DOTTORATO DI RICERCA IN

Ingegneria dell' informazione

CICLO XXXVIII - PNRR NextGenerationEU

Artificial intelligence methods and techniques for text understanding healthcare claims management

Settore Scientifico Disciplinare

ING-INF/05

Dottorando

Dott. Matteo Marulli

Supervisore

Prof. Marco Bertini,

Prof.ssa Vilma Pinchi

Coordinatore

Prof. Stefano Ricci

UNIVERSITÀ
DEGLI STUDI
FIRENZE

Università degli Studi di Firenze
Dipartimento di Ingegneria dell'Informazione (DINFO)
Corso di Dottorato in Ingegneria dell'Informazione
Curriculum: Computer engineering

Artificial Intelligence methods and techniques for text understanding healthcare claims management

Candidate
Matteo Marulli

Supervisors
Prof. Marco Bertini

Prof.ssa Vilma Pinchi

PhD Coordinator
Prof. Stefano Ricci

ciclo XXXVIII, 2022-2025

Università degli Studi di Firenze, Dipartimento di Ingegneria
dell'Informazione (DINFO).

I dedicate this thesis to anyone who works for the good of humanity.

Acknowledgments

I never thought that one day I would undertake a PhD. Although it has not always been a positive experience, there have been days and sometimes months that have really put me to the test. I often asked myself why I was continuing, and I found the answer along the way: I discovered that I like the world of research. It has allowed me to be a free thinker and to develop knowledge critically and independently.

I tend to be an introvert and, in the past, I mainly worked as a manual labourer to earn a living. However, this journey has changed me profoundly: it has made me realise that I can face the world and its challenges without fear, express my opinion and put myself out there even when I don't achieve the desired result. In this sense, I very much identify with Edison's thoughts on the light bulb: 'I have not failed; I have just found 1999 ways that won't work.' Failure and mistakes, if approached with the right critical spirit, allow us to learn and move forward, step by step, towards the right path. Embracing failure and learning from your mistakes is the best way to achieve the right result.

For this reason, I would like to express my deepest gratitude to my supervisor, Professor Marco Bertini, who welcomed me into his laboratory during my second year of doctoral studies. He has always supported and believed in me, and I will be forever grateful to him.

I would also like to thank Professor Shoen, who, despite not knowing me, understood a critical situation: without his intervention, I would most likely have abandoned my PhD or become a deeply repressed person.

Sincere thanks also go to Professor Pinchi, who, despite the difficulties encountered in the research, was very understanding towards me.

I would like to thank my colleagues at DISIT-lab; Marco, Enrico, Alessandro, Piero and Felix and also all my numerous colleagues at my new MICC laboratory for their kind support.

I would particularly like to thank Andrea, Quishi, Pavan and, although only visiting Micc for a short time, I would also like to thank Aki Shigesawa because, after many years of not studying Japanese, he rekindled my desire to discover the language and the country, and I must visit or live in Japan before I die (lol).

I would also like to thank Valerio and Chiara: I should have listened to them and taken them as an example much earlier, but, as they say, better late than never.

A special thought goes to my lifelong friends — Matteo, Federico, Ar-j, Manuel, Silvio, Luca, Nicole, Filippo and Antonio — who have supported me and offered me a valuable outlet during these three long years.

I cannot conclude these acknowledgements without mentioning my extraordinary family, who have always supported me despite their many doubts about my choice. Without them, I would be nothing: I will always be grateful for the love, support and education they have given me.

Contents

Chapter 1

Introduction

1.1 The objective and contributions

Document analysis is an interdisciplinary field of research that lies at the intersection of computer vision and natural language processing (NLP), with the aim of extracting, interpreting and processing information contained in structured and unstructured documents. These documents can take many forms, including digital texts, scanned images, historical documents or legal documents, often characterised by a high degree of heterogeneity from both a visual and semantic point of view. Document analysis is therefore a complex problem that requires the integration of visual and linguistic techniques to obtain an accurate and usable representation of the information content.

In recent years, this field has undergone profound changes thanks to the advent of Large Language Models (LLMs), capable of understanding and generating natural language with a high level of accuracy, and subsequently Vision-Language Models (VLMs), which extend these capabilities by integrating visual and textual information within a single model. These advances have led to a paradigm shift from traditional approaches, enabling complex tasks such as information extraction, classification, topic modelling and question answering to be tackled in a more robust and effective manner. Document analysis techniques, while having different objectives, are highly interconnected and are frequently used in conjunction with each other within complex pipelines aimed at transforming raw documents into structured knowledge.

The development of such pipelines is particularly important in highly

complex application contexts, such as medical-legal ones. In the healthcare sector, hospitals have to manage large volumes of documentation, especially in cases of alleged medical malpractice, which involve the production and analysis of numerous clinical and legal documents. Similarly, the legal domain is characterised by a considerable amount of documents, such as compensation letters, party-appointed technical consultants (CTP), court-appointed technical consultants (CTU), defence briefs, expert reports and judgements. Manual analysis of these documents is costly, prone to errors and difficult to scale, making it necessary to adopt automatic support systems.

Equipping hospitals and relevant institutions with advanced document analysis systems not only supports the management and resolution of medical malpractice cases, but also creates structured digital archives of cases handled. These archives are a valuable resource, as they allow data to be prepared and organised for subsequent quantitative and qualitative analysis, facilitating the application of data mining and text mining techniques. This approach also opens up the possibility of identifying recurring patterns, supporting epidemiological and legal studies, and improving data-driven decision-making processes.

Another critical aspect of document analysis concerns the quality of the documents themselves. Real-world documents are often affected by imperfections such as blurring, noise, scanning artefacts or degradation of the original medium, which compromise the performance of optical character recognition (OCR) systems and, more generally, information extraction algorithms. For this reason, document restoration is a fundamental step in document analysis pipelines, as it improves the visual quality of documents and, consequently, the reliability of the information extracted.

In this context, this thesis aims to address the problem of document analysis in the medical-legal field through an integrated approach. In particular, document restoration algorithms are developed to reduce the effects of blurring and noise; a complete document analysis pipeline applied to the forensic domain is designed and implemented; a dataset for topic modelling based on civil and criminal judgments of the Italian supreme court is also constructed, with the aim of analysing recurring themes. Finally, the thesis presents the development of a multimodal graph-based information retrieval system capable of retrieving relevant information from a ruling and automatically responding to queries of interest.

1.2 Contributions and Publications

This thesis is organized around the following research contributions. For each contribution, we report the related publication and the chapter where it is presented.

- **C3 — An AI-based system for medico-legal document analysis** *Related publication:* [57].
- **C4 — DocWaveDiff: A Predict-and-Refine approach for Document Image Enhancement with Wavelet U-Nets and Diffusion models** *Related publication:* [55].
- **C5 — A document processing pipeline for the construction of a dataset for topic modeling based on the judgments of the Italian Supreme Court** *Related publication:* [56].
- **C4 — A graph based multimodal model for paragraph Retrieval task in Italian Supreme Court judgments.**

A complete list of publications is reported in Chapter/Section A.

Chapter 2

Literature review

2.1 Pipeline of document analysis for medical-legal and legal field

The literature on automatic document analysis is extensive and interdisciplinary, including contributions ranging from computer vision to natural language processing, up to the most recent multimodal models. In recent years, attention has gradually shifted from isolated individual modules (e.g., OCR or text classification) to the development of integrated document analysis pipelines capable of transforming heterogeneous and unstructured documents into semantically rich representations that can be used for real-world applications. This section analyses the main lines of research relevant to the development of document analysis pipelines, with a particular focus on document layout analysis, OCR, multimodal integration, and applications in the medical and legal domains.

2.1.1 Document Layout Analysis as a Structural Backbone

Document Layout Analysis (DLA) is one of the fundamental pillars of modern document analysis pipelines, as it provides the spatial and logical structure necessary to correctly organise textual content [10]. The introduction of large-scale annotated datasets, such as PubLayNet [118] and DocLayNet [68], has enabled the training of object detection and instance seg-

mentation models based on deep architectures, including Faster R-CNN [23], Mask R-CNN [30], and YOLO [73]. These approaches have achieved performance close to human agreement in the segmentation of elements such as paragraphs, headings, tables and images, making DLA a reliable component for automated pipelines. In parallel, benchmarks such as FUNSD [26] and RVL-CDIP [28] have provided standard references for evaluating the structural understanding of documents. In modern pipelines, DLA is no longer a simple pre-processing step, but acts as a structural backbone that guides the subsequent stages of OCR, semantic segmentation, and text modelling, especially in long and complex documents such as medical reports and court rulings.

2.1.2 Evolution of OCR within Document Analysis Pipelines

OCR has undergone a profound evolution, moving from rule-based and artisanal systems to neural models and, more recently, to large-scale multimodal models. Traditional engines such as Tesseract [84], based on heuristic segmentation and character classification, have proven effective on clean documents but are severely limited in the presence of complex layouts and visual degradation. The advent of deep learning has introduced two-stage pipelines for text detection and text recognition, with models such as CRAFT [6] and DBNet [43] for text localisation, and Transformer-based architectures for recognition [51]. TrOCR [41] represented a turning point, unifying vision and language in a single encoder-decoder model. More recently, Vision-Language Models (VLMs) and Multimodal Large Language Models (MLLMs) such as Florence-2 [106], Qwen2-VL [101], PaliGemma [8], and Mistral-OCR [61] have further blurred the line between OCR and document understanding. These models enable text extraction, semantic understanding, and visual reasoning within a single architecture, making OCR a central and integrated component of Document AI pipelines.

2.1.3 Integrated Document Analysis Pipelines in Medical and Legal Domains

The application of document analysis pipelines in regulated domains such as legal-medicine and law has received increasing attention due to the need to manage large volumes of heterogeneous documents, often scanned and of suboptimal quality. In the medical field, several studies have proposed pipelines

that combine OCR, layout analysis, and NLP for the automatic extraction of clinical information. Hsu et al. [34] and Ma et al. [52] show how integrating document structure significantly improves extraction performance, while Sriram et al. [90] highlight the importance of layout segmentation for the automatic de-identification of clinical reports. Industrial solutions such as Wisedocs [38] and HealthMemo [31] confirm the application interest in OCR+NLP pipelines in the medical-legal context, although such systems are often poorly documented from a scientific point of view.

2.1.4 Privacy-Aware as requirement

An important aspect of designing these pipelines in the legal and medico-legal fields, where required, is compliance with privacy regulations. Systems based on NER [45, 81] and generalist models such as GLiNER [111] show how anonymisation can be effectively integrated into document processing flows, while remaining highly dependent on the domain and data quality.

2.2 Document image restoration and enhancement

Document image enhancement or restoration [119] focuses on improving the visual quality and readability of document images affected by degradations such as blurring, noise, or background artifacts. Traditional approaches typically employ deep learning models trained with pixel-wise losses (e.g., L1 or L2), aiming to optimize quantitative metrics like PSNR [11]. However, these metrics often fail to reflect human perceptual quality, resulting in visually subpar outputs. This limitation is particularly pronounced in GAN-based methods: although they can produce sharper edges, they are prone to mode collapse, especially in high-resolution settings.

2.2.1 Document image inpainting and text removal

Image inpainting [109] aims to restore missing or undesired regions in an image by filling them with visually plausible content. Text removal represents a specific case of this task and typically relies on text detection to accurately localize the regions to be erased. Early approaches adopted a

two-stage pipeline that combined traditional text detection techniques with standard inpainting algorithms.

With the rise of deep learning, end-to-end encoder-decoder networks have become the dominant paradigm for inpainting, including text removal. Recent methods can be broadly divided into two categories: single-stage and two-stage approaches. Single-stage models, such as EnsNet [115] and EraseNet [48], directly remove text without relying on external detectors. These architectures often follow a coarse-to-fine design and leverage local discriminators to maintain visual consistency. In contrast, two-stage methods like MTRNet [96] and CTRNet [47] integrate external text detection modules to guide the inpainting process, enhancing both the precision and robustness of the restoration.

2.2.2 Document image deblurring

Blur is a common degradation that severely reduces document readability, hindering both human interpretation and OCR performance. Early deblurring methods relied on estimating the blur kernel and applying non-blind deconvolution. With the advent of convolutional neural networks (CNNs), it became possible to perform blind deconvolution without explicitly modeling the blur kernel, leading to substantial improvements.

Later, GAN-based and transformer-based models—such as DeblurGAN [40], DEGAN [89], and DocEnTr [87]—further advanced performance. However, these models still struggle with artifacts, especially around edges, due to their reliance on pixel-level losses. More recently, diffusion-based approaches like DocDiff [110] have demonstrated the ability to generate sharper reconstructions and significantly enhance OCR accuracy.

2.2.3 Wavelet in neural networks

Wavelet transforms [112] provide a way to analyze signals, such as images or audio, by decomposing them into components at multiple scales and locations. Unlike the Fourier transform [91], which operates only in the frequency domain, wavelets retain spatial or temporal information, making them particularly suited for localized signal analysis.

In recent years, wavelet transforms have been increasingly adopted in neural networks due to their ability to capture multi-scale and multi-orientation features. This integration has yielded notable improvements in tasks such

as super-resolution [27], dehazing [108], and deblurring [42]. Within convolutional architectures like U-Net [49], wavelets enhance the preservation of high-frequency details, especially edges, leading to more accurate restoration results.

Moreover, wavelet transforms offer an efficient alternative to learned pooling operations. They enable parameter-free downsampling by replacing traditional strided convolutions with fixed, mathematically grounded operations [107]. Similarly, inverse wavelet transforms allow for upsampling without learnable parameters, further simplifying the design of encoder-decoder architectures.

2.2.4 Diffusion-based Image-to-image

Probabilistic diffusion models (DPMs) [15] have transformed generative modeling by producing highly diverse and realistic outputs. Models such as SR3 and Palette [80] showcase strong conditional generation capabilities across tasks like super-resolution, inpainting, and JPEG restoration. Whang et al. [103] introduced a "predict-and-refine" framework in Deblurring via Stochastic Refinement, where a deterministic prediction is first produced and then refined using a stochastic diffusion model. This approach enables the generation of multiple plausible reconstructions from a single blurred image, addressing the intrinsic uncertainty of the deblurring problem. Their framework laid the foundation for later models such as DocDiff [110] and NAF-DPM [12]. More recently, Shang et al. proposed ResDiff [83], which extends the predict-and-refine paradigm using frequency-domain modeling and innovative loss functions, achieving state-of-the-art results in natural image super-resolution.

Despite these advances, the combination of wavelet transforms with predict-and-refine architectures remains largely unexplored in document image restoration. In this work, we propose a novel method that integrates wavelet-based neural networks within a predict-and-refine framework. By leveraging wavelet transforms, our approach enhances the separation and reconstruction of fine image details, enabling effective restoration across tasks such as inpainting and deblurring, and resulting also in improved OCR performance.

2.3 Multimodal information retrieval for paragraph retrival

2.3.1 Information Retrieval in legal domain

Information Retrieval (IR) [78] deals with the study of methods to bridge the gap between a user request, expressed through a query, and the relevant information contained in a collection of documents. The goal is to identify and rank documents based on their likelihood of being relevant. In the legal domain [58], IR has historically been used successfully to retrieve case law, regulations and doctrine, thanks to the highly textual and structured nature of the sources, as well as the practical need to quickly identify precedents or regulatory references. However, the application of IR to the legal context, especially in languages other than English, presents significant challenges. For example, Italian legal language is highly specialised and often articulated in a way that is not very accessible [44], and documents can be very long. All this leads to a need to extract relevant content within the judgement that can answer general legal questions such as: "What was the outcome of the appeal?", "Which rules were referred to in the judgement?", "Which parties are involved and in what legal capacity?", and so on.

2.3.2 Paragraph-level Retrieval in Legal Information Retrieval

Recent years have seen increasing attention to fine-grained retrieval tasks in the legal domain, where the goal is to identify specific passages—rather than full documents—that respond to legal queries. This reflects a shift from traditional document-level retrieval systems toward more targeted approaches capable of isolating relevant fragments such as articles, clauses, or paragraphs.

To improve granularity and contextual awareness, several works have explored the use of graph-based representations of legal texts. Paragraphs or sentence blocks are often modelled as nodes in a graph, with edges encoding layout structure, citation links, or semantic similarity. Graph Neural Networks (GNNs), and in particular Graph Attention Networks (GAT) [99], have been employed to propagate contextual information across document structure, enabling better relevance estimation at the segment level [82].

The integration of multiple modalities—such as textual content, document layout, and visual cues—has also gained traction, especially in the analysis of legal PDFs and scanned materials. Multimodal models combine language encoders (e.g., BERT variants trained on legal corpora) with visual representations extracted from document regions or rendered layouts, improving robustness in the presence of noise, formatting irregularities, or ambiguous phrasing.

Training such systems presents a challenge due to the scarcity of fine-grained annotations in the legal domain. To overcome this, recent research has adopted contrastive learning objectives—such as Triplet Loss [19] or InfoNCE [3]—to learn discriminative embeddings for queries and passages. These objectives have shown promise in legal QA benchmarks, particularly when combined with weak supervision or synthetic query generation [54].

Overall, the combination of GNNs for structural modeling, multimodal encoders for layout-aware representation, and contrastive objectives for retrieval training represents a promising and actively explored direction in legal information retrieval research.

Chapter 3

An AI-based system for medico-legal document analysis

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 cases (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.

3.1 The objective

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.

Hospitals are among the main stakeholders interested in these technologies, as they must manage a considerable amount of documentation in the context of hospital claims.

When a patient submits a claim for compensation through lawyers or specialized agencies a formal procedure is activated. At this stage, the legal medicine department is responsible for retrieving all clinical and expert documentation, analyzing it, and providing a technical assessment to the 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, 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 chapter, 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).

3.2 Dataset

3.2.1 Data origin

To develop the system, we received a selection of real documents from Careggi University Hospital, which are routinely used by the forensic medicine department and the claims management committee when analyzing cases of disputes between the hospital and patients.

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;

Overall, the corpus provided comprises 60 documents. In Figure 3.1, we show two sample pages randomly extracted from the dataset, by way of example. However, some of these documents have quality defects, as they are the result of multiple successive scans of the same originals.

3.2.2 Data pre-processing and annotation

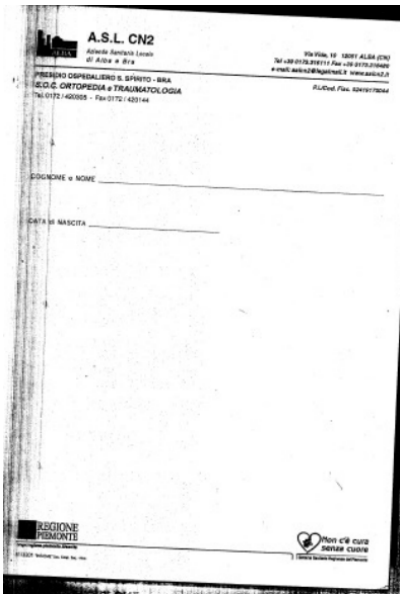
Converting PDF documents into images

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.

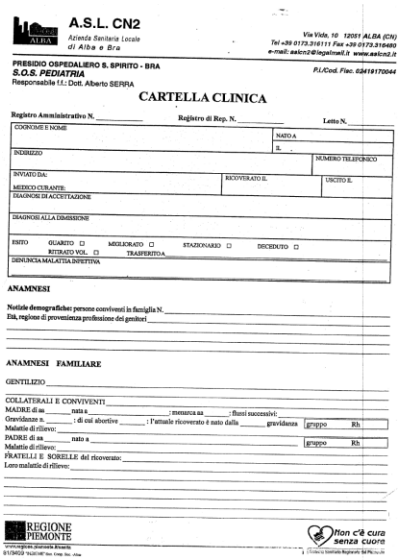
Annotations for the document layout analysis assignment

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.



(a) A page of a document extracted from medical examiner's report.



(b) A page of a document extracted from clinical notes report.

Figure 3.1: For privacy reasons, the image shown here is for illustrative purposes only and does not represent the data in our possession.

Since there are several datasets for DLA in the literature, such as DocLayNet [68], 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 subfigures 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: a paragraph of text;
- 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 3.2 shows some examples of annotated documents in our dataset.

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

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

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 3.3 shows just a few examples that make up the text extraction dataset.

[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.

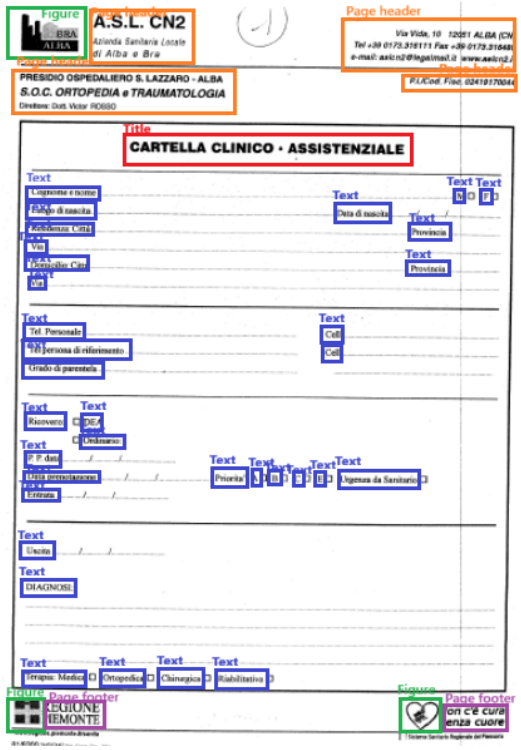


Figure 3.2: Example of document annotations for the DLA task. For privacy reasons, the image shown here is for illustrative purposes only and does not represent the data in our possession.

3.3 Method

3.3.1 Proposed pipeline

Figure 3.4 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;

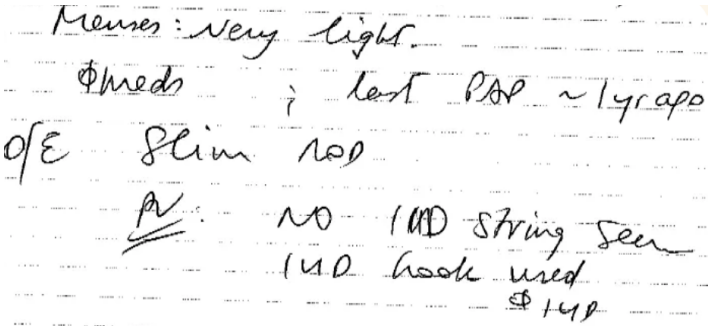


Figure 3.3: Example of text to be transcribed. For privacy reasons, the image shown here is for illustrative purposes only and does not represent the data in our possession.

- OCR with Qwen3-VL

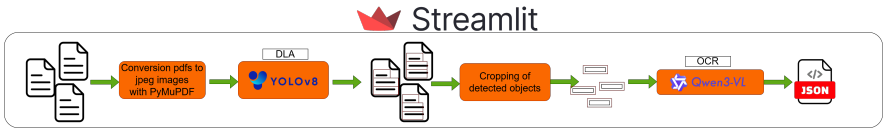


Figure 3.4: Our pipeline for the document analysis system for Careggi Hospital. The DLA module is implemented with a YOLOv8 model, while the OCR module is implemented with a Qwen3-VL model. The system is accompanied by a user interface that can be used by archivists and hospital staff.

3.3.2 Document Layout Analysis with object detection

As discussed in chapter 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.

This phase is based on Ultralytics' YOLOv8 architecture [97], the v8 series representing one of Ultralytics' most proven and reliable object detector solutions.

Since models pre-trained on the COCO dataset [46] 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 3.1.

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) [79] 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 3.1: 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

3.3.3 Optical character recognition with VLMs

The OCR phase uses the Qwen3-VL model [14], 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 per-

forms 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) [20] 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 [35]. 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 3.2 reports the hyperparameters used in our experiments.

[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.

Table 3.2: 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 3.3.2) 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.

3.4 Analysis

3.4.1 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:

$$Precision = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}. \tag{3.1}$$

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

$$Recall = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}. \tag{3.2}$$

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

$$\text{AP} = \sum_{i=1}^{n-1} (r_{i+1} - r_i)\, p_{interp}(r_{i+1}), \tag{3.3}$$

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

$$p_{interp}(r) = \max_{r' \geq r} p(r'). \tag{3.4}$$

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

$$\text{mAP} = \frac{1}{K} \sum_{i=1}^{K} \text{AP}_i. \tag{3.5}$$

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:

$$\text{CER} = \frac{\text{S} + \text{D} + \text{I}}{N}, \tag{3.6}$$

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:

$$\text{WER} = \frac{\text{S} + \text{D} + \text{I}}{M}, \tag{3.7}$$

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.

3.4.2 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 3.5 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.

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 combina-

tion of the mAP metrics:

$$f = 0.1 \times \text{mAP@50} + 0.9 \times \text{mAP@50–95} \tag{3.8}$$

The performance of the fitness function is illustrated in Figure 3.5.

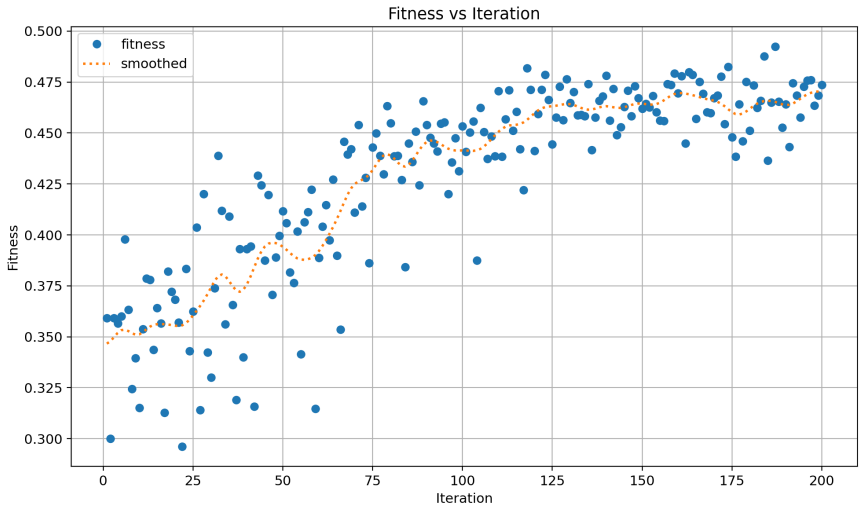


Figure 3.5: Fitness evolution during the 200 iterations of hyperparameter optimization. The smoothed curve (orange) highlights the convergence trend.

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 3.6 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 3.4.

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 3.5 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 3.6, which reports the overall and class-specific results, demonstrates that this hyperparameter configuration allowed us to get the most out of our available resources, maximizing the performance of the DLA module.

3.4.3 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 3.7 presents the baseline performance of Tesseract OCR and three Qwen3-VL-Instruct variants (2B, 4B, 8B parameters) using Word Error Rate (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 3.8 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 3.8 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.

3.5 Conclusion

In this chapter, 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 real cases (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.

