Università di Genova

PhD Program in Science and Technology for Electronic and Telecommunication Engineering

Personalized Medicine and Process Optimization: Analysis and Implementation of Intelligent Tools to Support the Clinical Process

Alaa WEHBE

Thesis submitted for the degree of *Doctor of Philosophy* (37° cycle)

March 2026

Prof. Silvana Dellepiane — Tutor
Prof. Irene Minetti — Co-tutor
Prof. Maurizio Valle — Head of the PhD program

Università di Genova
DITEN DIPARTIMENTO DI INGEGNERIA NAVALE, ELETTRICA, ELETTRONICA E DELLE TELECOMUNICAZIONI

Alaa WEHBE

Personalized Medicine and Process Optimization: Analysis and Implementation of Intelligent Tools to Support the Clinical Process

xv+108 p.

I would like to dedicate this thesis to:
My family — especially my mother and father — whose sacrifices, love, and unwavering encouragement have guided me to reach this milestone.
My brothers, for their constant help and unconditional support through every step of this journey.
My friends and all those who accompanied me along the way, sharing both the challenges and the beautiful moments.
And finally, to the special people who added beauty, depth, and meaning to this experience — who helped me grow, understand myself better, and see life in a new light.
This PhD has been more than an academic path; it has been a journey of self-discovery, revealing hidden parts of who I am and teaching me the true essence of balance and perseverance.
It can be best captured in the words of Rumi:
"The universe turns differently when fire loves water."

Acknowledgements

I would like to express my deepest gratitude to my supervisor, **Prof. Silvana Dellapiane**, for her invaluable guidance, constant encouragement, and insightful discussions throughout the course of my doctoral research. Her mentorship has been instrumental in shaping my scientific thinking and professional growth.

My sincere thanks go to **Prof. Irene Minetti** and **Prof. Valle** for their support, constructive feedback, and collaboration, which enriched both the technical depth and clarity of this work. I am also grateful to the members of the **RAISE project**, funded by the European Union–NextGenerationEU, for providing a stimulating research environment and the opportunity to contribute to impactful interdisciplinary initiatives.

A special acknowledgment goes to my colleagues and friends from the **University of Genoa** and the **AGH University of Cracow**, whose collaboration, discussions, and shared challenges have made this journey both intellectually rewarding and personally fulfilling.

I also wish to thank the **Welcome Office of the University of Genoa** for their support and assistance during my stay in Italy.

Finally, I am profoundly grateful to my family and loved ones for their endless patience, understanding, and belief in me. Their unconditional support has been my greatest source of strength and motivation throughout this journey.

Abstract

Artificial Intelligence (AI) has profoundly transformed medical imaging, enabling new frontiers in disease detection, staging, and diagnostic decision support. Yet, despite remarkable progress, most AI tools remain siloed—disconnected from Picture Archiving and Communication Systems (PACS) and Radiology Information Systems (RIS)—and often constrained by closed vocabularies that limit interpretability and clinical integration. This thesis addresses these challenges by designing a sequence of retrieval-driven, explainable AI frameworks that progressively bridge detection, retrieval, and multimodal reasoning for medical imaging, with a specific focus on lung cancer and thoracic radiography.

In the first stage, lung cancer is adopted as a case study for closed-vocabulary analysis. A 3D volumetric pipeline was developed using the YOLOv8 architecture for simultaneous detection and subtype classification of Adenocarcinoma (ADC), Squamous Cell Carcinoma (SCC), and Small Cell Lung Cancer (SCLC) on CT scans. Experimental results demonstrated a mean Average Precision (mAP) of **97.1%**, with the YOLOv8-Small variant achieving **96.1% precision**, a recall of **0.91**, and a detection speed of **0.22 seconds**—surpassing two-stage models such as Faster R-CNN and YOLOv7. The extracted features were subsequently used to train a custom **TNMClassifier**, achieving an overall accuracy of **98%** in staging classification. This integration of subtype and stage detection established a strong foundation for clinically relevant representation learning.

