An AI-based system for medico-legal document analysis*

Matteo Marulli[1,*], Vilma Pinchi[2], Simone Grassi[2,3], Martina Focardi[2] and Marco Bertini[1]

[1]Media integration communication center (MICC), DINFO, University of Florence, viale Morgagni, 65, Firenze, Italy
[2]Section of Forensic Medical Sciences, Department of Health Sciences, University of Florence, 50134 Florence, Italy
[3]Laboratory on the Management of Healthcare Incidents (LOGOS), Department of Health Sciences, University of Florence, Florence, Italy.

Abstract

Medico-legal documents produced in hospital compensation claim procedures are heterogeneous, non-standardized, and often of poor visual quality, making manual analysis slow and error-prone. In this work, we present an end-to-end Document AI pipeline that integrates Document Layout Analysis (DLA) and Optical Character Recognition (OCR) to support the forensic medicine department of Careggi University Hospital in Florence. We construct a real-world dataset of 60 legal medicine documents (809 pages) and adopt the DocLayNet taxonomy to annotate structural components. For DLA, we fine-tune YOLOv8m on the annotated corpus, achieving an mAP@50 of 0.633 and mAP@50–95 of 0.452. For OCR, we compare Tesseract with three Qwen3-VL multimodal models and show that, after lightweight LoRA fine-tuning, Qwen3-VL reduces WER and CER by at least 74.0% and 90.9%, respectively, significantly outperforming the baseline. The resulting pipeline produces structured, machine-readable representations of medico-legal documents and lays the foundation for downstream tasks such as case retrieval, de-identification, and decision support. This study demonstrates the feasibility and benefits of applying modern Document AI and vision–language models to medico-legal workflows in real-world settings.

Keywords

Document layout analysis, Optical character recognition, Document AI, legal medicine, medical malpractice,

1. Introduction

Automatic document analysis is an emerging field at the intersection of computer vision and natural language processing. This has led to the development of end-to-end Document AI systems, designed to analyze and extract information contained in documents in a structured manner.

Natural language processing is emerging in healthcare mainly for electronic health records screening, risk prediction, and text classification [1]. However, no application of automatic document analysis in the field of medico-legal management of malpractice claims has been reported until now.

When a patient submits a compensation claim for alleged medical malpractice either personally or through lawyers or specialized agencies a formal evaluation procedure is initiated. At this stage, some Countries like Italy, physicians specialized in legal medicine are responsible for retrieving all clinical and expert documentation, analyzing it, and providing a technical assessment to a claims management committee, which will decide whether to accept or reject the claim.

These documents include medico-legal reports, technical expert reports (CTP), compensation letters, and clinical reports. This content is heterogeneous and highly sensitive, often distributed in complex,

IRCDL 2026: 22nd Conference on Information and Research Science Connecting to Digital and Library Science, February 19-20, 2026, Modena, Italy

*Corresponding author.
matteo.marulli@unifi.it (M. Marulli); vilma.pinchi@unifi.it (V. Pinchi); vilma.pinchi@unifi.it (S. Grassi); vilma.pinchi@unifi.it (M. Focardi); marco.bertini@unifi.it (M. Bertini)
https://www.micc.unifi.it/people/matteo-marulli/ (M. Marulli); https://cercachi.unifi.it/p-doc2-0-0-A-3f2b342a362f2b.html (V. Pinchi); https://cercachi.unifi.it/p-doc2-0-0-A-3f2c362d3b2829-0.html (S. Grassi); https://cercachi.unifi.it/p-doc2-2023-0-A-2c2a3b293327-0.html (M. Focardi); https://www.micc.unifi.it/people/marco-bertini/ (M. Bertini)
0009-0002-9425-8485 (M. Marulli); 0000-0002-6124-3298 (V. Pinchi); 0000-0002-2420-6658 (S. Grassi); 0000-0002-2908-7117 (M. Focardi); 0000-0002-1364-218X (M. Bertini)

non-standardized layouts and only partially digitized. A significant portion is acquired through scans of old photocopies, resulting in degraded visual quality.

Despite these challenges, automatically understanding such documents is essential to accelerate medico-legal evaluation processes and enable further NLP and Information Retrieval tasks. These include text classification, the construction of databases of similar cases, and even the use of large language models (LLMs) for content interpretation. These tools can contribute to the broader process of knowledge discovery, leveraging data mining and topic modeling techniques.

Traditionally, information extraction from this documentation was done manually by experts, in a process that was costly, slow, and prone to human error. However, recent advances in deep learning-based Document Layout Analysis (DLA) and Optical Character Recognition (OCR) have made it possible to build end-to-end pipelines capable of automatically segmenting, recognizing, and structuring content even under low visual quality conditions.

In this paper, we propose an end-to-end system for the analysis of medico-legal documents, which integrates layout analysis and OCR modules in order to support the needs of the forensic medicine department of the Careggi University Hospital (Florence).

2. Related work

2.1. Document layout analysis

In recent years, research on Document Layout Analysis (DLA) [2] has made significant progress, driven by the evolution of deep learning architectures and the availability of large annotated datasets.

Datasets such as PubLayNet [3] and DocLayNet [4] have made millions of annotated pages available for layout segmentation, paving the way for training object detection and instance segmentation networks such as Faster R-CNN [5], Mask R-CNN [6], and YOLO [7].

These models have achieved performance close to human agreement in the division of paragraphs, titles, images, and tables, significantly improving the quality of automatic document analysis.