Table 3.3: 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

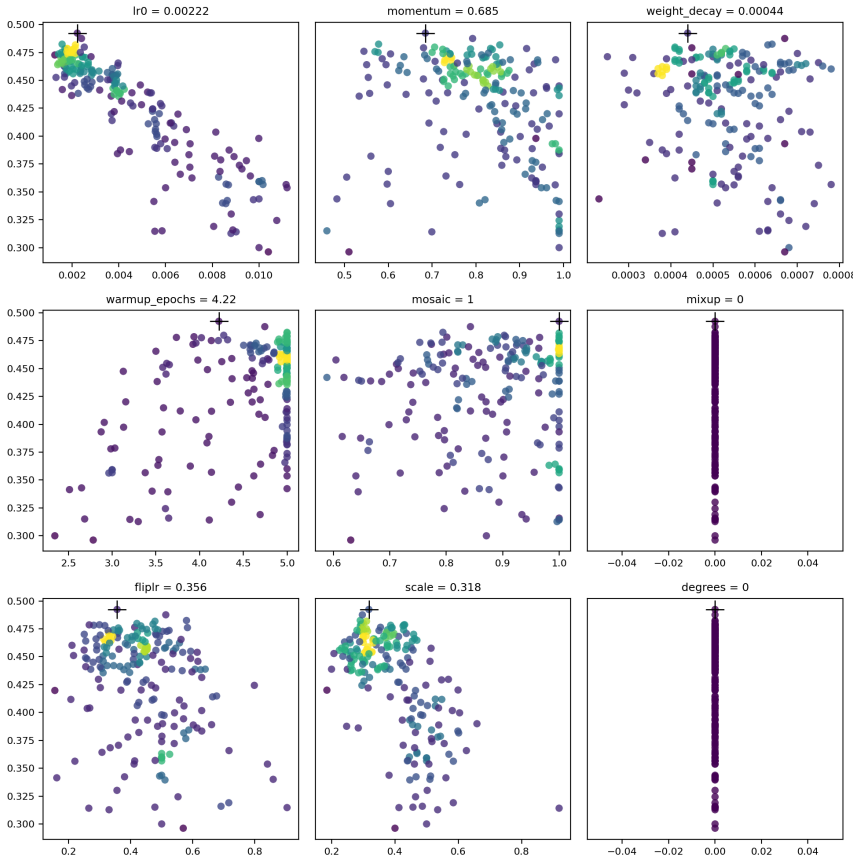


Figure 3.6: Scatter plot matrix of the hyper-parameter space of YOLO. Each plot represents a hyperparameter, and on the X-axis of each plot we find the value of the hyperparameter, while on the Y-axis we find the value of the fitness function. By analyzing the scatter matrix, it is possible to identify regions of values that maximize the fitness function (the regions are identified by the change in color of the points).

Table 3.4: Hyperparameters obtained after hyperparameter search for YOLOv8.

Hyper-parameter	**Value**
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

Table 3.5: 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 3.6: Results of the optimized YOLOv8m model on the seed test set 113. The arrows indicate how to read the metric.

Class	**Img.**	**Inst.**	**Prec.** (↑)	**Rec.** (↑)	**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

Table 3.7: 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 3.8: 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

Chapter 4

DocWaveDiff: A Predict-and-Refine approach for Document Image Enhancement with Wavelet U-Nets and Diffusion models

OCR and document layout analysis algorithms are essential components of AI-based document systems, yet they are typically trained on clean, degradation-free images. When applied to degraded documents such as blurred scans or pages spoiled by undesired handwritten text their performance drops significantly. To address this issue, we propose DocWaveDiff, a novel document restoration method based on a predict-and-refine diffusion framework incorporating wavelet U-Nets. Given a degraded image patch and optionally its prior features, our Early Predictor generates an initial restoration, which is then refined by a Denoiser Refiner that estimates the residual image. The combination of these outputs yields the final restored result. We evaluate DocWaveDiff on multiple public benchmarks and demonstrate its strong performance across various document degradation scenarios, including deblurring and handwriting removal. Our results confirm that integrating wavelet transforms into the predict-and-

Figure 4.1: Examples of degraded document images: (left) blurred printed text; (right) handwritten annotations over printed Chinese text.
refine framework enhances restoration quality and supports more robust document understanding systems.

4.1 Introduction

Document images often suffer from various types of degradation, such as blurring, textual artifacts (see Figure 4.1), which significantly reduce their readability and hinder automatic processing through optical character recognition (OCR) systems [62] and document layout algorithms [10]. Effective document restoration methods are therefore essential for enhancing image quality, improving human readability, and enabling accurate automated analysis.

Traditional document restoration techniques typically rely on heuristic approaches or classical image processing methods, focusing on thresholding algorithms and kernel-based convolution. However, these approaches frequently struggle to handle complex degradations, often resulting in suboptimal restoration outcomes. Recent advances in deep learning have significantly improved restoration performance, particularly with convolutional neural networks (CNNs) and generative adversarial networks (GANs) [24]. Nevertheless, these methods still exhibit limitations, particularly in retaining fine textual details and preserving high-frequency features critical for OCR applications.

To address these challenges, we propose DocWaveDiff, a novel predict-and-refine framework specifically tailored for document image enhancement and restoration. Our approach uniquely combines multi-scale wavelet de-

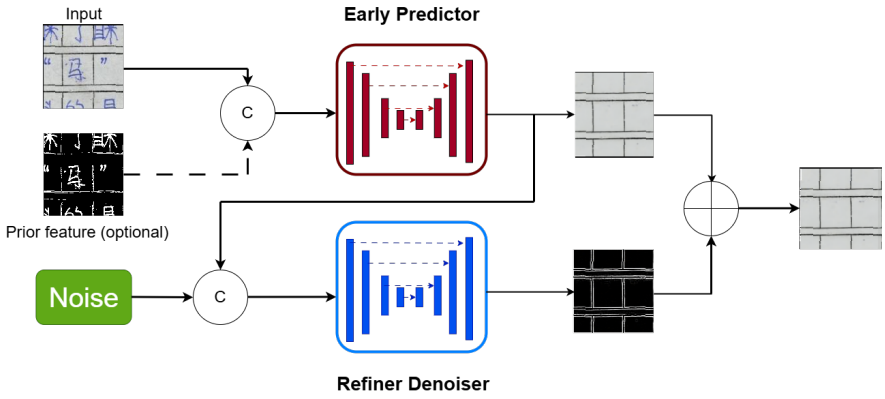


Figure 4.2: General schema of a predict and refine diffusion model.

composition, which effectively captures spatial-frequency information, with diffusion models, renowned for their generative capabilities. The resulting architecture leverages the strengths of wavelet-based U-Net networks [77] to accurately restore degraded images while preserving crucial high-frequency details.

In this chapter, we thoroughly evaluate DocWaveDiff across several benchmark tasks, including document image inpainting and deblurring. Our method achieves state-of-the-art performance, significantly surpassing existing solutions in both quantitative metrics and qualitative visual fidelity, and effectively addresses common document restoration challenges improving the performance in tasks like OCR. We conclude by identifying potential areas for future development and further application domains, setting a foundation for future work in document image restoration.

4.2 Methodology

The architecture of the proposed method is composed of two neural networks arranged sequentially, and is shown in Figure 4.2. The first one, termed the early predictor, reconstructs the low-frequency components and captures the overall structure of the degraded image. The second one, called the refiner net, is implemented as a diffusion model and estimates the residual signal—primarily the high-frequency details such as edges and textures.

The final output is obtained by summing the predictions of both networks. Training is performed jointly in an end-to-end fashion, allowing the two modules to complement each other throughout the learning process.

4.2.1 Early Predictor

Given a degraded input patch $I \in \mathbb{R}^{c \times m \times n}$, the goal of the early predictor E is to produce an initial restoration, defined as:

$$\hat{I} = E_\theta(I) \tag{4.1}$$

The early predictor is designed to restore the low-frequency content of the image and remove undesirable artifacts. This step is particularly effective for document images, where it reconstructs background regions and fills the interiors of text characters with coherent content. However, E alone struggles to recover high-frequency information, resulting in blurred edges and missing fine details especially in the deblurring task. To address this limitation, a subsequent refinement network is introduced to recover the lost high-frequency components.

4.2.2 Diffusion Refinement Network

To recover the missing high-frequency details, we introduce a refinement network based on a conditional probabilistic diffusion model. Following the strategy proposed in [12, 110], the network learns to model the distribution of residuals between the ground truth image and the initial prediction.

Formally, the residual signal is defined as:

$$I_0 = I_{res} = I_{GT} - \hat{I} \tag{4.2}$$

By learning to denoise samples progressively toward I_{res} the diffusion model refines the initial estimate $\hat{I}$, effectively restoring fine details such as edges and textures that are often lost in the early prediction.

Forward noise-adding process The forward diffusion process defines a Markov chain that progressively adds Gaussian noise to a clean residual sample I_O. At each time step t, the noisy sample I_t is obtained from the previous sample I_{t-1} through a Gaussian transition kernel:

$$q(I_t \mid I_{t-1}) := \mathcal{N}(I_t; \sqrt{\alpha_t} I_{t-1}, (1 - \alpha_t)\mathbf{I}) \tag{4.3}$$

This defines a complete Markov process over T steps:

$$q(I_{1:T} \mid I_0) = \prod_{t=1}^{T} q(I_t \mid I_{t-1}) \tag{4.4}$$

For efficiency, the marginal distribution at any timestep t can also be computed directly from the initial sample I_O as:

$$q(I_t \mid I_0) = \mathcal{N}(I_t; \sqrt{\bar{\alpha}_t} I_0, (1 - \bar{\alpha}_t)\mathbf{I}) \tag{4.5}$$

Here, $\{\alpha_t\}_{t=1}^{T}$ denotes the cumulative product.

By applying the reparameterization trick, Equation 4.5 can be expressed as:

$$I_t = \sqrt{\bar{\alpha}_t} I_0 + \sqrt{1 - \bar{\alpha}_t}, \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{4.6}$$

This formulation allows direct sampling of noisy inputs without sequentially running the full Markov chain.

Reverse denoising process: The generative mechanism of the model is governed by a reverse denoising process, which constitutes its core component. This process is modeled as an inverse Markov chain, starting from a Gaussian noise sample $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and iteratively producing a sequence of latent variables $\mathbf{x}_{T-1}, \mathbf{x}_{T-2}, \ldots, \mathbf{x}_0$ that progressively converge toward the target data distribution.

Each reverse step is defined as a conditional Gaussian distribution:

$$q(I_{t-1} \mid I_t, I_0) = \mathcal{N}\left(I_{t-1}; \mu_t(I_t, I_0), \beta_t(I_t, I_0)\mathbf{I}\right) \tag{4.7}$$

where both the mean μ_t and variance β_t depend on the current state I_t depend on the current state and the clean reference I_O.

Following the approach of [86, 110], we adopt a deterministic version of the reverse process by setting the variance to zero:

$$\beta_t(I_t, I_0) = 0 \tag{4.8}$$

Under this assumption, the mean simplifies to:

$$\mu_t(I_t, I_0) = \sqrt{\bar{\alpha}}_{t-1} I_0 + \sqrt{1 - \bar{\alpha}}_{t-1} \cdot \frac{I_t - \sqrt{\bar{\alpha}_t} I_0}{\sqrt{1 - \bar{\alpha}_t}} \tag{4.9}$$

The term inside the fraction corresponds to the noise component ϵ, recovered from I_t and I_O.

To make the process learnable, we introduce a denoising network D_θ, which estimates the residual $\hat{I}_{res}$ from a noisy input I_t, its timestep t, and the initial prediction $\hat{I} = E_\theta(I)$. This results in the parameterized reverse distribution:

$$p_\theta(I_{t-1} \mid I_t, \hat{I}) = q(I_{t-1} \mid I_t, D_\theta(I_t, t, \hat{I})) \tag{4.10}$$

The predicted residual is given by:

$$\hat{I}_{\text{res}} = D_\theta\left(I_t, t, E_\theta(I)\right) \tag{4.11}$$

Using $\hat{I}_{res}$ in place of the ground truth residual $\hat{I}_O$, the final deterministic reverse step becomes:

$$I_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{I}_{\text{res}} + \sqrt{1-\bar{\alpha}_{t-1}} \cdot \frac{I_t - \sqrt{\bar{\alpha}_t}\hat{I}_{\text{res}}}{\sqrt{1-\bar{\alpha}_t}} \tag{4.12}$$

This formulation of the reverse denoising process is similar to that of diffuse implicit models for noise reduction (DDIMs) [86] and allows the model to iteratively reconstruct the residual signal without stochastic sampling, enabling faster inference and stable predictions.

4.2.3 2D Wavelet Transform

The two dimensional Discrete Wavelet Transform (2D DWT) is a mathematical technique that decomposes an image into subbands capturing both spatial and frequency information across multiple scales. Unlike traditional downsampling operations such as max pooling or strided convolutions, the DWT is theoretically invertible and lossless, allowing for the preservation of high frequency details that are critical in reconstruction tasks like dehazing, denoising, and super-resolution.

A single level 2D DWT decomposes an input image patch I into four frequency components: I_{LL}: low frequency (approximation) component and I_{LH}, I_{HL}, I_{HH}: high frequency components representing horizontal, vertical, and diagonal details respectively.

This decomposition is performed by convolving the image with a pair of one dimensional wavelet filters: a scaling function $\phi(x)$ (low-pass) and a wavelet function $\psi(x)$ (high-pass). These filters are applied along both image axes:

$$I_{LL}(x,y) = \phi(x)\phi(y) * I(x,y), \tag{4.13}$$

$$I_{LH}(x,y) = \phi(x)\psi(y) * I(x,y), \tag{4.14}$$

$$I_{HL}(x,y) = \psi(x)\phi(y) * I(x,y), \tag{4.15}$$

$$I_{HH}(x,y) = \psi(x)\psi(y) * I(x,y), \tag{4.16}$$

followed by a downsampling operation along both dimensions.

By separating low and high frequency content, the DWT enables better structural representation and noise isolation. Its integration into deep neural networks provides clear advantages: using DWT instead of pooling preserves edge and texture information during resolution reduction, while the inverse transform (IDWT) enables exact reconstruction of feature maps.

This property supports fully differentiable and reversible architectures, making wavelet-based networks well suited for dense prediction tasks in document and natural image restoration.

Haar Filter Bank: In this work, we adopt the 2D Haar wavelet transform due to its simplicity, efficiency, and exact invertibility. The transform employs a set of fixed 2×2 filters to compute the four subbands:

$$LL = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \qquad LH = \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix},$$

$$HL = \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}, \qquad HH = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$

These filters are orthogonal and separable, and are applied via convolution followed by a downsampling operation with stride 2. Each filter extracts different frequency components: LL captures the low-frequency approximation, while LH, HL, and HH extract horizontal, vertical, and diagonal high-frequency details, respectively.

The inverse transform (IDWT) uses the same filters to reconstruct the original signal exactly, without any learnable parameters. This property makes the Haar wavelet transform a compelling alternative to pooling operations, offering resolution reduction with no information loss—an essential advantage in image restoration tasks.

4.2.4 Wavelet U-net

The early predictor E and the denoising refiner D are implemented using a Wavelet U-Net architecture enhanced with residual and attention mechanisms, specifically tailored for document restoration tasks. We adopt a modified U-Net that replaces standard downsampling and upsampling operations with wavelet-based transforms. This design choice is motivated by

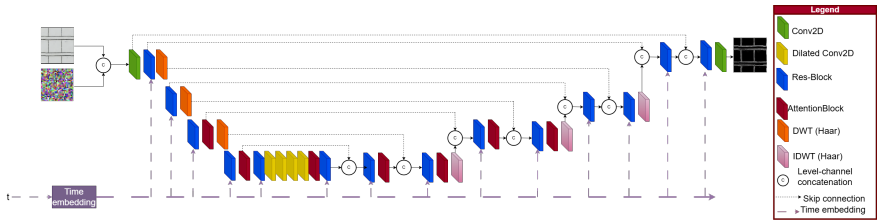


Figure 4.3: The wavelet U-Net architecture for the Refiner Denoiser. The design is accompanied by a legend illustrating the blocks that make up the network. Details of the design of the ResBlock, Attention Block, and time embedding are provided in the supplementary materials.

the ability of wavelet transforms to preserve both spatial and frequency-domain information—an essential property for recovering sharp edges and fine textures in degraded document images.

The architecture follows a symmetric encoder–decoder layout, where feature maps are downsampled using the 2D Haar Wavelet Transform and upsampled using its inverse (IDWT). Each resolution level in both paths is processed by a `ResidualBlock`, optionally followed by an `AttentionBlock`. An overview of the architecture is shown in Figure 4.3. For the sake of space, in Figure 4.3 we report only the architecture of the Denoiser Refiner diffusion model, which is richer in detail than the Early Predictor network, which shares many implementation details of the Denoiser Refiner. In addition to the network, diagrams of the time embedding and Res-blocks underlying the Denoiser Refiner are given.

In the encoder, the input image patch is first decomposed into four subbands (LL, LH, HL, HH) using the Haar transform. These components are concatenated along the channel dimension, allowing the model to jointly process both low-frequency structure and high-frequency detail. The resulting tensor is then passed through a sequence of `DownBlocks`, each composed of a residual unit and, if enabled, a self-attention module. Temporal embeddings are injected within the residual blocks to condition the model on the diffusion timestep.

To maintain computational efficiency, the self-attention mechanism is only activated at the last two resolution levels, where the spatial size of feature maps is reduced. This selective application allows the model to capture global context where it is most beneficial, without incurring the high memory cost associated with attention on large feature maps.

At the network's bottleneck lies a `MiddleBlock`, which employs dilated convolutions with increasing dilation rates (2, 4, 8, 16) to enlarge the receptive field without deepening the architecture. This block also includes an attention module and an additional residual unit, refining intermediate features before reconstruction.

In the decoder path, feature maps are progressively upsampled using the inverse wavelet transform. At each resolution level, the upsampled tensor is concatenated with the corresponding encoder feature map via skip connections and processed by an `UpBlock`, which mirrors its downsampling counterpart with residual and optional attention components.

4.2.5 Training

We now describe the training objectives and procedure for the Early Predictor E_θ and the Denoiser Refiner D_θ, which are jointly optimized in an end-to-end fashion.

Frequency Separation Training: Following the strategy proposed by [110], we adopt Frequency Separation Training (FST) to enhance the reconstruction quality of document images. This technique explicitly decomposes the learning objective into low-frequency and high-frequency components, allowing the network to handle structural and textural details separately during training.

To achieve this, we apply a low pass filter ϕ_H and a Laplacian high pass filter ϕ_L to both predicted outputs and training targets. The low frequency loss encourages accurate reconstruction of global image structure:

$$\mathcal{L}_{\text{low}} = \mathbb{E}\|\phi_L * (\hat{I} - I_{\text{GT}})\|_2 \tag{4.17}$$

while the high-frequency loss focuses on the recovery of fine details from the residual prediction produced by the denoising network:

$$\mathcal{L}_{\text{high}} = \mathbb{E}\|\phi_H * \Big(I_0 - D_\theta(I_t, t, \hat{I})\Big)\|_2. \tag{4.18}$$

This separation improves the network's ability to preserve sharp edges and textures while maintaining global consistency, which is particularly beneficial for document restoration tasks.

Pixel-loss: To guide the early predictor E in recovering the coarse structure of the degraded image, we minimize a standard pixel wise reconstruction

loss. This objective encourages the network to accurately restore the low-frequency content and suppress undesired artifacts:

$$\mathcal{L}_{\text{pixel}} = \mathbb{E}\|I_{\text{GT}} - \hat{I}\|_2 \tag{4.19}$$

where $\hat{I} = E(I)$ is the initial prediction produced by the early predictor, and I_{GT} is the corresponding ground truth image.

Denoiser loss: The training objective for the denoising network is to minimize the discrepancy between the model's predicted distribution $p_\theta(x_{t-1}|x_t, x_C)$ and the true reverse posterior $q(I_{t-1}|I_t, I_{0)}$. This is achieved by minimizing the following L2 reconstruction loss:

$$\mathcal{L}_{\text{DM}} = \mathbb{E}\|I_0 - D_\theta\left(\sqrt{\bar{\alpha}_t} I\text{res} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t, \hat{\hat{I}}\right)\|_2 \tag{4.20}$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\hat{\hat{I}}$ is a detached clone of the early prediction $\hat{I}$. During backpropagation, gradients are allowed to flow only through the residual input I_{res} and the denoising network D_θ while $\hat{\hat{I}}$ remains fixed and does not participate in gradient updates. As a result, the loss propagates indirectly to the early predictor E_ϵ through its contribution to $I_{res} = I_0 - \hat{I}$.

Final loss: To enable end to end training of both the early predictor and the denoiser refiner E_θ, we define the total training objective as a weighted combination of the pixel level loss, the diffusion model loss, and the frequency-separated losses. The overall objective function is:

$$\mathcal{L}_{\text{total}} = \beta_1\left(\mathcal{L}_{\text{pixel}} + \beta_0\mathcal{L}_{\text{low}}\right) + \left(\mathcal{L}_{\text{DM}} + \beta_0\mathcal{L}_{\text{high}}\right) \tag{4.21}$$

where β_0 and β_1 are scalar hyperparameters that control the relative weight of low and high-frequency objectives with respect to the primary reconstruction losses. In our experiments, we set $\beta_0 = 2$ and $\beta_0 = 0.5$, following the configuration proposed in [110].

By minimizing $\mathcal{L}_{\text{total}}$ via gradient descent, the model jointly optimizes both components: the early predictor is encouraged to focus on low-frequency structure and artifact removal, while the denoising refiner learns to enhance high-frequency content such as edges and textures.

The complete training and inference procedures are summarized in Algorithm 4.4.

Algorithm 1 DocWaveDiff: Training and Sampling

Require: E_θ: Early predictor, D_θ: Denoiser refiner, I: degraded patch, ϕ_{H}: high-frequency filter, I_{gt}: clean patch, $\alpha_{0:T}$: noise schedule

Training:
repeat
  Sample $(I, I_{\mathrm{gt}}) \sim q(I, I_{\mathrm{gt}}), t \sim \mathrm{Uniform}\{1, \ldots, T\}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\hat{I} \leftarrow E_\theta(I)$
  $I_t \leftarrow \sqrt{\bar{\alpha}_t}(I_{\mathrm{gt}} - \hat{I}) + \sqrt{1 - \bar{\alpha}_t}\,\epsilon$
  $\hat{I}_{\mathrm{res}} \leftarrow D_\theta(I_t, t, \hat{I})$
  Take a gradient step on $\mathcal{L}_{\mathrm{total}}(\hat{I}, \hat{I}_{\mathrm{res}}, I_{\mathrm{gt}}, \phi_{\mathrm{H}}; E_\theta, D_\theta)$
    // Frequency separation loss; see Eq. 21
until converged

Sampling:
$\hat{I} \leftarrow E_\theta(I)$
$I_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T$ **to** 1 **do**
  $\hat{I}_{\mathrm{res}} \leftarrow D_\theta(I_t, t, \hat{I})$
  $I_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}}\,\hat{I}_{\mathrm{res}} + \frac{\sqrt{1-\bar{\alpha}_{t-1}}\left(I_t - \sqrt{\bar{\alpha}_t}\,\hat{I}_{\mathrm{res}}\right)}{\sqrt{1-\bar{\alpha}_t}}$
end for
return $\hat{I} + I_0$

Figure 4.4: DocWaveDiff: Training and Sampling. The pseudo code has been transformed to resolve incompatibility issues with the pseudo codes in other chapters that use specific packages.

4.3 Experiments and Results

4.3.1 Datasets and Implementation details

DocWaveDiff[1] is designed to address multiple types of common degradations in document image restoration.

Document image inpainting: For the document image inpainting task, we selected the EnsExam dataset [36], which contains 545 exam sheet images, split into 430 training and 115 test samples. Each image is annotated with visual cues that define meaningful regions for removal, making the dataset suitable for supervised inpainting experiments.

For this task we will use a prior feature, a binary mask that will help the model focus on the pixels to be edited and ignore the others to eliminate handwritten text.

However, EnsExam does not provide binary masks to explicitly indicate the regions to be restored. We generate the masks exactly as the authors of the EnsExam dataset did. We compute the absolute pixel-wise difference between the corrupted image I and the ground truth I_{GT} , and apply Otsu's thresholding method T_O to obtain a binary mask:

$$I_M = T_O(|I - I_{GT}|) \tag{4.22}$$

where the resulting mask $I_M \in \{0,1\}^{m \times n}$ highlights the pixels corresponding to missing or corrupted content.

These masks are called coarse stroke masks by the original authors, who emphasize that these masks are not perfect because there is noise that can confuse the models.Finally, for both training and test sets, we divide I, I_{GT} and I_M into non-overlapping patches of size 128×128, which are used as inputs for the inpainting model.

Document image deblurring: For the document image inpainting task, we used the TDD dataset [33], composed of 66,743 training pairs and 1,601 test pairs of clean and blurry image patches, each of size 300×300 pixels.

4.3.2 Evaluation Metrics

To evaluate and compare the performance of the proposed method, we adopt supervised evaluation metrics tailored to each specific restoration task. The

[1]Source code and checkpoints: `https://github.com/miccunifi/DocWaveDiff`

Table 4.1: Document inpainting results on the test set of EnsExam dataset. The comparison is at the patch level (mask-guided), and the best results are in **bold**.

Model	**MSE** (↓)	**MS-SSIM** (↑)	**PSNR** (↑)
Pix2Pix [39]	0.16	0.8979	29.01
EnsNet [115]	0.07	0.9551	33.54
EraseNet [48]	0.07	0.9355	33.54
MTRNet++ [95]	0.10	0.9234	31.23
Net_{refine} [36]	0.06	0.9647	35.44
DocWaveDiff (our)	**0.0003**	**0.9817**	**40.37**

evaluation is conducted as follows:

Document Image Inpainting: we use the Multi-Scale Structural Similarity Index Measure (MS-SSIM) [102], Mean Squared Error (MSE) and PSNR [32] to assess perceptual and pixel-wise accuracy of the reconstructed content.

Document Image Deblurring: we employ the Learned Perceptual Image Patch Similarity (LPIPS) [114] and the standard Structural Similarity Index Measure (SSIM) [65], which jointly evaluate perceptual similarity and structural fidelity and PSNR.

In addition, for the deblurring task, we evaluate the performance of OCR on the restored images in terms of character error rate (CER) [67] and word error rate (WER) [4].

Document image inpainting

We assess the effectiveness of DocWaveDiff on the document image inpainting task using the EnsExam dataset, which consists of exam sheet images annotated for visually meaningful text removal. Quantitative results are reported in Table 4.1.

DocWaveDiff achieves state-of-the-art results, significantly outperforming previous methods in all evaluated metrics. Specifically, it reaches an MSE of 0.0003, a MS-SSIM of 0.9817, and a PSNR of 40.37 dB, highlighting its superior ability to reconstruct complex textual structures and retain high-frequency details that are often lost in conventional architectures.