Building upon these embeddings, the second stage introduced a **Content-Based Image Medical Retrieval (CBIMR)** system that leverages YOLOv8-derived features to retrieve clinically similar cases across cancer groups and TNM stages. The retrieval framework achieved a **precision of 0.961**, **recall of 0.945**, and **mAP@0.5 of 0.971**, effectively linking detection and retrieval pipelines. This detection-driven CBIR model demonstrated how region-aware embeddings can enhance interpretability, consistency, and case-based reasoning—key prerequisites for integration into PACS/RIS environments.

The final stage transitions toward open-vocabulary and multimodal learning through the introduction of **MedFL** (Medical Florence), a unified framework for medical image retrieval and report generation. Built upon the Florence-2 Vision–Language Model (VLM), MedFL employs parameter-efficient fine-tuning via LoRA on the MIMIC-CXR dataset and introduces a novel prompt-conditioned feature fusion strategy combining three complementary representations: `CAPTION`, `DETAILED_CAPTION`, and `OD`. The fused 2304-dimensional embeddings enable cross-modal retrieval and automated report generation within a unified space. Extensive experiments demonstrate that MedFL achieves **Recall@10 = 0.930** for retrieval and **BLEU-4 = 0.3363** for report generation, outperforming strong baselines such as CLIP, BioMedCLIP, and GatorTron-CLIP. Furthermore, MedFL attains **mAP = 0.35 @0.5 IoU** on VinDr-CXR, confirming robust localization and visual–textual alignment.

Collectively, this thesis advances the paradigm of AI in medical imaging from detection to understanding—evolving from closed-vocabulary CT-based detection to open-vocabulary, multimodal reasoning in radiology. By integrating visual and textual representations into a unified, explainable, and retrieval-driven architecture, the proposed

frameworks lay the groundwork for clinically deployable systems that can assist physicians in case retrieval, diagnostic interpretation, and automated reporting, ultimately supporting more transparent, personalized, and efficient medical decision-making.

Keywords: Medical imaging, Content-Based Image Retrieval (CBIR), YOLOv8, TNM staging, Vision–Language Models, MedFL, Multimodal learning, Florence-2, Explainable AI, Radiology.

Contents

List of Figures

List of Tables

Notation

I_a, I_b	Input images a and b used for similarity computation
$f(I)$	Feature embedding of image I
$d(I_a, I_b)$	Distance measure between embeddings of I_a and I_b
$\text{sim}(I_a, I_b)$	Cosine similarity between embeddings of I_a and I_b
Q, K, V	Query, Key, and Value matrices in Transformer self-attention
W_Q, W_K, W_V	Learned weight matrices for Q, K, V projections
d_k	Dimensionality of key vectors in attention mechanism
$\text{Atten}(Q, K, V)$	Scaled dot-product attention function
$\mathbf{y}^T$	Detection output vector containing confidence, bounding box, and class probabilities
P_c	Confidence score of an object detection prediction
(b_x, b_y, b_h, b_w)	Bounding box parameters (center, height, width)
c_i	Class probability for category i in detection tasks
IoU	Intersection over Union, ratio of overlap to union of predicted and ground-truth boxes
TP, FP, FN	True Positives, False Positives, False Negatives
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
$F1$	Harmonic mean of precision and recall
mAP	Mean Average Precision, averaged across classes and IoU thresholds
$\mathcal{L}$	Loss function (e.g., negative log-likelihood, contrastive loss)
$\mathcal{L}_{\text{contrastive}}$	Contrastive loss for aligning image–text embeddings
v_i, t_i	Image and text embeddings for sample i
τ	Temperature parameter in contrastive learning
Recall@K	Fraction of relevant items among the top-K retrieved results
BLEU-n	n-gram overlap metric for evaluating generated text against references

