD) and Table Structure Recognition (TSR) (25).

Before the widespread adoption of DL, TR was primarily addressed through heuristic or rule-based approaches. These methodologies leverage classical computer vision algorithms and a set of predefined rules, often tailored to a specific document domain, to identify and parse tables based on explicit visual cues such as lines, whitespace, and text alignment. The core principle of this paradigm is to exploit the predictable formatting conventions of structured documents. Several key techniques form the foundation of heuristic TR systems, such as coordinate-based cropping, colour-based segmentation, line detection, contour analysis and Hough transform.

More recently, there has been a paradigm shift in computer vision, with data-driven DL models supplanting many classical, rule-based systems. This revolution has extended to document analysis and table recognition, reframing the problem from one of rule engineering to one of learning feature representations directly from data. Instead of relying on handcrafted features like lines and whitespace, DL models, particularly CNNs, learn to identify the salient visual patterns that define a table through training on vast datasets of examples.

An alternative approach frames the problem as one of semantic segmentation, where the goal is to classify every pixel in the image as belonging to a specific class (e.g., table, column, row, or background). This is often accomplished using Fully Convolutional Networks (FCNs), such as the U-Net architecture. These models produce a pixel-wise mask that can precisely delineate tables of arbitrary shapes. Influential systems like *DeepDeSRT* employ this

strategy, using separate FCNs to generate masks for the table region and for the column separators within it, which are then combined to reconstruct the table structure (26).

Some research has focused on developing unified, end-to-end models that perform both table detection and structure recognition within a single network. *TableNet*, for instance, uses an encoder-decoder architecture where the encoder learns a shared feature representation of the input image, and two separate decoder branches use this representation to simultaneously predict a mask for the table region and a mask for the columns (27).

The success of this data-driven paradigm is critically dependent on two key factors. First is the availability of large-scale, publicly annotated datasets. Datasets like *Marmot* and, more recently, the massive *PubTabNet*—containing over 500,000 table images with corresponding structured HTML representations—have been instrumental in training and benchmarking these data-hungry models (28). Second is the concept of transfer learning. Given the relative scarcity of annotated document images compared to natural scene images, a common and highly effective strategy is to initialize a model with weights pre-trained on a large natural image dataset (like *ImageNet*) and then fine-tune it on the smaller, domain-specific table dataset. This allows the model to leverage general visual features learned from millions of images and adapt them to the specific task of document analysis, significantly improving performance and reducing training data requirements.

The choice between a heuristic and a DL approach for table recognition represents a fundamental trade-off between specialization and generalization. A critical analysis reveals that neither paradigm is universally superior; rather, **the optimal methodology is dictated by the specific constraints and characteristics of the problem domain.**

DL models represent the state-of-the-art in terms of raw performance on diverse and complex documents. Their core strength is generalization: the ability to accurately process documents with varied and previously unseen

layouts without modification. This robustness comes at a significant cost. **DL models are computationally intensive**, often requiring specialized hardware like GPUs for acceptable performance. They are data-dependent, necessitating large, meticulously labelled datasets for training. Furthermore, their decision-making processes are often opaque, earning them the label of "black box" models, which can make debugging and interpretation challenging.

On the other hand, **heuristic approaches are computationally lightweight**, exceptionally fast on standard CPUs, require no training data, and their rule-based logic is highly interpretable ("white box"). Their well-documented weakness is brittleness; they are rigidly designed for a specific format and will fail if the layout deviates from the expected rules.

This analysis leads to a crucial conclusion regarding the design of the data extraction pipeline for this thesis. The input documents—DXA reports generated by a specific Hologic® system over a defined five-year period—are highly standardized and template-based. They exhibit minimal to no variation in the layout or position of the key data tables. In such a constrained environment, the primary value proposition of the DL approach—its ability to handle layout diversity—is nullified. The problem is not one of general "document understanding" but of deterministic "data location."

Therefore, employing a complex, resource-intensive DL model for this task would be an inefficient allocation of computational resources, expending significant effort to solve a problem (locating the tables) that is already solved by the inherent consistency of the input data. Conversely, the primary weakness of the heuristic approach—its brittleness—is completely neutralized by this same structural consistency. In this context, a well-designed heuristic pipeline, such as the one developed for this project, is not a compromise or a technically inferior choice. It is the most pragmatic and efficient engineering solution, one that is intentionally specialized and optimized for the specific constraints of the task at hand.

## 4.Design and development

We aimed to design a pipeline capable of taking one or more raw DXA images as input and producing a Comma Separated Value (CSV) file in long format as output, containing the target information from the tables of interest. A summary of the steps followed by the pipeline are depicted in Figure 3.

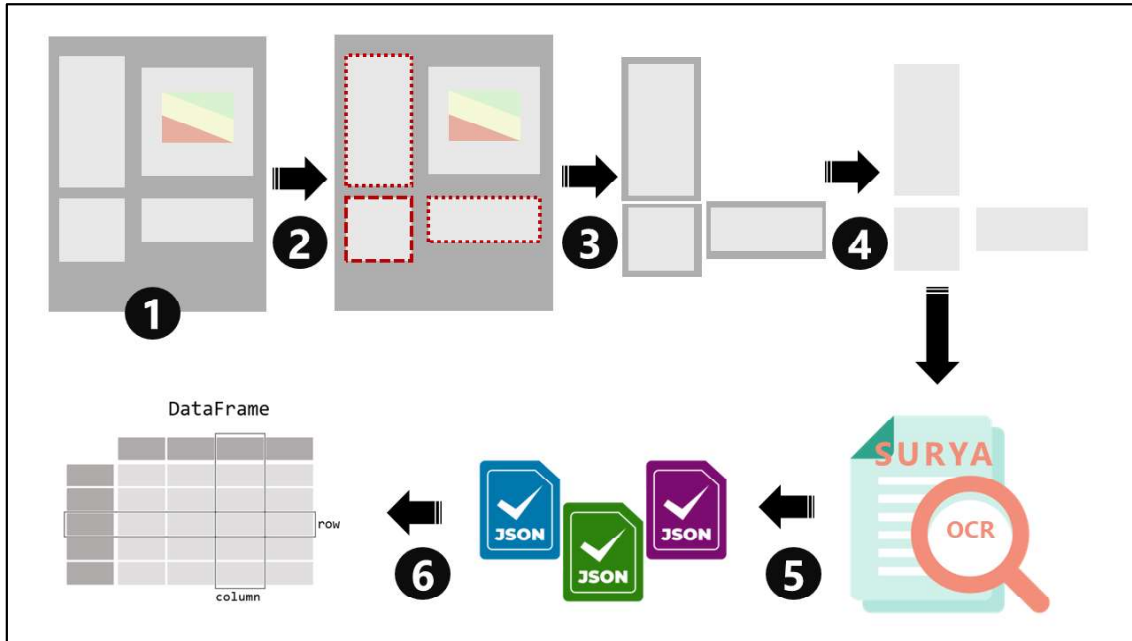


**Figure 3. Summary of the pipeline.** The numbers represent the sequential steps. Steps 1 and 2 are only performed for the development of the pipeline.

The pipeline was developed on a **local computer** with the following specifications: Operative System *Windows 11*, *Python 3.13.5*, *PyTorch 2.8.0+cu128*, *CUDA 12.8* and GPU *NVIDIA GeForce RTX 4060* with 8.6 GB VRAM.

This project obtained the **approval of the Clinical Research Ethics Committee** (CEIm for its acronym in Spanish) from the Hospital Universitari i Politècnic La Fe (Registry Number 2025-0493-1). The raw DXA images were requested to the institution's Clinical Documentation and Admissions Unit (UDCA for its acronym in Spanish), responsible for the management of the clinical-care information system.

The UDCA provided a primer of **100 de-identified DXA images** for developing and testing the pipeline. Specifically, the training dataset comprised 70

lumbar spine DXA images. The test dataset was made of 30 lumbar spine DXA images that were never used for the development of the pipeline and only employed for performance evaluation. As shown in Table 3, the proportion of reports for the different years is kept between the training and test dataset.

| <b>Year</b>                     | <b>2025</b>   | <b>2024</b>   | <b>2023</b>   | <b>2022</b> | <b>2021</b> | <b>2020</b> | <b>Total</b> |
|---------------------------------|---------------|---------------|---------------|-------------|-------------|-------------|--------------|
| <b>Training</b><br><b>(70%)</b> | 17<br>(24.3%) | 27<br>(38.6%) | 12<br>(17.2%) | 6<br>(8.6%) | 5<br>(7.2%) | 3<br>(4.3%) | <b>70</b>    |
| <b>Test</b><br><b>(30%)</b>     | 7<br>(23.3%)  | 11<br>(36.6%) | 5<br>(17%)    | 3<br>(10%)  | 2<br>(6.7%) | 2<br>(6.7%) | <b>30</b>    |

*Table 3. Number of images according to acquisition year and dataset origin.*

#### 4.1. Step 1. Inspection of images: Identification of ROIs.

During the first phase of the project design (Figure 3, step 1), a small subset of 10 DXA report images (2 reports per year) were analysed by our team to manually identify the layout and the location of the regions of interest (ROIs). The densitometry images in the dataset were 24-bit RGB colour images in JPEG (.jpg) format and uniform dimensions (1420x1200 pixels) with an estimated size of 200-300 KB per image. The whole image background was dark grey (RGB 128,128,128). Table borders and cell backgrounds were uniformly light grey (RGB 222,222,222). An example of DXA report image is shown in Figure 4.