In addition to patch-level evaluation, we also perform full-document restora-

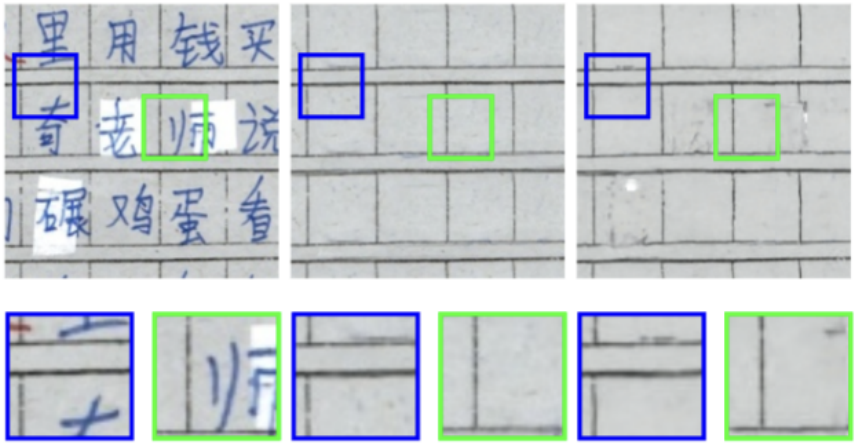


Figure 4.5: Example 41 from the test-set of EnsExam: *Left)* Input; *Mid)* Ground truth; *Right)* Output. The blue and red handwritten text has been completely removed, and the white regions of the document have been restored.

tion. As shown in Figure 4.5, DocWaveDiff removes handwritten annotations and reconstructing background areas with high visual fidelity.

Table 4.2 reports the corresponding quantitative results on whole-document inference, confirming that DocWaveDiff maintains its superiority even at larger input scales.

4.3.3 Document image deblurring

We evaluate DocWaveDiff on the Document Image Deblurring task using the TDD dataset, which provides paired samples of clean and blurred document patches. The quantitative results are reported in Table 4.3.

DocWaveDiff achieves state-of-the-art performance in perceptual quality (LPIPS of 0.0211) and structural similarity (SSIM of 0.9733), while also reaching a competitive PSNR of 26.28 dB. These results confirm the model's ability to preserve both text sharpness and document layout integrity, which are essential for subsequent OCR and document understanding tasks.

Figure 4.6 presents visual comparisons, showing that DocWaveDiff successfully reconstructs fine details—such as edges, strokes, and character

Table 4.2: Document inpainting results on the test set of EnsExam dataset. The comparison is at the image document level (mask-guided), and the best results are in **bold**.

Model	**MSE** (↓)	**MS-SSIM** (↑)	**PSNR** (↑)
Pix2Pix [39]	0.16	0.8954	28.99
EnsNet [115]	0.07	0.9493	33.87
EraseNet [48]	0.07	0.9369	33.54
MTRNet++ [95]	0.08	0.9264	32.77
Net_{refine} [36]	0.05	0.9659	36.05
DocWaveDiff	**0.0002**	**0.9849**	**39.54**

Table 4.3: Document deblurring results on the TDD test set. ↓ lower is better, ↑ higher is better. Best in **bold**.

Model	Parameters	**LPIPS** (↓)	**SSIM** (↑)	**PSNR** (↑)
DE-GAN [89]	31 M	0.0843	0.9226	22.24
Text-DIAE [88]	—	—	—	23.58
DocEnTr [87]	67 M	0.9130	0.2225	21.28
DocDiff [110]	8.2 M	0.0307	0.9505	23.28
DocRes [113]	—	—	0.9723	**27.35**
DocWaveDiff	8.2 M	**0.0211**	**0.9733**	26.28

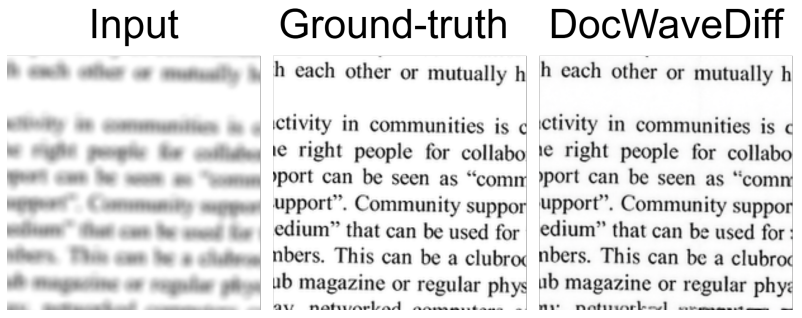


Figure 4.6: Qualitative comparison on the 0000014_n_00_blur image of the TDD test set. The corrupted input, ground truth, the image restored by DocDiff and DocWaveDiff are reported.

boundaries which are often blurred or distorted in conventional approaches. By leveraging wavelet-based decomposition and diffusion-based refinement, the model overcomes the limitations of standard regression-based techniques, which typically struggle with texture preservation and edge clarity.

To assess the practical benefits of DocWaveDiff, we conducted OCR experiments using Tesseract and a deep learning OCR on the restored images of the TDD dataset. OCR methods generally perform poorly on blurred documents, resulting in unreliable text extraction (CER of 0.7874 and WER of 0.9349). However, DocWaveDiff significantly improves image clarity, drastically reducing Tesseract errors (0.1316 CER and 0.6495 WER) compared to DocDiff (0.3710 CER and 0.6623 WER), and with similar performance for the deep learning OCR. Table 4.5 summarizes the OCR performance improvements obtained by DocWaveDiff.

Ablation Study

We conducted ablations to assess the contribution of each component on inpainting (EnsExam) and deblurring (TDD).

Inpainting (whole image): The mask is decisive: removing it drastically degrades performance (MS-SSIM 0.9849 → 0.9427; PSNR 39.54 → 27.22 dB; MSE 0.0003 → 0.0026). Disabling attention yields only marginal changes (MS-SSIM 0.9849→0.9835; PSNR 39.54 → 39.18 dB). Replacing wavelets with strided/transposed convolutions has negligible impact (MS-SSIM 0.9849→0.9831; PSNR 39.54→39.53 dB). Notably, removing the denoiser refiner leaves results essentially unchanged (MS-SSIM 0.9849; PSNR 39.53dB; MSE 0.0002), suggesting that on the EnsExam dataset the early predictor already solves most of the task and the refiner offers little headroom at these near-ceiling scores.

Deblurring: In contrast, wavelets are important: without them, SSIM and PSNR decrease and LPIPS worsens (SSIM 0.9733 → 0.9608; PSNR 26.28→24.63 dB; LPIPS 0.0211→0.0299). Attention is also crucial for blur removal: its ablation causes the greatest degradation (SSIM 0.9366, PSNR 22.44 dB, LPIPS 0.0468), indicating that features that look at both local and global detail are needed to best restore high and low frequencies. Finally, removing the denoiser-refiner increases the PSNR (26.28→27.09 dB) but significantly worsens perception (LPIPS 0.0211→0.0447), revealing the usual

trade-off between L2 and perception: the refiner restores high-frequency and visually salient details that improve LPIPS/SSIM at the cost of slight pixel misalignment.

The architecture is adaptable to the task: for inpainting, the mask guidance is indispensable, while attention/wavelet/refiner have a minor impact at the reported operating point; for deblurring, wavelets and attention are crucial for recovering edges and features, and the refiner improves perceptual quality even when PSNR is slightly lower. We report the results in Table 4.4.

4.4 Conclusions

We present DocWaveDiff, a predict-and-refine approach for document image enhancement. We use wavelet Unet as an alternative to classic Unet for both modules. On EnsExam (inpainting), we achieve SOTA: (MS-SSIM 0.9849, PSNR 39.54 dB). On TDD (deblurring), we achieve LPIPS 0.0211, SSIM 0.9733, PSNR 26.28 dB, which is competitive (second only to DocRes). The benefits carry over to OCR. With deep learning OCR: CER 0.1316, WER 0.3204. The ablations explain the choices. For inpainting, the mask is essential; attention and wavelet add little near the ceiling; the refiner gives minimal margins. For deblurring, wavelets are crucial; attention captures the global layout: removing them degrades more. The refiner improves perceptual quality at the cost of a slight PSNR drop, in line with the perception-distortion trade-off.

Table 4.4: Table of ablation study. Each experiment is evaluated *only* with the metrics of its corresponding task.

Ablation	Task	Dataset	**SSIM** (↑)	**MS-SSIM** (↑)	**PSNR** (↑)	**MSE** (↓)	**LPIPS** (↓)
No attention	Inpainting	EnsExam	-	0.9835	39.18	0.0002	-
No mask	Inpainting	EnsExam	-	0.9427	27.22	0.0026	-
No wavelet	Inpainting	EnsExam	-	0.9831	39.53	0.0002	-
No denoiser-refiner	Inpainting	EnsExam	-	0.9849	39.53	0.0002	-
No wavelet	Deblurring	TDD	0.9608	-	24.63	-	0.0299
No attention	Deblurring	TDD	0.9366	-	22.44	-	0.0468
No denoiser-refiner	Deblurring	TDD	0.9693	-	27.09	-	0.0447

Table 4.5: Results for OCR on the test-set of TDD dataset. ↓ indicates that lower values are better. Best results in **bold**.

Model	**CER** (↓)	**WER** (↓)
Blurred + Tesseract-OCR	0.7874	0.9349
DocDiff + Tesseract-OCR [110]	0.3710	0.6623
DocWaveDiff + Tesseract-OCR	**0.3554**	**0.6495**
Blurred + (db-resnet50 + crnn-vgg16bn)	0.6422	0.8071
DocDiff + (db-resnet50 + crnn-vgg16bn) [110]	0.1320	0.3214
DocWaveDiff + (db-resnet50 + crnn-vgg16bn)	**0.1316**	**0.3204**

Chapter 5

A document processing pipeline for the construction of a dataset for topic modeling based on the judgments of the Italian Supreme Court

Topic modeling in the context of Italian jurisprudence is limited by the lack of publicly available datasets, which hampers the analysis of legal themes covered by the Italian Supreme Court. To overcome this challenge, we developed a document processing pipeline that generates an anonymized dataset of legal judgments optimized for topic modeling. The pipeline combines document layout analysis, optical character recognition, and text anonymization. The DLA module, based on YOLOv8x, achieved a mAP@50 of 0.964 and a mAP@50–95 of 0.800. The OCR text detector based on YOLOv8x reached a mAP@50–95 of 0.9022, while the text recognizer (TrOCR) obtained a Character Error Rate (CER) of 0.0047 and a Word Error Rate (WER) of 0.0248. The dataset enabled topic modeling with a diversity score of 0.6198 and a coherence score of 0.6638, surpassing traditional OCR-only approaches. We applied BERTopic to extract topics and used Large Language Models (LLMs) to generate labels and sum-

maries. Outputs were evaluated against domain expert interpretations, showing comparable semantic quality. Specifically, Claude Sonnet 3.7 achieved a BERTScore F1 of 0.8119 for topic labeling and 0.9130 for topic summarization.

5.1 Introduction

Supervised learning techniques [29], such as neural networks [98], support vector machines (SVMs) [13], and decision trees (DTs) [94], are widely applied in legal AI for classification and regression tasks, making them essential for applications such as legal document categorization and case outcome prediction. In contrast, unsupervised methods [85], including clustering [69] and topic modeling [1], facilitate exploratory data analysis by uncovering latent structures in large textual datasets, such as legal topic discovery. However, their adoption remains limited in the NLP context for the Italian legal domain due to the scarcity of datasets explicitly designed for exploratory purposes and the need for expert involvement to contextualize results.

Legal topic modeling datasets for Italian case law are particularly rare, making it difficult to identify recurring themes in legal documents, extract knowledge, and analyze large volumes of normative and jurisprudential texts. This limitation hinders the development of advanced tools for information retrieval and automated judgment categorization, increasing reliance on manual analysis and expert work.

Furthermore, legal datasets contain sensitive information about individuals or companies, imposing restrictions on their sharing due to privacy and data protection concerns. Limited access to these resources hampers the creation of open benchmarks and the reproducibility of studies, slowing progress in the application of topic modeling to legal texts. Additional constraints arise from complex data protection regulations, such as the GDPR (General Data Protection Regulation) [100] in Europe, which requires anonymization strategies to safeguard privacy

Building a dataset is a complex and costly process that requires significant investment in time and resources. Computer vision emerges as a promising approach for automatically generating datasets for natural language processing (NLP) applications, leveraging advanced methods for analyzing and processing textual documents. Optical character recognition (OCR) and document layout analysis (DLA) techniques [93], [10] are examples of such

methods.

To address these challenges, we propose a document processing pipeline to create a dataset for topic modeling using rulings from the Italian Supreme Court[1] as a source.

Our pipeline initially employs computer vision methods to analyze the structure of judgments, enabling the accurate extraction and segmentation of textual content. This step improves OCR accuracy and ensures that the extracted text is properly structured for subsequent NLP tasks.

To ensure privacy compliance, we integrate a transformer-based Named Entity Recognition (NER) model capable of detecting and anonymizing sensitive information, such as personal names, addresses, emails, tax codes, and dates, in line with GDPR guidelines.

After constructing the dataset, we apply topic modeling to identify key topics. Subsequently, we employ large language models (LLMs) [117] to interpret the extracted topics by generating summaries and labels for each. However, due to their inherent biases and lack of legal expertise, we will systematically compare LLM-generated interpretations with those provided by legal experts, ensuring a critical assessment of their reliability in a judicial context.

In this paper, we introduce a document processing pipeline that combines computer vision and NLP algorithms to create a novel dataset for topic modeling in the Italian jurisprudence domain. Our main contributions are:

- The construction of a new dataset of Italian Supreme Court judgments through document processing techniques and its impact in topic modeling.

- The application of topic modeling in the legal domain and in the Italian language.

- The application of LLMs to interpret topics by generating labels and summaries describing the topics and comparing them with the labels and summaries generated by a domain expert.

[1] Authors' note: In this contribution, we refer to the Italian Supreme Court as the Corte di Cassazione. It is the highest judicial body in civil and criminal matters in the Italian legal system.

5.2 Dataset

5.2.1 Data source

The Italian Supreme Court publishes its judgments, ordinances, and decrees annually on the Italgiure website[2], providing access to legal documents in PDF format. These documents are generally of high quality, as most scans are free from defects such as skewness, blurring, photometric alterations, compression artifacts, and low resolution. However, minor imperfections can occasionally be found. Handwritten scribbles or corrections may appear in some cases, while ink defects affecting entire pages or lines are rare but not entirely absent.

Given the extensive availability and high quality of these documents, we selected a subset of judgments for analysis. Specifically, we randomly chose 161 civil judgments and 146 criminal judgments, resulting in a total of 307 cases.

After assembling the corpus of judgments, we performed a quantitative analysis to examine its internal structure. As a first step, we assessed the overall length of the documents, uncovering significant differences between civil and criminal judgments.

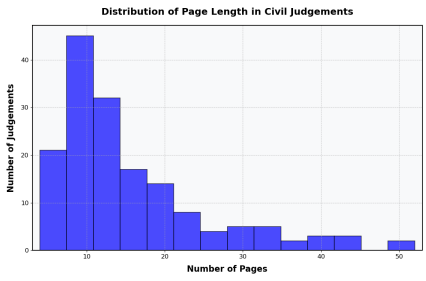


Figure 5.1: Distribution of page length of civil judgments of our dataset.

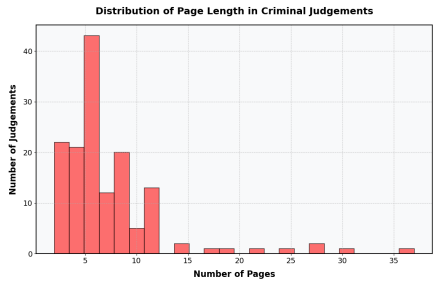


Figure 5.2: Distribution of page length of criminal judgments of our dataset.

Civil judgments tend to be longer than criminal judgments, as illustrated in Figure 5.1. This difference is particularly evident in the distribution of page counts: civil judgments exhibit a longer right tail, indicating the presence of significantly lengthier cases. In contrast, the distribution of criminal

[2] `https://www.italgiure.giustizia.it/sncass/`

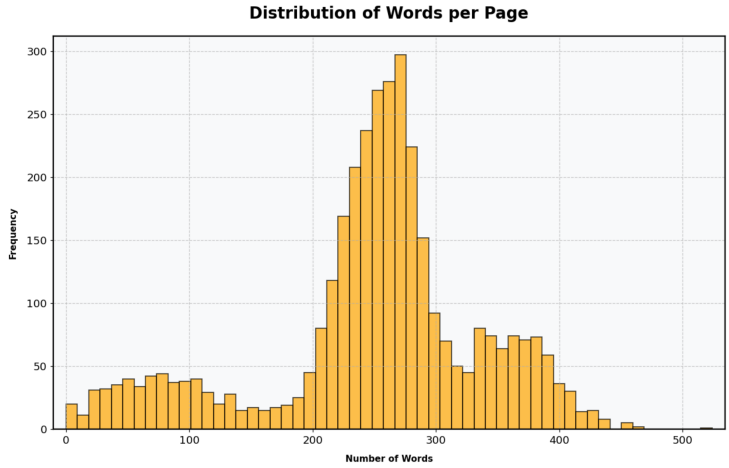


Figure 5.3: The distribution of the number of words per page of the judgments from our dataset.

judgments, shown in Figure 5.2, features a shorter and more dispersed right tail, suggesting greater variability in document length.

The distribution of the number of words per page indicates that most sentences are densely packed with text, averaging around 250 words per page and occasionally exceeding 400, as shown in Figure 5.3. This high word density poses a challenge for NLP models, as judgment pages contain extensive information that requires careful preprocessing. Without such preprocessing, models like BERT may struggle to fully leverage the content due to their maximum context length of 512 tokens.

While BERT's context length may seem sufficient to process the content of a single page, several factors complicate its practical application. Transformer models like BERT and GPT are primarily trained on English corpora and later adapted to other languages, including Italian. This adaptation affects tokenization, as BERT's WordPiece algorithm [105] constructs a vocabulary based on its training data. Consequently, many Italian words are split into multiple tokens if they are not present in the pre-built dictionary. As a result, numerous pages surpass the 512-token limit, causing truncation and potential loss of critical information, which may reside in the omitted portion. To address this issue, our work proposes a solution. In the following subsections, we describe the dataset annotation process.

5.2.2 Dataset for document layout analysis

To build a dataset for the Document Layout Analysis (DLA) task, modeled as an object detection problem, we downloaded an additional 59 judgments. We then annotated the layout of these documents using Label Studio [92], following the annotation scheme of the DocLayNet dataset [68]. This structured approach enabled us to leverage transfer learning, transferring knowledge from DocLayNet to our dataset.

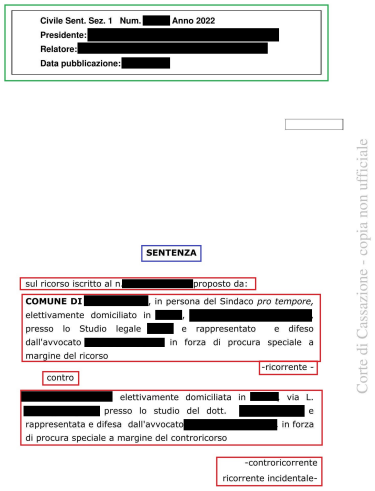
Ensuring accurate text segmentation was a crucial step. We verified that text-related bounding boxes aligned with paragraphs, as paragraph-level segmentation offers a balanced trade-off between processing efficiency and information completeness. Segmenting entire pages as text risks truncation due to BERT's 512-token limit, whereas finer segmentation at the line or sentence level can lead to overly fragmented text and increased computational costs.

The object classes we reused from DocLayNet are:

- Text: Regular paragraphs;
- Section-header: Any kind of heading in the text, except the overall judgment title;
- Page-footer: Repeating elements, such as page numbers at the bottom, outside of the normal text flow;
- Title: The overall title of a judgment, (almost) exclusively on the first page and typically appearing in a large font.

Figures 5.4 and 5.5 illustrate our document annotation approach. Paragraphs are enclosed in red bounding box, while section headers are marked as blue bounding box. Page footers are highlighted as orange bounding box, and title are distinguished by green bounding box.

We conclude this subsection by presenting Tables 5.1 and 5.2, which summarize key statistics of our dataset. Table 5.1 reports the overall size of the image dataset, while Table 5.2 details the number of annotations per class along with their percentage frequencies, providing insights into the dataset's composition.


Civile Sent. Sez. 1 Num. [redacted] Anno 2022
Presidente: [redacted]
Relatore: [redacted]
Data pubblicazione: [redacted]

SENTENZA

sul ricorso iscritto al n [redacted] proposto da:

COMUNE DI [redacted], in persona del Sindaco *pro tempore*, elettivamente domiciliato in [redacted], presso lo Studio legale [redacted] e rappresentato e difeso dall'avvocato [redacted] in forza di procura speciale a margine del ricorso

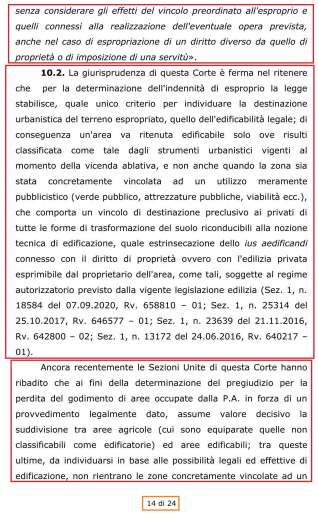
-ricorrente -

contro

[redacted], elettivamente domiciliata in [redacted] via L. [redacted] presso lo studio del dott. [redacted] e rappresentata e difesa dall'avvocato [redacted] in forza di procura speciale a margine del controricorso

-controricorrente ricorrente incidentale-

Figure 5.4: Front page of a annotated civil judgment of the Supreme Court.


senza considerare gli effetti del vincolo preordinato all'esproprio e quelli connessi alla realizzazione dell'eventuale opera prevista, anche nel caso di espropriazione di un diritto diverso da quello di proprietà o di imposizione di una servitù».

10.2. La giurisprudenza di questa Corte è ferma nel ritenere che per la determinazione dell'indennità di esproprio la legge stabilisce, quale unico criterio per individuare la destinazione urbanistica del terreno espropriato, quello dell'edificabilità legale; di conseguenza un'area va ritenuta edificabile solo ove risulti classificata come tale dagli strumenti urbanistici vigenti al momento della vicenda ablativa, e non anche quando la zona sia stata concretamente vincolata ad un utilizzo meramente pubblicistico (verde pubblico, attrezzature pubbliche, viabilità ecc.), che comporta un vincolo di destinazione preclusivo ai privati di tutte le forme di trasformazione del suolo riconducibili alla nozione tecnica di edificazione, quale estrinsecazione dello *ius aedificandi* connesso con il diritto di proprietà ovvero con l'edilizia privata esprimibile dal proprietario dell'area, come tali, soggette al regime autorizzatorio previsto dalla vigente legislazione edilizia (Sez. 1, n. 18584 del 07.09.2020, Rv. 658810 – 01; Sez. 1, n. 25314 del 25.10.2017, Rv. 646577 – 01; Sez. 1, n. 23639 del 21.11.2016, Rv. 642800 – 02; Sez. 1, n. 13172 del 24.06.2016, Rv. 640217 – 01).

Ancora recentemente le Sezioni Unite di questa Corte hanno ribadito che ai fini della determinazione del pregiudizio per la perdita del godimento di aree occupate dalla P.A. in forza di un provvedimento legalmente dato, assume valore decisivo la suddivisione tra aree agricole (cui sono equiparate quelle non classificabili come edificatorie) ed aree edificabili; tra queste ultime, da individuarsi in base alle possibilità legali ed effettive di edificazione, non rientrano le zone concretamente vincolate ad un

14 di 24

Figure 5.5: A page of the same annotated civil judgment

Dataset Total	Training-set	Validation-set	Test-set
799	634 (80%)	83 (10%)	82 (10%)

Table 5.1: Distribution of document images in our DLA dataset, split according to an 80-10-10 ratio.

Label	Count	Percentage
Text	3,671	79.4%
Page-footer	663	14.2%
Title	229	4.9%
Section-header	59	1.2%

Table 5.2: Distribution of annotations for each class in the dataset

5.2.3 Dataset for OCR

In addition to annotations for layout analysis, we also created a dataset for OCR.

Annotations for text line detection

To train the OCR text detection module, we extracted a crop dataset from paragraphs labeled as "Text" in the DLA dataset. Each crop was annotated with bounding boxes corresponding to text lines, represented by green bounding boxes, as illustrated in Figure 5.6. This dataset serves as the foundation for text line detection, ensuring precise localization of textual content. Table 5.3 provides an overview of the dataset size and details the distribution across different splits.

Dataset Total	Training-set	Validation-set	Test-set
10,442	8,341 (80%)	1,023 (10%)	1,078 (10%)

Table 5.3: Distribution of text line images in our OCR dataset, split according to an 80-10-10 ratio.

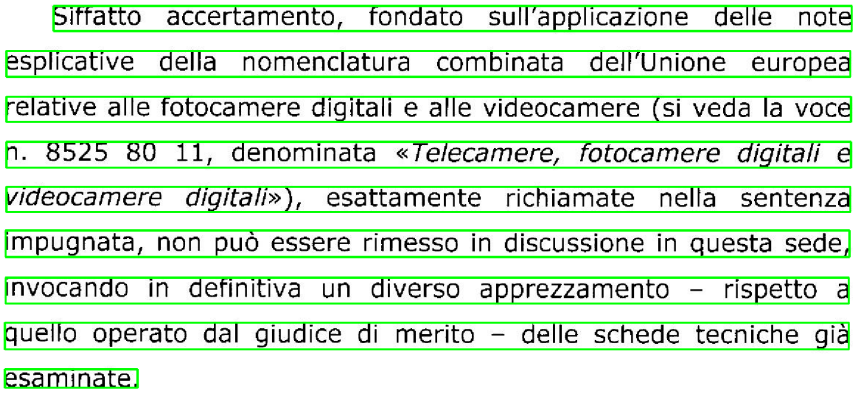


Figure 5.6: As an example, an annotation case for the text detection task is shown. The green bounding boxes represent ground truth and identify a row within the crop.

Annotations for text recognition

The text extraction module is trained on pairs of cropped images and their corresponding transcriptions. To construct this dataset, we generated line-level text crops and manually transcribed the content, as illustrated in Figure 5.7. This process resulted in a total of 27,707 annotated samples extracted from the judgments. Table 5.4 provides a detailed breakdown of the dataset size across different splits.

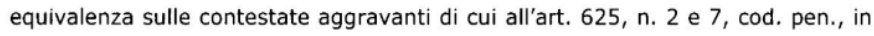


Figure 5.7: Line of text extracted from one of the sentences.

Dataset Total	Training-set	Validation-set	Test-set
27 707	22 175 (80%)	2 754 (10%)	2 778 (10%)

Table 5.4: Distribution of text line images in our OCR dataset, split according to an 80-10-10 ratio.

5.3 Methodology

5.3.1 Overview of the pipeline

Figure 5.8 provides a comprehensive overview of our workflow for converting legal judgments from PDFs into anonymized textual data ready for further analysis. The pipeline consists of six sequential steps:

- Preprocessing: converting PDFs into JPEG images.
- Document Layout Analysis: detecting structural components using YOLOv8 [73].
- Text Line Detection: identifying individual lines of text with YOLOv8.
- Text Recognition: extracting machine-readable text using TrOCR [41].
- Text Anonymization: masking sensitive information via GLiNER [111].
- Information Extraction: storing document content in a structured, machine-readable format.

Each of these steps is crucial for ensuring high-quality text extraction and structured representation. In the following subsections, we provide a detailed discussion of the models, training procedures, and hyperparameters used at each stage.