Acronyms

AI	Artificial Intelligence
ADC	Adenocarcinoma
CAD	Computer-Aided Detection
CBIR	Content-Based Image Retrieval
CBMIR	Content-Based Medical Image Retrieval
CLIP	Contrastive Language–Image Pretraining
CNN	Convolutional Neural Network
CT	Computed Tomography
FN	False Negative
FP	False Positive
FPS	Frames Per Second
HOG	Histogram of Oriented Gradients
HVP	Hanging Viewer Protocol
ILD	Interstitial Lung Disease
IoU	Intersection over Union
LBP	Local Binary Patterns
LLM	Large Language Model
mAP	Mean Average Precision
ML	Machine Learning
MLM	Masked Language Modeling
MRI	Magnetic Resonance Imaging
NSIP	Nonspecific Interstitial Pneumonia
OVOD	Open-Vocabulary Object Detection
PACS	Picture Archiving and Communication System
PBR	Pathology-Bearing Region
PET	Positron Emission Tomography
PPV	Positive Predictive Value
R-CNN	Region-based Convolutional Neural Network
RIS	Radiology Information System
RNN	Recurrent Neural Network
ROI	Region of Interest
SCC	Squamous Cell Carcinoma
SCLC	Small Cell Lung Cancer
SIFT	Scale-Invariant Feature Transform
SSD	Single Shot Multibox Detector
STBi-YOLO	Stochastic-Pooling-Based Spatial Pyramid Pooling Bidirectional Feature Pyramid YOLO
TNM	Tumor, Node, Metastasis
TP	True Positive
ViT	Vision Transformer
VGG	Visual Geometry Group Network
VLM	Vision–Language Model
YOLO	You Only Look Once

CHAPTER 1

Introduction

1.1 General introduction and motivation

Artificial Intelligence (AI) has profoundly reshaped science and technology, with medicine emerging as one of its most transformative domains. From symbolic logic systems to today's multimodal, large-scale frameworks, the evolution of AI reflects advances in computational power, data availability, and algorithmic innovation [1, 2, 3]. In healthcare, where precision, interpretability, and ethical responsibility are critical, AI has demonstrated its potential to improve diagnosis, prognosis, and workflow optimization [4, 5].

Early AI systems, such as rule-based expert systems like MYCIN, attempted to emulate clinical reasoning through symbolic rules [6]. While promising, they lacked adaptability to complex, real-world data variability. The subsequent rise of statistical learning methods in the 1990s, such as Support Vector Machines (SVMs) and Random Forests, enabled data-driven modeling in tasks such as epidemiology and imaging-based disease classification [7, 8, 9].

The breakthrough came in the 2010s with deep learning, particularly Convolutional Neural Networks (CNNs), which enabled hierarchical feature extraction directly from raw imaging data. CNNs revolutionized radiology, driving advances in tumor detection, segmentation, and classification in Computed Tomography (CT), Magnetic Resonance (MRI) , and ultrasound [3, 10, 11]. Parallel developments in Recurrent Neural Networks (RNNs) extended applications to sequential data, such as electrocardiograms [12, 13]. With the rise of transformers and Vision-Language Models (VLMs), multimodal integration of visual, textual, and structured data became feasible, enabling new forms of intelligent diagnostic support [14, 15, 16].

In clinical practice, radiologists are often required to make rapid yet highly accurate diagnostic decisions based on complex imaging data. Achieving diagnostic precision depends not only on the visual interpretation of current scans but also on the effective use of prior cases, imaging history, and structured clinical information available within

hospital information systems. In this context, *Content-Based Image Retrieval (CBIR)* has emerged as a paradigm to complement computer-aided diagnosis by enabling physicians to retrieve and compare new patient cases with visually and semantically similar historical cases [17, 18]. Such systems can support diagnostic reasoning, provide decision references from confirmed pathologies, and ultimately guide more personalized treatment planning.

However, most CBIR frameworks remain disconnected from the core clinical infrastructure, namely the *Picture Archiving and Communication System (PACS)* and the *Radiology Information System (RIS)*. PACS serve as centralized repositories for storing, retrieving, and visualizing medical images across modalities (e.g., CT, MRI, X-ray), while RIS manages complementary clinical and administrative data, including patient demographics, radiology reports, and scheduling [19, 20]. Together, PACS and RIS form the backbone of radiological workflows by linking imaging data with patient context and prior medical history. Despite their central role, current CBIR implementations often operate as standalone systems, limiting their clinical utility and preventing physicians from leveraging all available patient data in a unified diagnostic environment.