Our team collaborated with rheumatologists with valuable experience on the field of bone health and DXA interpretation to identify the most important areas for data extraction. Reports from the past five years (2020 – 2025) were analysed, so that variations in the layouts (if any) were considered. Major position changes of the ROIs within the image were not present for any of the reports examined. Only minor variations in x and y coordinates of the ROIs could be found between reports of different years.

Eventually, three tables were identified as targets for their identification and extraction: ‘patient\_table’, ‘exam\_table’ and ‘dmo\_table’. A fourth region common to all DXA reports (‘graph’ in Figure 2) consisted of a rectangular area with a miniature image depicting the anteroposterior aspect of the scanned lumbar segments (left) plus a colour map indicating the patient’s risk of fracture according to age and BMD values (right). Our team was advised not to extract this graph area because its sole purpose was to improve the patient’s legibility of the document. Besides, this information was redundant because it was detailed within the dmo\_table (Figure 4). For this reason, this ‘graph’ area was excluded from the pipeline.

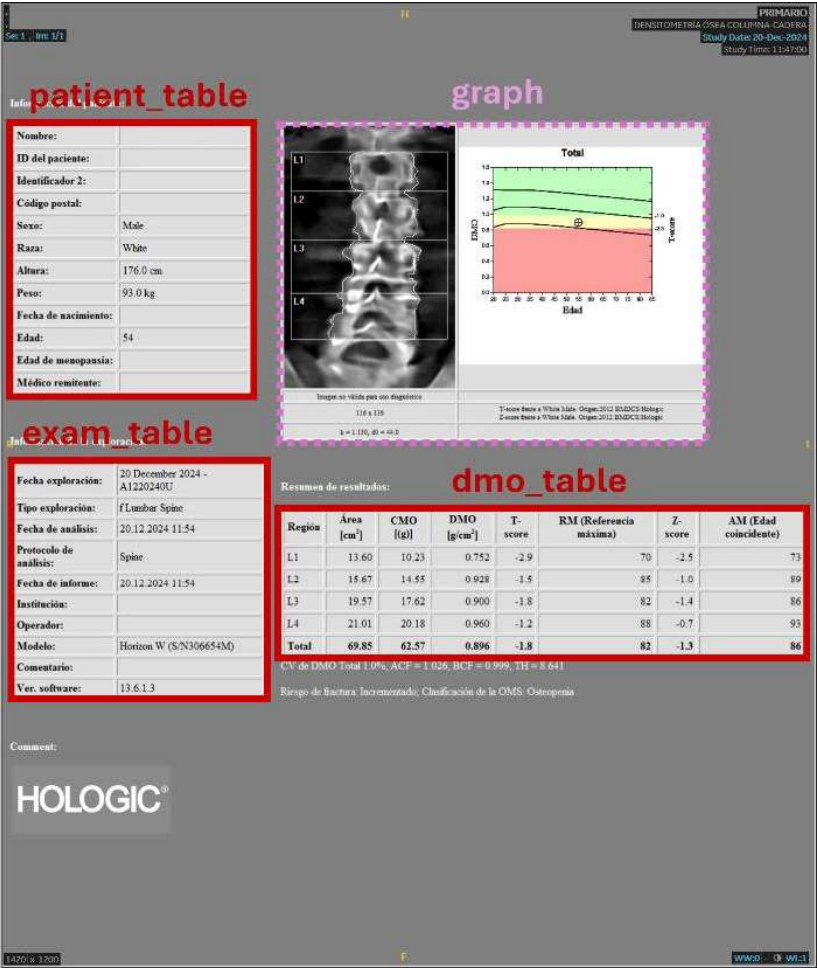


**Figure 4.** Example of one raw DXA report containing the three tables of interest for extraction (solid red borders) and the merely informative graph area (dotted pink border). Sensitive patient data has been removed.

The patient table contains demographic information about the patient. It is a 2x2 table, containing key-value (left – right) pairs. The fields in this table invariably were:

1. Name. Patient's name and surname in a unique field.
2. Patient's ID. Unique identifier for the patient in the whole demographic area.
3. Second ID. Second optional unique identifier.
4. Postal Code of the patient.
5. Gender.
6. Race.
7. Height in cm.
8. Weight in kg.
9. Date of birth in dd.mm.yyyy format.
10. Age.
11. Age of menopause
12. Referring doctor.

Some fields were always empty in the training dataset: ID 2, Postal Code and Referring Doctor. Besides, the expert group considered this information unnecessary, so these fields were excluded from the extraction pipeline. The final set of target variables for extraction were nine: surname, id, sex, race, age, date of birth, weight, height, age of menopause.

The order of the rows in this table was invariable in the reports examined for training. However, it is important to note that there were variations in the language of the keys and values. Most of the reports were in Spanish, but a few of them were displayed in English. This had to be taken into consideration when designing a solution for detecting and extracting the text (discussed in Chapter 4).

The exam table contains information about the acquisition system and the circumstances of the exam. It is also a 2x2 table, containing key-value (left – right) pairs. The fields in this table are always:

1. Scan Date. Date of the scan in dd.mm.yyyy hh:mm format.
2. Exploration Type. Scan location: Spine, wrist or hip.
3. Analysis Date. Same as Scan Date and Report Date.
4. Analysis Protocol. Same value as Exploration Type.
5. Report Date. Same as Scan Date and Analysis Date.
6. Institution.
7. Operator.
8. Model
9. Comments
10. Software version

Many fields in the 'exam\_table' were redundant. For example, there were no different dates for Scan, Analysis or Report (they all pointed to the same date and time of data acquisition). The final fields selected as of interest from this table were three: Exploration Type, Analysis Date and Model. The fields Institution, Operator and Comments were always empty in the training dataset and thus were excluded from the pipeline.

The dmo ('*densidad mineral ósea*' or bone mass density) table is the most important as it contains information about the measurements of bone mass and its comparison with healthy (T-score) and same age (Z-score) cohorts.

Unlike the others, this table has a variable number of 3 to 6 rows (header, vertebral levels and total) but a fixed number of 8 columns. In most cases, a complete exam comprises 6 rows and 8 columns. The variation in the number of rows depends on the number of vertebral levels analysed. The standard protocol involves measuring the bone mineral density (BMD) from L1 to L4 vertebrae, while L5 is typically excluded as recommended by the International Society for Clinical Densitometry (29). However, some vertebral levels can be missing from the report if the patient has metallic implants, previous fracture or significant degenerative changes at a given vertebral level. In these cases, the affected vertebrae are excluded from analysis by the operator because falsely high BMD measurements can be taken. In this table,

the variables of interest are the columns, so that for each vertebral level we have:

1. **Region**. 'L' for lumbar + number of vertebrae.
2. Area (**A**) of the vertebrae in anteroposterior view in cm<sup>2</sup>.
3. Bone Mineral Content (**BMC**) in grams. Total amount of mineral content in the vertebrae.
4. Bone Mineral Density (**BMD**). Principal measure. It is the result of BMC / A. Represents the mineral density of the bone.
5. **T-Score**. See Introduction for details.
6. Maximum Reference (**RM**). Represents the age-matched equivalent or "Bone Age" for that specific vertebral level.
7. **Z-Score**. See Introduction for details.
8. Age-matched (**AM**). Age-match against the opposite gender's reference population.

The target fields commonly used in clinical practice are region, area, BMC, BMD, T-score and Z-Score. Comparisons with the reference population are taken from T and Z-scores, and not from age-matched variables (RM and AM). These two variables were not excluded from the pipeline, but no optimization for their extraction or parsing was performed.

## 4.2. Step 2. Annotating the ROIs.

The first step in the pipeline (see Figure 3, step 2) began with manual annotation of the ROI using *LabelImg*, an open-source graphical image annotation tool (30). Initially, 20 DXA reports were manually segmented to obtain the coordinates of the three ROIs or tables.

Each image was annotated using *PascalVOC* XML format, generating bounding box coordinates (xmin, ymin, xmax, ymax) for each table region. To ensure comprehensive table capture across all images in the dataset, universal coordinates were derived by calculating the extreme values from all annotated samples:

- Patient Data Table: xmin=8, xmax=401, ymin=169, ymax=637

- Examination Data Table: xmin=8, xmax=391, ymin=670, ymax=1087
- DMO Data Table: xmin=401, xmax=1190, ymin=707, ymax=984

These universal coordinates were selected using a conservative approach (minimum xmin/ymin, maximum xmax/ymax) to prevent any table content from being cropped during the extraction process. This was necessary because, although all images in the dataset had identical dimensions, there were slight (few pixels) variations in the layout structure.

### **4.3. Step 3. Cropping the tables.**

The original densitometry images were processed using *OpenCV* to extract the three table regions based on the universal coordinates (Figure 3, step 3). For each input image, three separate ROI crops were generated using array slicing operations. This step produced initial table crops that contained the target tables but often included surrounding background areas and potential noise, like table titles or footnotes (Figure 5).