5.3.2 Document Preprocessing

To facilitate layout analysis and text extraction, PDF documents were converted into JPEG images with a resolution of 200 DPI, ensuring optimal text readability. Each judgment was organized into a dedicated folder, with pages sequentially numbered to maintain a clear mapping to the original document. This structured organization preserves document integrity and simplifies downstream processing.

5.3.3 Document Layout Analysis

Document Layout Analysis was performed using the X version of YOLOv8 from Ultralytics. We selected YOLO due to its superior accuracy and efficiency, with version X ranking among one of most powerful detector models.

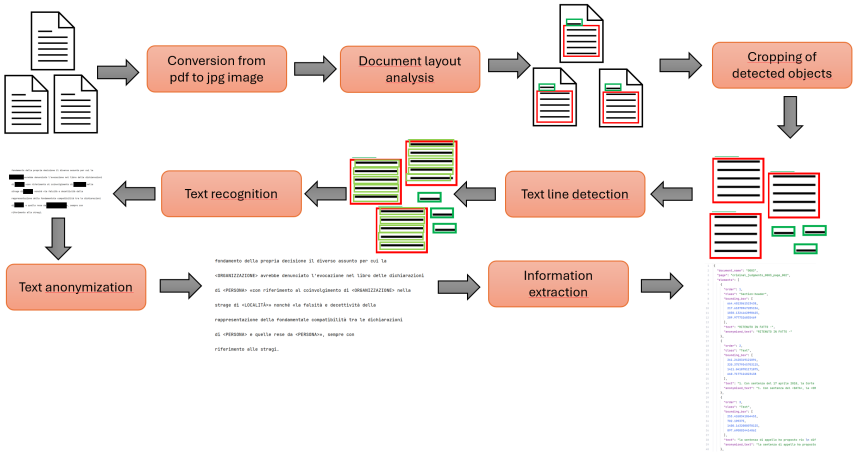


Figure 5.8: Overview of our pipeline for document content analysis and extraction.

Hyper-parameter	epoch	patience	batch-size	imgsz	freeze	worker
Value	50	5	8	1024	22	20

Table 5.5: Hyperparameters used for YOLOv8X fine-tuning on our dataset.

To leverage transfer learning, we initially trained YOLO on the DocLayNet dataset [68], aligning our annotation strategy with DocLayNet's labeling rules. This approach ensured compatibility and enabled more effective model adaptation. Given the limited size of our dataset, we mitigated overfitting by freezing the first 22 layers of YOLO during training.

The model was trained for 50 epochs, with early stopping set at 5 and a batch size of 8. We set the input resolution to 1024×1024 pixels, prioritizing high spatial fidelity to detect even small structural elements, such as page footers.

Table 5.5 summarizes the hyperparameters used for fine-tuning. The trained model outputs bounding boxes, which are sorted to reflect the Western reading order. These ordered segments are then cropped and further processed for text extraction.

The model outputs bounding boxes that delineate different structural

Hyper-parameter	epoch	patience	batch-size	imgsz	freeze	worker
Value	25	5	8	1024	0	20

Table 5.6: Hyperparameters used for YOLOv8X fine-tuning on text line dataset detection.

components of the document such as: title, section-header, text, and page-footer. These bounding boxes are then sorted according to the Western reading order, ensuring a coherent sequence. Once ordered, the extracted regions undergo a cropping operation to facilitate accurate text line detection and text recognition.

5.3.4 OCR

The OCR module was designed by dividing it into two components:

- Text line detection by YOLOv8.
- Text extraction by TrOCR.

Text line detection

Extracted document components were analyzed to identify individual lines of text. For this task, we employed YOLOv8 version X, initializing it with the pre-trained DocLayNet model. Given the ample size of our dataset, we opted to train all network layers, allowing the model to fully adapt to our domain. The hyperparameters remained unchanged from Table 5.5, except for the number of epochs, set to 25, and the number of trainable layers, as detailed in Table 5.6.

Once detected, the text lines were sorted according to the Western reading order and cropped to fit the text extraction module, ensuring a structured and coherent representation.

Text extraction

For the crucial task of converting detected text line into a raw text, we used TrOCR [41], a powered transformer based text recognition model. Unlike traditional approaches that rely on CNN backbones and RNN-based

decoders, TrOCR harnesses the power of pretrained Transformers for vision and language in an end to end architecture.

The model processes each clipped text line using an image transformer encoder, which segments the input into 16×16 pixel patches and extracts visual features. These features are then passed to a text transformer, which decodes them into word-level text.

For this task, we employed the TrOCR-Small configuration, which integrates BEiT [7] as the vision encoder and an auto-regressive decoder initialized with RoBERTa weights [50]. This setup demonstrated superior performance in document text recognition, making it the optimal choice for our pipeline.

A key advantage of TrOCR in our use case is its ability to leverage context for accurate text generation. Unlike conventional OCR models, its pretrained decoder incorporates strong language understanding, enabling it to exploit contextual cues while reconstructing text. This capability allows TrOCR to effectively handle variations in font styles and sizes, which are common in legal documents, while maintaining high recognition accuracy.

To further adapt the model to the Italian legal domain, we fine-tuned TrOCR-Small using Microsoft's pretrained version 'microsoft/trocr-small-printed' on our dataset. The hyperparameters used for fine-tuning are detailed in Table 5.7.

Hyper-param.	**epoch**	**batch**	**max-len**	**beams**	**len-pen.**	**no-repeat**
Value	7	8	64	8	2	3

Table 5.7: Hyperparameters used for TrOCR fine-tuning on our dataset.

As stated in Subsection 5.2.1, the quality of the judgments received is satisfactory; however, printing defects, such as ink omissions, may sporadically occur.

Consequently, we implemented a data augmentation technique to enhance TrOCR's generalization capabilities by simulating these ink defects. For illustrative purposes, Figures 5.9, 5.10, and 5.11 depict examples of this data augmentation.

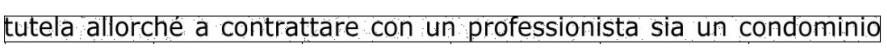


Figure 5.9: An example of our data augmentation

Figure 5.10: An example of our data augmentation

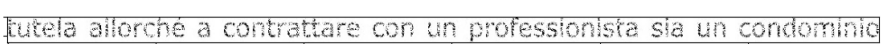


Figure 5.11: An example of our data augmentation

5.3.5 Text anonymization

Although Supreme Court judgments are public and written without confidentiality constraints, ensuring the privacy and dignity of the subjects involved remains a priority. Anonymization not only protects sensitive information but also mitigates bias, preventing NLP models from memorizing overly specific patterns embedded in the data.

To achieve this, we focus on identifying and masking key entities, including parties, witnesses, companies, dates, and locations. These entities have been grouped into five categories, as detailed in Table 5.8.

For the crucial task of text anonymization, we rely on GLiNER, a Named Entity Recognition (NER) model designed to detect and classify entities in legal documents. GLiNER employs a BERT-like bidirectional transformer encoder, which processes text in both directions to capture rich contextual information.

The model generates embeddings for entity types, encodes text into a latent space, and aggregates tokens into span representations, enabling the recognition of multi-word entities. These representations are then evaluated by a classifier that identifies which text segments correspond to the specified entity types.

A key feature of GLiNER is its zero-shot classification capability, allowing it to recognize entity types never seen during training. This makes the model highly adaptable across different domains, including legal contexts with diverse and unpredictable entity forms.

In our work, we use an Italian-optimized version of GLiNER, fine-tuned to detect Personally Identifiable Information (PII), available from the Hugging Face Hub [16]. This configuration enables us to effectively identify and redact sensitive entities commonly found in legal documents.

5.3.6 Information extraction

After applying all processing steps, we store each page's information in a JSON document, enabling seamless integration with downstream NLP applications, such as topic modeling. Each JSON file contains the page name and an "elements" list, where every document component is described by both spatial and semantic attributes. These include the bounding box coordinates in Pascal VOC format [3] $[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$, the element class, the plain text, and the corresponding anonymized text.

To make the pipeline accessible to non-expert users, we developed an interactive Streamlit front end, shown in Figures 5.12 and 5.13. The application allows users to load models and documents with ease, and to adjust key hyperparameters, such as YOLO's confidence threshold and IoU, or GLiNER's entity confidence threshold, to fine-tune model behavior.

Additionally, users can manually edit the JSON content, enabling them to review and correct the plain text or anonymized fields as needed, ensuring both flexibility and control over the final output.

5.3.7 A novel topic modeling dataset

After processing all available judgments, we compiled the final dataset by aggregating the generated JSON files.

As shown in Figure 5.14, the topic modeling pipeline begins by extracting key elements from the JSON files, including the document name, page number, anonymized text, and content class. This information is then structured into a Pandas [60] DataFrame, enabling further enrichment with additional metadata, such as text length.

To refine the dataset for analysis, we filtered out text belonging to the Title, Section-header, and Page-footer classes, as these elements were not relevant to our study.

5.3.8 Topic modeling

Topic modeling was performed using BERTopic [25], chosen for its high modularity, which enables customization by replacing or adding components based on task-specific requirements.

[3] In the Pascal VOC format, x_{min} and y_{min} are the coordinates of the top left corner of the bounding box, while x_{max} and y_{max} represent the coordinates of the bottom right corner.

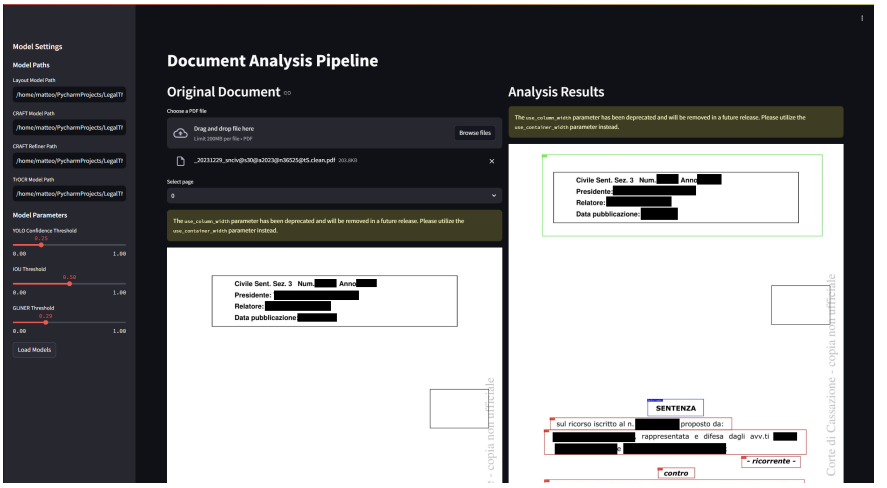


Figure 5.12: The streamlit app for our pipeline.

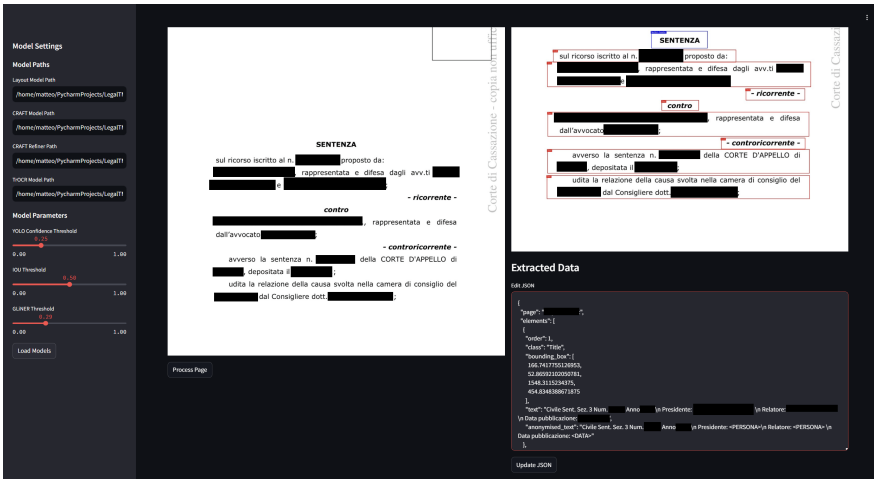


Figure 5.13: The streamlit app for our pipeline.

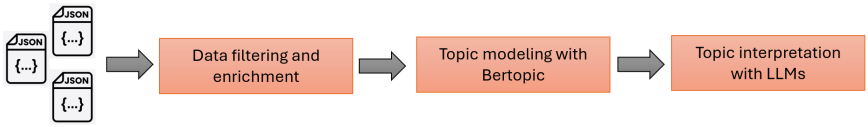


Figure 5.14: The pipeline for topic modeling.

To generate text embeddings, we replaced the standard transformer with Distilled Legal-Italian BERT [44], a model specialized for the Italian legal domain. This choice was driven by the linguistic characteristics of our dataset, ensuring better representation of legal terminology.

The resulting embeddings, typically high-dimensional (768 dimensions) and dense, were then processed with UMAP [22], a dimensionality reduction technique. While measures like cosine distance can help mitigate high-dimensionality issues, clustering algorithms still struggle to extract meaningful structures from such large feature spaces. Reducing dimensionality improves computational efficiency and enhances cluster separability.

Finally, we applied HDBSCAN [59] to cluster the reduced embeddings. As an advanced variant of DBSCAN, HDBSCAN offers greater speed and robustness, particularly in datasets with variable-density clusters, making it well-suited for text data.

BERTopic identifies the characteristic words of each cluster through a structured three-step process:

1. Stop word removal: Eliminating frequent, non-informative terms such as “courts,” “lawyers,” “firms,” “judgments,” and anonymized words.

2. Bag-of-Words (BoW) model construction [71], representing text as token frequency distributions.

3. c-TF-IDF computation [25], a cluster-level adaptation of TF-IDF that enhances word importance estimation.

In particular, we employed the c-TF-IDF-bm25 variant, which is optimized for small datasets and offers improved handling of stop words, ensuring better topic representation.

The hyperparameters used in the BERTopic pipeline are listed in Table 5.9.

5.3.9 Interpretation of topics with LLMs

Below, we present the interpretation of topics obtained with BERTopic, enhanced by the use of Large Language Models (LLMs). Since OpenAI introduced the widely popular ChatGPT, interest in Generative Artificial Intelligence has surged across academia, industry, and the general public. This growing attention is driven by the remarkable capabilities of these models in both interpreting and generating content in response to user-defined prompts.

Integrating LLMs with topic modeling techniques such as BERTopic enables a deeper analysis of extracted topics, providing contextualized interpretations of patterns within legal documents. In addition to these powerful LLMs, we also use LLaMA 3.1 8B [21] in this research. LLaMA 3.1 8B can run nimbly locally on machines with modest hardware and can understand and generate Italian text.

Given that our data were anonymized, we also integrated powerful closed-source LLMs, such as Claude 3.7 Sonnet [5] and GPT-4 [2]. In these cases, anonymization is crucial, as these models operate on private servers located outside Italy and the European Union.

To conduct our analysis, we designed two prompts in Italian. The first prompt generates topic labels by analyzing the most representative keywords and a set of sample documents:

> *""" Sei un esperto di giurisprudenza e di diritto italiano. Rispondi in modo professionale alle richieste. Ho un topic descritto dalle seguenti keywords: [KEYWORDS] In questo topic, i seguenti documenti sono un sottoinsieme piccolo ma rappresentativo di tutti i documenti dell'argomento: [REPR_DOCS]. Sulla base delle informazioni di cui sopra, fornisci una breve label a questo topic. Assicurati di riportare solo la label e nient'altro. """*

The second prompt is designed to generate a brief descriptive summary of each topic:

> *""" Sei un esperto di giurisprudenza e di diritto italiano. Rispondi in modo professionale alle richieste. Ho un topic descritto dalle seguenti keywords: [KEYWORDS] In questo topic, i seguenti documenti sono un sottoinsieme piccolo ma rappresentativo di tutti i documenti dell'argomento: [REPR_DOCS]. Sulla base delle*

informazioni di cui sopra, fornisci una descrizione di questo topic nel seguente formato: topic: < descrizione > """

Within BERTopic, keywords refer to the terms automatically extracted by the model to define each topic. Conversely, Representative Documents correspond to exemplar texts that best encapsulate the core characteristics of each topic.

For text generation, Claude-Sonnet 3.7 and GPT-4 were used with default hyperparameters, while LLaMA 3.1 was fine-tuned with the configurations listed in Table 5.10, both for label generation and text summarization tasks.

To conclude this subsection and the entire section, we present the pseudocode of our topic modeling algorithm in Algorithm 5.15.

5.4 Evaluation

5.4.1 Metrics

In this section we report metrics to evaluate our pipeline, topic modeling metrics, and text generation metrics for LLMs.

Object detection metrics

Object detection performance is typically assessed using a combination of metrics that evaluate both localization quality (bounding box) and classification accuracy of the detected instances.

To evaluate the accuracy of our DLA model, we primarily use mean Average Precision (mAP), computed at a 50% Intersection over Union (IoU) [4] threshold (mAP50). Additionally, we assess performance across multiple IoU thresholds ranging from 50% to 95% (mAP50-95), following the standard procedure in [66].

Precision and Recall Before describing mAP, it is helpful to introduce *precision* and *recall*, which measure how accurate and complete the model's predictions are:

[4] Given two rectangles, R1 (representing the ground truth) and R2 (the model's prediction), Intersection over Union (IoU) measures the area of intersection between R1 and R2. If the two rectangles do not overlap, the IoU is 0; if they overlap perfectly, the IoU is 1.

- Precision is the fraction of predictions that are correct (i.e., true positives[5]) among all predicted instances (the sum of true positives and false positives [6]). Formally:

$$Precision = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}. \tag{5.1}$$

- Recall is the fraction of objects that are correctly identified (true positives) among all instances that should have been detected (the sum of true positives and false negatives [7]). Formally:

$$Recall = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}. \tag{5.2}$$

Both metrics are crucial in object detection: high precision indicates a low number of spurious detections (i.e., few false positives), while high recall means the model misses few objects (i.e., few false negatives).

In object detection tasks, a prediction is typically considered a true positive if the IoU between the predicted bounding box and the ground-truth bounding box exceeds a given threshold (e.g., 50%).

Average Precision (AP) To combine precision and recall into a single metric, Average Precision (AP) is used. It is computed as the area under the precision-recall curve for a single class.

An AP value close to 1 indicates that, across various model confidence thresholds, precision remains high over a wide range of recall values. Typically, AP is computed for each class and then averaged across all classes of interest, yielding the mean Average Precision (mAP)."

mAP50 and mAP50-95 The calculation of mAP depends on the IoU threshold(s) used to consider a detection correct:

- mAP50 is computed by evaluating the Average Precision for each class at a single IoU threshold (commonly 0.5) and then averaging these values. This metric is less stringent, as an IoU of at least 0.5 is enough to count a detection as correct.

[5] True positives are model predictions that correspond to the truth class of the detected object and satisfy the IoU threshold.

[6] False positives are detections that do not correspond to the class of the detected object or that do not exceed the IoU threshold.

[7] False negative are detections that are not taken

- mAP50-95 is computed by averaging the AP values over multiple IoU thresholds (usually from 0.5 to 0.95 in increments of 0.05). This is more demanding, as it requires more precise overlap between the predicted bounding box and the ground-truth object.

Formally, mAP can be expressed as:

$$\text{mAP} = \frac{1}{K} \sum_{i=1}^{K} \text{AP}_i, \tag{5.3}$$

where K is the number of classes and AP_i is the Average Precision for the i-th class.

OCR metrics for text recognition task

For the text recognition model, we employ two metrics [63] commonly used in OCR: Character Error Rate (CER) and Word Error Rate (WER).

Character Error Rate (CER) Character Error Rate (CER) is defined as the ratio of the total number of character-level errors (substitutions, deletions, and insertions) to the total number of characters in the ground truth:

$$\text{CER} = \frac{\text{S} + \text{D} + \text{I}}{N} \tag{5.4}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the total number of characters in the reference text. A lower CER indicates better accuracy at the character level.

Word Error Rate (WER) Word Error Rate (WER) measures the proportion of word-level errors in the recognized text:

$$\text{WER} = \frac{\text{S} + \text{D} + \text{I}}{M} \tag{5.5}$$

where S, D, and I are the number of substitutions, deletions, and insertions at the word level, and M is the total number of words in the reference text. A lower WER indicates better accuracy at the word level.

These metrics allow us to quantify the performance of the OCR module at both character and word level, and for both metrics, values close to 0 indicate that the text recognition model makes few errors and is reliable.

Topic modeling metrics

We evaluate the quality of the topics generated by our topic model using topic coherence and topic diversity.

Topic coherence C_v Topic coherence measures the degree of semantic similarity between high-probability words in a topic. Among the various coherence measures, we employ the C_v metric [75], which is known to correlate well with human interpretability of topics. The C_v coherence score is based on the normalized pointwise mutual information (NPMI) between pairs of top words in a topic.

The NPMI score between two words x_i and x_j is defined as follows:

$$\text{NPMI}(x_i, x_j) = \frac{\log \frac{p(x_i,x_j)+\epsilon}{p(x_i)p(x_j)}}{-\log p(x_i, x_j) + \epsilon} \tag{5.6}$$

where $p(x_i, x_j)$ represents the co-occurrence probability of words x_i and x_j in a reference corpus, and ϵ is a smoothing constant used to avoid division by zero.

The coherence score for a topic t_k is then calculated by averaging the NPMI values for all word pairs within the top words of the topic.

To obtain a single coherence score for the entire model, the C_v score is computed as the mean coherence score across all topics K:

$$C_v = \frac{1}{T} \sum_{i=1}^{T} \cos(\mathbf{v}_{\text{NPMI}}(x_i), \mathbf{v}_{\text{NPMI}}(x_j)) \tag{5.7}$$

where V_{NPMI} represents the NPMI score vectors for each topic word.

A C_v score close to 1 indicates that the keywords effectively describe the topics and are semantically related. Conversely, a value near 0 indicates that the keywords do not accurately represent the topics and are not semantically related.

Topic diversity Topic diversity [18] is a metric that quantifies how semantically diverse the obtained topics are. It is defined as follows:

$$TD = \frac{|\bigcup_k T_j|}{K \times N} \tag{5.8}$$

where T_j represents the set of the first words in topic j, K is the total number of topics, and N is the number of words considered for each topic. The numerator represents the number of unique words across all topics.

The topic diversity metric ranges from 0 to 1, where 0 indicates an exact match between all topics in terms of their primary lexical elements, while 1 denotes complete divergence in vocabulary across all topics.

A topic diversity value close to 1 suggests that the keywords describing the topics are highly diverse, meaning that none or very few topics share keywords. Conversely, a value close to 0 indicates that many topics share a significant number of keywords.

Trade-off between topic diversity and topic coherence Topic diversity and topic coherence have an inverse relationship [9], [104]. Maximizing one of these metrics typically results in minimizing the other: models that maximize topic coherence tend to generate topics with highly similar main words, leading to reduced keyword diversity. As a result, topics become too similar.

Conversely, a model that maximizes topic diversity will produce topics that are very different from each other, incorporating a broader range of keywords but potentially reducing the interpretability of the generated topics. Therefore, it is essential to find a balance between the two metrics."

Text generation metrics

To evaluate labels and summaries for topics, we chose the BERTScore metric [116] for LLMs. Unlike traditional metrics based on n-gram counts, such as ROUGE or BLEU, BERTScore can capture deeper semantic similarities between texts.

By leveraging BERT's contextual embeddings, this metric can recognize synonyms, paraphrases, and reformulations that preserve meaning while using different words. Additionally, BERTScore has shown a stronger correlation with human judgments than classical metrics, making it particularly suitable for evaluating generative outputs such as summaries and labels, where expressive flexibility is as important as content accuracy.

BERTScore computes the similarity between a generated text and a reference text using contextualized embeddings. Given a generated sentence $X = \{x_1, x_2, \ldots, x_m\}$ and a ground truth sentence $Y = \{y_1, y_2, \ldots, y_n\}$,

their corresponding embeddings are obtained using a pre-trained transformer model:

$$\mathbf{x}_i = \text{BERT}(x_i), \quad \mathbf{y}_j = \text{BERT}(y_j) \tag{5.9}$$

where $\mathbf{x}_i$ and $\mathbf{y}_j$ are the embedding vectors for tokens x_i and y_j, respectively.

BERTScore measures the pairwise cosine similarity between token embeddings:

$$s_{ij} = \frac{\mathbf{x}_i \cdot \mathbf{y}_j}{\|\mathbf{x}_i\| \|\mathbf{y}_j\|} \tag{5.10}$$

where s_{ij} represents the cosine similarity between the embedding of the i-th token in the generated text and the j-th token in the reference text.

BERTScore aggregates these similarity scores using precision, recall, and F1-score.

BERTScore: Precision BERTscore's Precision measures how aligned the sentence generated by the model is with the ground truth sentence.

$$P = \frac{1}{m} \sum_{i=1}^{m} \max_j s_{ij} \tag{5.11}$$

A value close to 1 indicates that the generated sentence is able to capture the meaning of the ground truth sentence with high accuracy. While a value close to 0 indicates a significant semantic distance between the generated sentence and the ground truth sentence.

BERTScore: Recall BERTscore's recall measures how much of the meaning of the ground truth sentence is present in the sentence generated by the model.

$$R = \frac{1}{n} \sum_{j=1}^{n} \max_i s_{ij} \tag{5.12}$$

A value close to 1 indicates that almost all the information in the ground truth sentence has been captured in the generated sentence. While a recall value close to 0 indicates the generated sentence lacks much content present in the ground truth sentence.

BERTScore: F1-score BERTscore's F1-score is the harmonic mean of precision and recall, often used as a final score.

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{5.13}$$

A value close to 1 indicates that the generated sentence contains all the essential information present in the ground truth sentence and is very close in wording to that of ground truth, without adding irrelevant information. While a value close to 0 indicates that the generated sentence does not contain all the essential information present in the ground truth sentence and is not very close in wording to the ground truth sentence, without adding irrelevant information.

5.4.2 Hardware requirements

The following results were obtained with a computer that has the following hardware configuration:

- Processor: 13th Gen Intel(R) Core(TM) i9-13900H 2.60 GHz;
- GPU: NVIDIA RTX 4090 mobile, 16GB GDDR6;
- RAM: 32.0 GB

5.4.3 Results

Our comprehensive evaluation, conducted across multiple performance metrics, offers valuable insights into the effectiveness of each component within the pipeline. The results highlight both the strengths of individual modules and the areas that require further improvement, guiding future refinements of the system.

Object detection evaluations

Evaluation on the test set confirms the strong performance of our document layout analysis model across all content classes. The model consistently detects all structural components of the documents with a high average recall of 0.960, indicating its ability to identify nearly all relevant elements. Precision is also strong, reaching 0.933, as incorrect classifications are rare. The predicted bounding boxes align closely with the actual layout components,

as reflected by an overall mAP@50–95 of 0.800, underscoring the model's accuracy in localization.

A class-wise analysis reveals particularly noteworthy results. The "Title" class achieves a perfect recall of 1.000, meaning all titles are correctly detected, paired with a precision of 0.902 and the highest mAP@50–95 score of 0.959 among all classes. These metrics highlight the model's exceptional ability to both recognize and precisely localize document titles.

In the case of the "Text" class, which dominates the dataset (411 out of 509 instances), the model delivers balanced and robust performance, with a precision of 0.954, recall of 0.951, and mAP@50–95 of 0.892. This confirms the model's reliability in identifying the main textual content of legal documents.