In parallel, benchmarks such as FUNSD [8] and RVL-CDIP [9] have provided a standard reference for evaluating structural understanding and document classification capabilities.

2.2. Optical character recognition

The field of Optical Character Recognition (OCR) has evolved significantly over the past decades, transitioning from rule-based and handcrafted systems to deep learning and, more recently, multimodal architectures. Traditional OCR engines such as Tesseract OCR [10] relied on heuristic text line segmentation, connected component analysis, and character classification. While effective on clean, printed documents, these systems struggled with complex layouts, handwritten text, and noisy backgrounds.

The advent of deep learning introduced a two-stage paradigm in which text detection and recognition were modeled separately. Architectures such as CRAFT [11] and DBNet [12] considerably improved text localization accuracy, even under challenging spatial distortions. Meanwhile, recognition networks evolved from CNN–RNN hybrids to Transformer-based encoder–decoder architectures [13]. In particular, TrOCR [14] unified vision and language processing within a single Transformer model, achieving character error rates below 5% on standard benchmarks and establishing a new baseline for neural OCR.

Building on these advances, recent research has shifted toward large-scale *Vision–Language Models* (VLMs) and *Multimodal Large Models* (MLLMs) that perform end-to-end text extraction and reasoning. Models such as Florence-2 [15], Qwen2-VL [16], PaliGemma [17], and Mistral-OCR [18] integrate visual and textual modalities into a shared embedding space, enabling unified handling of detection, transcription, and semantic understanding. This new generation of instruction-tuned multimodal models can generalize across domains and tasks—including document transcription, layout-aware extraction, and visual question answering—blurring the boundaries between OCR and document understanding. As a result, OCR has evolved from a specialized perception task into a central component of comprehensive document intelligence systems.

2.3. Document analysis in the medical field

In recent years, several studies have addressed the automation of digitization and extraction of clinical information from scanned documents, which are often characterized by complex layouts and suboptimal visual quality.

Hsu et al. (2022) [19] proposed a pipeline for the automatic conversion of sleep medical reports, combining OCR with Tesseract and ClinicalBERT for the extraction of clinical parameters. The results show that the inclusion of the layout structure significantly improves classification and information extraction performance.

Similarly, Ma et al. (2023) [20] designed an end-to-end system for laboratory reports. After the OCR phase, a CRF model extracts clinical entities such as analyses, values, units of measurement, and reference ranges. On a real dataset, the system achieved over 93% OCR accuracy and an average F1-score of 0.86.

Finally, Sriram et al. (2025) [21] focused on the automatic de-identification of scanned reports. Their system uses YOLOv3 for layout segmentation and an NLP module for PHI (Protected Health Information) recognition, achieving an F1-score of 0.97 in the detection of sensitive entities.

2.4. Document analysis in the legal fields

In the legal domain, attention has recently shifted toward integrated pipelines that combine layout analysis, OCR, and semantic text understanding.

Bogdanović et al. (2025) [22] developed LexBERT, a BERT-based language model optimized for automatic correction of OCR errors in Serbian legal texts. The system corrected 88% of the incorrect words, demonstrating the effectiveness of integrating OCR and contextual models in improving the quality of text extracted from scanned documents.

In the Italian context, Marulli et al. (2025) [23] proposed a complete pipeline for Supreme Court rulings, combining YOLOv8 for layout detection and TrOCR for text recognition. The system achieved a mAP@50 of 0.964 for text area detection and a CER of 0.47%, generating a structured and anonymized dataset of judgments. Although designed for the legal domain, the methodology is also easily adaptable to medico-legal reports and expert opinions, which present similar challenges in terms of extracting sensitive content from non-standardized layouts.

Finally, OCR+NLP applications for the management of medico-legal documentation have also emerged in the industrial sector. Platforms such as Wisedocs [24] and HealthMemo [25] assist law firms and insurance companies in the automatic review of reports and expert opinions, achieving accuracies of over 90% and drastically reducing the time required to analyze cases.

3. Dataset

3.1. Data origin

To develop the system, after approval from the Ethics Committee (code: “n.24059_oss, date 19/03/2023”), we selected cases of medical malpractice claims of Careggi University Hospital (Florence, Italy) from 1 January 2010 to 30 May 2024. Cases have been anonymized and overall, the corpus provided comprises 60 documents.

These documents reflect the typical clinical-legal documentation handled in compensation claim proceedings and exhibit considerable variety. For each case, the collection generally includes:

- Letter of compensation written by the law firm representing the patient;
- Hospital medical examiner’s report;
- Report by the party-appointed expert (CTP);
- Clinical examinations;
- Memories;

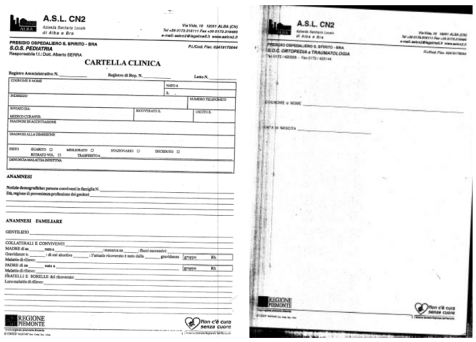

Figure 1: For privacy reasons, the image shown here is for illustrative purposes only and does not represent the data in our possession.On left a page of a document extracted from clinical notes report. On right a page of a document extracted from medical examiner's report.