### **4.4. Step 4. Refining the table boundaries.**

To precisely delineate table boundaries and remove extraneous background, a colour-based refinement step was implemented (Figure 3, step 4). Through visual inspection of the densitometry reports, the table background was identified as a consistent light grey colour with RGB values approximately (222, 222, 222), with slight variations in the range (215-235, 215-235, 215-235).

|                      |            |
|----------------------|------------|
| Nombre:              |            |
| ID del paciente:     |            |
| Identificador 2:     |            |
| Código postal:       |            |
| Sexo:                | Male       |
| Raza:                | White      |
| Altura:              | 175.0 cm   |
| Peso:                | 89.0 kg    |
| Fecha de nacimiento: | 12.05.1946 |
| Edad:                | 77         |
| Edad de menopausia:  |            |
| Médico remitente:    |            |

|                                |                              |
|--------------------------------|------------------------------|
| Información de la exploración: |                              |
| Fecha exploración:             | 28 February 2024 - A02282406 |
| Tipo exploración:              | f Lumbar Spine               |
| Fecha de análisis:             | 28.02.2024 09:26             |
| Protocolo de análisis:         | Spine                        |
| Fecha de informe:              | 28.02.2024 09:37             |
| Institución:                   |                              |
| Operador:                      |                              |
| Modelo:                        | Horizon W (S/N306654M)       |
| Comentario:                    |                              |
| Ver. software:                 | 13.6.1.3                     |

| Resumen de resultados:                                     |                         |           |                          |         |                        |         |                       |
|------------------------------------------------------------|-------------------------|-----------|--------------------------|---------|------------------------|---------|-----------------------|
| Región                                                     | Área [cm <sup>2</sup> ] | CMO [(g)] | DMO [g/cm <sup>3</sup> ] | T-score | RM (Referencia máxima) | Z-score | AM (Edad coincidente) |
| L1                                                         | 14.86                   | 13.64     | 0.918                    | -1.4    | 86                     | -0.4    | 95                    |
| L2                                                         | 16.54                   | 17.59     | 1.063                    | -0.3    | 97                     | 0.8     | 109                   |
| L3                                                         | 12.63                   | 15.73     | 1.246                    | 1.3     | 113                    | 2.4     | 127                   |
| L4                                                         | 15.16                   | 16.66     | 1.099                    | 0.1     | 101                    | 1.2     | 114                   |
| Total                                                      | 59.19                   | 63.62     | 1.075                    | -0.1    | 99                     | 1.0     | 111                   |
| CV de DMO Total 1.0%, ACF = 1.026, BCF = 0.999, TH = 9.245 |                         |           |                          |         |                        |         |                       |

**Figure 5. Resulting tables after applying coordinate cropping.** Mind the dark-grey background areas surrounding the tables, some with titles and footnotes. Patient table was de-identified.

The refinement process involved:

- Colour mask generation. Converting cropped images to RGB colour space and applying `cv2.inRange()` to create binary masks identifying light grey table regions.
- Contour detection. Using `cv2.findContours()` with `RETR_EXTERNAL` flag to identify table boundaries.
- Boundary box calculation. Selecting the largest contour (corresponding to the main table) and computing its bounding rectangle using `cv2.boundingRect()`.
- Final cropping. Extracting the refined table region using the computed boundaries.

This colour-based refinement effectively removed non-table background areas, resulting in tightly cropped table images with minimal extraneous content (Figure 6).

|                      |            |
|----------------------|------------|
| Nombre:              |            |
| ID del paciente:     |            |
| Identificador 2:     |            |
| Código postal:       |            |
| Sexo:                | Male       |
| Raza:                | White      |
| Altura:              | 175.0 cm   |
| Peso:                | 89.0 kg    |
| Fecha de nacimiento: | 12.05.1946 |
| Edad:                | 77         |
| Edad de menopausia:  |            |
| Médico remitente:    |            |

|                        |                              |
|------------------------|------------------------------|
| Fecha exploración:     | 28 February 2024 - A02282406 |
| Tipo exploración:      | f Lumbar Spine               |
| Fecha de análisis:     | 28.02.2024 09:26             |
| Protocolo de análisis: | Spine                        |
| Fecha de informe:      | 28.02.2024 09:37             |
| Institución:           |                              |
| Operador:              |                              |
| Modelo:                | Horizon W (S/N306654M)       |
| Comentario:            |                              |
| Ver. software:         | 13.6.1.3                     |

| Región | Área [cm <sup>2</sup> ] | CMO [(g)] | DMO [g/cm <sup>2</sup> ] | T-score | RM (Referencia máxima) | Z-score | AM (Edad coincidente) |
|--------|-------------------------|-----------|--------------------------|---------|------------------------|---------|-----------------------|
| L1     | 14.86                   | 13.64     | 0.918                    | -1.4    | 86                     | -0.4    | 95                    |
| L2     | 16.54                   | 17.59     | 1.063                    | -0.3    | 97                     | 0.8     | 109                   |
| L3     | 12.63                   | 15.73     | 1.246                    | 1.3     | 113                    | 2.4     | 127                   |
| L4     | 15.16                   | 16.66     | 1.099                    | 0.1     | 101                    | 1.2     | 114                   |
| Total  | 59.19                   | 63.62     | 1.075                    | -0.1    | 99                     | 1.0     | 111                   |

**Figure 6. Refined tables after cropping.** Note that grey surrounding areas have been removed.

4.5. Step 5. Image to Text conversion with OCR.

After table identification and extraction, the next fundamental challenge involves converting the characters of the cells into editable text (Figure 3, step 5). For this purpose, this project leverages OCR, the technology that converts images of typed, handwritten, or printed text into machine-encoded text. The origins of OCR trace back to the early 20th century (31), but it was not until the advent of digital computing in the 1970s and 1980s that OCR became commercially viable, when companies like Kurzweil Computer Products introduced systems capable of recognising printed text in various fonts (32).

Contemporary approaches leverage deep NNs—particularly transformer-based architectures derived from models like Donut (Document Understanding Transformer). These systems can handle multiple languages, varying fonts, and complex layouts with remarkable accuracy. In the context of medical densitometry reports, OCR serves a crucial function: It transforms the visual representation of numerical measurements, patient demographics,

and acquisition system parameters into structured, machine-readable text that can be processed programmatically. Table 4 offers an overview of the features of the most popular OCR engines available in Python.

| Feature                 | Pytesseract                                                                                           | EasyOCR                                                                                                                                 | PaddleOCR                                                                                                     | Surya                                                                                                          |
|-------------------------|-------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------|
| <b>Core Technology</b>  | Python wrapper for Google's Tesseract engine (C++). Uses LSTMs (DL).                                  | Pure Python library using CRNN (DL) models.                                                                                             | Full DL toolkit by Baidu. Has models for detection, direction, and recognition.                               | Two-step DL pipeline: a layout detection model and a separate text recognition model.                          |
| <b>Installation</b>     | Complex                                                                                               | Very Easy                                                                                                                               | Easy                                                                                                          | Easy                                                                                                           |
| <b>Accuracy</b>         | Good. A solid baseline but can struggle with noisy images or unusual fonts compared to newer models.  | Very Good. Generally, more accurate and robust than Pytesseract out-of-the-box.                                                         | Excellent. Often state-of-the-art accuracy. Very strong in multilingual contexts.                             | Excellent. State-of-the-art, especially when layout is complex.                                                |
| <b>Layout Detection</b> | Basic. Can detect lines and blocks of text (using PSMs) but has no advanced table or column analysis. | Text Detection Only. Excellent at finding bounding boxes for lines of text, but not at identifying higher-level structures like tables. | Advanced. Includes dedicated models for table structure recognition, making it a full document analysis tool. | Primary Feature. Has a powerful, dedicated model for identifying text lines, paragraphs, tables, figures, etc. |
| <b>Language support</b> | Excellent (100+)                                                                                      | Very Good (80+)                                                                                                                         | Excellent (80+)                                                                                               | Excellent (90+)                                                                                                |
| <b>Best for...</b>      | Simple OCR tasks, projects with existing Tesseract dependencies                                       | Quick projects, beginners, pure Python dependencies.                                                                                    | High-accuracy applications, especially for multilingual documents                                             | Complex layout documents (tables, columns, headers)                                                            |

**Table 4. Main OCR engines available in Python.**

For our project, *PyTesseract*, *EasyOCR* and *Surya* were tested on the refined tables for character recognition. *PyTesseract* yielded the worst results regarding accuracy, and no position information could be obtained from the extracted texts. *EasyOCR* performed well on character recognition, but errors were also common and, furthermore, no information about the position of the bounding box in the table could be retrieved (Appendix, Table S1). Of all three, *Surya* yielded the best accuracy, and the precise report on the coordinates of the bounding boxes could be used to reproduce the table structure in the next steps.

#### **4.6. Step 6. Parsing OCR output & Data frame creation.**

The final component of the pipeline addresses the data wrangling challenge, i.e., the process of transforming and mapping raw data into a format suitable for analysis (Figure 3, step 6). Even with successful layout detection and text extraction, the output from OCR engines typically consists of unstructured or semi-structured text with positional information (bounding boxes) that must be parsed, validated, and organised into meaningful variables.

This parsing phase involves several complex tasks: identifying key-value pairs (such as "Patient ID: 1811967"), extracting measurements from tabular structures, handling variations in formatting and terminology, and managing potential OCR errors or ambiguities. The challenge is compounded by the need to maintain associations between related data points—ensuring that a T-score is correctly linked to its corresponding skeletal region (L1, L2, L3, or L4), for instance.

The parsing logic was consolidated within the so named *multiparser.py* script, which orchestrates the extraction of patient demographics, examination metadata and densitometry measurements into a unified pandas DataFrame structure.

The fundamental innovation of the *MultiParser* system lies in its use of spatial coordinates provided by *Surya OCR* to reconstruct the logical structure of the document. Each text element from OCR includes:

- text: The recognised text string
- bbox: Bounding box coordinates [x\_min, y\_min, x\_max, y\_max]
- confidence: OCR confidence score
- polygon: Four-corner polygon for precise text location

This positional information transforms raw OCR output into structured data by enabling the parsers to understand not just what text is present, but where it appears—crucial for maintaining the semantic relationships.