The "Page-footer" class shows high precision (0.954), but slightly lower recall (0.926), indicating that footers are occasionally missed. Interestingly, while the mAP@50 is very high (0.982), the mAP@50-95 drops to 0.612, suggesting that the general location of footers is well predicted, but the precision of the selection rectangle can be improved. This result on the "Page-footer" class suggests that our model has problems with document objects being small. Fortunately, the "Page-footer" text of the judgments does not contain important information.

For the "Section-header" class, both precision (0.923) and recall (0.960) are in line with the global average. However, the mAP@50–95 score of 0.738 indicates that there is still room for improvement in precisely localizing these elements.

Overall, these results confirm the effectiveness and robustness of our layout analysis model, Table 5.11. It delivers consistently high performance across all structural classes, with particularly strong results for titles and main text, which are crucial for downstream NLP tasks.

OCR evaluations

The OCR module was developed by integrating two key components: text detection and text recognition. To assess the system's overall effectiveness, we evaluated each component separately, focusing on their individual contributions to the OCR task.

The text detection model, based on YOLOv8x, demonstrates exceptional performance, achieving a precision of 0.9864 and a recall of 0.9834. These results indicate that the model accurately identifies nearly all textual content

in the documents, while minimizing false positives. The quality of localization is confirmed by a mAP@50 of 0.9901 and a mAP@50–95 of 0.9022, reflecting the precision of the predicted bounding boxes.

The text recognition module, implemented with TrOCR-Small, also achieves high transcription accuracy. The model records a Character Error Rate (CER) of 0.0047, showing that character-level recognition errors are minimal. Likewise, the Word Error Rate (WER) of 0.0248 confirms that the vast majority of words are correctly transcribed.

The combination of these two modules results in a robust and reliable OCR system, capable of both accurate localization and high-fidelity transcription of legal text. Performance metrics for the OCR pipeline are reported in Table 5.12.

Table 5.12 presents the results on the test set for both the text detection and text recognition tasks.

The impact of the pipeline on Topic modeling

In this sub section, we analyze the impact of the pipeline on the performance of BERTopic in topic modeling. To this end, we prepared two versions of the dataset for comparison. The first consists of documents processed exclusively with the OCR module, without further refinement. The second was generated using our full pipeline, then filtered to remove nonessential content—such as titles, page footers, section headers, and short or uninformative paragraphs, which often contain run-on or fragmented text.

The filtered dataset retains high-quality paragraphs, specifically those in the right-hand tail of the logarithmic distribution of tokenized paragraph lengths, ensuring a focus on content-rich text, Figure 5.16. Both datasets underwent anonymization to ensure compliance with privacy constraints.

For each version, we conducted a series of experiments to assess topic modeling performance. In particular, we varied the number of topics K between 2 and 50, and computed two key metrics: topic diversity and topic coherence, as defined by the BERTopic framework.

During the experiments, we used the distil-ita-legal-bert embedding model as the backbone for BERTopic. When applied to the dataset generated through our segmentation pipeline, BERTopic demonstrated stable performance across a range of topics from K = 2 to 15. Both topic diversity and coherence remained within desirable ranges, often approaching the upper bounds of the respective metrics and exhibiting a strong positive correla-

tion.

In contrast, when BERTopic was applied to the unsegmented dataset produced solely through the OCR module, the model frequently generated topics with excessively high diversity values, often surpassing the optimal threshold. Although topic coherence entered the acceptable range starting from K = 18, it tended to decline as the number of topics increased, suggesting reduced interpretability and topic quality at higher values of K.

These findings indicate that our segmentation strategy plays a critical role in improving the quality of topic modeling. By structuring the dataset around coherent and content-rich textual units, the pipeline enables BERTopic to discover a greater number of diverse and consistent topics, enhancing both the granularity and reliability of the extracted topic structure, Figure 5.17.

To further validate our approach, we repeated the experiment using another general-purpose Italian embedding model, sentence-bert-base-italian-uncased [64]. Once again, the dataset generated through our segmentation pipeline yielded superior results, confirming the robustness and generalizability of the method across different embedding strategies.

A detailed summary of the experimental outcomes is presented in Table 5.13, highlighting the consistent performance gains achieved through structured preprocessing.

Text anonymization evaluations

Unlike the DLA and OCR modules, the GLiNER-based anonymization component currently lacks a labeled dataset for quantitative performance evaluation. Building such a dataset—following the methodology used for the other components of the pipeline—would require a substantial investment of time and resources, involving the manual annotation of a large corpus of legal documents and the precise definition of sensitive entity types.

To address this challenge, we adopted a pragmatic approach, leveraging a pre-trained GLiNER template available on Hugging Face, specifically oriented toward the Italian language and the detection of Personally Identifiable Information (PII).

This strategy enabled us to rapidly deploy a functional anonymization solution, ensuring practical utility within our pipeline. However, we acknowledge the limitations inherent in this approach, particularly the lack of quantitative validation, which constrains the ability to measure performance

systematically.

Despite the absence of a quantitative evaluation framework, we observed that data anonymization has a substantial impact on topic modeling outcomes. In particular, applying BERTopic to non-anonymized data often results in the identification of keywords and representative documents that differ markedly from those produced using anonymized input.

Without anonymization, the model tends to elevate named entities—such as people, locations, and companies to the status of informative keywords, since they frequently appear in the corpus and are interpreted as semantically meaningful, see Figure 5.18. As a result, many topics are defined around these sensitive identifiers, rather than around the actual substance of the legal text.

By contrast, when the data is anonymized through GLiNER, these entities are replaced with generic placeholders (e.g., ⟨PERSONA⟩, ⟨ORGANIZAZZIONE⟩, ⟨LOCALITÀ⟩). Because these placeholders recur across multiple documents, their c-TF-IDF scores are low, and BERTopic does not consider them topic-defining. Instead, the model focuses on semantically richer and contextually relevant terms, as illustrated in Figure 5.19.

This observation confirms the dual benefit of anonymization: it safeguards sensitive information while simultaneously improving the semantic clarity and interpretability of the topics discovered.

5.4.4 Analysis of the detected topics from a legal perspective

The analysis of the Supreme Court's rulings revealed a number of recurring legal topics pertaining to both procedural law (civil and criminal) and different areas of substantive law. The main topics identified are examined below, retaining the original numerical identifier for each and summarizing its legal content in a manner consistent with the doctrinal and jurisprudential tradition. Normative references are given where appropriate, so as to contextualize each topic with reference to the current legal framework.

Topic 0 - Appeal to the Court of Cassation on Evidence and Reasoning Concerns appeals to the Court of Cassation based on errors in the evaluation of evidence or in the reasoning of judgments on the merits. In particular, it refers to the case where the appellate court has failed to examine a decisive fact in the dispute (ground of appeal provided for in Article

360 No. 5 of the Code of Civil Procedure). The Supreme Court can annul the appealed judgment if it finds serious flaws in the reasoning (missing or illogical reasoning) or violations of procedural rules, but it cannot freely review the merits of the case: that is, it does not go into the merits of the evidence assessed by the judges of first and second instance, unless there are manifest errors. In summary, the issue highlights the limits of the Supreme Court's legitimacy review of judgments on the merits, distinguishing errors of law (which can be censured in the Supreme Court) from evaluations of fact (which remain unquestionable).

Topic 1 - Surety bonds and void contract clauses Addresses the validity and interpretation of clauses in surety (guarantee for others' debts) contracts, with attention to potentially void clauses and the limits of review in Supreme Court on contractual interpretation. It is clarified that if some clauses in a contract are found to be void (e.g., standard anti-competitive clauses in bank surety contracts), the court must consider whether the rest of the contract can remain valid without them (partial nullity, Art. 1419 Civil Code). The principle is reiterated that the interpretation of the contract is a factual determination reserved for the court of merit (court or court of appeals); the Supreme Court can intervene only in the case of a clear violation of the legal criteria of interpretation (Art. 1362 et seq. Civil Code) or nonexistent/illogical reasoning. In other words, the Supreme Court does not uphold the appeal just because the appellant proposes a different interpretation of the contract, but intervenes if the judgment on the merits ignored the rules of contractual hermeneutics or if its explanation is so deficient as to constitute an error of law.

Topic 6 - Appeals in Tax Litigation: new grounds and "self-sufficiency" Concerns procedural rules specific to tax litigation and, more generally, to the Supreme Court. The first principle is the prohibition of new exceptions on appeal: in tax litigation, Article 57 of Legislative Decree 546/1992 prohibits the introduction in the second instance of issues or exceptions that were not raised in the first instance. This means that in tax appeals the parties may discuss only the points already dealt with in the first instance, without adding new grounds for dispute. The second principle is that of self-sufficiency of the appeal to the Supreme Court: a person appealing to the Supreme Court must file an appeal that contains within it all the ele-

ments necessary to understand the issue, including the relevant parts of the previous court documents. In particular, if the appellant complains that the trial judges did not consider a certain exception or evidence, the appeal must indicate where and how that issue was raised in the previous stages (e.g., cite the notice of appeal or the record of the hearing in which it was raised). If the appellant raises a new issue in the Supreme Court that does not appear in the appealed judgment or in the previous acts, the Supreme Court will declare it inadmissible. In summary, in the tax process (but the principle applies in general) one cannot change the argument or add new issues at the last moment, and in Cassation the appeal must be valid by reporting everything necessary to assess the errors of law complained.

Topic 7 - Reasoning of Judgments and Limits of Review in the Supreme Court (Press Defamation) This topic highlights the extent to which the Supreme Court can review the reasoning of a judgment on the merits, with reference to cases of press defamation. According to Article 111 of the Constitution, the judgment must be reasoned: however, the Supreme Court only verifies compliance with a "constitutional minimum" of reasoning. This means that the Supreme Court will annul the appealed decision only if the appellate judgment lacks motivation, or if the motivation is merely apparent, irremediably contradictory, or so illogical as to be incomprehensible. In other cases, the assessment of facts and evidence remains reserved for the trial court. In the context of defamation by the press (e.g., an offensive newspaper article), this principle implies that aspects such as reconstructing the facts, evaluating the phrases found to be defamatory, and ascertaining whether the author of the article could invoke the exemption of the right to report or criticism (i.e., writing out of a duty to report without defaming) are considered factual determinations left to the judges of the merits. The Supreme Court cannot question these assessments if they are supported by logical reasoning. Its review is limited to whether the judges took into account the basic criteria (e.g., were the incriminated sentences relevant to the public interest of the news story? Were the facts narrated true? Was the language measured?) and that they explained their conclusions consistently. In short, the Supreme Court does not decide whether an article is defamatory or not on the merits, but it checks that the judgment on the merits is reasonably reasoned and that essential elements of assessment are not missing altogether.

Topic 14 - Computer Fraud and Computer System Misuse and Access Covers the criminal law of information technology, distinguishing various cases of computer crime. It starts with the difference between computer fraud (Article 640-ter of the Criminal Code) and misuse of stolen payment cards. It is explained that when someone uses a stolen credit or ATM card to withdraw money, it might look like fraud, but technically it falls under a different crime (regulated by Art. 55 co.9 Legislative Decree 231/2007, concerning the use of other people's payment cards) if there is no actual "hacking" of the system. Computer fraud, on the other hand, occurs when the deception is aimed directly at a computer system: for example, manipulating software, introducing false data or altering an electronic system to obtain an undue profit, rather than deceiving a specific person. The topic emphasizes that in computer fraud the deception affects the functionality of the system (the computer, the server, the ATM), while traditional fraud or misuse of cards affects the human victim who is misled. In addition, the crime of abusive access to a computer system (Art. 615-ter of the Criminal Code) is examined, pointing out that it can be committed not only by those who are completely outside the system, but also by those who were initially authorized to access it: for example, an employee of a company who has credentials for the company system but uses them beyond the permitted limits or for purposes other than those authorized. In practice, if a person enters a protected computer system by violating the conditions set by the owner (perhaps by accessing confidential data without permission while having an account), he or she may still be liable for abusive access. In summary, the topic sheds light on the different types of computer crime, explaining the subtle differences: using a stolen card vs. computer fraud vs. abusive access, and how case law defines the boundaries between these cases.

Topic 26 - Easements over condominium property and "condominium" status Concerns a question of real estate law: whether the holder of an easement over a common part of a condominium should be considered a condominium owner with rights and duties equal to the other owners of the building. A servient easement is, for example, the right of way over someone else's driveway, or the right to use a yard to run pipes, in favor of a neighboring property. In this case, let us imagine that a person from outside the condominium has the right to pass through the hallway or courtyard of a condominium to access his or her property: this person does not

own anything in the condominium, he or she only has a right of easement. The legal question is: does having this easement make her a condominium owner (i.e., part of the condominium)? The answer, clarified by the Supreme Court, is negative. The holder of an easement over condominium property does not automatically become a condominium owner and therefore does not acquire co-ownership of that property nor is he or she required, as a rule, to contribute to general condominium expenses. The condominium owner's rights (e.g., voting in the assembly, or owning a share of the common parts such as stairs, roof, etc.) derive from ownership of a property unit in the building, not from a mere right of easement. Similarly, a condominium owner's obligations (such as paying expenses for elevator maintenance, roof maintenance, etc.) do not apply to those outside the condominium property. In our example, the holder of an easement of passage will only contribute to the specific expenses associated with the exercise of that easement, if any (e.g., he or she may have to contribute to keeping the driveway he or she uses in good condition, according to Article 1069 of the Civil Code), but he or she will not pay the expenses for the elevator or the stair light because he or she is not a condominium owner. In summary, the recognition of an easement in favor of an outsider does not bring him into the "condominium": he remains a third party who enjoys a limited right, governed by the rules on easements, and is not part of the condominium community unless otherwise agreed by contract.

Topic 28 - Appeal to the Supreme Court for trial error In Italian civil law, "misrepresentation of evidence" refers to when a judge misunderstands evidence, basing the decision on something other than what that evidence actually indicates. A distinction is made between errors of perception and errors of interpretation: an error of perception occurs when the judge "missees" the content of evidence (e.g., misreading a date on a document, or mistaking a picture of a car for a picture of a river), while an error of interpretation occurs when the judge correctly understands the evidence but gives it the wrong meaning or weight. This distinction is important because it determines whether and how the error can be rectified. A misrepresentation due to an obvious oversight (factual error) can be corrected by seeking a reversal of the judgment, that is, by reopening the trial to correct that factual error. In contrast, an error in the interpretation of evidence normally cannot be challenged in the Supreme Court, since the Supreme Court does

not review the facts. Only in exceptional cases where the error is so serious as to amount to a failure to consider decisive evidence does the law allow an appeal to the Court of Cassation (Article 360, paragraph 1, no. 5 of the Code of Civil Procedure, relating to failure to examine a decisive fact) to correct the misrepresentation.

We conclude this subsection by presenting Figure 5.20, which visualizes all detected topics within the 2D embedding space of the documents in the dataset.

LLMs evaluations

After analyzing the topic structure of our legal corpus, we investigated the ability of LLMs to interpret and describe the topics identified by BERTopic. Our focus was on assessing their performance in two key tasks: generating descriptive labels and producing concise summaries, which we compared against content produced by a legal domain expert.

To establish a reference standard, we collaborated with a legal expert who manually interpreted and summarized each topic. This allowed us to carry out a dual evaluation—both quantitative and qualitative—of the outputs produced by the LLMs.

In the label generation task, all models achieved comparable results, with LLaMA 3.1 (8B) performing slightly better, attaining an F1 score of 0.8256. However, the differences between models were minor, suggesting similar effectiveness in producing accurate topic labels.

Although the LLAMA model achieved the highest score for label generation in terms of F1, we saw its outputs and were not satisfied because the model frequently ignored the prompt by providing a summary rather than a short label. Claude and GPT's outputs, on the other hand, in addition to being semantically correct, turned out to be true labels.

For the summary generation task, Claude 3.7 was the top performer, with an F1 score of 0.9130, marginally surpassing GPT-4o. Both models generated high-quality summaries, clearly outperforming LLaMA 3.1, which often produced generic or less insightful content. These results are expected, given the training scale and capabilities of Claude and GPT-4o.

In addition to this quantitative evaluation (see Table 5.14), we conducted a qualitative analysis by embedding all labels and summaries—including those generated by the expert—and projecting them into 2D space using UMAP. The resulting visualizations (Figures 5.21 and 5.22) show that the

LLM outputs cluster closely around expert-generated content, indicating strong semantic alignment.

5.5 Conclusion

In this chapter, we presented a document processing pipeline that transforms raw judicial documents (PDFs) into a dataset optimized for topic modeling, ensuring compliance with privacy regulations for litigants involved in cases published on the Italgiurie website by the Italian Supreme Court.

Our pipeline integrates multiple computer vision techniques, including YOLOv8X for layout analysis, which was fine-tuned on our dataset, achieving solid performance on the test set (mAP50: 0.964, mAP50-95: 0.8). The OCR module was implemented using the YOLOv8X algorithm for line-level text detection (mAP@50-95: 0.9022), combined with TrOCR-small for text recognition, which was also fine-tuned on our dataset, reaching high accuracy (CER: 0.0047, WER: 0.0248).

The results demonstrate that topic modeling techniques such as BERTopic benefit significantly from our pipeline, as it improves text segmentation, enabling the extraction of more diverse and coherent topics (topic diversity: 0.6198, topic coherence (C_v): 0.663) compared to datasets processed without segmentation.

A thorough BERTopic analysis of the dataset revealed a complex and layered semantic structure within the Italian legal domain. Key legal topics identified include computer fraud, press Defamation, financial intermediation, appeal to the Supreme Court for trial error and appeals related to Article 606 of the Code of Criminal Procedure.

To evaluate topic interpretability, we compared LLM-generated outputs against human expert annotations. Our findings indicate that Claude 3.7 and GPT4o proved to be quite promising in generating summaries (F1 Score: 0.9130 for Claude 3.7 and F1 Score: 0.9122 for GPT) and also at generating a label to the topic (F1 Score: 0.8119 for Claude 3.7 and F1 Score: 0.7995 for GPT). In contrast LLaMA 3.1 (8B) proved to be very mixed, in fact they generated summaries for topics that were semantically adequate but not as rich in detail as those done by Claude 3.7 and GPT4o, and on the topic label generation task it proved to ignore the prompt several times by generating a summary instead of label.

Despite the reliable performance of our pipeline, some limitations should

be acknowledged. In particular, the GLiNER anonymization module was used in zero-shot mode due to time constraints in dataset preparation for document layout detection and text recognition. Although our thresholding-based anonymization approach proved effective, it occasionally produced false negatives, which could be mitigated by a more tailored training process.

A potential improvement could involve adapting GLiNER to a customized dataset of Italian legal documents or expanding the dataset by collecting additional judgments from Italgiurie. This enhancement would strengthen the anonymization process, resulting in a more robust and legally compliant pipeline.

Our future works will focus on refining this aspect to further improve the reliability and scalability of our approach to legal document processing and extend this work to other documents of Supreme Court like the ordinances and decrees.

Entity	**Tagging**	**Description**
Organization	⟨ORGANIZAZZIONE⟩	Organizations, companies, agencies, institutions
Person	⟨PERSONA⟩	Names of parties, lawyers, judges, family, experts, witness
Location	⟨LOCALITÀ⟩	Place of birth, residence, addresses, countries, cities, states
Email	⟨EMAIL⟩	emails, PECs
DATE	⟨DATA⟩	Date of birth, marriage, death, date of events
ID	⟨ID⟩	Tax Codes, number plates, VAT numbers

Table 5.8: Entity name and description of the 5 entities to be detected

Algorithm	**Hyper-parameters**
Embedding Model	id_model = dlicari/distil-ita-legal-bert max_seq_length = 512 batch_size = 32
UMAP	n_components = 5 min_dist = 0.0 metric = cosine n_neighbors = 5
Count Vectorizer	ngram_range = (1, 2) stop_words = stopwords
HDBSCAN	min_cluster_size = 5 metric = euclidean min_samples = 5
Topic Extraction	top_n_words = 15 diversity = 0.35

Table 5.9: Configuration of hyperparameters for the models that make up BERTopic.

Task	**max_new_tokens**	**temperature**	**repetition_penalty**
Label Generation	50	0.1	1.1
Text Summary	2048	0.1	1.1

Table 5.10: Hyperparameters for Llama 3.1 (8B) model tasks

Algorithm 1 Topic Modeling Algorithm

1: **Input:**

- Let $D = \{\{p_{11}, p_{12}, \ldots, p_{1m_1}\}, \ldots, \{p_{n1}, p_{n2}, \ldots, p_{nm_n}\}\}$ a set of anonymized textual data

2: **Output:**

- Identified topics with labels and description

3: **Steps:**

1. **Embedding Generation with Distiled Italian-legal BERT**
   - $\mathbf{e}_{ij} = M(p_{ij}) \quad \forall i = 1, 2, \ldots, n \quad \wedge \quad \forall j = 1, 2, \ldots, m_i$
2. **Dimensionality Reduction with UMAP**
   - $E = \{\{\mathbf{e}_{11}, \mathbf{e}_{12}, \ldots, \mathbf{e}_{1m_1}\}, \{\mathbf{e}_{21}, \mathbf{e}_{22}, \ldots, \mathbf{e}_{2m_2}\}, \ldots, \{\mathbf{e}_{n1}, \mathbf{e}_{n2}, \ldots, \mathbf{e}_{nm_n}\}\}$
   - $\mathbf{z}_{ij} = \text{UMAP}(\mathbf{e}_{ij}) \quad \forall i = 1, 2, \ldots, n \quad \wedge \quad \forall j = 1, 2, \ldots, m_i$
3. **Clustering embeddings with HDBSCAN**
   - $Z = \{\{\mathbf{z}_{11}, \mathbf{z}_{12}, \ldots, \mathbf{z}_{1m_1}\}, \{\mathbf{z}_{21}, \mathbf{z}_{22}, \ldots, \mathbf{z}_{2m_2}\}, \ldots, \{\mathbf{z}_{n1}, \mathbf{z}_{n2}, \ldots, \mathbf{z}_{nm_n}\}\}$
   - $C = \text{HDBSCAN}(Z)$
   - $C = \{C_1, C_2, \ldots, C_k, -1\}$
4. **Topic Representation with using c-TF-IDF**
   - $w_{x,C_i} = \text{tf}_{x,C_i} \times \log\left(1 + \frac{A - f_x + .5}{f_x + .5}\right)$
   - w_{x,C_i} represents the weight of the term x in cluster C_i,
   - tf_{x,C_i} is the frequency of the term x within cluster C_i,
   - f_x is the frequency of the term x across all clusters,
   - A is the average number of words per cluster (documents).
5. **Generate labels using a LLMs**
6. **Generate text-summarization using a LLMs**

Figure 5.15: Topic modelling algorithm for the analysis of legal topics. The pseudo code has been transformed to resolve incompatibility issues with the pseudo codes in other chapters that use specific packages.

Table 5.11: The table shows the results of the document layout analysis model developed to examine the layout of judgments on test-set. Having treated the layout analysis problem as an object detection problem, evaluation metrics from that domain are used to evaluate the model. The arrow indicates how the metrics should be read. The higher the better.

Class	Images	Instances	Precision ↑	Recall ↑	mAP50 ↑	mAP50-95 ↑
All	82	509	0.933	0.960	0.964	0.800
Page-footer	67	67	0.954	0.926	0.982	0.612
Section-header	24	25	0.923	0.960	0.905	0.738
Text	81	411	0.954	0.951	0.975	0.892
Title	6	6	0.902	1.000	0.995	0.959

Table 5.12: The table reports the results of the OCR model developed to detect and extract the text of sentences on test-set. It consists of two rows in the table. In the first row are the results related to the text detection module, and in the second row are the results of the text recognition module. The direction of the arrow helps in the interpretation of the metrics.

Model	Task	Precision ↑	Recall	mAP50 ↑	mAP@50-95 ↑	CER ↓	WER ↓
YOLOV8x	Text detection	0.9864	0.9834	0.9901	0.9022	-	-
TrOCR-Small	Text recognition	-	-	-	-	0.0047	0.0248

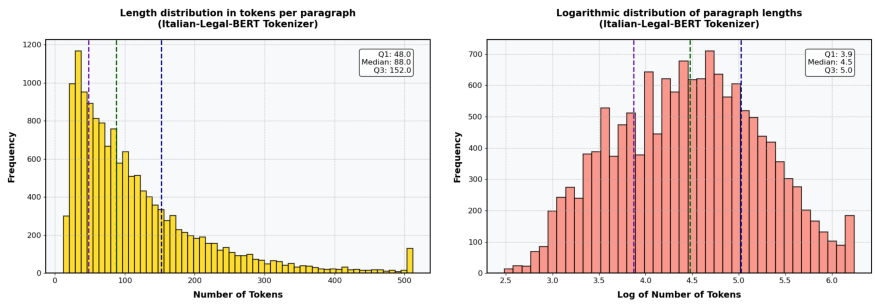


Figure 5.16: Paragraph length distributions in the analyzed documents. Left: absolute frequency distribution of paragraph lengths tokenized with distil-italian-legal-bert. Right: logarithmic distribution of lengths, showing a more normalized pattern. The dotted lines indicate the first quartile (purple), median (green) and third quartile (blue) of the distribution, respectively.