Overall, the corpus provided comprises 60 documents. In Figure 1, we show a couple of illustrative examples of document images that make up our dataset. However, some of these documents have quality defects, as they are the result of multiple successive scans of the same originals.

3.2. Data pre-processing and annotation

3.2.1. Converting PDF documents into images

The construction of the system begins with the data pre-processing phase. Starting from the original PDF documents, we generate rasterized images using PyMuPDF, obtaining a total of 809 pages.

3.2.2. Annotations for the document layout analysis assignment

In this section, we describe the annotation protocol adopted for Document Layout Analysis (DLA).

For efficiency reasons, we decided to model the task as an object detection problem.

Since there are several datasets for DLA in the literature, such as DocLayNet [4], we chose to reuse its classes and annotation guidelines, with the aim of exploiting transfer learning.

The following list shows the classes used, accompanied by a brief definition:

- Caption: caption describing a figure or table;
- Footnote: note at the bottom of the page (additional text at the bottom of the page, often with references in subscript);
- List item: a single item in a list (not the entire list);
- Page footer: lower page header (e.g., page number, copyright notices, recurring footers);
- Page header: top header of the page (e.g., section/chapter title, recurring headers);
- Picture: graphic object (figures, photos, diagrams). Related sub-figures should be grouped into a single Picture;
- Section header: section header on its own line; highlighted text (bold/italic) at the beginning of a paragraph is not a section header unless it appears alone on a line;
- Table: note the table area, including the graphic lines and minimizing the white margin;
- Text: paragraph of "normal" text (discursive content) that does not fall into the above categories;
- Title: Main title (e.g., title of the document or page).

The only class not reused is Formula, as it is absent from our medico-legal documents, where mathematical expressions are rare or non-existent.

Figure 2 shows some examples of annotated documents in our dataset.

The annotations were made using the open-source tool Label Studio.

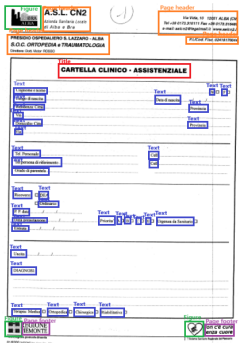

Figure 2: Example of document annotations for the DLA task. For privacy reasons, the image shown here is for illustrative purposes only and does not represent the data in our possession.

Once the dataset was labeled, we created five versions of the dataset [1] that were used to evaluate the DLA module.

3.2.3. Annotation for OCR

In addition to annotations for layout analysis, we also created a dataset for OCR and several versions of this dataset[2]. This task has now made great strides, and previously it was necessary to prepare text detection annotations to train a text detector and text extraction annotations for a text recognizer. With VLMs, only annotations for the text extraction task are needed. We therefore prepared a dataset consisting of various excerpts from different documents accompanied by their transcriptions. Figure 3 shows just a few examples that make up the text extraction dataset.

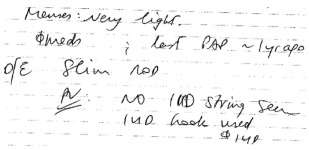

Figure 3: Example of text to be transcribed. For privacy reasons, the image shown here is for illustrative purposes only and does not represent the data in our possession.

4. Method

4.1. Proposed pipeline

Figure 4 shows the Document AI end-to-end pipeline for analysis and information extraction optimized for medico-legal documents. The pipeline consists of the following steps:

- Preprocessing: converting PDFs into JPEG images;
- Document Layout Analysis: detecting structural components using YOLOv8;
- OCR with Qwen3-VL

4.2. Document Layout Analysis with object detection

As discussed in the previous section 2, the layout analysis phase aims to identify the structural components of each page, providing a segmentation of the document that guides the OCR module to better extract textual and tabular information.

[1] We created 5 versions of the dataset using the seeds: 43, 113, 179, 187, 189.
[2] We created 3 versions of the dataset using the seeds: 32, 229, 271.

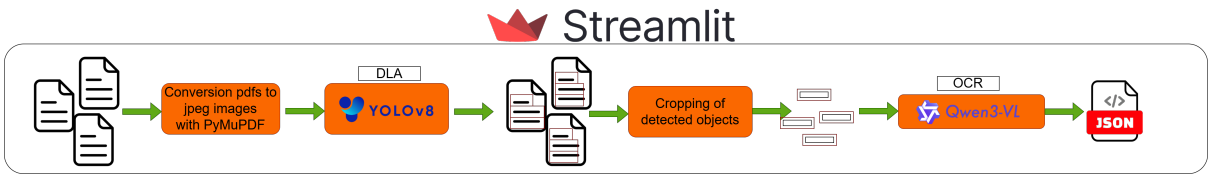

Figure 4: Our pipeline for the document analysis system for Careggi Hospital. The DLA module is implemented with a YOLOv8 model, while the OCR module is implemented with a Qwen3-VL model. The system is accompanied by a user interface that can be used by archivists and hospital staff.

This phase is based on Ultralytics' YOLOv8 architecture [26], the v8 series representing one of Ultralytics' most proven and reliable object detector solutions.

Since models pre-trained on the COCO dataset [27] are not optimized for the document domain, the detector is initialized from weights pre-trained on the DocLayNet dataset, a large-scale benchmark designed specifically for document layout analysis.

To adapt the model to the medico-legal corpus, given the small size of the dataset, transfer learning is applied by freezing the backbone and optimizing only the detection head. This approach preserves high-level structural representations while specializing the model to the distribution of the target documents. Pre-training is conducted using the hyperparameters summarized in Table 1.