The parsers operate on a pixel-based coordinate system where:

- X-axis: Horizontal position (0 = left edge)
- Y-axis: Vertical position (0 = top edge)
- Thresholds: Configurable tolerances for grouping elements

Key threshold parameters used throughout:

- ROW\_THRESHOLD: 10-15 pixels (vertical grouping tolerance)
- COLUMN\_GAP\_THRESHOLD: 200-300 pixels (maximum horizontal gap for key-value pairs)
- VERTICAL\_OVERLAP\_THRESHOLD: 0.7 (70% overlap required for same-line detection)
- HORIZONTAL\_ALIGNMENT\_THRESHOLD: 0.5 (50% overlap for vertical field association).

Custom adjustments were made to each of the three parsers for each specific type of table. For example, for the patient's table, weight and height normalization, multiple language recognition (English and Spanish) or date parsing were implemented. The singularities for the engineering of each parser are described in Chapter 5. The code for the *MultiParser* and the individual parsers are provided as Supplementary Materials.

## 5. Experiments and results

### 5.1. Experiments

During the training phase, all images obtained after cropping were manually reviewed to detect accidental trimming of essential information within tables that would prevent the pipeline from recognizing the text in downstream steps. Because more images were now used for training than for manual segmentation (70 vs 20), slight changes to the initial coordinates had to be made as some table cells were truncated. Coordinates were slightly adjusted after visual inspection to the following (in bold the modified coordinates relative to previously set), so that no table was incorrectly cropped.

- Patient Data Table: xmin=8, xmax=401, ymin=169, ymax=637
- Examination Data Table: xmin=8, xmax=391, ymin=670, **ymax=1095**
- DMO Data Table: xmin=401, xmax=1190, **ymin=690, ymax=1000**

Visual examination after refinement of all images yielded a 100% of accuracy, meaning that all tables were correctly delimited and no cells had their text trimmed.

Before selecting the cropping and refining method as final, we performed tests with other available libraries with different approaches.

#### 5.1.1. Cropping with OpenCV Template matching

*OpenCV's* template matching is a method for finding a smaller image (the template) within a larger one (the source image). The template image slides over every possible location of the source image, and for each position, a calculation is performed to see how well the template "matches" that specific area of the source. When implemented in our project, most of the tables were correctly identified and extracted. However, there were some important drawbacks that prevented us from using this method:

- Some images can get missed, and these heavily rely on the template given. There may be other areas sharing more common features with

the template than the real target, so the ROI can mistakenly be pulled out (Figure 7).

- This method crops the images to the resolution given as input template, so even if the table is correctly identified, it can become truncated.
- It is comparative slower than when coordinates are hard-coded.

Because these errors are avoided when fixed coordinates are provided and efficiency is also improved, we chose this latter approach.

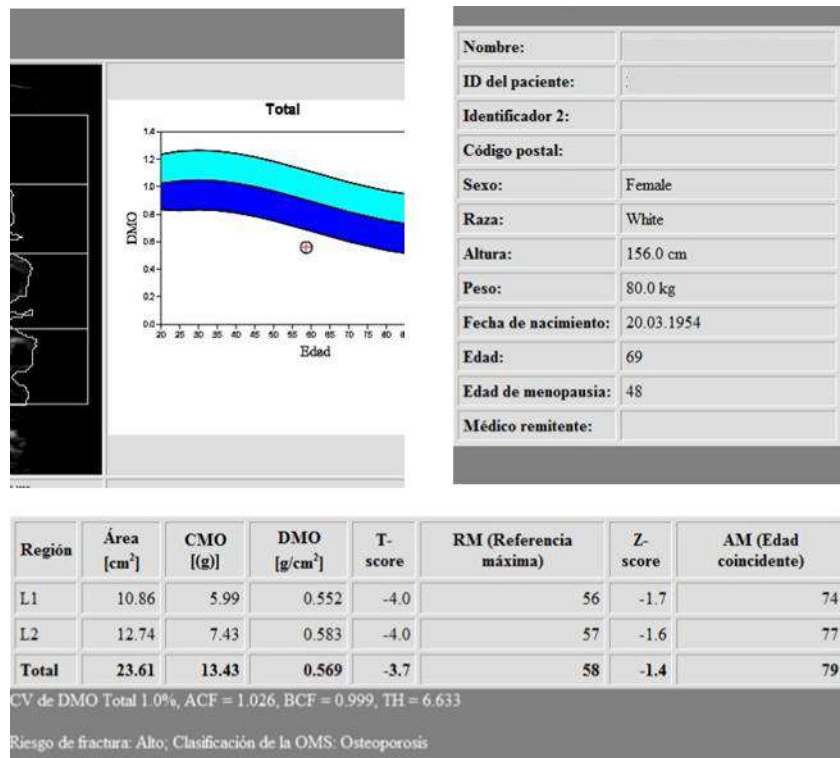


**Figure 7. Results of obtaining the tables using template matching.** Some ROIs are wrongly identified, like the colour part in some of the DXAs mistaken as a patient table (upper left). Other patient tables are not cropped accurately (upper right, bottom)

### 5.1.2. Table layout detection with img2Table

The tool *img2table* uses classical computer vision techniques from the OpenCV library to detect tables, rather than a DL approach. Using the information from either line detection or contour analysis, the library reconstructs the complete table structure. It creates a data representation of the grid, defining the exact coordinates (bounding box) for each individual cell. This grid of cell locations is then ready to be sent to an OCR engine to

extract the text from each cell. However, when test in our images, many of the cells were not properly identified (Figure 8), so we abandoned this approach.

| Región | Área [cm <sup>2</sup> ] | CMO [(g)] | DMO [g/cm <sup>2</sup> ] | T-score | RM (Referencia máxima) | Z-score | AM (Edad coincidente) |
|--------|-------------------------|-----------|--------------------------|---------|------------------------|---------|-----------------------|
| L1     | 14.42                   | 11.55     | 0.801                    | -1.7    | 81                     | 0.1     | 101                   |
| L2     | 14.01                   | 13.11     | 0.936                    | -0.8    | 91                     | 1.2     | 116                   |
| L3     | 13.35                   | 12.68     | 0.950                    | -1.2    | 88                     | 0.9     | 112                   |
| L4     | 15.61                   | 15.85     | 1.015                    | -0.4    | 96                     | 1.8     | 124                   |
| Total  | 57.39                   | 53.19     | 0.927                    | -1.1    | 89                     | 1.0     | 113                   |

CV de DMO Total 1.0%, ACF = 1.026, BCF = 0.999, TH = 7.105

**Figure 8. Table Layout detection using *img2table*.** Some cells are not correctly identified as shown by the solid red line. Some cells in T-score and AM columns are not identified as such.

## 5.2. Results

### 5.2.1. Performance on the *patient\_data* table

The overall performance of the pipeline on the patient table is excellent. All target rows are accurately captured (100%). The parser outputs a data frame in wide format, each row representing one patient and each column representing one field of the demographic data. All NA values come from the ‘menopause’ field. These values are already missing on the original images, as the input of this field is operator dependant. No true NA values (present in the image but missed by the pipeline) are found in this table.

|                   | Training dataset (n = 70) |
|-------------------|---------------------------|
| Expected rows     | 70                        |
| Detected rows (%) | 70 (100%)                 |
| Total fields      | 560                       |
| Wrong fields      | 0                         |
| Total NAs (%)     | 34 (6%)                   |
| True NAs (%)      | 0 (0.0%)                  |

**Table 5. Performance of the pipeline on the *patient\_data* tables in the training dataset.**

|                          | <b>Test dataset (n = 30)</b> |
|--------------------------|------------------------------|
| <b>Expected rows</b>     | 30                           |
| <b>Detected rows (%)</b> | 30 (100%)                    |
| <b>Total fields</b>      | 240                          |
| <b>Wrong fields</b>      | 0                            |
| <b>Total NAs (%)</b>     | 15 (6.25%)                   |
| <b>True NAs (%)</b>      | 0 (0.0%)                     |

*Table 6. Performance of the pipeline on the patient\_data tables in the test dataset*

The fundamental approach of the parser relies on spatial coordinate analysis to reconstruct the logical relationships between field labels and their corresponding values. The script uses the knowledge of the table layout, where information is presented in a left-aligned label format followed by right-aligned values. OCR processing fragments this visual structure into individual text elements with coordinate positions. The script rebuilds these relationships by analysing the horizontal positioning of text elements, using a coordinate threshold to distinguish between labels appearing on the left side of the document and values positioned to the right.

The spatial separation mechanism employs a horizontal coordinate threshold of 120 pixels to categorise text elements. Elements positioned to the left of this threshold are treated as potential field labels, whilst those to the right are considered candidate values.

The script accommodates both English and Spanish words. The key mapping system includes not only direct translations but also recognises truncated text fragments that commonly result from OCR processing limitations. This flexible matching mechanism proves important when dealing with complex field names that frequently span multiple lines in source documents. Terms like "Fecha de nacimiento" (birth date) often appear split across separate text lines due to formatting constraints or OCR segmentation decisions. The script addresses this through text merging logic that recognises partial key phrases and reconstructs complete field identifiers from fragmented components.

The implementation includes specific handling for commonly problematic fields such as birth date and menopause age, where Spanish terminology frequently exceeds single-line capacity in document layouts. The algorithm searches for partial matches like "fecha de" or "naci" and employs contextual analysis to determine whether subsequent text lines contain the completion of these field names. This approach significantly improves extraction accuracy for documents with variable formatting or OCR quality issues.

Patient name extraction presents unique challenges because names frequently span multiple text elements and may appear in various cultural formats. The script implements specialised logic for name processing that recognises the common practice of presenting surnames and given names as separate elements. The algorithm attempts to identify both components and reconstruct complete names while handling cases where only single name elements are successfully captured by OCR. The name processing logic employs proximity-based matching to associate name components that appear near "Nombre" or "Name" field labels. When multiple potential name values are detected, the algorithm prioritises the value closest to the field label based on vertical coordinate analysis. This approach accommodates documents where names appear in varying positions relative to their labels whilst maintaining accuracy in field association.