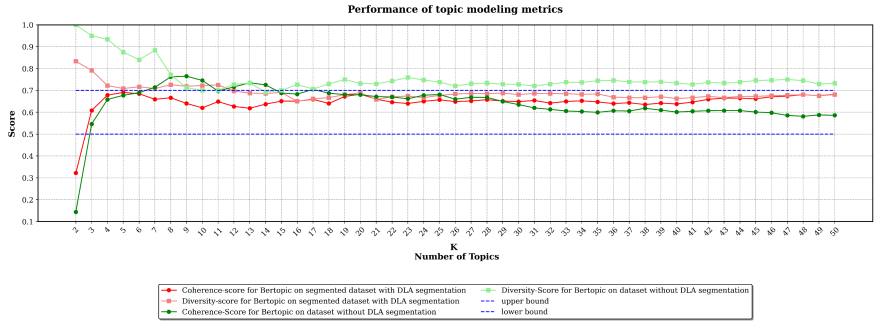
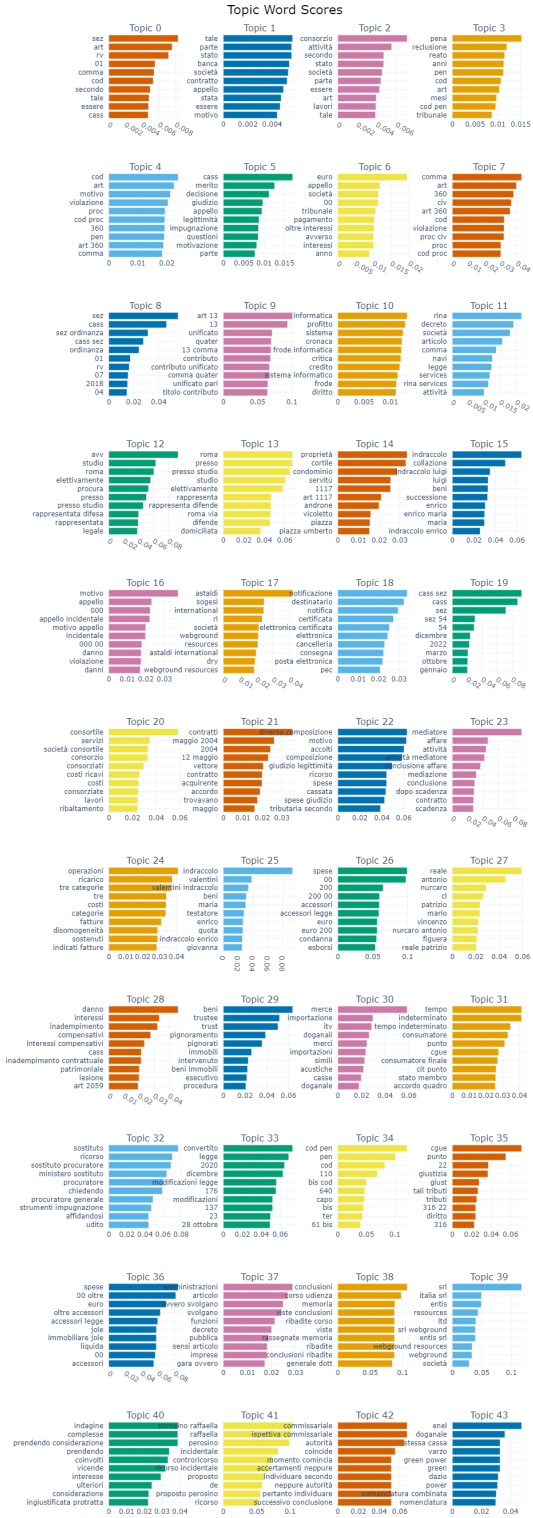
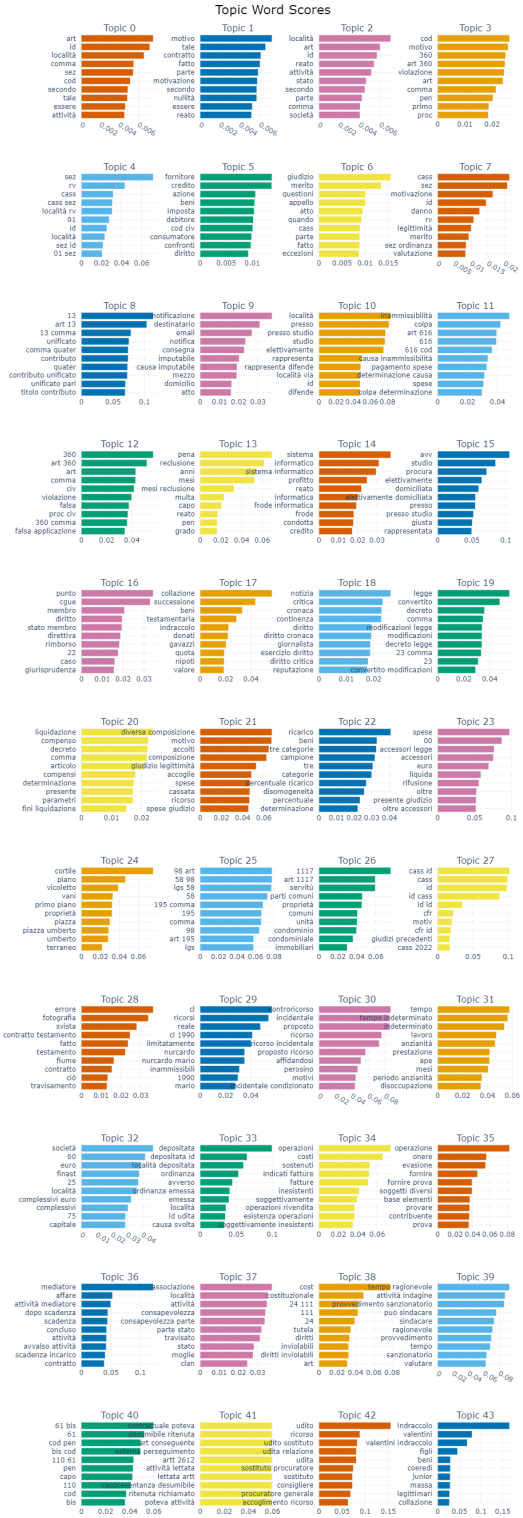


Figure 5.17: The Figures illustrates the performance trend of BERTopic as a function of the number of topics, using the distil-ita-legal-bert embedding model. The dashed blue lines represent the search boundaries for topic diversity and topic coherence values. The dark red (topic coherence) and light red (topic-diversity) lines represent BERTopic performance on the dataset created with our pipeline, while the dark green (topic coherence) and light green (topic-diversity) lines represent BERTopic performance on data obtained without the use of our pipeline.

Embedding model	**DLA Seg.**	**K** ↑	**TD** ↑	C_v ↑
distil-ita-legal-bert	Yes	48	0.6803	0.6809
distil-ita-legal-bert	No	15	0.7	0.6878
sentence-bert-base-italian-uncased	Yes	47	0.6956	0.6506
sentence-bert-base-italian-uncased	No	12	0.6061	0.6280

Table 5.13: Comparison of topic modeling results using different versions of BERTopic with and without DLA segmentation.


Topic Word Scores
Topic 0
sez
art
rv
01
comma
cod
secondo
tale
essere
cass
Topic 1
tale
parte
stato
banca
società
contratto
appello
stata
essere
motivo
Topic 2
consorzio
attività
secondo
stato
società
parte
essere
art
lavori
tale
Topic 3
pena
reclusione
reato
anni
pen
cod
art
mesi
cod pen
tribunale
Topic 4
cod
art
motivo
violazione
proc
cod proc
360
pen
art 360
comma
Topic 5
cass
merito
decisione
giudizio
appello
legittimità
impugnazione
questioni
motivazione
parte
Topic 6
euro
appello
società
00
tribunale
pagamento
oltre interessi
avverso
interessi
anno
Topic 7
comma
art
360
civ
art 360
cod
violazione
proc civ
proc
cod proc
Topic 8
sez
cass
sez ordinanza
cass sez
ordinanza
01
rv
07
2018
04
Topic 9
art 13
13
unificato
quater
13 comma
contributo
contributo unificato
comma quater
unificato pari
titolo contributo
Topic 10
informatica
profitto
sistema
cronaca
frode informatica
critica
credito
sistema informatico
frode
diritto
Topic 11
rina
decreto
società
articolo
comma
navi
legge
services
rina services
attività
Topic 12
avv
studio
roma
elettivamente
procura
presso
presso studio
rappresentata difesa
rappresentata
legale
Topic 13
roma
presso
presso studio
studio
elettivamente
rappresenta
rappresenta difende
roma via
difende
domiciliata
Topic 14
proprietà
cortile
condominio
servitù
1117
art 1117
androne
vicoletto
piazza
piazza umberto
Topic 15
indraccolo
collazione
indraccolo luigi
luigi
beni
successione
enrico
enrico maria
maria
indraccolo enrico
Topic 16
motivo
appello
000
appello incidentale
motivo appello
incidentale
000 00
danno
violazione
danni
Topic 17
astaldi
sogesi
international
rl
società
webground
resources
astaldi international
dry
webground resources
Topic 18
notificazione
destinatario
notifica
certificata
elettronica certificata
elettronica
cancelleria
consegna
posta elettronica
pec
Topic 19
cass sez
cass
sez
sez 54
54
dicembre
2022
marzo
ottobre
gennaio
Topic 20
consortile
servizi
società consortile
consorzio
consorziati
costi ricavi
costi
consorziate
lavori
ribaltamento
Topic 21
contratti
maggio 2004
2004
12 maggio
vettore
contratto
acquirente
accordo
trovavano
maggio
Topic 22
diversa composizione
motivo
accolti
composizione
giudizio legittimità
ricorso
spese
cassata
spese giudizio
tributaria secondo
Topic 23
mediatore
affare
attività
società mediatore
conclusione affare
mediazione
conclusione
dopo scadenza
contratto
scadenza
Topic 24
operazioni
ricarico
tre categorie
tre
costi
categorie
fatture
dismogeneità
sostenuti
indicati fatture
Topic 25
indraccolo
valentini
valentini indraccolo
beni
maria
testatore
enrico
quota
indraccolo enrico
giovanna
Topic 26
spese
00
200
200 00
accessori
accessori legge
euro
euro 200
condanna
esborsi
Topic 27
reale
antonio
nurcaro
cl
patrizio
mario
vincenzo
nurcaro antonio
figuera
reale patrizio
Topic 28
danno
interessi
inadempimento
compensativi
interessi compensativi
cass
inadempimento contrattuale
patrimoniale
lesione
art 2059
Topic 29
beni
trustee
trust
pignoramento
pignorati
immobili
intervento
beni immobili
esecutivo
procedura
Topic 30
merce
importazione
itv
doganali
merci
importazioni
simili
acustiche
casse
doganale
Topic 31
tempo
indeterminato
tempo indeterminato
consumatore
punto
cgue
consumatore finale
clt punto
stato membro
accordo quadro
Topic 32
sostituto
ricorso
sostituto procuratore
ministero sostituto
procuratore
chiedendo
procuratore generale
strumenti impugnazione
affidandosi
udito
Topic 33
convertito
legge
2020
dicembre
modificazioni legge
176
modificazioni
137
23
28 ottobre
Topic 34
cod pen
pen
cod
110
bis cod
640
capo
bis
ter
61 bis
Topic 35
cgue
punto
22
giustizia
giust
tali tributi
tributi
316 22
diritto
316
Topic 36
spese
00 oltre
euro
oltre accessori
accessori legge
jole
immobiliare jole
liquida
00
accessori
Topic 37
amministrazioni
articolo
avvero svolgano
svolgano
funzioni
decreto
pubblica
sensi articolo
imprese
gara ovvero
Topic 38
conclusioni
corso udienza
memoria
note conclusioni
ribadite corso
viste
assegnate memoria
ribadite
conclusioni ribadite
generale dott
Topic 39
srl
italia srl
entis
resources
ltd
srl webground
entis srl
webground resources
webground
società
Topic 40
indagine
complesse
prendendo considerazione
prendendo
coinvolti
vicende
interesse
ulteriori
considerazione
ingiustificata protratta
Topic 41
commissariale
ispettiva commissariale
perosino
incidentale
controricorso
avverso incidentale
proposto
de
proposto perosino
ricorso
Topic 42
enel
doganale
stessa cassa
varzo
green power
green
dazio
power
nomenclatura combinata
nomenclatura
Topic 43


Topic Word Scores
Topic 0
art
id
località
comma
sez
cod
secondo
tale
essere
attività
Topic 1
motivo
tale
contratto
fatto
parte
motivazione
secondo
nullità
essere
reato
Topic 2
località
art
id
reato
attività
stato
secondo
parte
comma
società
Topic 3
cod
motivo
360
art 360
violazione
art
comma
pen
primo
proc
Topic 4
sez
rv
cass
cass sez
località rv
01
id
località
sez id
01 sez
Topic 5
fornitore
credito
azione
beni
imposta
debitore
cod civ
consumatore
confronti
diritto
Topic 6
giudizio
merito
questioni
appello
atto
quando
cass
parte
fatto
eccezioni
Topic 7
cass
sez
motivazione
id
danno
rv
legittimità
merito
sez ordinanza
valutazione
Topic 8
13
art 13
13 comma
unificato
comma quater
contributo
quater
contributo unificato
unificato pari
titolo contributo
Topic 9
notificazione
destinatario
email
notifica
consegna
imputabile
causa imputabile
mezzo
domicilio
atto
Topic 10
località
presso
presso studio
studio
elettivamente
rappresenta
rappresenta difende
località via
id
difende
Topic 11
inammissibilità
colpa
art 616
616
616 cod
causa inammissibilità
pagamento spese
determinazione causa
spese
colpa determinazione
Topic 12
360
art 360
art
comma
civ
violazione
falsa
proc civ
360 comma
falsa applicazione
Topic 13
pena
reclusione
anni
mesi
mesi reclusione
multa
capo
reato
pen
grado
Topic 14
sistema
informatico
sistema informatico
profitto
reato
informatica
frode informatica
condotta
credito
Topic 15
avv
studio
procura
elettivamente
domiciliata
elettivamente domiciliata
presso
presso studio
frode
giusta
rappresentata
Topic 16
punto
cque
membro
diritto
stato membro
direttiva
rimborso
22
caso
giurisprudenza
Topic 17
collazione
successione
beni
testamentaria
indraccolo
donati
gavazzi
quota
nipoti
valore
Topic 18
notizia
critica
cronaca
continenza
diritto
diritto cronaca
giornalista
esercizio diritto
diritto critica
reputazione
Topic 19
legge
convertito
decreto
comma
modificazioni legge
modificazioni
decreto legge
23 comma
23
convertito modificazioni
Topic 20
liquidazione
compenso
decreto
comma
articolo
compensi
determinazione
presente
parametri
fini liquidazione
Topic 21
diversa composizione
motivo
accolti
composizione
giudizio legittimità
accoglie
spese
cassata
ricorso
spese giudizio
Topic 22
ricarico
beni
tre categorie
campione
tre
categorie
percentuale ricarico
disomogeneità
percentuale
determinazione
Topic 23
spese
00
accessori legge
accessori
euro
liquida
rifusione
oltre
presente giudizio
oltre accessori
Topic 24
cortile
piano
vicoletto
vani
primo piano
proprietà
piazza
piazza umberto
umberto
terraneo
Topic 25
98 art
98 98
lgs 58
58
195 comma
195
comma
98
art 195
lga
Topic 26
1117
art 1117
servitù
parti comuni
proprietà
comuni
unità
condominio
condominiale
immobiliari
Topic 27
cass id
cass
id
id cass
id id
cfr
motivi
cfr id
giudizi precedenti
cass 2022
Topic 28
errore
fotografia
svista
contratto testamento
fatto
testamento
fiume
contratto
ciò
travisamento
Topic 29
cl
ricorsi
reale
cl 1990
limitatamente
nurcardo
nurcardo mario
inammissibili
1990
mario
Topic 30
controricorso
incidentale
proposto
ricorso
ricorso incidentale
proposto ricorso
affidandosi
perosino
motivi
incidentale condizionato
Topic 31
tempo
tempo indeterminato
indeterminato
lavoro
anzianità
prestazione
ape
mesi
periodo anzianità
disoccupazione
Topic 32
società
80
euro
finast
25
località
complessivi euro
complessivi
75
capitale
Topic 33
depositata
depositata id
località depositata
ordinanza
avverso
ordinanza emessa
emessa
località
id udita
causa svolta
Topic 34
operazioni
costi
sostenuti
indicati fatture
fatture
inesistenti
soggettivamente
operazioni rivendita
esistenza operazioni
soggettivamente inesistenti
Topic 35
operazione
onere
evasione
fornire
fornire prova
soggetti diversi
base elementi
provare
contribuente
prova
Topic 36
mediatore
affare
attività mediatore
dopo scadenza
scadenza
concluso
attività
avvalso attività
scadenza incarico
contratto
Topic 37
associazione
località
attività
consapevolezza
consapevolezza parte
parte stato
travisato
stato
moglie
clan
Topic 38
cost
costituzionale
24 111
111
24
tutela
diritti
inviolabili
diritti inviolabili
art
Topic 39
tempo
tempo ragionevole
attività indagine
provvedimento sanzionatorio
può sindacare
sindacare
ragionevole
provvedimento
tempo
sanzionatorio
valutare
Topic 40
61 bis
61
cod pen
bis cod
110 61
pen
capo
110
cod
bis
Topic 41
udito
ricorso
udito sostituto
udita relazione
udita
sostituto procuratore
sostituto
consigliere
procuratore generale
accoglimento ricorso
Topic 42
indraccolo
valentini
valentini indraccolo
figli
beni
coeredi
junior
massa
legittimari
collazione
Topic 43

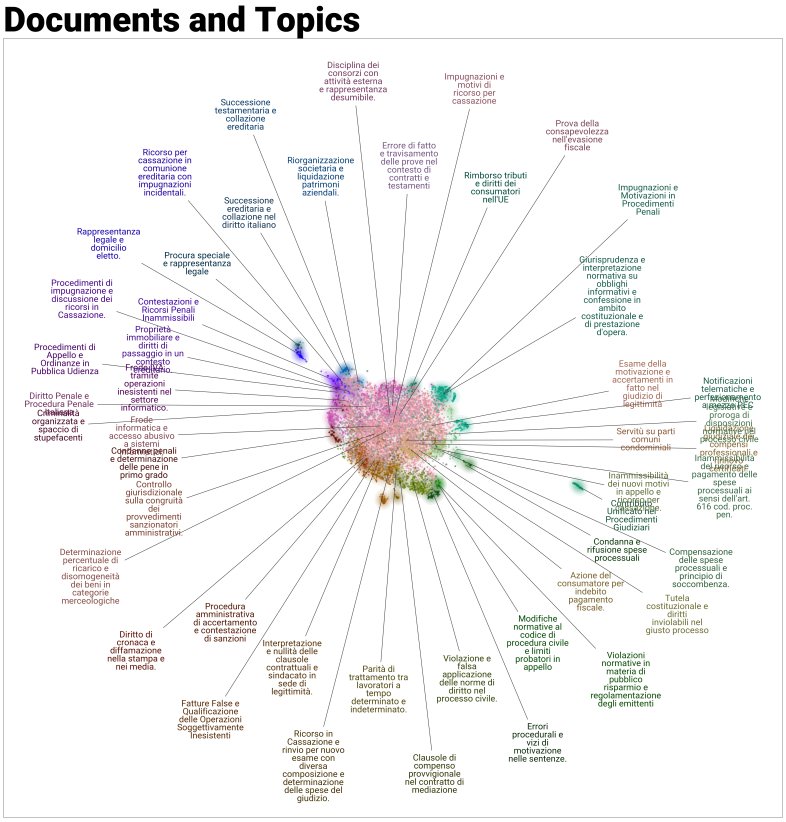


Figure 5.20: Documents and Topics in Italian Legal Procedures: A 2D scatter plot showing the complex dataset structure. The color-coded clusters represent different legal topic (administrative law, criminal procedure, digital authentication, and COVID-19 related regulations).

Model	Topic-Label Generation			Topic Summarization		
	Prec. ↑	**Rec. ↑**	**F1 ↑**	**Prec. ↑**	**Rec. ↑**	**F1 ↑**
Claude-Sonnet3.7	0.8220	0.8046	0.8119	0.9058	**0.9205**	**0.9130**
GPT4	0.7895	**0.8153**	0.7995	0.7995	0.9180	0.9122
Llama3.1 (8B)	**0.8558**	0.8010	**0.8256**	**0.9092**	0.8817	0.8951

Table 5.14: Comparison of BERTScore metrics for Topic-Label Generation and Topic Summarization tasks across different models. Best results are highlighted in **bold**.

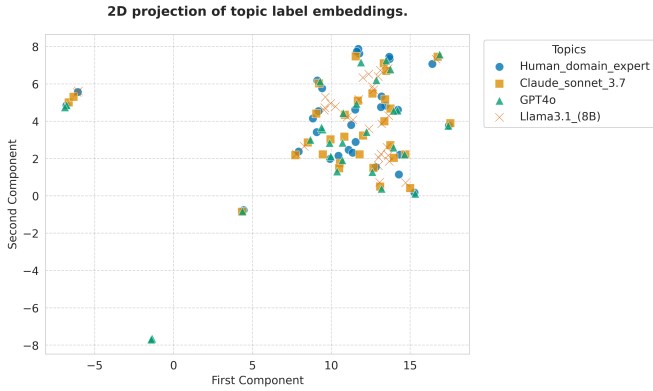


Figure 5.21: Two-dimensional projection of topic label embeddings generated by different language models (Claude-Sonnet 3.7, GPT4o, Llama 3.1 (8B)) and a human expert. The visualization shows significant spatial overlap between the labels generated by the different LLMs, suggesting strong semantic agreement.

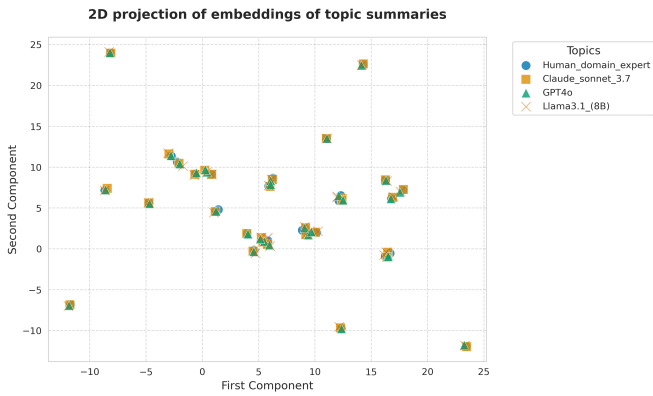


Figure 5.22: Two-dimensional projection of topic text summary embeddings generated by different language models (Claude-Sonnet 3.7, GPT4, Llama 3.1(8B)) and a human expert. The visualization shows significant spatial overlap between the summaries generated by the different LLMs, suggesting strong semantic agreement.

Chapter 6

A graph based multimodal model for paragraph Retrieval task in Italian Supreme Court judgments

This chapter addresses the problem of paragraph retrieval in Italian Supreme Court judgments, where the goal is to identify the most relevant passages in response to legally significant queries. We propose a multimodal graph-based model that represents each document as a graph of paragraphs and integrates textual, visual, and structural information, trained with contrastive learning. Experimental results show stable and competitive performance (MAP ≈ 0.83–0.84, *MRR* > 0.91*), with ablation studies highlighting the dominant role of textual features in the current setup.*

6.1 Introduction

The judgments of the Italian Supreme Court are one of the most authoritative sources for the interpretation and application of law in Italy. However, these documents are often multi-page and dense with arguments, citations and regulatory references: quickly identifying key information (e.g. final de-

cision, parties involved, legal costs or principles of law) requires time and specific skills. For this reason, automatic *information retrieval* tools for legal documents can provide concrete support to law firms and researchers, reducing consultation costs and speeding up analysis.

In this work, we address a paragraph retrieval task on Italian Supreme Court rulings: given a query in natural language, the goal is to retrieve the most relevant paragraphs of the document to answer the question. To make the problem operational, we define a set of seven queries (q0–q6) covering central aspects of understanding a judgement, such as the identification of the parties, the presence of citations of precedents, the final decision, the legal costs, the articles cited, the grounds for appeal and the statement of the principle of law.

A distinctive feature of our setting is the structured nature of the document: the relevance of a paragraph depends not only on its textual content, but also on its location and relationships with adjacent sections. To capture this aspect, we model each document as a graph of paragraphs, constructing connections based on the spatial proximity of regions and introducing links between consecutive pages to maintain continuity throughout the entire document. On this representation, we propose a neural model that combines (i) textual embeddings based on a BERT [17] encoder specialised for legal Italian [44], (ii) visual embeddings extracted from RGB crops of the paragraphs, and (iii) a graph encoding module (GATv2, [99]) to propagate context along the layout structure. The query-paragraph relevance estimate is performed using a cross-attention mechanism, while training is done with contrastive learning (InfoNCE [3] and Triplet loss [19]).

Since we do not have annotations curated by domain experts, we use an LLM model (GPT-4o mini, [37]) to associate the most relevant query with each paragraph, while also maintaining confidence in alternative classes. This choice allows us to build a supervised dataset and define mining strategies for contrastive learning, enabling controlled experiments on the contribution of different components.

The main contributions of this chapter can be summarised as follows:

- We define a paragraph retrieval task on Italian Supreme Court rulings using a set of legally significant queries;
- We propose a graph-based formulation of the document, which integrates textual, visual and structural information;

- we conduct an experimental evaluation with standard ranking metrics, comparing InfoNCE and Triplet loss and analysing the impact of the modules through ablation studies.

6.2 Dataset

The judgments of the Italian Supreme Court are published annually on the Italgiure website, which provides access to legal documents in PDF format. The quality of these documents is generally high, as most of the digitised documents do not have defects such as skewing, blurring, photometric alterations, compression artefacts or poor resolution. Nevertheless, minor imperfections may occasionally occur. In some cases, handwritten notes or corrections appear, while ink defects affecting entire pages or lines are rare but not entirely absent. In this work, we downloaded several civil and criminal judgments and processed them with a document analysis pipeline built with YOLOv8 [97] and TrOCR [41] in order to segment the documents into paragraphs and extract their content.

Judgments are a type of multi-page legal document dense with information, so extracting and retrieving information from these documents is indeed a useful activity that simplifies the study of these documents by law firms and legal scholars who need to study legal judgments.

For each document, the pipeline will generate a series of json files corresponding to the pages that make up the document. Each json file reports the parsing of the document, in which the bounding boxes and extracted text are reported for each paragraph. This information is essential for evaluating the use of a multimodal approach.

To build an information retrieval system for these documents, we have formulated questions that are useful for understanding the judgement. These questions are:

- q0: 'Who are the parties involved (roles: plaintiff, counterparty, defendant, third parties, etc.)?'
- q1: 'Does the block contain citations of previous Supreme Court rulings (e.g., "Cass. [date], no. [number]")?'
- q2: 'What is the final decision of the Court of Cassation (confirmation, rejection, annulment with referral, annulment without referral)?'

- q3: 'Who is charged with the legal costs?'
- q4: 'Which articles of law or regulations are explicitly cited in the text?'
- q5: 'Does the paragraph contain an indication of a ground for appeal?' q6: 'Does the paragraph contain an indication of the principle of law or the maxim of the judgement?'

Since we did not have a domain expert who could tell us which paragraphs answer these questions, we used GPT-4o mini for this purpose and leveraged its knowledge of legal matters. For each paragraph, we have GPT-4o mini's choice and also the confidences of the other questions, which allows us to use approaches based on contrastive learning.

6.2.1 Construction of incidence matrices

The dataset does not provide a predefined graphical representation. Therefore, for each page, we construct an incidence matrix based on the position of the regions identified through layout analysis. Let

$$P = \{p_0, \dots, p_{m-1}\}$$

be the set of pages in a document and let

$$C_p = \{c_0, \dots, c_{n_{p-1}}\}$$

the centres of the bounding boxes of page p. We use these points to perform a neighbourhood query in 2D Euclidean space using a KD-tree [70], obtaining a local incidence matrix for each page

$$A_p \in \{0,1\}^{n_p \times n_p}.$$

Algorithm 6.1 describes the procedure.

Since many documents are multi-page, we combine the local matrices $\{A_p\}$ to obtain a global incidence matrix

$$G \in \{0,1\}^{N \times N}, \qquad N = \sum_{p=0}^{m-1} n_p,$$

where the diagonal blocks correspond to pages and a link is added between the last node of A_p and the first node of A_{p+1}, preventing the formation

Algorithm 3: Adjacency Matrix Construction via k-Nearest Neighbors

Input: Node centers $C = \{c_1, \ldots, c_n\}$, number of neighbors k
Output: Adjacency matrix $A \in \{0,1\}^{n \times n}$

```
if n = 0 then
    return empty matrix
k ← min(k, n);                    // limit number of neighbors
Train a k-NN model on C;
Retrieve N_k(i) for each node c_i;
Initialize A ← 0_{n×n};
for i = 1 to n do
    foreach j ∈ N_k(i) \ {i} do
        A_ij ← 1;
return A
```

Figure 6.1: Algorithm for constructing the local adjacency matrix and then deriving the list of edges. The pseudo code has been transformed to resolve incompatibility issues with the pseudo codes in other chapters that use specific packages.

of disconnected components. Alg. 6.2 formalises the process and Figure 6.3 shows an example of global adjacency matrix for a judgement .