The detector predicts bounding boxes and class labels corresponding to relevant layout elements, including page header, paragraph, list item, and other elements. Non-maximum suppression (NMS) [28] with an intersection over union (IoU) threshold of 0.5 is applied to eliminate redundant detections.

During inference, all predictions are exported as JSON files containing bounding box coordinates, class labels, and confidence scores. These structured outputs serve as input for the subsequent OCR stage, enabling region-level text extraction and semantic analysis.

Table 1
Training hyperparameters used for fine-tuning on the *DocLayNet* dataset.

Parameter	Value
Training epochs	40
Image size ($imgsz$)	1024
Batch size ($batch$)	8
Mosaic augmentation ($mosaic$)	1.0

4.3. Optical character recognition with VLMs

The OCR phase uses the Qwen3-VL model [29], a state-of-the-art multimodal language model (MLLM) capable of jointly processing visual and textual inputs. These models have been trained on a multitude of data and tasks, including OCR, making them highly versatile. In the pipeline, Qwen3-VL receives the components detected by the DLA module and performs text and table extraction.

Model architecture: Qwen3-VL adopts a dynamic-resolution design and introduces three key architectural innovations that improve visual understanding and text-image alignment. First, it employs an *Interleaved-MRoPE* positional encoding scheme, which interleaves temporal (t), height (h), and width (w) dimensions across all frequency bands. This yields a more uniform positional representation, enhancing spatial coherence and long-sequence comprehension. Second, Qwen3-VL integrates a DeepStack feature fusion mechanism, which aggregates multi-level representations from the Vision Transformer (ViT) [30] backbone. Visual tokens extracted from several ViT layers are injected across multiple layers of the language model, enabling finer-grained perception of both low-level text details and high-level semantic context. Third, the model refines its temporal alignment strategy via a text–timestamp mechanism that

interleaves textual and frame-based representations. These innovations enable Qwen3-VL to surpass the performance of the previous model, Qwen2-VL.

Fine-tuning for domain adaptation: To adapt Qwen3-VL to the domain of medico-legal documents, we performed lightweight fine-tuning using the Unsloth library with the LoRA (Low-Rank Adaptation) method [31]. This approach efficiently updates a small subset of parameters while preserving the general multimodal capabilities of the pretrained model, thus reducing GPU memory usage during optimization. A specialized system prompt was employed to align the model's behavior with the OCR task:

> *[ENG] System prompt:* "You are an expert assistant in extracting text from medical document images.[3]"
> *[ENG] User prompt:* "Extract all visible text from this image.[4]"

This prompt-based instruction tuning encourages consistent transcription across heterogeneous visual regions, including printed paragraphs, handwritten annotations, stamped signatures and tables. Table 2 reports the hyperparameters used in our experiments.

Table 2
Hyperparameters for Qwen3-VL.

Hyper-parameter	Value
learning rate	0.000004
lora dropout	0.05
rank	32
alpha	32
epochs	10
batch size	1

Inference: During inference, each cropped region from the layout detector (see Section 4.2) is passed to the fine-tuned Qwen3-VL model. Inference is performed through the Hugging Face interface with `bfloat16` precision to optimize GPU utilization. The transcribed outputs are stored in structured JSON files containing textual content, bounding-box coordinates, and page indices. These enriched annotations form the foundation for subsequent semantic analysis and structured information extraction. Compared to conventional OCR systems such as Tesseract or docTR, the proposed VLM-based approach demonstrates improved robustness to degraded scans, uneven lighting, and handwritten elements—owing to its multimodal contextual reasoning and deep cross-layer visual-textual alignment.

5. Analysis

5.1. Metrics

To evaluate the performance of the *Document Layout Analysis* (DLA) and *Optical Character Recognition* (OCR) modules of the proposed system, standard metrics were adopted in their respective fields of reference.

Document Layout Analysis: The performance of the YOLOv8X model fine-tuned on our dataset was evaluated in terms of Precision, Recall, and Mean Average Precision (mAP). Precision measures the proportion of correctly detected objects relative to all predicted objects:

[3]The Italian version: *[ITA] System prompt:* Sei un assistente esperto in estrazione del testo da immagini di documenti medici
[4]The italian version: *[ITA] User prompt:* "Estrai tutto il testo visibile nell'immagine.

$$Precision = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}. \tag{1}$$

Recall quantifies the model's ability to identify all objects actually present in the document:

$$Recall = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}. \tag{2}$$

Average Precision (AP) represents the area under the Precision–Recall curve and is calculated as:

$$\text{AP} = \sum_{i=1}^{n-1} (r_{i+1} - r_i)\, p_{interp}(r_{i+1}), \tag{3}$$

where $p_{interp}(r)$ is the interpolated precision defined as:

$$p_{interp}(r) = \max_{r' \geq r} p(r'). \tag{4}$$

Finally, the Mean Average Precision (mAP) corresponds to the average AP calculated across all K classes in the dataset:

$$\text{mAP} = \frac{1}{K} \sum_{i=1}^{K} \text{AP}_i. \tag{5}$$

Optical Character Recognition: Two commonly used indicators were adopted to evaluate the OCR: the Character Error Rate (CER) and the Word Error Rate (WER).

The CER quantifies the Levenshtein distance between the predicted text and the reference text, relative to the total length of the ground truth string:

$$\text{CER} = \frac{\text{S} + \text{D} + \text{I}}{N}, \tag{6}$$

where S, D, and I represent the number of substitutions, deletions, and insertions needed to transform the predicted text into the correct text, respectively, and N is the total number of characters in the ground truth.