For situations where traditional key-value pairing fails, the script includes fallback detection mechanisms that search for characteristic values throughout the document text. Sex and race information, for example, can be identified through value-based pattern matching when explicit field labels are not successfully recognised. This dual-approach strategy significantly improves data completeness when dealing with documents that have irregular formatting or poor OCR quality.

Numeric fields like height and weight undergo intelligent unit conversion that recognises when values appear in unexpected units. Heights exceeding 250 are assumed to be recorded in millimetres rather than centimetres and

are automatically converted, whilst weights over 300 are treated as gram measurements requiring conversion to kilograms.

Date processing employs flexible pattern matching that accommodates multiple date formats. The algorithm searches for various delimiter patterns including dots, slashes, and hyphens, whilst handling both two-digit and four-digit year representations.

Age fields receive careful validation to ensure numeric consistency, with the script converting text representations to appropriate integer formats whilst preserving missing data indicators for incomplete records. Special handling for menopause age acknowledges that this field may be irrelevant or unavailable for certain patient populations, maintaining data integrity through nullable integer conversion.

The coordinate-based matching algorithm includes proximity validation that ensures field labels and values are reasonably associated before creating data pairs. The twenty-pixel tolerance threshold prevents erroneous associations between spatially distant text elements whilst accommodating minor positioning variations that commonly result from document scanning or OCR processing inconsistencies.

### 5.2.2. Performance on the exam\_data table

The overall performance of the pipeline on the patient table is also excellent. The parser outputs a data frame in wide format, each row representing one patient and each column representing one field of the exam data.

|                   | Training dataset (n = 70) |
|-------------------|---------------------------|
| Expected rows     | 70                        |
| Detected rows (%) | 70 (100%)                 |
| Total fields      | 240                       |
| Wrong fields      | 0                         |
| Total NAs (%)     | 0 (0%)                    |

*Table 7. Performance of the pipeline on the exam\_data tables in the training dataset.*

|                   | Test dataset (n = 30) |
|-------------------|-----------------------|
| Expected rows     | 30                    |
| Detected rows (%) | 30 (100%)             |
| Total fields      | 90                    |
| Wrong fields      | 0                     |
| Total NAs (%)     | 0 (0%)                |

*Table 8. Performance of the pipeline on the exam\_data tables in the test dataset.*

The parser represents a more focused approach to OCR data processing, specifically targeting three critical metadata fields. Unlike the patient parser, which handles variable numbers of fields with complex validation requirements, the exam parser operates with a precisely defined scope targeting scan type, analysis date and model information.

The parser uses the same coordinate-based matching principle as the patient’s table parser. The algorithm employs a multi-stage approach to value association that begins with same-line extraction for cases where keys and values appear in traditional delimiter-separated formats. When this approach fails to identify values, the spatial analysis extends to examine nearby text elements based on vertical coordinate proximity. A five-pixel tolerance threshold is used in this case.

This parser also employs bilingual pattern matching. The date processing operates under the assumption that examination dates follow the European dd.mm.yyyy convention found in most images. The conversion process employs pandas datetime functionality to ensure proper date validation and standardisation. The algorithm discards time information that may be present but is irrelevant for our case.

### 5.2.3. Performance on the dmo\_data table

The result of applying *Surya OCR* directly to the refined tables yields excellent results (Table 9). The resulting data frame from the parser outputs a data frame in long format, each row representing one region. All tables in the training dataset contained information on 4 lumbar vertebrae and the

total (5 rows). Only one table had information on two vertebral levels (L1 and L2). Thus, the expected number of rows for the resulting data frame is 348.

|                          | <b>Training dataset (n = 70)</b> |
|--------------------------|----------------------------------|
| <b>Expected rows</b>     | 348                              |
| <b>Detected rows (%)</b> | 348 (100%)                       |
| <b>Total fields</b>      | 2784                             |
| <b>Wrong fields</b>      | 0                                |
| <b>Total NAs (%)</b>     | 18 (0.6%)                        |

**Table 9. Summary performance of the pipeline for the 'dmo\_table' containing the BMD values.**

After OCR, the parser script cleans the text from the resulting JSON file by removing HTML tags and standardising the name of the vertebral levels. Then, the script employs a spatial analysis approach to reconstruct tabular structure from the fragmented text elements produced by OCR. *Surya OCR* returns individual text fragments with bounding box coordinates rather than maintaining any inherent table structure. The challenge lies in reassembling these scattered pieces back into meaningful rows and columns that preserve the original clinical data relationships.

The row detection algorithm operates on the principle that text elements appearing at similar vertical positions (Y-coordinates) belong to the same conceptual row. The script addresses vertical coordinates variations through a tolerance-based clustering mechanism that treats any text elements within 10 pixels vertically as belonging to the same row. When a new element's Y-coordinate falls within the tolerance range of an established row, it joins that collection. Otherwise, a new row is created. This approach handles variable numbers of text elements per image whilst maintaining structural integrity.

Once rows are established, the script reconstructs column ordering through horizontal coordinate sorting. Each row's constituent text elements are arranged according to their X-coordinates, effectively recreating the left-to-right reading sequence that defines column structure in the original densitometry table.

This horizontal sorting proves particularly crucial for densitometry data, where the sequence of measurements follows a standardised convention. The region identifier typically appears first, followed by area measurements, then various density calculations including BMC, BMD, T-scores, reference means, Z-scores, and age-matched percentiles.

Besides, the script is designed to specifically target rows beginning with vertebral identifiers such as L1 through L4, plus summary totals, filtering out extraneous content that would compromise the final dataset. By restricting processing to rows with vertebral identifiers, the script maintains focus on quantitative clinical measurements. When the OCR successfully captures all elements, the validation process proceeds through standard assignment pathways, mapping each detected text element to its appropriate measurement category according to positional sequence. However, OCR processing occasionally produces incomplete row captures due to image quality issues, text recognition failures, or formatting irregularities.

As seen in Table 9, the *NAs* represent uncaptured data that were present in the original image (Figure 9). All these missing values pertained to the same column (the ‘am’ column or last column) (Figure 10).

These missing values were traced back in the pipeline to find the point where they were lost. We identified the OCR engine as unable to recognize these pieces of text, probably because of image quality issues or proximity to the edge of the image. To overcome this, some experiments were performed to improve the recognition of the text, such as scaling the image up to twice its original size, applying Otsu’s binarization, Gaussian and morphological kernels and background removal techniques (Figure 11).

| Región | Area[cm <sup>2</sup> ] | BMC[(g)] | BMD[g/cm <sup>2</sup> ] | T-score | RM (Referencia máxi | Z-score | EC (Edad coincident |
|--------|------------------------|----------|-------------------------|---------|---------------------|---------|---------------------|
| L1     | 11.48                  | 6.14     | 0.535                   | -4.1    | 54                  | -3.0    | 62                  |
| L2     | 9.89                   | 5.58     | 0.564                   | -4.2    | 55                  | -3.0    | 63                  |
| L3     | 10.98                  | 6.30     | 0.574                   | -4.6    | 53                  | -3.3    | 61                  |
| L4     | 11.91                  | 6.90     | 0.579                   | -4.4    | 55                  | -3.0    | 64                  |
| Total  | 44.26                  | 24.92    | 0.563                   | -4.4    | 54                  | -3.1    | 62                  |

*Figure 9. The original sample 'dmo\_table' after refinement.*

| Región | Area[cm2] | BMC[(g)] | BMD[g/cm2] | T-score | RM (Referencia máxi | Z-score | EC (Edad coincident |    |
|--------|-----------|----------|------------|---------|---------------------|---------|---------------------|----|
| L1     | 11.48     | 6.14     | 0.535      | -4.1    |                     | 54      | -3.0                | 62 |
| L2     | 9.89      | 5.58     | 0.564      | -4.2    |                     | 55      | -3.0                |    |
| L3     | 10.98     | 6.30     | 0.574      | -4.6    |                     | 53      | -3.3                |    |
| L4     | 11.91     | 6.90     | 0.579      | -4.4    |                     | 55      | -3.0                |    |
| Total  | 44.26     | 24.92    | 0.563      | -4.4    |                     | 54      | -3.1                | 62 |

*Figure 10. Native Surya-OCR image recognition with default parameters. The last column has missing values.*

| Región | Area[cm <sup>2</sup> ] | BMC[(g)] | BMD[g/cm <sup>2</sup> ] | T-score | RM (Referencia máxi | Z-score | EC (Edad coincident |
|--------|------------------------|----------|-------------------------|---------|---------------------|---------|---------------------|
| L1     | 11.48                  | 6.14     | 0.535                   | -4.1    | 54                  | -3.0    | 62                  |
| L2     | 9.89                   | 5.58     | 0.564                   | -4.2    | 55                  | -3.0    | 63                  |
| L3     | 10.98                  | 6.30     | 0.574                   | -4.6    | 53                  | -3.3    | 61                  |
| L4     | 11.91                  | 6.90     | 0.579                   | -4.4    | 55                  | -3.0    | 64                  |
| Total  | 44.26                  | 24.92    | 0.563                   | -4.4    | 54                  | -3.1    | 62                  |

*Figure 11. Sample 'dmo\_table' after morphological kernel and background removal. Contrast is somewhat enhanced to the human eye, but the OCR engine was unable to properly recognize all the characters.*

The latter combines morphological closing and Gaussian blur to create a robust background estimation that is then subtracted from the original image. The morphological closing operation (dilation followed by erosion) using a large 25×25 elliptical kernel "fills in" small foreground objects (text and numbers) while preserving large background structures, creating a background template that captures textural patterns and gradual intensity variations in the document. Simultaneously, Gaussian blur with a 51×51 kernel creates a smooth background approximation by averaging local neighbourhoods, which handles irregular lighting and paper artifacts. The

weighted combination (60% morphological + 40% Gaussian) leverages the strengths of both approaches: the morphological component captures structured background patterns while the Gaussian component handles smooth gradients. When this composite background is subtracted from the original grayscale image and normalized, the result eliminates background interference while preserving foreground text with enhanced contrast.