Since the global incidence matrix represents the entire document, nodes are indexed continuously in global space, taking values in

$$\{0, \ldots, N-1\}.$$

Consequently, the index of the ground truth node, initially defined at the page level, must also be remapped to global space.

Let $local_{idx}$ be the local index of the correct node on page p and let n_j be the number of nodes on page j. The corresponding global index is obtained by

$$global_{idx} = offset[p] + local_{idx},$$

where the cumulative offset is defined as

$$offset[p] = \begin{cases} 0, & \text{if } p = 0, \\ \sum_{j=0}^{p-1} n_j, & \text{if } p \geq 1. \end{cases}$$

Algorithm 4: Global Adjacency Matrix Construction Across Document Pages

Input: Dictionary of page-wise adjacency matrices
$\mathcal{A} = \{A_1, \ldots, A_P\}$
Output: Global incidence matrix $G \in \{0,1\}^{N \times N}$

if $\mathcal{A}$ *is empty* **then**
  return empty matrix
$N \leftarrow 0$;
foreach *page* p *in sorted*($\mathcal{A}$) **do**
  $n_p \leftarrow |A_p|$; // nodes in page
  Assign node IDs in $[N, N + n_p)$;
  $N \leftarrow N + n_p$;
Initialize $G \leftarrow \mathbf{0}_{N \times N}$;
$i \leftarrow 0$, $j \leftarrow 0$;
for $p = 1$ **to** P **do**
  $A \leftarrow A_p$;
  Copy A into $G[i : i + |A|,\ j : j + |A|]$;
  if $p < P$ **then**
    $G[i + |A| - 1,\ j + |A|] \leftarrow 1$
    ; // inter-page link
  $i \leftarrow i + |A|,\ j \leftarrow j + |A|$;
Remove diagonal self-loops: $G[x, x] \leftarrow 0$;
return G

Figure 6.2: Algorithm for constructing the global adjacency matrix and then deriving the list of edges. The pseudo code has been transformed to resolve incompatibility issues with the pseudo codes in other chapters that use specific packages.

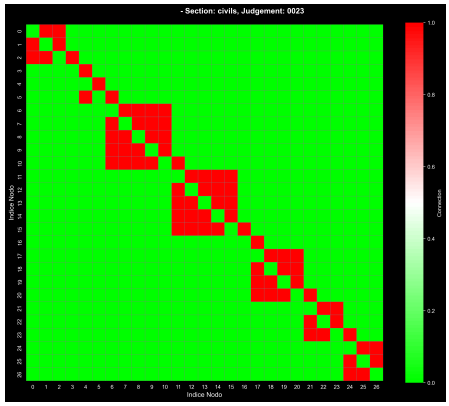


Figure 6.3: An example of global adjacency matrix for a judgement. The colour red indicates that there is a connection between two nodes in the graph, while green indicates that there is no connection.

The inverse mapping, necessary to retrieve the local index from the global one, is simply given by

$$local_{idx} = global_{idx} - offset[p].$$

6.3 Proposed approach

In this section, we present our method. We have decided to exploit all the features that the dataset has to offer, so our method will be multimodal. In Figure 6.4, we present our model.

6.3.1 Image encoder

Our model accepts RGB image crops of documents as input. This choice is central to the proposed architecture and is well justified in the literature. Pre-trained Vision Transformers such as CLIP [72] ViT-B/32 [20] operate on fixed resolutions (224×224 pixels): processing an entire document at this resolution would result in a critical loss of textual information. By extracting crops corresponding to the bounding boxes of individual elements, each region maintains sufficient resolution to preserve many details. Each crop is transformed into an embedding $v_i \in \mathbb{R}^{d_v}$ and then aligned so that it has the same size as the text embeddings.

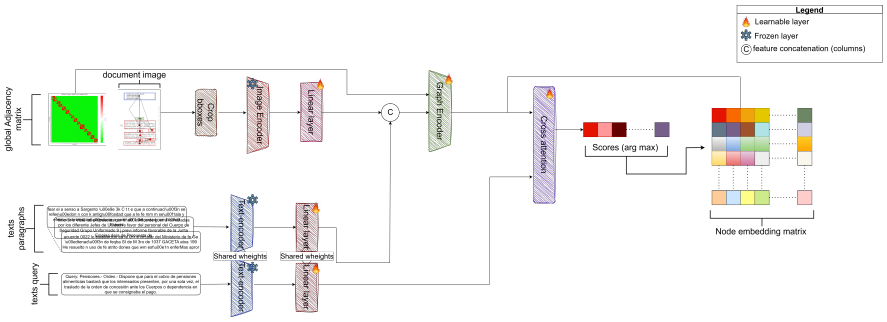


Figure 6.4: Architecture of our multimodal model. The various components will be discussed in the appropriate subsections.

6.3.2 Text encoder

The query and paragraphs of the document are processed by a BERT-type text encoder optimised for Italian legal language, Legal-italian-BERT. This BERT model is perfect for solving classification, NER and sentence embedding tasks for the legal field, so we will use it as a frozen backbone to create embeddings for our task.

6.3.3 Graph Encoder

Once multimodal representations of individual document fragments have been obtained, the next step is to integrate the structural information derived from the layout. In the context of historical documents, the meaning of a fragment does not depend solely on its local content, but also on its position and relationship with surrounding elements, such as titles, adjacent paragraphs, or consecutive sections.

To model these structural dependencies, we represent each document as an undirected graph

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}),$$

where each node $v_i \in \mathcal{V}$ corresponds to a fragment of the document, while the edges $e_{ij} \in \mathcal{E}$ are defined from the global incidence matrix G described in the previous section. The edges encode spatial proximity relationships within the same page and sequential links between adjacent pages.

Given the set of node embeddings

$$H^{(0)} = \{h_1^{(0)}, \ldots, h_N^{(0)}\}, \qquad h_i^{(0)} = [\, v_i \,||\, t_i \,]$$

obtained by concatenating the visual and textual representations, we apply a Graph Neural Network based on Graph Attention Networks (GATv2). This model allows us to dynamically weight the contribution of adjacent nodes during the aggregation process, making the graph sensitive to both the structure and content of the fragments.

In each layer of the graph encoder, the embedding of a node is updated according to:

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(l)} W^{(l)} h_j^{(l)}\right)$$

where $\mathcal{N}(i)$ denotes the set of neighbours of node i, $\alpha_{ij}^{(l)}$ represents the attention coefficient learned on arc (i, j) and $W^{(l)}$ is a linear projection matrix. The graph encoder consists of two GATv2 layers, with multi-head attention in the first layer and final aggregation in the second.

The main effect of this module is the propagation of structural context throughout the document: each fragment enriches its own representation by integrating information from neighbouring fragments. In this way, the model is able to capture semantic-structural patterns distributed across multiple regions, making it particularly robust in the presence of textually ambiguous or partially degraded fragments.

Fig. 6.5 illustrates the role of the graph encoder within the overall architecture.

6.3.4 Cross-attention to estimate query-paragraph relevance

After multimodal fusion and contextualisation via graph encoder, we obtain a contextual representation for each fragment of the document:

$$Z = \{z_0, \ldots, z_{N-1}\}, \qquad z_i \in \mathbb{R}^d,$$

where N is the total number of nodes (fragments) in the document and d is the common dimension. In parallel, the natural language query is encoded into an embedding

$$q \in \mathbb{R}^d.$$

The retrieval task can be interpreted as a *query-conditioned scoring* problem: given q and the set of nodes Z, we want to estimate a score s_i that quantifies how relevant fragment i is to the query.

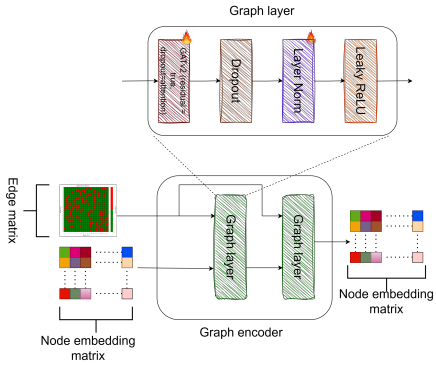


Figure 6.5: Schematic diagram of the graph encoder based on GATv2. Node embeddings are updated by aggregating information from adjacent fragments according to the document layout structure.

For this purpose, we introduce a *cross-attention* module that uses the query as **Query** and the nodes as **Key/Value**. The intuition is that the query should dynamically "interrogate" the document fragments, assigning greater weight to nodes that are semantically and structurally consistent with the topic expressed by the query.

In particular, we project queries and nodes into the same latent space $\mathbb{R}^d$ and apply a multi-head attention:

$$Q = W_q q \in \mathbb{R}^d, \qquad K = W_k Z \in \mathbb{R}^{N \times d}, \qquad V = W_v Z \in \mathbb{R}^{N \times d}.$$

The attention coefficients (normalised) are obtained as:

$$\alpha = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) \in \mathbb{R}^N,$$

where α_i represents the weight assigned to node i with respect to the query. In our model, we directly use these weights as ranking scores:

$$\text{score}_i = \alpha_i.$$

This scheme has two main advantages. First, the scores are query-adaptive: the same node can receive different weights depending on the query, allowing the model to select different regions of the document for

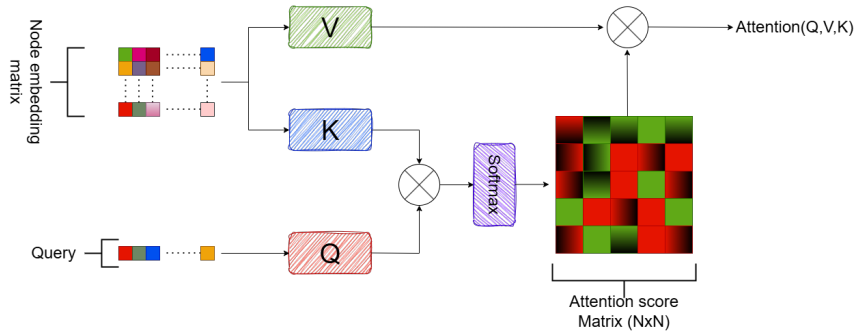


Figure 6.6: Cross attention module to calculate the relevance of paragraphs for queries.

different topics. Second, cross-attention operates on embeddings already contextualised by the graph: the score of a fragment depends not only on its local content, but also on the structural context propagated by the graph encoder.

Fig. 6.7 illustrates how the cross-attention module works: the query is used as a guide vector to weight the set of fragments and produce a relevance distribution on the document.

6.3.5 Triplet loss

Our training objective is to learn a latent space in which the query embedding is closer to the correct (positive) fragment than to an irrelevant (negative) fragment. To this end, we adopt a cosine distance-based triplet loss, widely used in metric learning and retrieval contexts.

For each training instance, given the query embedding $q \in \mathbb{R}^d$ and the document node embeddings $\{z_0, \ldots, z_{N-1}\}$ produced by the model, we construct a triplet

$$(q,\ z_p,\ z_n),$$

where p is the index of the positive node (ground-truth) and n is the index of a negative node randomly selected from the set of non-positive nodes. In our case, the positive index p corresponds to the fragment with maximum similarity in the document, while the negative ones are defined as all other fragments.

We define the distance between two embeddings as:

$$d_{\cos}(a, b) = 1 - \cos(a, b) = 1 - \frac{a^\top b}{\|a\|_2 \, \|b\|_2}.$$

The triplet margin loss is therefore:

$$\mathcal{L}_{\text{triplet}} = \max\Big(0,\ d_{\cos}(q, z_p) - d_{\cos}(q, z_n) + \gamma\Big),$$

where γ is a positive margin set to $\gamma = 0.3$.

In other words, the loss penalises the model when a negative fragment is closer to the query than the correct fragment, or when the separation between positive and negative does not exceed the margin γ. This formulation encourages the formation of a discriminative latent space for retrieval: for the same query, relevant fragments are attracted to the query, while irrelevant ones are rejected.

In our setup, we use GPT4o-mini's knowledge to write a mining algorithm to construct triples. The algorithm is shown in the pseudo-code below.

6.3.6 Data Augmentation

To improve model generalisation and prevent overfitting, three data augmentation techniques were implemented, each of which can be independently disabled for ablation studies.

Query Paraphrasing

To increase the model's robustness with respect to the linguistic variability of queries, each original query q_i was associated with a set of 5 semantically equivalent paraphrases $\mathcal{P}(q_i) = \{q_i^{(1)}, \ldots, q_i^{(5)}\}$. During training, a different paraphrase is selected pseudo-deterministically at each epoch e:

$$\tilde{q}_i^{(e)} = \mathcal{P}(q_i)\,[\,h(i, e) \mod |\mathcal{P}(q_i)|\,]$$

where $h(i, e)$ is a hash function that depends on the query index and epoch. This ensures that:

- The same query in the same epoch always uses the same paraphrase (reproducibility)
- Different queries in the same epoch can use different paraphrases (variety)

Algorithm 5: Compute Triplets

Data: D (data collection), Q (queries)
Result: *results* (list of query information)

```
results ← ∅;
for each q ∈ Q do
    P ← {i | D[i].label = q.id};
    N_easy ← ∅;
    N_semi ← ∅;
    N_hard ← ∅;
    for each element d ∈ D do
        if d.label ≠ q.id then
            C ← confidences of d;
            for each pair (k, v) ∈ C do
                if k matches q.id then
                    if v < 0.25 then
                        add index to N_easy;
                    else if v ≥ 0.25 and v < 0.5 then
                        add index to N_semi;
                    else if v ≥ 0.5 then
                        add index to N_hard;
                end
            end
        end
    end
    N ← N_easy ∪ N_semi ∪ N_hard;
    info ← (q.id, q.text, P, N, N_easy, N_semi, N_hard);
    add info to results;
end
return results;
```

Figure 6.7: Algorithm for generating triples for triplet-loss support mining. The pseudo code has been transformed to resolve incompatibility issues with the pseudo codes in other chapters that use specific packages.

- The same query in different epochs uses different paraphrases (complete exposure)

The paraphrases were created manually to preserve semantic meaning, adapting the linguistic register and syntactic structure. For example, for the query "*What is the final decision of the Court of Cassation?*"", paraphrases include variants such as "*How did the Court rule: acceptance, rejection or cassation?*"" and "*Did the Supreme Court accept or reject the appeal?*"".

Edge Dropout (DropEdge)

DropEdge [76] is a regularisation technique for GNNs that prevents oversmoothing and improves generalisation. During training, a random fraction p of the edges are removed from the graph:

$$\tilde{\mathcal{E}} = \{e \in \mathcal{E} \mid \xi_e > p\}, \quad \xi_e \sim \text{Uniform}(0, 1)$$

where $\mathcal{E}$ is the original set of edges and p is the dropout rate (default $p = 0.15$).

This technique forces the model not to depend excessively on individual connections in the document graph, making it more robust to variations in layout structure. During evaluation, all edges are retained.

6.4 Evaluation

6.4.1 Evaluation metrics

We evaluate the proposed retrieval model using a set of standard ranking metrics [74] commonly adopted in information retrieval and question answering tasks. Given a query q and a document represented as a graph with N nodes (paragraphs), the model produces a relevance score $s_i \in \mathbb{R}$ for each node i. The nodes are sorted in descending order of score, obtaining a ranking

$$\pi = (\pi_1, \pi_2, \ldots, \pi_N).$$

Let $R \subseteq \{1, \ldots, N\}$ be the set of ground truth relevant nodes for query q, with $|R|$ relevant fragments.

Cutoff metrics: We calculate Precision@K, Recall@K, and F1@K for $K \in \{1, 3, 5, 10\}$. Precision@K measures the fraction of relevant nodes among the first K results returned:

$$\text{Precision@K} = \frac{|\{\pi_1, \ldots, \pi_K\} \cap R|}{K}.$$

Recall@K quantifies the fraction of relevant nodes retrieved within the first K results:

$$\text{Recall@K} = \frac{|\{\pi_1, \ldots, \pi_K\} \cap R|}{|R|}.$$

F1@K is defined as the harmonic mean of Precision@K and Recall@K:

$$\text{F1@K} = \frac{2 \cdot \text{Precision@K} \cdot \text{Recall@K}}{\text{Precision@K} + \text{Recall@K}}.$$

These metrics reflect the quality of the results at the top of the ranking and are particularly relevant in application scenarios where the user can only inspect a limited number of fragments.

Mean Reciprocal Rank (MRR): To measure how quickly the system retrieves at least one relevant fragment, we calculate the Reciprocal Rank:

$$\text{RR}(q) = \frac{1}{\text{rank}_q},$$

where rank_q indicates the position of the first relevant node in the ranking associated with query q. The Mean Reciprocal Rank is obtained by averaging across all queries:

$$\text{MRR} = \frac{1}{|Q|} \sum_{q \in Q} \text{RR}(q).$$

MRR emphasises the early retrieval of relevant content and is therefore suitable for contexts where quickly identifying at least one correct answer is crucial.

Mean Average Precision (MAP): MAP evaluates the overall quality of the ranking by considering the precision at each position where a relevant node appears. For a single query q, Average Precision is defined as:

$$\text{AP}(q) = \frac{1}{|R|} \sum_{k:\, \pi_k \in R} \text{Precision@k}.$$

Mean Average Precision is therefore calculated as:

$$\text{MAP} = \frac{1}{|Q|} \sum_{q \in Q} \text{AP}(q).$$

MAP rewards systems that place relevant fragments at the top of the ranking and penalises the dispersion of relevant results throughout the sorted list. For this reason, MAP is used as the main metric for model selection.

Normalised Discounted Cumulative Gain (nDCG): To take into account the importance of position in the ranking, we also report the nDCG. The Discounted Cumulative Gain at cutoff K is defined as:

$$\text{DCG@K} = \sum_{i=1}^{K} \frac{\text{rel}(\pi_i)}{\log_2(i+1)},$$

where $\text{rel}(\pi_i) \in \{0, 1\}$ indicates a binary relevance. The DCG is normalised using the Ideal DCG (IDCG), obtained from a perfect ranking:

$$\text{nDCG@K} = \frac{\text{DCG@K}}{\text{IDCG@K}}.$$

The nDCG assigns greater weight to the correct positioning of relevant fragments in the top positions, which is particularly important in legal retrieval scenarios.

Aggregation: All metrics are calculated for each pair (document graph, query) and then averaged over the entire dataset. In addition to global averages, we also report per-query metrics in order to analyse the relative difficulty of different types of legal questions.

6.4.2 Results

In this section, we report the experimental results on the task of paragraph retrieval in legal judgements. Given a document d segmented into N paragraphs (nodes) and a query q, the model produces a ranking of the nodes based on their relevance to q. We evaluate performance using the metrics described in Section 6.4.1. Since the test set consists of only 44 documents, to reduce the effect of stochastic variability, we report the mean and standard deviation over three seeds.

Table 6.1: Global metrics (mean ± std over 3 seeds) for InfoNCE and Triplet loss.

Metrica	**InfoNCE**	**Triplet**
MAP	0.8332 ± 0.0140	$\mathbf{0.8338 \pm 0.0095}$
MRR	0.9108 ± 0.0134	$\mathbf{0.9218 \pm 0.0070}$
nDCG (full)	0.9075 ± 0.0100	$\mathbf{0.9103 \pm 0.0061}$

6.4.3 Comparison of loss functions: InfoNCE vs Triplet

We compare two contrastive learning objectives: *InfoNCE* and *Triplet loss*. Table 6.1 reports the global metrics, while Table 6.2 shows the @K metrics.

On average, the two losses obtain similar values in terms of MAP, while Triplet loss shows a more evident improvement on MRR and nDCG (both at cutoff and on the entire ranking), suggesting a better ability to place at least one relevant fragment in the top positions.

6.4.4 Ablation study

Given that Triplet loss has overall performance comparable to InfoNCE and is more consistent on MRR and nDCG, we use the Triplet loss-based configuration as the reference model in subsequent ablation experiments.

To understand the contribution of the different architectural modules, we conduct an ablation study by removing or replacing specific components: (i) removal of visual features, (ii) removal of the cross-attention-based scorer, (iii) replacement of the graph encoder (GATv2) with an multi layer perceptron (MLP), and (iv) removal of textual features from the nodes. The results are shown in Table 6.3.

Three main observations emerge from the ablation study.

(1) Textual features are crucial. Removing textual features from nodes causes a drastic drop in performance (MAP from ~ 0.834 to ~ 0.605, F1@5 from ~ 0.571 to ~ 0.421). This result indicates that, in our setup, linguistic information is the dominant factor in discriminating the relevance of paragraphs with respect to the query.

(2) The contribution of the visual component is not always positive. Removing visual features results in a slight average improvement (MAP +0.012 compared to the full model). A plausible interpretation is

Table 6.2: Metrics @K (mean ± std over 3 seeds) for InfoNCE and Triplet loss.

Metrica	InfoNCE	Triplet
P@1	0.8502 ± 0.0160	$\mathbf{0.8685 \pm 0.0132}$
R@1	0.2627 ± 0.0177	$\mathbf{0.2677 \pm 0.0155}$
F1@1	0.3530 ± 0.0181	$\mathbf{0.3600 \pm 0.0166}$
nDCG@1	0.8502 ± 0.0160	$\mathbf{0.8685 \pm 0.0132}$
P@3	0.7390 ± 0.0106	0.7337 ± 0.0027
R@3	0.5585 ± 0.0254	0.5531 ± 0.0214
F1@3	0.5467 ± 0.0176	0.5403 ± 0.0098
nDCG@3	0.8500 ± 0.0206	0.8506 ± 0.0119
P@5	0.6288 ± 0.0031	$\mathbf{0.6326 \pm 0.0064}$
R@5	0.7013 ± 0.0192	$\mathbf{0.7064 \pm 0.0146}$
F1@5	0.5669 ± 0.0100	$\mathbf{0.5711 \pm 0.0056}$
nDCG@5	0.8554 ± 0.0168	$\mathbf{0.8616 \pm 0.0092}$
P@10	0.4622 ± 0.0036	0.4608 ± 0.0074
R@10	0.8672 ± 0.0144	0.8678 ± 0.0125
F1@10	0.5181 ± 0.0036	0.5171 ± 0.0053
nDCG@10	0.8782 ± 0.0130	0.8807 ± 0.0066

that RGB crops, while preserving layout details, introduce noise (rendering variations, repeated headers, non-informative elements) or do not provide additional information compared to the text already extracted with OCR.

(3) The advantage of the graph encoder, in this configuration, is limited. Replacing GATv2 with an MLP produces comparable results (very small differences and within the variability between seeds). Similarly, removing the cross-attention scorer does not cause any noticeable deterioration. These results suggest that, with the current graph construction (kNN on bounding boxes + inter-page links) and the current training regime, the graph structure is not adding a sufficiently strong signal beyond the textual representations.

Modello	MAP	MRR	nDCG (full)	F1@5
Full (Triplet)	0.834 ± 0.009	0.922 ± 0.007	0.910 ± 0.006	0.571 ± 0.006
No visual features	0.846 ± 0.013	0.924 ± 0.014	0.917 ± 0.008	0.574 ± 0.004
No cross-attention scorer	0.833 ± 0.004	0.917 ± 0.010	0.909 ± 0.003	0.569 ± 0.006
Replace GATv2 with MLP	0.837 ± 0.005	0.921 ± 0.009	0.912 ± 0.004	0.569 ± 0.004
No text features for nodes	0.605 ± 0.022	0.752 ± 0.031	0.766 ± 0.017	0.421 ± 0.021

Table 6.3: Ablation study (mean± std over 3 seeds).

6.4.5 Per-query analysis

In this section, we analyse the performance of the complete model trained with *Triplet loss* at the single query level. We consider the run shown in Table 6.4 (seed 456), which achieves a total of **MAP = 0.8297**, **MRR = 0.9137**, and **F1@5 = 0.5724**. These values indicate a generally reliable ranking: the high MRR indicates that, in most cases, a relevant paragraph appears in the top positions.

Queries with the best performance: Queries **q0** (parties involved) and **q2** (final decision) show the highest performance (**MAP** 0.8941 and 0.8955, respectively; **F1@5** 0.7151 and 0.8037). A plausible interpretation is that this information is often expressed in portions of the judgement with relatively recurring formulations and strong semantic consistency, making it easier to align queries and paragraphs in the latent space. **q5** (indication of a ground for appeal) also achieves very solid results (**MAP** 0.9294), with very high precision (P@5 = 0.9666): when the model retrieves the correct paragraphs, it tends to do so with few false positives among the top results.

Queries of intermediate difficulty: Queries **q1** (citations of precedents) and **q3** (legal costs) fall in the intermediate range (MAP 0.7864 and 0.7794). In particular, q3 has high **Recall@5** (0.7333) but lower precision (0.6931), suggesting that information relating to costs may appear in multiple places or with partially ambiguous textual signals, causing the model to include "similar" but not always correct candidates.

More difficult queries: Query **q6** (principle of law / maxim) is the most complex (**MAP** 0.6713, **F1@5** 0.4476), followed by **q4** (cited articles of law), which, while maintaining a decent MAP (0.8168), shows limited **recall** (R@5

Query	P@5	R@5	F1@5	nDCG	MAP
q0	0.94 ± 0.03	0.63 ± 0.05	0.74 ± 0.04	0.96 ± 0.01	0.92 ± 0.02
q1	0.76 ± 0.03	0.58 ± 0.04	0.63 ± 0.03	0.90 ± 0.02	0.82 ± 0.02
q2	0.91 ± 0.02	0.72 ± 0.05	0.79 ± 0.03	0.96 ± 0.01	0.89 ± 0.01
q3	0.78 ± 0.08	0.77 ± 0.04	0.74 ± 0.06	0.90 ± 0.05	0.84 ± 0.05
q4	0.88 ± 0.03	0.40 ± 0.02	0.55 ± 0.02	0.90 ± 0.02	0.79 ± 0.02
q5	0.92 ± 0.04	0.58 ± 0.06	0.69 ± 0.04	0.96 ± 0.02	0.89 ± 0.03
q6	0.61 ± 0.03	0.42 ± 0.04	0.45 ± 0.04	0.79 ± 0.02	0.66 ± 0.02

Table 6.4: Per-query performance of the full model with Triplet loss, reported as mean and standard deviation over three seeds (36, 123, 456).

= 0.4063). This difficulty is consistent with the nature of the content: (i) the "maxim" can be expressed in highly variable formulations and is not always marked by stable lexical patterns; (ii) regulatory references can be numerous and distributed, sometimes appearing in lists or argumentative passages where the association with the query is less direct. In other words, the model can identify semantically similar paragraphs, but not necessarily those that maximise the correspondence with the operational definition adopted in the annotations.

Visualisation of the embedding space (t-SNE): To qualitatively inspect the structure of the latent space, in Fig. 6.8 we show a t-SNE [53] projection of the embeddings in the test set, colouring each point according to the query. The visualisation shows a tendency for *clusters* to form for some queries (particularly those with more specific and repetitive signals), while other categories show greater overlap. This overlap is consistent with the difficulty observed in the metrics: when the semantics of the query are "diffuse" or less standardised, the relevant paragraphs may be located in regions of the space closer to general argumentative content, reducing the separability between classes. We emphasise that t-SNE is an exploratory tool (sensitive to parameters and not directly interpretable in terms of global distance), but it is useful for highlighting qualitative patterns consistent with per-query performance.