The **WER** measures the error at the word level and is calculated as:

$$\text{WER} = \frac{\text{S} + \text{D} + \text{I}}{M}, \tag{7}$$

where M is the number of words in the reference text. This metric is more stringent than CER, as it penalizes errors that change entire words rather than single characters.

5.2. Performance analysis of the document layout analysis module

This section describes the performance of the Document Layout Analysis (DLA) module, with the aim of evaluating its accuracy in detecting different semantic regions within documents.

The first phase of the analysis compares all YOLOv8 models pre-trained on DocLayNet. Each variant was trained with the default hyperparameters of the V8 series, using the different datasets generated in our study. Table 5 shows the complete evaluation metrics: precision, recall, mAP@50, mAP@50–95, and number of model parameters.

The results reveal a non-monotonic relationship between architectural complexity and performance. YOLOv8m achieved the highest scores in both mAP@50 (0.720 ± 0.046) and mAP@50–95 (0.475 ± 0.043), despite having only 3.78 million parameters. In comparison, YOLOv8x—the largest model—showed an 8.6% performance penalty in mAP@50. Compared to the smallest model (YOLOv8n), YOLOv8m outperformed by 5.9% in mAP@50 and by 4.2% in mAP@50–95.

Table 3
Average performance (± standard deviation) of YOLOv8 models with different backbone sizes (n, s, m, l, x) on the validation set. The metrics reported are Precision, Recall, mAP@50, and mAP@50–95 calculated on the "all" class. The direction of the arrow indicates how to read the metric, with the best results for the mAP@50 and mAP@50–95 metrics shown in bold.

Model	Precision (↑)	Recall (↑)	mAP@50 (↑)	mAP@50–95 (↑)	learnable parameters
YOLOv8n	0.643 ± 0.080	0.659 ± 0.041	0.680 ± 0.047	0.461 ± 0.022	753,457
YOLOv8s	0.629 ± 0.073	0.675 ± 0.028	0.697 ± 0.045	0.472 ± 0.034	2,120,305
YOLOv8m	0.672 ± 0.062	0.661 ± 0.053	**0.720 ± 0.046**	**0.475 ± 0.043**	3,782,065
YOLOv8l	0.647 ± 0.071	0.637 ± 0.076	0.688 ± 0.041	0.456 ± 0.031	5,591,281
YOLOv8X	0.662 ± 0.060	0.623 ± 0.045	0.658 ± 0.039	0.438 ± 0.030	8,728,561

Each variant exhibited different tradeoffs between precision and recall. YOLOv8s achieved the highest recall (0.675 ± 0.028), while YOLOv8m achieved the best precision (0.672 ± 0.062). This suggests that YOLOv8m provides a better balance in the detection and classification pipeline. Metric variability remained low for almost all models, except for YOLOv8l which showed a high standard deviation in recall (±0.076), indicating possible instability in localization.

The second phase involved hyperparameter optimization. We chose to focus the optimization on the YOLOv8m variant and the dataset generated with seed 113, which represents the median case with respect to the mAP@50–95 metric. For the optimization, we used an evolutionary algorithm based on genetic mutations, performing 200 iterations with random parameter perturbations at each epoch.

The fitness function adopted to guide the search is a weighted combination of the mAP metrics:

$$f = 0.1 \times \text{mAP@50} + 0.9 \times \text{mAP@50–95} \tag{8}$$

The performance of the fitness function is illustrated in Figure 5.

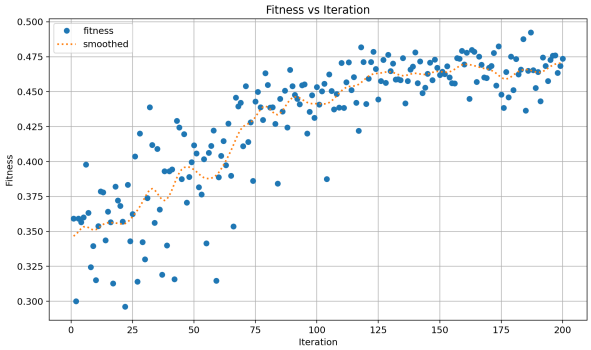

Figure 5: Fitness evolution during the 200 iterations of hyperparameter optimization. The smoothed curve (orange) highlights the convergence trend.

During the first 125 iterations, the curve shows an increasing trend, indicating that the optimization process was successful. In subsequent iterations, the function stabilizes, suggesting convergence. Figure 6 shows the distribution of fitness function values in relation to the hyperparameters. Some data augmentation parameters, such as mixup and degrees, are irrelevant in our case.

The best identified hyperparameters are reported in Table 4.

With these parameters, we re-evaluated the model on the validation and test sets. On the validation set, we observed improvements of 15.3% in precision, 11.4% in recall, and 4.7% in mAP@50–95. On the test set, we observed a 5.6% improvement in precision, a 4.3% decrease in recall, a 4.1% increase in mAP@50, and a 1% increase in mAP@50–95. Table 5 reports these results.