The background removal method was the most effective transformation that allowed for 100% identification of the ‘am’ column (see Figure 12).

| Región | Area[cm2] | \vert BMC[(g)] | \vert BMD[g/cm2] | T-score | RM (Referencia máxi | Z-score | EC (Edad coincident |
|--------|-----------|----------------|------------------|---------|---------------------|---------|---------------------|
| L1     | 11.48     | 6.14           | 0.535            | -4.1    | 54                  | -3.0    | 62                  |
| L2     | 9.89      | 5.58           | 0.564            | -4.2    | 55                  | -3.0    | 63                  |
| L3     | 10.98     | 6.30           | 0.574            | -4.6    | 53                  | -3.3    | 61                  |
| L4     | 11.91     | 6.90           | 0.579            | -4.4    | 55                  | -3.0    | 64                  |
| Total  | 44.26     | 24.92          | 0.563            | -4.4    | 54                  | -3.1    | 62                  |

**Figure 12. Surya OCR character recognition after background removal of sample ‘dmo\_table’.** The ‘am’ column was properly read. However, this change on the other images made some characters unrecognizable.

This improvement came with a cost: The “L1” vertebral level was not identified for some images and then only 139/150 rows were detected. This was because the string L1 was read by the OCR engine as ‘LI’, ‘LQ’ or even as ‘m’ in some images. The parser script could be tailored to accommodate these variations, but because other critical values were also misread, and because it was impossible to predict all the variations that OCR may produce for the “L1” vertebral level, we deemed this transformation as less accurate than the original one abandoned this branch of development. This decision was made according to the objective of the project, considering that the ‘am’ variable was not a principal target of the pipeline.

Results of the pipeline on the test dataset are shown in Table 10. Again, the missing values come from the ‘am’ variable in the DMO table.

|                   | Test dataset (n = 30) |
|-------------------|-----------------------|
| Expected rows     | 150                   |
| Detected rows (%) | 150 (100%)            |
| Total fields      | 1200                  |
| Wrong fields      | 0                     |
| Total NAs         | 7 (0.58%)             |

*Table 10. Performance metrics for 'dmo\_table' in the test dataset.*

#### 5.2.4. Resulting unified data frame

After conversion of each table to their respective data frames, the last step of the *MultiParser* was executing a left join of the three tables, so that each row in the final data frame represents a region (L1-L4 and total) from the 'dmo\_table'. The common ID for the join is a 'patient\_number' variable that the parsers create for each individual data frame and is taken from the name of the input image. The first 20 entries of the resulting unified table are shown in the Appendix (Table S2).

## 6. Discussion

The primary advantages of the selected architecture for the pipeline are its exceptional processing speed on standard CPU hardware, minimal computational resource requirements, and significantly lower total cost of ownership (TCO). By leveraging the known consistency of the document layout, the pipeline avoids the substantial computational overhead inherent in DL-based layout analysis models. This targeted approach eliminates entire stages of complex processing, resulting in a system that is not only orders of magnitude faster but also more reliable, easier to implement, and more transparent to debug and maintain.

The foundational premise of this analysis rests on the uniform nature of the source documents. These reports, generated by specialised medical equipment, are highly standardised. They are not free-form text but are produced from a fixed template. Consequently, key information appears in consistent and predictable locations across all reports originating from the same system.

This structural invariance is the single most important constraint influencing the design of an optimal data extraction pipeline. It fundamentally changes the nature of the problem. Instead of a general "document understanding" task, where the system must learn to identify elements in novel and varied layouts, the problem is reduced to a more constrained and deterministic "data location" task. The challenge is not to find the tables, but to efficiently extract the data from their known positions. This distinction is paramount, as solutions designed to handle layout diversity inherently carry a computational cost that may be unnecessary in a fixed-layout environment (33). Table 11 synthesises the main features of the most popular solutions available for document layout recognition and table extraction. It highlights the trade-offs between generality and efficiency, positioning the proposed pipeline as the most specialised and resource-efficient solution for the defined task.

The most significant efficiency gain in our pipeline comes from using fixed coordinates to perform an initial crop of the ROIs. This is the simplest form of localisation. It involves using pre-determined pixel coordinates to slice the main image array and extract a smaller sub-image that contains the table. In *OpenCV*, this is a basic *NumPy* array. This method is only viable when the target object appears in the exact same location (or within a slight variation) in every image, which is the case for DXA standardised reports. This makes it superior to dynamic detection algorithms.

Beyond layout and pattern recognition, the choice of the OCR engine is critical to the overall accuracy of the pipeline. *Surya-OCR* was selected because it represents the pinnacle of open-source OCR technology, balancing high accuracy with the essential feature of providing detailed positional data.

A key requirement for this pipeline is that the OCR engine must return not just the recognised text, but also the precise bounding box coordinates for each text element (word or line). *Surya* excels at this, providing detailed (x1, y1, x2, y2) coordinates for each detected line of text. This positional information is the raw material for the final structure reconstruction stage. Without it, the extracted text would be just an unordered collection of words, devoid of its tabular meaning.

The final step is to algorithmically reconstruct the table's grid structure from this data. This is a well-understood problem in document analysis and does not require a DL model.

The fundamental principle is to group text elements based on their spatial proximity and alignment. A common and effective method involves analysing the x and y coordinates of the bounding boxes.

By combining row grouping (based on y-coordinates) and column clustering (based on x-coordinates), a logical grid can be constructed. Each piece of text can then be assigned to a specific (row, column) cell, transforming the unstructured OCR output into a fully structured table. This final step is purely algorithmic, deterministic, and computationally fast, completing the

efficient data extraction process. Other approaches also include intersecting the detected row and column bounding boxes to define the final cell structure.

Overall, the architecture of our pipeline is not purely heuristic nor purely DL; it is a pragmatic hybrid that intelligently allocates computational resources only where they are most needed. It uses simple, deterministic methods for the "solved" part of the problem—locating the tables—and a state-of-the-art DL model integrated in *Surya* for the "hard" part—recognising the text.

The conclusion is stark: the proposed architecture bypasses a detection step that costs hundreds of milliseconds (on GPU) or potentially many seconds (on CPU), replacing it with a near-instantaneous heuristic. For any application requiring low latency or high throughput on standard CPU-based infrastructure, this difference is decisive. The general-purpose DL models expend a vast amount of computational effort to solve a problem (where are the tables?) that is already solved by the very nature of the input data.

The most common criticism of heuristic or rule-based systems is their "brittleness"—their inability to adapt to changes in input format. A DL model can gracefully handle a new invoice layout; our rule-based parser will fail if the document layout is changed or if a field moves. This must be taken into consideration, since the pipeline was designed to identify the tables in a closed time span (2020-2025) and for a specific machine vendor. If DXA scans were to be analysed previous or posterior to these years, or even if the vendor updates the format of the layout, our model would be completely “blind” to these tables and would have to be rewritten to adapt to these changes. Because this pipeline was framed within a larger project that aims to study the DXA reports in this period and given that the HOLOGIC system was constat during those years, our approach is still valid for the final goal.

| Feature                          | Developed Pipeline                                         | img2table (hybrid)                                           | PaddleOCR (PP-Structure)                               | Surya                                                 |
|----------------------------------|------------------------------------------------------------|--------------------------------------------------------------|--------------------------------------------------------|-------------------------------------------------------|
| <b>Localisation method</b>       | Fixed coordinates & colour segmentation (heuristic)        | Hough transform & contour analysis (classical cv)            | SLANET / SLANEXT (DL cnn/transformer)                  | EfficientVit-based model (DL transformer)             |
| <b>Primary strength</b>          | Extreme speed and efficiency for fixed layouts             | Fast processing of bordered tables with some layout variance | High accuracy on complex and varied table structures   | State-of-the-art accuracy on diverse document layouts |
| <b>Ideal use case</b>            | Invariable, template-based documents (e.g., medical forms) | Structured documents with clear borders (e.g., invoices)     | Unstructured documents with inconsistent table layouts | General-purpose document understanding at scale       |
| <b>Localisation time (CPU)</b>   | < 10 ms                                                    | ~50-100 ms                                                   | ~40-500 ms                                             | ~157,000 ms (157 s)                                   |
| <b>Localisation time (GPU)</b>   | not applicable                                             | not applicable                                               | ~20-85 ms                                              | ~130-270 ms                                           |
| <b>Requires GPU for speed?</b>   | No                                                         | No                                                           | Recommended                                            | Yes, essential                                        |
| <b>Model size (localisation)</b> | Negligible (few lines of code)                             | N/A (algorithm-based)                                        | ~7-351 mb                                              | >1 GB (part of a larger toolkit)                      |
| <b>Interpretability</b>          | High ("white box")                                         | Medium                                                       | Low ("black box")                                      | Low ("black box")                                     |

*Table 11. Frameworks for document layout and table recognition.*

## 7. Conclusions and future work

In summary, the presented pipeline fulfils the primary objective of converting one or more DXA image reports into a data frame with relevant patient, exam and bone specific measurements. The generated data structure can be readily used for research and/or merged with other existing datasets. Besides, in a clinical environment, this pipeline enables physicians to **automatically transfer** the information of DXA report into EHRs.