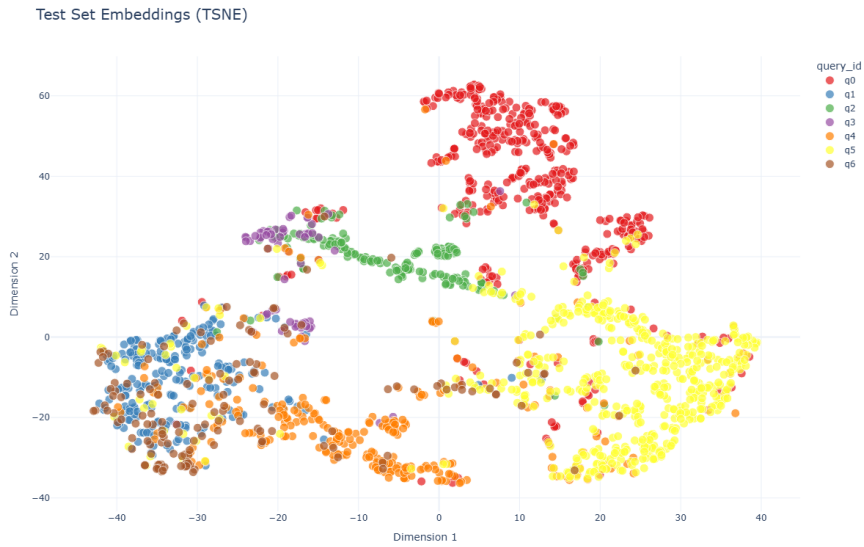


Figure 6.8: t-SNE visualisation of the embedding space on the test set for the complete model with Triplet loss. Each point represents a paragraph of a document, coloured according to the query. Distinct clusters can be observed for some queries (e.g. q0, q2, q5), while overlaps between other groups reflect greater semantic ambiguity.

6.5 Conclusion

In this work, we studied the problem of paragraph retrieval in Italian Supreme Court rulings, showing that it is possible to formulate the task as a graph retrieval problem, in which the paragraphs of the document are modelled as nodes connected according to structural relationships derived from the layout.

We proposed a multimodal model that integrates textual, visual, and structural representations, trained with contrastive learning. The experimental results show stable and competitive performance on all ranking metrics, despite the small size of the test set, with MAP values around 0.83–0.84 and MRR values above 0.91. The comparison between InfoNCE and Triplet loss shows similar overall performance, with a slight advantage for Triplet loss in terms of MRR and nDCG, suggesting a better ability to quickly po-

sition relevant fragments.

The ablation study highlights that textual information is the dominant signal for the task: removing textual features causes a significant degradation in performance. In contrast, the contribution of the visual component and the graph encoder is limited in the current configuration, with only minor differences compared to simpler variants. This result indicates that, although graph modelling is conceptually appropriate for multi-page legal documents, the effectiveness of GNNs critically depends on the quality of the structural relationships encoded in the graph.

Overall, the work demonstrates the feasibility of a neural and structured approach to legal information retrieval and provides a basis for future developments, where large collections of richer and semantically informed graphs can fully leverage the potential of Graph Neural Networks in the legal field.

Chapter 7

Conclusion

This doctoral thesis addressed the problem of document analysis in complex real-world scenarios, with particular reference to medical-legal and judicial documents. Integrating computer vision, natural language processing and multimodal learning techniques, the work proposed a comprehensive view of Document AI systems covering the entire processing pipeline, from document restoration to semantic understanding and information retrieval.

A central contribution of the thesis is the development and evaluation of end-to-end pipelines for document analysis, designed for operational contexts characterised by heterogeneous and visually degraded data. Unlike many previous works, which were mainly evaluated on clean or synthetic benchmarks, the proposed approaches have been validated on real medical-legal and judicial documents. The experimental results demonstrate that the combination of document layout analysis modules and vision-language models finely tuned for OCR allows the extraction of reliable and automatically readable representations, which are fundamental for subsequent applications such as structured archiving, legal analysis, and decision support systems.

In terms of image processing, the thesis introduced DocWaveDiff, a predict-and-refine framework for document image restoration based on U-Net architectures enriched with wavelet representations. The results obtained on inpainting and deblurring benchmarks show that the use of wavelets is particularly effective in preserving both textual details and the overall structure of the document. It is particularly relevant that improvements in the perceptual quality of images translate into increased OCR performance, reinforcing the role of document restoration as an enabling component of the

entire pipeline and not as a simple pre-processing step.

At the semantic level, the thesis addressed the large-scale thematic analysis of legal texts through the construction of a dataset compliant with privacy protection requirements, based on the rulings of the Supreme Court of Cassation, and through the design of a layout-aware processing pipeline. The results confirm that accurate structural segmentation significantly improves the consistency and diversity of the extracted topics. Furthermore, the evaluation of Large Language Models for the generation of summaries and topic labels highlights their potential as tools to support legal interpretation, while also highlighting some limitations related to domain adaptation and compliance with the instructions provided.

Finally, the thesis explored the problem of fine-grained information retrieval within large legal documents, formulating paragraph retrieval as a multimodal graph retrieval problem. The proposed approach integrates textual, visual, and structural information and achieves competitive performance despite the limited size of the training dataset. Ablation analysis shows that textual information is the dominant signal for retrieval performance, while the contribution of visual and structural components strongly depends on the quality of the structural relations encoded in the document graph. Overall, this work supports a unified view of document analysis as a multi-level problem, in which improvements at each stage of restoration, layout understanding, text extraction, semantic modelling and information retrieval jointly contribute to the realisation of document analysis systems for real-world scenarios.

7.1 Limitations

Despite the contributions presented, this work has several limitations that should be acknowledged.

Regarding the medico-legal document analysis pipeline (Chapter 3), the dataset used for training consists of 60 real cases (809 pages) from a single hospital (Careggi University Hospital). While this reflects genuine operational conditions, the limited size and single-institution provenance may reduce the generalisability of the trained models to other healthcare facilities, where document templates, scanning devices, and degradation patterns may differ substantially, as each hospital has its own standards for documentation. Furthermore, the Document Layout Analysis module, despite achieving

solid performance on core classes such as Text and List-item, showed lower accuracy on less frequent categories, suggesting that class imbalance remains an open issue.

Concerning DocWaveDiff (Chapter 4), the proposed framework was evaluated exclusively on two established benchmarks (EnsExam for inpainting and TDD for deblurring). Although the results are competitive, the evaluation does not cover other common types of document degradation, such as shadowing or warping. The perception-distortion trade-off observed in the ablation study also suggests that jointly optimising for perceptual quality and pixel-level fidelity remains a challenge.

For the Italian Supreme Court topic modeling pipeline (Chapter 5), the GLiNER anonymisation module was employed in zero-shot mode due to time constraints, which occasionally produced false negatives. Although the thresholding-based approach proved effective in most cases, incomplete anonymisation represents a risk in privacy-sensitive legal contexts, although the judgments issued by the Italian Supreme Court are final and issued in the name of the Italian people and therefore public. Additionally, the evaluation of Large Language Models for summary and label generation revealed that smaller models (e.g., LLaMA 3.1 8B) struggled with instruction following, sometimes generating summaries instead of labels, indicating that the reliability of LLM-based annotation tools varies significantly with model size and architecture.

Finally, regarding the multimodal graph-based paragraph retrieval model (Chapter 6), the ablation study revealed that the visual and structural components provided only marginal improvements over text-only baselines in the current configuration. This suggests that the graph construction strategy, which relies on layout-derived structural relations, may not yet capture sufficiently rich semantic connections between paragraphs. Moreover, the test set was relatively small, which limits the statistical robustness of the reported results and makes it difficult to draw definitive conclusions about the relative importance of each modality.

7.2 Future work

The findings and limitations discussed above open several promising directions for future research.

A natural extension of the medical-legal pipeline would involve expand-

ing the training dataset by integrating documents from multiple hospitals and healthcare institutions, thereby improving the diversity of document templates and degradation conditions. However, although this is an obvious goal, it may not be easy to achieve as it requires agreements between different parties. Exploring domain adaptation and few-shot learning techniques could further reduce the annotation burden when deploying the system in new operational contexts. Furthermore, integrating the pipeline with downstream decision support tools, such as automated claim categorization or risk assessment forms, would increase its practical utility.

For DocWaveDiff, future work could extend the framework to handle a broader range of degradation types beyond inpainting and deblurring, including binarisation, ink bleed-through removal, and geometric distortion correction.

Regarding the legal topic modeling pipeline, a key improvement would be the fine-tuning of the GLiNER anonymisation module on a dedicated dataset of Italian legal documents to reduce false negatives and strengthen privacy compliance. The dataset could also be expanded by collecting additional document types from the Italian Supreme Court, such as ordinances and decrees, to provide a more comprehensive view of the Italian legal landscape. Moreover, fine-tuning open-source LLMs on domain-specific legal corpora could improve their reliability for summary generation and topic labelling, particularly for smaller models that currently struggle with instruction adherence.

For the graph-based paragraph retrieval model, future developments should focus on enriching the graph structure with semantically informed edges, for instance by exploiting cross-reference links, citation networks, or rhetorical relations between paragraphs, rather than relying solely on layout proximity. Training on larger collections of annotated legal documents would improve the statistical significance of the results and allow the model to better leverage multimodal signals. Investigating pre-training strategies on large-scale legal corpora, combined with contrastive or self-supervised objectives, could also help the model learn more transferable representations.

More broadly, the integration of all the proposed components into a unified, end-to-end Document AI system represents a long-term goal. Such a system would seamlessly combine document restoration, layout analysis, text extraction, semantic understanding, and information retrieval within a single framework, enabling fully automated processing of complex institutional

documents. Exploring cross-domain transfer, where models trained on legal documents are adapted to medical or administrative domains, is another direction that could amplify the impact of this research.

7.3 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.

Appendix A

Publications

International Journals

Submitted

1. **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.org*, 2025. **(Submitted to Artificial intelligence and law and waiting for revision)**

International Conferences and Workshops

1. **M. Marulli**, M. Bertini. "DocWaveDiff: A Predict-and-Refine approach for Document Image Enhancement with Wavelet U-Nets and Diffusion models", in *IEEE/CVF Winter Conference on Applications of Computer Vision 2026*, Tucson (USA), 2026. **(Paper accepted in conference)**

National Conferences

1. **M. Marulli**, V. Pinchi, S. Grassi, M. Focardi, M. Bertini. "An AI-based system for medico-legal document analysis", in *Italian Research Conference on Digital Libraries (IRCDL)*, Modena, Italy, 2026. **(Paper accepted in conference)**

Bibliography

[1] A. Abdelrazek, Y. Eid, E. Gawish, W. Medhat, and A. Hassan, "Topic modeling algorithms and applications: A survey," *Information Systems*, vol. 112, p. 102131, 2023.

[2] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[3] L. Aitchison and S. Ganev, "Infonce is variational inference in a recognition parameterised model," *arXiv preprint arXiv:2107.02495*, 2021.

[4] A. Ali and S. Renals, "Word error rate estimation for speech recognition: e-wer," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics (ACL), 2018, pp. 20–24.

[5] Anthropic, "Claude 3 family announcement," https://www.anthropic.com/news/claude-3-family, ultimo accesso: 18 febbraio 2025.

[6] Y. Baek, B. Lee, D. Han, S. Yun, and 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. [Online]. Available: https://arxiv.org/abs/1904.01941

[7] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[8] 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.

[9] I. M. Bilal, B. Wang, M. Liakata, R. Procter, and A. Tsakalidis, "Evaluation of thematic coherence in microblogs," *arXiv preprint arXiv:2106.15971*, 2021.

[10] G. M. Binmakhashen and S. A. Mahmoud, "Document layout analysis: a comprehensive survey," *ACM Computing Surveys (CSUR)*, vol. 52, no. 6, pp. 1–36, 2019.

[11] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6228–6237.

[12] G. Cicchetti and D. Comminiello, "Naf-dpm: A nonlinear activation-free diffusion probabilistic model for document enhancement," *arXiv preprint arXiv:2404.05669*, 2024.

[13] B. Clavié and M. Alphonsus, "The unreasonable effectiveness of the baseline: discussing svms in legal text classification," in *Legal Knowledge and Information Systems*. IOS Press, 2021, pp. 58–61.

[14] A. Cloud, "Qwen3-vl: The next generation vision-language model," https://qwen.ai/blog?id=99f0335c4ad9ff6153e517418d48535ab6d8afef, 2025.

[15] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 45, no. 9, pp. 10 850–10 869, 2023.

[16] DeepMount00, "GLiNER PII ITA," https://huggingface.co/DeepMount00/GLiNER_PII_ITA, ultimo accesso: 18 febbraio 2025.

[17] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[18] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "Topic modeling in embedding spaces," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 439–453, 2020. [Online]. Available: https://aclanthology.org/2020.tacl-1.29/

[19] X. Dong and J. Shen, "Triplet loss in siamese network for object tracking," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 459–474.

[20] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[21] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[22] B. Ghojogh, A. Ghodsi, F. Karray, and M. Crowley, "Uniform manifold approximation and projection (umap) and its variants: tutorial and survey," *arXiv preprint arXiv:2109.02508*, 2021.

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

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[25] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

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

[27] T. Guo, H. Seyed Mousavi, T. Huu Vu, and V. Monga, "Deep wavelet prediction for image super-resolution," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)*, 2017, pp. 104–113.

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

[29] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, R. Tibshirani, and J. Friedman, "Overview of supervised learning," *The elements of statistical learning: Data mining, inference, and prediction*, pp. 9–41, 2009.

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

[31] HealthMemo AI, "Healthmemo: Intelligent processing of medical and insurance documents," https://www.healthmemo.ai/, 2025, accessed November 2025.

[32] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *2010 20th international conference on pattern recognition*. IEEE, 2010, pp. 2366–2369.

[33] M. Hradiš, J. Kotera, P. Zemcık, and F. Šroubek, "Convolutional neural networks for direct text deblurring," in *Proc. of the British Machine Vision Conference (BMVC)*, vol. 10, no. 2, 2015.

[34] E. Hsu, I. Malagaris, Y.-F. Kuo, R. Sultana, and 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.

[35] 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*, vol. 1, no. 2, p. 3, 2022.

[36] L. Huang, B. Chen, C. Liu, D. Peng, W. Zhou, Y. Wu, H. Li, H. Ni, and L. Jin, "Ensexam: a dataset for handwritten text erasure on examination papers," in *Proc. of the International Conference on Document Analysis and Recognition (ICDAR)*. Springer, 2023, pp. 470–485.

[37] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[38] W. Inc., "Wisedocs: Ai-powered medical document review platform," 2025, toronto, Canada. Accessed November 2025. [Online]. Available: https://www.wisedocs.ai/

[39] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[40] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "Deblurgan: Blind motion deblurring using conditional adversarial networks," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8183–8192.

[41] M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei, "Trocr: Transformer-based optical character recognition with pre-trained models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 13 094–13 102.

[42] Z. Lian and H. Wang, "An image deblurring method using improved u-net model based on multilayer fusion and attention mechanism," *Scientific Reports*, vol. 13, no. 1, p. 21402, 2023.

[43] M. Liao, Z. Zou, Z. Wan, C. Yao, and X. Bai, "Real-time scene text detection with differentiable binarization and adaptive scale fusion," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 919–931, 2022.

[44] D. Licari and G. Comandè, "Italian-legal-bert models for improving natural language processing tasks in the italian legal domain," *Computer Law & Security Review*, vol. 52, p. 105908, 2024.

[45] D. Licari, M. F. Romano, G. Comandé *et al.*, "Automatic anonymization of italian legal textual documents using deep learning," in *JADT 2022 Proceedings of the 16th International Conference on Statistical Analysis of Textual Data*, vol. 2. 2022 Vadistat press in coedizione Edizioni Erranti, 2022, pp. 552–557.

[46] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[47] C. Liu, L. Jin, Y. Liu, C. Luo, B. Chen, F. Guo, and K. Ding, "Don't forget me: accurate background recovery for text removal via modeling local-global context," in *Proc. of the European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 409–426.

[48] C. Liu, Y. Liu, L. Jin, S. Zhang, C. Luo, and Y. Wang, "Erasenet: End-to-end text removal in the wild," *IEEE Transactions on Image Processing*, vol. 29, pp. 8760–8775, 2020.

[49] P. Liu, H. Zhang, W. Lian, and W. Zuo, "Multi-level wavelet convolutional neural networks," *IEEE Access*, vol. 7, pp. 74 973–74 985, 2019.

[50] Y. Liu, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, vol. 364, 2019.

[51] S. Long, X. He, and C. Yao, "Scene text detection and recognition: The deep learning era," *International Journal of Computer Vision*, vol. 129, no. 1, pp. 161–184, 2021.

[52] M.-W. Ma, X.-S. Gao, Z.-Y. Zhang, S.-Y. Shang, L. Jin, P.-L. Liu, F. Lv, W. Ni, Y.-C. Han, and H. Zong, "Extracting laboratory test information from paper-based reports," *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, p. 251, 2023.

[53] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[54] J. Martinez-Gil, "A survey on legal question–answering systems," *Computer Science Review*, vol. 48, p. 100552, 2023.

[55] M. Marulli, Matteo; Bertini, "Docwavediff: A predict-and-refine approach for document image enhancement with wavelet u-nets and diffusion models," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2016, pp. 8511–8520.

[56] M. Marulli, G. Panattoni, and 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.

[57] M. Marulli, V. Pinchi, S. Grassi, M. Foccardi, and M. Bertini, "An ai-based system for medico-legal document analysis," in *Proceedings of IRCDL 2025*, 2026.

[58] K. T. Maxwell and B. Schafer, "Concept and context in legal information retrieval," in *Legal Knowledge and Information Systems*. IOS Press, 2008, pp. 63–72.

[59] L. McInnes, J. Healy, S. Astels *et al.*, "hdbscan: Hierarchical density based clustering." *J. Open Source Softw.*, vol. 2, no. 11, p. 205, 2017.

[60] W. McKinney *et al.*, "pandas: a foundational python library for data analysis and statistics," *Python for high performance and scientific computing*, vol. 14, no. 9, pp. 1–9, 2011.

[61] Mistral AI, "Mistral ocr: High-performance multilingual text recognition model," https://mistral.ai/news/mistral-ocr/, 2025, accessed November 2025.

[62] R. Mittal and A. Garg, "Text extraction using ocr: a systematic review," in *2020 second international conference on inventive research in computing applications (ICIRCA)*. IEEE, 2020, pp. 357–362.

[63] C. Neudecker, K. Baierer, M. Gerber, C. Clausner, A. Antonacopoulos, and S. Pletschacher, "A survey of ocr evaluation tools and metrics," in *Proceedings of the 6th International Workshop on Historical Document Imaging and Processing*, 2021, pp. 13–18.

[64] nickprock, "sentence-bert-base-italian-uncased," https://huggingface.co/nickprock/sentence-bert-base-italian-uncased, ultimo accesso: 18 febbraio 2025.

[65] J. Nilsson and T. Akenine-Möller, "Understanding ssim," *arXiv preprint arXiv:2006.13846*, 2020.

[66] R. Padilla, S. L. Netto, and E. A. B. da Silva, "A survey on performance metrics for object-detection algorithms," in *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2020, pp. 237–242.

[67] C. Park, H. Kang, and T. Hain, "Character error rate estimation for automatic speech recognition of short utterances," in *2024 32nd European Signal Processing Conference (EUSIPCO)*, 2024, pp. 131–135.

[68] B. Pfitzmann, C. Auer, M. Dolfi, A. S. Nassar, and 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.

[69] P. Prince-Tritto and H. Ponce, "Exploring the challenges and limitations of unsupervised machine learning approaches in legal concepts discovery," in *Mexican International Conference on Artificial Intelligence*. Springer, 2023, pp. 52–67.

[70] O. Procopiuc, P. K. Agarwal, L. Arge, and J. S. Vitter, "Bkd-tree: A dynamic scalable kd-tree," in *International symposium on spatial and temporal databases*. Springer, 2003, pp. 46–65.

[71] W. A. Qader, M. M. Ameen, and B. I. Ahmed, "An overview of bag of words; importance, implementation, applications, and challenges," in *2019 international engineering conference (IEC)*. IEEE, 2019, pp. 200–204.

[72] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[73] J. Redmon, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[74] P. Riba, A. Molina, L. Gomez, O. Ramos-Terrades, and J. Lladós, "Learning to rank words: Optimizing ranking metrics for word spotting," in *International Conference on Document Analysis and Recognition*. Springer, 2021, pp. 381–395.

[75] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the eighth ACM international conference on Web search and data mining*, 2015, pp. 399–408.

[76] Y. Rong, W. Huang, T. Xu, and J. Huang, "Dropedge: Towards deep graph convolutional networks on node classification," *arXiv preprint arXiv:1907.10903*, 2019.

[77] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[78] A. Roshdi and A. Roohparvar, "Information retrieval techniques and applications," *International Journal of Computer Networks and Communications Security*, vol. 3, no. 9, pp. 373–377, 2015.

[79] R. Rothe, M. Guillaumin, and 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.

[80] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 45, no. 4, pp. 4713–4726, 2022.

[81] G. Salierno, R. Bertè, L. Attias, C. Morrone, D. Pettazzoni, and D. Battisti, "Giusberto: A legal language model for personal data de-identification in italian court of auditors decisions," *arXiv preprint arXiv:2406.15032*, 2024.

[82] J. Šavelka and K. D. Ashley, "Legal information retrieval for understanding statutory terms," *Artificial Intelligence and Law*, vol. 30, no. 2, pp. 245–289, 2022.

[83] S. Shang, Z. Shan, G. Liu, L. Wang, X. Wang, Z. Zhang, and J. Zhang, "Resdiff: Combining cnn and diffusion model for image super-resolution," in *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 8, 2024, pp. 8975–8983.

[84] R. Smith, "An overview of the tesseract ocr engine," in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, vol. 2, 2007, pp. 629–633.

[85] S. Solorio-Fernández, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, "A review of unsupervised feature selection methods," *Artificial Intelligence Review*, vol. 53, no. 2, pp. 907–948, 2020.

[86] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[87] M. A. Souibgui, S. Biswas, S. K. Jemni, Y. Kessentini, A. Fornés, J. Lladós, and U. Pal, "Docentr: An end-to-end document image enhancement transformer," in *Proc. of the International Conference on Pattern Recognition (ICPR)*. IEEE, 2022, pp. 1699–1705.

[88] M. A. Souibgui, S. Biswas, A. Mafla, A. F. Biten, A. Fornés, Y. Kessentini, J. Lladós, L. Gomez, and D. Karatzas, "Text-diae: a self-supervised degradation invariant autoencoder for text recognition and document enhancement," in *proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 2, 2023, pp. 2330–2338.

[89] M. A. Souibgui and Y. Kessentini, "De-gan: A conditional generative adversarial network for document enhancement," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 44, no. 3, pp. 1180–1191, 2020.

[90] 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*, vol. 7, no. 1, pp. 154–164, 2025.

[91] G. Strang, "Wavelet transforms versus fourier transforms," *Bulletin of the American Mathematical Society*, vol. 28, no. 2, pp. 288–305, 1993.

[92] L. Studio, "Label Studio – Open Source Data Labeling Tool," https://labelstud.io, ultimo accesso: 18 febbraio 2025.

[93] N. Subramani, A. Matton, M. Greaves, and A. Lam, "A survey of deep learning approaches for ocr and document understanding," *arXiv preprint arXiv:2011.13534*, 2020.

[94] S. Thammaboosadee, B. Watanapa, and N. Charoenkitkarn, "A framework of multi-stage classifier for identifying criminal law sentences," *Procedia Computer Science*, vol. 13, pp. 53–59, 2012.

[95] O. Tursun, S. Denman, R. Zeng, S. Sivapalan, S. Sridharan, and C. Fookes, "Mtrnet++: One-stage mask-based scene text eraser," *Computer Vision and Image Understanding*, vol. 201, p. 103066, 2020.

[96] O. Tursun, R. Zeng, S. Denman, S. Sivapalan, S. Sridharan, and C. Fookes, "Mtrnet: A generic scene text eraser," in *Proc. of the International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019, pp. 39–44.

[97] Ultralytics, "Yolov8: Real-time object detection and image segmentation," https://docs.ultralytics.com/, 2023, accessed November 2025.

[98] S. Undavia, A. Meyers, and J. E. Ortega, "A comparative study of classifying legal documents with neural networks," in *2018 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2018, pp. 515–522.

[99] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[100] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, vol. 10, no. 3152676, pp. 10–5555, 2017.

[101] 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.

[102] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The thrity-seventh asilomar conference on signals, systems & computers, 2003*, vol. 2. Ieee, 2003, pp. 1398–1402.

[103] J. Whang, M. Delbracio, H. Talebi, C. Saharia, A. G. Dimakis, and P. Milanfar, "Deblurring via stochastic refinement," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16 293–16 303.

[104] X. Wu, C. Li, Y. Zhu, and Y. Miao, "Short text topic modeling with topic distribution quantization and negative sampling decoder," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 1772–1782. [Online]. Available: https://aclanthology.org/2020.emnlp-main.138/

[105] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[106] B. Xiao, H. Wu, W. Xu, X. Dai, H. Hu, Y. Lu, M. Zeng, C. Liu, and 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.

[107] G. Xu, W. Liao, X. Zhang, C. Li, X. He, and X. Wu, "Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation," *Pattern recognition*, vol. 143, p. 109819, 2023.

[108] H.-H. Yang and Y. Fu, "Wavelet u-net and the chromatic adaptation transform for single image dehazing," in *Proc. of the IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 2736–2740.

[109] J. Yang and N. I. R. Ruhaiyem, "Review of deep learning-based image inpainting techniques," *IEEE Access*, 2024.

[110] Z. Yang, B. Liu, Y. Xxiong, L. Yi, G. Wu, X. Tang, Z. Liu, J. Zhou, and X. Zhang, "Docdiff: Document enhancement via residual diffusion models," in *Proc. of the ACM International Conference on Multimedia (ACM MM)*, 2023, pp. 2795–2806.

[111] U. Zaratiana, N. Tomeh, P. Holat, and T. Charnois, "Gliner: Generalist model for named entity recognition using bidirectional transformer," 2023.

[112] D. Zhang, "Wavelet transform," in *Fundamentals of image data mining: Analysis, Features, Classification and Retrieval*. Springer, 2019, pp. 35–44.

[113] J. Zhang, D. Peng, C. Liu, P. Zhang, and L. Jin, "Docres: a generalist model toward unifying document image restoration tasks," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 15 654–15 664.

[114] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[115] S. Zhang, Y. Liu, L. Jin, Y. Huang, and S. Lai, "Ensnet: Ensconce text in the wild," in *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 801–808.

[116] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[117] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[118] X. Zhong, J. Tang, and 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.

[119] Y. Zhou, S. Zuo, Z. Yang, J. He, J. Shi, and R. Zhang, "A review of document image enhancement based on document degradation problem," *Applied Sciences*, vol. 13, no. 13, p. 7855, 2023.

La borsa di dottorato è stata finanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4 – Componente 2 – "Dalla ricerca all'impresa" –
Investimento THE Tuscany Health Ecosystem" – *CUP B83C22003920001*
(inserire investimento e CUP appropriati)