Finally, we retrained the YOLOv8m model on the union of the training and validation sets of the 113-seeded dataset and evaluated it on the test set, using the optimized hyperparameters. Table 6, which reports the overall and class-specific results, demonstrates that this hyperparameter configuration

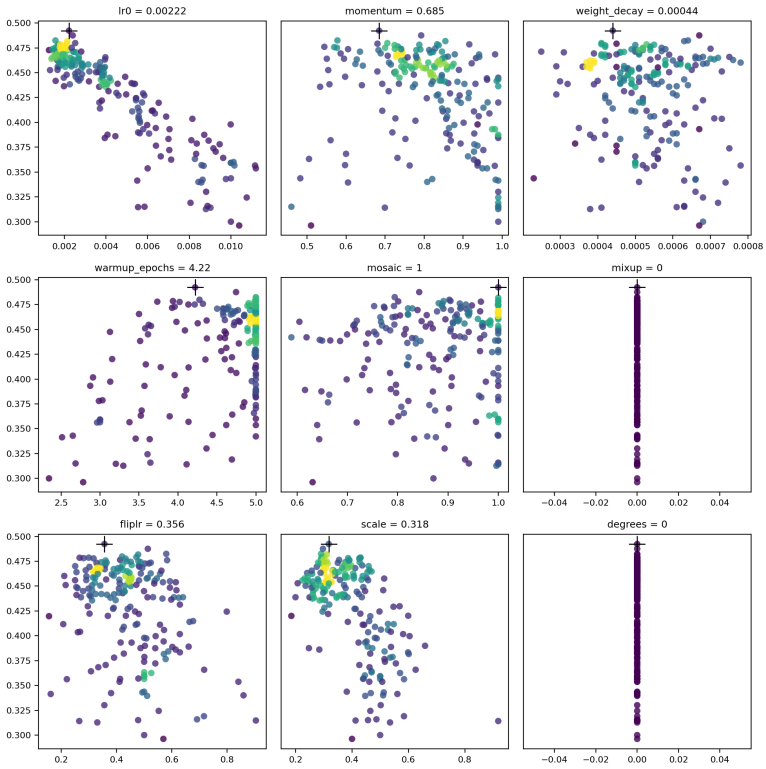

Figure 6: Scatter plot matrix of the hyper-parameter space of YOLO. Each plot represents a hyperparameter, and on the X-axis of each plot we find the value of the hyperparameter, while on the Y-axis we find the value of the fitness function. By analyzing the scatter matrix, it is possible to identify regions of values that maximize the fitness function (the regions are identified by the change in color of the points).

Table 4
Hyperparameters obtained after hyperparameter search for YOLOv8.

Hyper-parameter	Value
Epochs	50
Patience	5
lr0	0.00222
momentum	0.68472
weight_decay	0.00044
warmup_epochs	4.22022
mosaic	1.0
mixup	0.0
fliplr	0.35642
scale	0.3178
degrees	0.0

allowed us to get the most out of our available resources, maximizing the performance of the DLA module.

5.3. Performance analysis of the Optical character recognition module

This section evaluates the OCR module performance across different models and configurations, comparing baseline performance against fine-tuned results. Table 7 presents the baseline performance of Tesseract OCR and three Qwen3-VL-Instruct variants (2B, 4B, 8B parameters) using Word Error Rate

Table 5
Performance comparison of the YOLOv8m model on the seed dataset 113, in the subsequent stages of the optimization and retraining process. The reported metrics (Precision, Recall, mAP@50, mAP@50–95) are calculated on the "all" class. The *Best Params* and *Retrain All* columns indicate whether optimized hyperparameters were used and whether the model was retrained on the union of the training and validation sets. The last four columns report the percentage changes of the main metrics compared to the baseline configuration. The arrows indicate how to read the metric.

Set	Precision (↑)	Recall (↑)	mAP@50 (↑)	mAP@50–95 (↑)	Best Params	Retrain All	Δ_P (↑)	Δ_R (↑)	Δ_{mAP50} (↑)	$\Delta_{mAP50-95}$ (↑)
validation-set (seed 113)	0.588	0.621	0.694	0.470	no	no	–	–	–	–
validation-set (seed 113)	0.678	0.692	0.701	0.492	yes	no	15.3%	11.4%	1%	4.7%
test-set (seed 113)	0.569	0.628	0.585	0.389	no	no	–	–	–	–
test-set (seed 113)	0.601	0.601	0.609	0.396	yes	no	5.6%	−4.3%	4.1%	1.8%
test-set (seed 113)	0.612	0.632	0.633	0.450	yes	yes	–	–	–	–

Table 6
Results of the optimized YOLOv8m model on the seed test set 113. The arrows indicate how to read the metric.

Class	Images	Instances	Precision (↑)	Recall (↑)	mAP@50 (↑)	mAP@50-95 (↑)
all	110	773	0.612	0.632	0.633	0.452
Caption	3	5	0.237	0.255	0.333	0.156
Footnote	4	8	0.523	0.500	0.547	0.397
List-item	16	83	0.779	0.880	0.858	0.707
Page-footer	84	90	0.653	0.732	0.660	0.355
Page-header	20	26	0.748	0.577	0.572	0.441
Picture	24	33	0.645	0.727	0.774	0.593
Section-header	28	38	0.489	0.421	0.427	0.251
Table	6	9	0.750	0.556	0.686	0.575
Text	102	453	0.816	0.852	0.892	0.711
Title	25	28	0.479	0.821	0.576	0.338

(WER) and Character Error Rate (CER) metrics. All models exhibited notably higher error rates on the validation set compared to the test set (42% degradation for WER, 67% for CER), suggesting potential distribution differences between datasets. Interestingly, the pre-trained Qwen3-VL-2B achieved the best test CER (0.1279 ± 0.0349), outperforming larger variants and improving upon Tesseract's CER by 48%.