This pipeline can be further improved by **increasing the variability** of the training images, both temporarily and qualitatively. By **increasing the time span**, the algorithm could accommodate changes in the fixed coordinates of the tables before 2020 and after 2025. This strong dependency on the fixed layout of the report can be seen as a liability. If newer versions of the report are realised by the vendor, these must also be accounted for in the training of the pipeline, or else it will fail to execute. Also, by **increasing the number of images** from different vendors, the pipeline will be more capable to generalize and expand to other machines used in other healthcare institutions.

Regarding the pipeline's excellent performance, future work must focus on the optimization of the image for OCR recognition. Although some values were deemed as secondary (like the 'am' column), the algorithm must seek to **apply effective transformations to the raw images** that can grant their retrieval in case this information become valuable in future research.

One next step after the deployment of the presented pipeline consists in its incorporation into the main branch of the project. The resulting data frame from our pipeline is expected to be merged with radiomic features extracted from plain X-ray images of the lumbar spine along with other clinical features coming from EHR that will enable our team to **develop an AI-driven model** that aims to predict osteoporosis and fracture risk in our population

More pragmatically, future work will focus on **developing an API** that enables healthcare providers to automatically convert DXA images into plain text that can be further incorporated into EHR systems. Currently, physicians

need to manually input the information contained in the images, a task that is time consuming and prone to human error. Developing a user-friendly interface that enables users to drag-and-drop one or more image files to produce editable text may reduce consultation times and transcription errors.

Finally, AI offers a powerful solution to the **early and widespread screening for osteoporosis** by leveraging the vast number of routine medical images that are already being performed for countless other clinical indications. DL models have demonstrated a remarkable ability to detect subtle changes in bone texture and density that are often invisible to the human eye. By unlocking the hidden diagnostic potential within the millions of medical images generated daily, artificial intelligence provides the key to overcoming the long-standing barriers to effective osteoporosis management. This technological leap promises to shift the paradigm from reactive fracture treatment to proactive, population-scale prevention, ultimately safeguarding the health, mobility, and quality of life for aging populations worldwide.

## 8. References

1. GBD 2019 Fracture Collaborators. Global, regional, and national burden of bone fractures in 204 countries and territories, 1990-2019: a systematic analysis from the Global Burden of Disease Study 2019. *Lancet Healthy Longev.* 2021 Sept;2(9):e580–92.
2. Curtis EM, van der Velde R, Moon RJ, van den Bergh JPW, Geusens P, de Vries F, et al. Epidemiology of fractures in the United Kingdom 1988-2012: Variation with age, sex, geography, ethnicity and socioeconomic status. *Bone.* 2016 June;87:19–26.
3. Kanis JA, Norton N, Harvey NC, Jacobson T, Johansson H, Lorentzon M, et al. SCOPE 2021: a new scorecard for osteoporosis in Europe. *Arch Osteoporos.* 2021 June 2;16(1):82.
4. Johnell O, Kanis JA. An estimate of the worldwide prevalence and disability associated with osteoporotic fractures. *Osteoporos Int J Establ Result Coop Eur Found Osteoporos Natl Osteoporos Found USA.* 2006 Dec;17(12):1726–33.
5. WHO Scientific Group on the Prevention and Management of Osteoporosis (2000 : Geneva S. Prevention and management of osteoporosis : report of a WHO scientific group [Internet]. World Health Organization; 2003 [cited 2025 Aug 4]. Available from: <https://iris.who.int/handle/10665/42841>
6. World Osteoporosis Day | Simple ways to prevent osteoporosis [Internet]. [cited 2025 Aug 4]. Available from: <https://www.who.int/westernpacific/articles/item/simple-ways-to-prevent-osteoporosis>
7. Lippuner K, Johansson H, Kanis JA, Rizzoli R. FRAX assessment of osteoporotic fracture probability in Switzerland. *Osteoporos Int J Establ Result Coop Eur Found Osteoporos Natl Osteoporos Found USA.* 2010 Mar;21(3):381–9.
8. Professor Kanis JA, Melton LJ III, Christiansen C, Johnston CC, Khaltsev N. The diagnosis of osteoporosis. *J Bone Miner Res.* 1994 Aug 1;9(8):1137–41.
9. Kanis JA, Johnell O. Requirements for DXA for the management of osteoporosis in Europe. *Osteoporos Int.* 2005 Mar 1;16(3):229–38.
10. Pickhardt PJ, Pooler BD, Lauder T, del Rio AM, Bruce RJ, Binkley N. Opportunistic screening for osteoporosis using abdominal computed tomography scans obtained for other indications. *Ann Intern Med.* 2013 Apr 16;158(8):588–95.

11. Krishnaraj A, Barrett S, Bregman-Amitai O, Cohen-Sfady M, Bar A, Chettrit D, et al. Simulating Dual-Energy X-Ray Absorptiometry in CT Using Deep-Learning Segmentation Cascade. *J Am Coll Radiol JACR*. 2019 Oct;16(10):1473–9.
12. Benhamou CL, Poupon S, Lespessailles E, Loiseau S, Jennane R, Siroux V, et al. Fractal analysis of radiographic trabecular bone texture and bone mineral density: two complementary parameters related to osteoporotic fractures. *J Bone Miner Res Off J Am Soc Bone Miner Res*. 2001 Apr;16(4):697–704.
13. GBD 2021 Low Back Pain Collaborators. Global, regional, and national burden of low back pain, 1990–2020, its attributable risk factors, and projections to 2050: a systematic analysis of the Global Burden of Disease Study 2021. *Lancet Rheumatol*. 2023 June;5(6):e316–29.
14. Wang L, Yang R, Sha Z, Kuraszkiwicz AM, Leonik C, Zhou L, et al. Assessing Risk Factors for Cognitive Decline Using Electronic Health Record Data: A Scoping Review. *Res Sq*. 2024 Aug 9;rs.3.rs-4671544.
15. Górriz JL, Adán Gil F, Botana López MA, Buño Soto A, Campos Cabrera FJ, Cisneros A, et al. Mejora en la detección, diagnóstico y tratamiento temprano de la enfermedad renal crónica en España. Proyecto InterERKit. *Nefrología*. 2025 May 1;45(5):351–9.
16. Ghildayal N. Public Health Surveillance in Electronic Health Records: Lessons From PCORnet. *Prev Chronic Dis [Internet]*. 2024 [cited 2025 Aug 5];21. Available from: [https://www.cdc.gov/pcd/issues/2024/23\\_0417.htm](https://www.cdc.gov/pcd/issues/2024/23_0417.htm)
17. Hsieh CI, Zheng K, Lin C, Mei L, Lu L, Li W, et al. Automated bone mineral density prediction and fracture risk assessment using plain radiographs via deep learning. *Nat Commun*. 2021 Sept 16;12(1):5472.
18. Chaisen M, Sritara C, Chitrapazt N, Suppasilp C, Chamroonrat W, Promma S, et al. Opportunistic Screening for Osteoporosis by CT as Compared with DXA. *Diagnostics*. 2024 Jan;14(24):2846.
19. Gyftopoulos S, Pelzl CE, Chang CY. Quantifying the Opportunity and Economic Value of Bone Density Screening Using Opportunistic CT: A Medicare Database Analysis. *J Am Coll Radiol*. 2025 Mar 1;22(3):349–57.
20. Yamamoto N, Sukegawa S, Kitamura A, Goto R, Noda T, Nakano K, et al. Deep Learning for Osteoporosis Classification Using Hip Radiographs and Patient Clinical Covariates. *Biomolecules*. 2020 Nov;10(11):1534.
21. Hong N, Cho SW, Shin S, Lee S, Jang SA, Roh S, et al. Deep-Learning-Based Detection of Vertebral Fracture and Osteoporosis Using Lateral Spine X-Ray Radiography. *J Bone Miner Res*. 2023;38(6):887–95.

22. Sato Y, Yamamoto N, Inagaki N, Iesaki Y, Asamoto T, Suzuki T, et al. Deep Learning for Bone Mineral Density and T-Score Prediction from Chest X-rays: A Multicenter Study. *Biomedicines*. 2022 Sept 19;10(9):2323.
23. Jang R, Choi JH, Kim N, Chang JS, Yoon PW, Kim CH. Prediction of osteoporosis from simple hip radiography using deep learning algorithm. *Sci Rep*. 2021 Oct 7;11(1):19997.
24. Zhang B, Yu K, Ning Z, Wang K, Dong Y, Liu X, et al. Deep learning of lumbar spine X-ray for osteopenia and osteoporosis screening: A multicenter retrospective cohort study. *Bone*. 2020 Nov;140:115561.
25. Kasem M, Abdallah A, Berendeyev A, Elkady E, Abdalla M, Mahmoud M, et al. Deep learning for table detection and structure recognition: A survey [Internet]. *arXiv*; 2022 [cited 2025 Aug 18]. Available from: <http://arxiv.org/abs/2211.08469>
26. Schreiber S, Agne S, Wolf I, Dengel A, Ahmed S. DeepDeSRT: Deep Learning for Detection and Structure Recognition of Tables in Document Images. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) [Internet]. 2017 [cited 2025 Aug 18]. p. 1162–7. Available from: <https://ieeexplore.ieee.org/document/8270123>
27. Paliwal S, D V, Rahul R, Sharma M, Vig L. TableNet: Deep Learning model for end-to-end Table detection and Tabular data extraction from Scanned Document Images [Internet]. *arXiv*; 2020 [cited 2025 Aug 18]. Available from: <http://arxiv.org/abs/2001.01469>
28. *ibm-aur-nlp/PubTabNet* [Internet]. *ibm-aur-nlp*; 2025 [cited 2025 Aug 18]. Available from: <https://github.com/ibm-aur-nlp/PubTabNet>
29. Bao J, Zou D, Li W. Characteristics of the DXA Measurements in Patients Undergoing Lumbar Fusion for Lumbar Degenerative Diseases: A Retrospective Analysis of Over 1000 Patients. *Clin Interv Aging*. 2021 June 16;16:1131–7.
30. *labelImg* · PyPI [Internet]. [cited 2025 Aug 11]. Available from: <https://pypi.org/project/labelImg/>
31. Emanuel Goldberg – Computer Timeline [Internet]. [cited 2025 Aug 9]. Available from: <http://www.computer-timeline.com/timeline/emanuel-goldberg/>
32. Kurzweil Computer Products [Internet]. [cited 2025 Aug 9]. Available from: <https://www.kurzweiltech.com/kcp.html>
33. AI vs. Rule-Based PDF Parsing Tools | Parseur® [Internet]. 2025 [cited 2025 Aug 9]. Available from: <https://parseur.com/blog/ai-vs-rule-based-pdf-parsers>