Table 7
This table shows the performance of the OCR systems before fine-tuning. Performance was measured in terms of WER and CER, and for each model, an average value and standard deviation are provided, as these values summarize the different versions of the dataset for the OCR task. The arrows indicate how to read the metric.

Model	WER (val) (↓)	CER (val) (↓)	WER (test) (↓)	CER (test) (↓)
Tesseract OCR	0.2627±0.0607	0.4569±0.2828	0.1506±0.0522	0.2469±0.2422
Qwen3-VL-2B-Instruct	0.2820±0.0444	0.3458±0.1457	0.1713±0.0403	0.1279±0.0349
Qwen3-VL-4B-Instruct	0.2950±0.0324	0.3927±0.1037	0.1797±0.0626	0.2184±0.1582
Qwen3-VL-8B-Instruct	0.2895±0.0329	0.3860±0.1020	0.1742±0.0657	0.2115±0.1604

Table 8 shows the significant improvements achieved through fine-tuning with LoRA. Fine-tuning produced exceptional results, and all Qwen3-VL models outperform not only their out-of-the-box versions but also Tesseract in terms of both WER and CER. Focusing on the smallest version, Qwen3-VL-2B, we can see that the WER decreased by an average of 74.0% (from 0.1713 to 0.0446) and the CER by 90.9% (from 0.1279 to 0.0117). The standard deviations have decreased substantially (e.g., WER from ±0.040 to ±0.001), indicating more consistent performance across different document types. Furthermore, Table 8 shows that as the size of the Qwen3-vl model increases, so does its performance. From an efficiency standpoint, Qwen3-VL-2B is ideal for implementations with limited resources, but for optimal accuracy, Qwen3-VL-8B is recommended. However, it should be noted that the improvement is negligible considering that almost twice as many parameters need to be trained. These results demonstrate that once fine-tuned, vision-language models show dramatically superior OCR performance compared to

traditional and widely used OCR engines such as Tesseract, which are often inadequate when documents have visual defects that lower document quality.

Table 8
OCR performance on test set: comparison before and after fine-tuning. Δ values indicate percentage error reduction (improvement) achieved through fine-tuning. The arrows indicate how to read the metric.

Model	WER (↓) (before)	WER (↓) (after)	CER (↓) (before)	CER (↓) (after)	Δ_{WER} (↑)	Δ_{CER} (↑)	Learnable Parameters
Tesseract OCR	0.1506±0.0522	0.1506±0.0522	0.2469±0.2422	0.2469±0.2422	—	—	0
Qwen3-VL-2B	0.1713±0.0403	0.0446±0.0008	0.1279±0.0349	0.0117±0.0040	74.0%	90.9%	47,448,096
Qwen3-VL-4B	0.1797±0.0626	0.0434±0.0071	0.2184±0.1582	0.0128±0.0014	75.8%	94.1%	78,643,200
Qwen3-VL-8B	0.1742±0.0657	0.0382±0.0052	0.2115±0.1604	0.0116±0.0008	78.1%	94.5%	102,693,888

6. Conclusion

In this work, we presented an end-to-end Document AI pipeline for the analysis of medico-legal documents from compensation claim procedures at Careggi University Hospital. Starting from 60 medico legal documents (809 pages), we built a layout analysis dataset aligned with the DocLayNet taxonomy and a complementary OCR dataset of cropped regions and transcriptions, reflecting the heterogeneity and visual degradation of real-world medico-legal documentation.

For Document Layout Analysis, a YOLOv8m detector pre-trained on DocLayNet and adapted via transfer learning and hyperparameter optimization achieved an overall mAP@50 of 0.633 and mAP@50–95 of 0.452 on the test set, with strong performance on core classes such as Text, List-item, and Picture. For OCR, multimodal Qwen3-VL models, after lightweight LoRA fine-tuning, significantly outperformed Tesseract, reducing WER and CER by at least 74.0% and 90.9%, respectively, while the 2B variant offered the best trade-off between accuracy and efficiency.

These results indicate that combining a specialized DLA module with a fine-tuned VLM-based OCR yields reliable, machine-readable representations of complex medico-legal documents, suitable for downstream tasks such as case retrieval, structured database construction, and decision support.

From a medico-legal point of view, empowering a hospital and medical malpractice claims management committees with these technologies may have several implications. First, digitalization of health information is becoming more prevalent also because of accreditations systems like Joint Commission International Accreditation [32], and then improving analysis of electronic documentation can strongly improve time efficiency. Moreover, as future research perspective, these new technologies could also contribute to risk management, a core interest of hospitals in terms of identification of sources of error and implementation of safety policies, classifying and flagging terms that may relate to specific categories of failures, like sentinel events [33].

7. Acknowledgements and notes

Funding: This research and project was funded by NextGenerationEU PNRR 2022 funds, in the context of the project: THE Tuscany Health Ecosystem (The spoke 2, CUP B83C22003920001); subproject 2.1.3 “Predictive models for healthcare claims management”, P.I. Vilma Pinchi. **Privacy concerns:** Following the Ethics Committee approval (code: n.24059_oss) dated 19/03/2023, and in compliance with privacy regulations, the image shown is for illustrative purposes only and does not represent the actual data in our possession. The example images were sourced from the web and do not contain identifiable or sensitive information.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] F. Alafari, M. Driss, A. Cherif, Advances in natural language processing for healthcare: A comprehensive review of techniques, applications, and future directions, Computer Science Review 56 (2025) 100725.

[2] G. M. Binmakhashen, S. A. Mahmoud, Document layout analysis: a comprehensive survey, ACM Computing Surveys (CSUR) 52 (2019) 1–36.

[3] X. Zhong, J. Tang, A. J. Yepes, Publaynet: largest dataset ever for document layout analysis, in: 2019 International conference on document analysis and recognition (ICDAR), IEEE, 2019, pp. 1015–1022.

[4] B. Pfitzmann, C. Auer, M. Dolfi, A. S. Nassar, P. Staar, Doclaynet: A large human-annotated dataset for document-layout segmentation, in: Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining, 2022, pp. 3743–3751.

[5] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.

[6] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.

[7] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.

[8] J.-P. T. Guillaume Jaume, Hazim Kemal Ekenel, Funsd: A dataset for form understanding in noisy scanned documents, in: Accepted to ICDAR-OST, 2019.

[9] A. W. Harley, A. Ufkes, K. G. Derpanis, Evaluation of deep convolutional nets for document image classification and retrieval, in: International Conference on Document Analysis and Recognition (ICDAR), ????

[10] R. Smith, An overview of the tesseract ocr engine, in: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), volume 2, 2007, pp. 629–633. doi:10.1109/ICDAR.2007.4376991.

[11] Y. Baek, B. Lee, D. Han, S. Yun, H. Lee, Character region awareness for text detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9365–9374. URL: https://arxiv.org/abs/1904.01941. doi:10.1109/CVPR.2019.00959.

[12] M. Liao, Z. Zou, Z. Wan, C. Yao, X. Bai, Real-time scene text detection with differentiable binarization and adaptive scale fusion, IEEE transactions on pattern analysis and machine intelligence 45 (2022) 919–931.

[13] S. Long, X. He, C. Yao, Scene text detection and recognition: The deep learning era, International Journal of Computer Vision 129 (2021) 161–184.

[14] M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, F. Wei, Trocr: Transformer-based optical character recognition with pre-trained models, in: Proceedings of the AAAI conference on artificial intelligence, volume 37, 2023, pp. 13094–13102.

[15] B. Xiao, H. Wu, W. Xu, X. Dai, H. Hu, Y. Lu, M. Zeng, C. Liu, L. Yuan, Florence-2: Advancing a unified representation for a variety of vision tasks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 4818–4829.

[16] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, et al., Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, arXiv preprint arXiv:2409.12191 (2024).

[17] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello, et al., Paligemma: A versatile 3b vlm for transfer, arXiv preprint arXiv:2407.07726 (2024).
[18] Mistral AI, Mistral ocr: High-performance multilingual text recognition model, https://mistral.ai/news/mistral-ocr/, 2025. Accessed November 2025.
[19] E. Hsu, I. Malagaris, Y.-F. Kuo, R. Sultana, K. Roberts, Deep learning-based nlp data pipeline for ehr scanned document information extraction, [pre-print / journal if known] (2022). ArXiv:2110.11864; OCR with Tesseract, ClinicalBERT, layout inclusion.
[20] M.-W. Ma, X.-S. Gao, Z.-Y. Zhang, S.-Y. Shang, L. Jin, P.-L. Liu, F. Lv, W. Ni, Y.-C. Han, H. Zong, Extracting laboratory test information from paper-based reports, BMC Medical Informatics and Decision Making 23 (2023) 251.
[21] R. Sriram, et al., De-identification of protected health information in clinical document images using deep learning and pattern matching, Journal of Electronics, Electromedical Engineering, and Medical Informatics 7 (2025) 154–164.
[22] M. Bogdanović, M. Frtunić Gligorijević, J. Kocić, L. Stoimenov, Improving text recognition accuracy for serbian legal documents using bert, Applied Sciences 15 (2025) 615.
[23] M. Marulli, G. Panattoni, M. Bertini, A document processing pipeline for the construction of a dataset for topic modeling based on the judgments of the italian supreme court, arXiv preprint arXiv:2505.08439 (2025).
[24] W. Inc., Wisedocs: Ai-powered medical document review platform, 2025. URL: https://www.wisedocs.ai/, toronto, Canada. Accessed November 2025.
[25] HealthMemo AI, Healthmemo: Intelligent processing of medical and insurance documents, https://www.healthmemo.ai/, 2025. Accessed November 2025.
[26] Ultralytics, Yolov8: Real-time object detection and image segmentation, https://docs.ultralytics.com/, 2023. Accessed November 2025.
[27] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.
[28] R. Rothe, M. Guillaumin, L. Van Gool, Non-maximum suppression for object detection by passing messages between windows, in: Asian conference on computer vision, Springer, 2014, pp. 290–306.
[29] A. Cloud, Qwen3-vl: The next generation vision-language model, https://qwen.ai/blog?id=99f0335c4ad9ff6153e517418d48535ab6d8afef, 2025.
[30] A. Dosovitskiy, An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).
[31] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., Lora: Low-rank adaptation of large language models., ICLR 1 (2022) 3.
[32] D. Al Shawan, The effectiveness of the joint commission international accreditation in improving quality at king fahd university hospital, saudi arabia: a mixed methods approach, Journal of healthcare leadership (2021) 47–61.
[33] A. B. Hooker, A. Etman, M. Westra, W. J. Van der Kam, Aggregate analysis of sentinel events as a strategic tool in safety management can contribute to the improvement of healthcare safety, International journal for quality in health care 31 (2019) 110–116.