# 9.Appendix

**Table S1.** OCR performance for PyTesseract and EasyOCR on a sample ‘dmo\_table’.

| Región | Área<br>[cm²] | CMO<br>[(g)] | DMO<br>[g/cm²] | T-<br>score | RM (Referencia<br>máxima) | Z-<br>score | AM (Edad<br>coincidente) |
|--------|---------------|--------------|----------------|-------------|---------------------------|-------------|--------------------------|
| L1     | 14.42         | 11.55        | 0.801          | -1.7        | 81                        | 0.1         | 101                      |
| L2     | 14.01         | 13.11        | 0.936          | -0.8        | 91                        | 1.2         | 116                      |
| L3     | 13.35         | 12.68        | 0.950          | -1.2        | 88                        | 0.9         | 112                      |
| L4     | 15.61         | 15.85        | 1.015          | -0.4        | 96                        | 1.8         | 124                      |
| Total  | 57.39         | 53.19        | 0.927          | -1.1        | 89                        | 1.0         | 113                      |

```
=====
PYTESSERACT
=====
a Area CMO DMO +. RM (Referencia L AM (Edad

[em] fl [g/cm] score maxima) score coincidente)
Li 1442| 11.55 0801-17 si| 0.1 101
L2 14.01 13.11 0936-08 91 12 116
L3 13.35) 12.68 0980, -12 8809 112
L4 15.61 15.85 1015-04 96 «18 124

Total 57.39 53.19 0.927 -11 89 10 113
=====
EASYOCR
=====
Area
CMO
DMO
RM (Referencia
AM (Edad
Region
[cm"]
[(g)]
[gtcm"]
score
maxima)
score
coincidente)
L1
14.42
1155
0.801
-1.7
101
L2
14.01
13.11
0.936
116
13.35
12.68
0.950
-12
112
L4
15.61
15.85
1.015
0.4
124
Total
57.39
53.19
0.927
L
1.0
113
```

**Table S2.** View of the resulting unified data frame. Only the twenty first entries of the training data set are displayed. Personal data has been blurred for privacy purposes.

| #  | A tibble: 150 × 22 | patient_number | surname | patient_id | sex     | race    | age   | birth_date | height_cm | weight_kg | menopause | scan_type   | analysis_date | scanner_model | image_name          | region | area  | cmo   | dmo   | tscore |
|----|--------------------|----------------|---------|------------|---------|---------|-------|------------|-----------|-----------|-----------|-------------|---------------|---------------|---------------------|--------|-------|-------|-------|--------|
|    |                    | <dbl>          | <chr>   | <dbl>      | <chr>   | <chr>   | <dbl> | <date>     | <dbl>     | <dbl>     | <dbl>     | <chr>       | <date>        | <chr>         | <chr>               | <chr>  | <dbl> | <dbl> | <dbl> | <dbl>  |
| 1  |                    |                |         |            | Homb... | Blan... | 43    | 1976-12-11 | 172       | 61        | NA        | f Column... | 2020-02-03    | Discovery     | wi... spine_000...  | L1     | 15.3  | 13.2  | 0.863 | -1.9   |
| 2  |                    |                |         |            | Homb... | Blan... | 43    | 1976-12-11 | 172       | 61        | NA        | f Column... | 2020-02-03    | Discovery     | wi... spine_000...  | L2     | 15.3  | 13.2  | 0.86  | -2.1   |
| 3  |                    |                |         |            | Homb... | Blan... | 43    | 1976-12-11 | 172       | 61        | NA        | f Column... | 2020-02-03    | Discovery     | wi... spine_000...  | L3     | 16.3  | 13.0  | 0.797 | -2.8   |
| 4  |                    |                |         |            | Homb... | Blan... | 43    | 1976-12-11 | 172       | 61        | NA        | f Column... | 2020-02-03    | Discovery     | wi... spine_000...  | L4     | 18.2  | 12.9  | 0.71  | -3.5   |
| 5  |                    |                |         |            | Homb... | Blan... | 43    | 1976-12-11 | 172       | 61        | NA        | f Column... | 2020-02-03    | Discovery     | wi... spine_000...  | Total  | 65.0  | 52.2  | 0.803 | -2.6   |
| 6  |                    |                |         |            | Mujer   | Blan... | 63    | 1956-11-03 | 148       | 63        | 54        | fe Lumba... | 2020-01-02    | Explorer      | (S/... spine_000... | L1     | 10.1  | 7.94  | 0.787 | -1.3   |
| 7  |                    |                |         |            | Mujer   | Blan... | 63    | 1956-11-03 | 148       | 63        | 54        | fe Lumba... | 2020-01-02    | Explorer      | (S/... spine_000... | L2     | 9.76  | 8.41  | 0.862 | -1.5   |
| 8  |                    |                |         |            | Mujer   | Blan... | 63    | 1956-11-03 | 148       | 63        | 54        | fe Lumba... | 2020-01-02    | Explorer      | (S/... spine_000... | L3     | 12.0  | 9.76  | 0.815 | -2.4   |
| 9  |                    |                |         |            | Mujer   | Blan... | 63    | 1956-11-03 | 148       | 63        | 54        | fe Lumba... | 2020-01-02    | Explorer      | (S/... spine_000... | L4     | 10.9  | 9.72  | 0.891 | -2     |
| 10 |                    |                |         |            | Mujer   | Blan... | 63    | 1956-11-03 | 148       | 63        | 54        | fe Lumba... | 2020-01-02    | Explorer      | (S/... spine_000... | Total  | 42.7  | 35.8  | 0.839 | -1.9   |
| 11 |                    |                |         |            | Mujer   | Blan... | 58    | 1963-10-12 | 169       | 63        | NA        | f Column... | 2021-11-30    | Discovery     | wi... spine_000...  | L1     | 11.5  | 6.14  | 0.535 | -4.1   |
| 12 |                    |                |         |            | Mujer   | Blan... | 58    | 1963-10-12 | 169       | 63        | NA        | f Column... | 2021-11-30    | Discovery     | wi... spine_000...  | L2     | 9.89  | 5.58  | 0.564 | -4.2   |
| 13 |                    |                |         |            | Mujer   | Blan... | 58    | 1963-10-12 | 169       | 63        | NA        | f Column... | 2021-11-30    | Discovery     | wi... spine_000...  | L3     | 11.0  | 6.3   | 0.579 | -4.6   |
| 14 |                    |                |         |            | Mujer   | Blan... | 58    | 1963-10-12 | 169       | 63        | NA        | f Column... | 2021-11-30    | Discovery     | wi... spine_000...  | L4     | 11.9  | 6.9   | 0.579 | -4.4   |
| 15 |                    |                |         |            | Mujer   | Blan... | 58    | 1963-10-12 | 169       | 63        | NA        | f Column... | 2021-11-30    | Discovery     | wi... spine_000...  | Total  | 44.3  | 24.9  | 0.563 | -4.4   |
| 16 |                    |                |         |            | Male    | White   | 79    | 1943-10-12 | 170       | 60        | 50        | f Lubar...  | 2023-03-20    | Discovery     | wi... spine_000...  | L1     | 15.8  | 13.6  | 0.862 | -1.2   |
| 17 |                    |                |         |            | Male    | White   | 79    | 1943-10-12 | 170       | 60        | 50        | f Lubar...  | 2023-03-20    | Discovery     | wi... spine_000...  | L2     | 15.8  | 13.3  | 0.841 | -1.7   |
| 18 |                    |                |         |            | Male    | White   | 79    | 1943-10-12 | 170       | 60        | 50        | f Lubar...  | 2023-03-20    | Discovery     | wi... spine_000...  | L3     | 17.1  | 15.7  | 0.921 | -1.5   |
| 19 |                    |                |         |            | Male    | White   | 79    | 1943-10-12 | 170       | 60        | 50        | f Lubar...  | 2023-03-20    | Discovery     | wi... spine_000...  | L4     | 18.0  | 14.6  | 0.811 | -2.3   |
| 20 |                    |                |         |            | Male    | White   | 79    | 1943-10-12 | 170       | 60        | 50        | f Lubar...  | 2023-03-20    | Discovery     | wi... spine_000...  | Total  | 66.6  | 57.1  | 0.858 | -1.7   |

# i 130 more rows  
 # i 3 more variables: rm <dbl>, zscore <dbl>, am <dbl>  
 # i Use 'print(n = ...)' to see more rows