A key opportunity lies in the integration of CBIR systems with *Adaptive Hanging Viewer Protocols (HVPs)*, which can dynamically present both current and retrieved cases within the radiologist's reading interface [21, 22]. By embedding retrieval-driven insights directly into PACS viewers, clinicians could benefit from context-aware visualization of similar cases, historical outcomes, and associated reports, thus enhancing diagnostic accuracy and treatment planning. This thesis is motivated by the need to close this gap—designing retrieval-driven, explainable AI tools that seamlessly integrate into radiological workflows, bridging retrospective knowledge with real-time clinical decision-making.

1.2 AI in Medical Imaging: Opportunities and Challenges

Artificial Intelligence (AI) has become a transformative force in medical imaging, enabling advances in early disease detection, precision diagnosis, and workflow automation [4, 5, 21]. In thoracic imaging, for instance, AI-based detection systems have achieved remarkable sensitivity in identifying small pulmonary nodules on CT scans—sometimes surpassing radiologists in recall for early-stage lung cancer [10, 11, 23]. Automated staging

algorithms further support oncologists by providing consistent Tumor, node, metastasis (TNM) classifications, improving reproducibility in treatment planning [24, 25].

Beyond detection and classification, Vision–Language Models (VLMs) have introduced a new paradigm in radiology by jointly analyzing images and reports, enabling explainable, text-driven retrieval and visual grounding [15, 16]. AI has also demonstrated broad clinical relevance: improving fibrosis staging in hepatology through ultrasound biomarkers [26, 27], and facilitating rehabilitation monitoring through markerless motion analysis [28, 29]. Together, these examples highlight AI's potential to make medical imaging more quantitative, consistent, and patient-centered.

However, real-world deployment in hospitals remains limited. The challenge is no longer algorithmic accuracy, but **clinical integration** — ensuring that AI tools complement radiologists' workflows and leverage the full potential of existing infrastructure. Radiologists make diagnostic decisions by combining current images with historical data, prior cases, and structured clinical information stored in **Picture Archiving and Communication Systems (PACS)** and **Radiology Information Systems (RIS)**. Yet, most AI and retrieval systems operate in isolation, detached from these clinical repositories.

This disconnect underscores the need for **Content-Based Image Retrieval (CBIR)** systems that can retrieve relevant past cases directly within PACS/RIS environments. Such systems can assist radiologists in verifying rare findings, comparing disease progression, or validating uncertain diagnoses by surfacing similar, previously annotated cases. When integrated with adaptive hanging protocols, CBIR becomes not just a retrieval engine but a clinical reasoning assistant—bridging AI-driven analysis with radiological expertise.

Despite this promise, several barriers persist:

Workflow integration: Current AI tools rarely interface with PACS/RIS, limiting their usability during diagnostic review.

Interpretability: Deep models often behave as "black boxes," providing limited transparency and reducing clinical trust.

Scalability: Closed-vocabulary detection frameworks cannot adapt to evolving medical terminology or new disease patterns.

Interoperability: Imaging data, reports, and clinical metadata often remain siloed, impeding structured retrieval and case linking.

This thesis addresses these challenges by developing retrieval-driven, explainable AI systems that integrate seamlessly into radiological workflows. Specifically, it advances

CBIR methodologies that leverage both visual and textual embeddings to deliver semantically relevant, clinically interpretable results within PACS/RIS.

Figure 1.1 illustrates the conceptual CBIR workflow, consisting of an *offline phase*, where case embeddings are generated and indexed, and an *online phase*, where a query image is processed and matched to relevant cases. Feature extraction has evolved from handcrafted descriptors (e.g., SIFT, HOG) to convolutional representations (e.g., VGG, ResNet), and more recently, to transformer-based architectures (e.g., ViT, Swin, Florence), forming the foundation of this thesis.

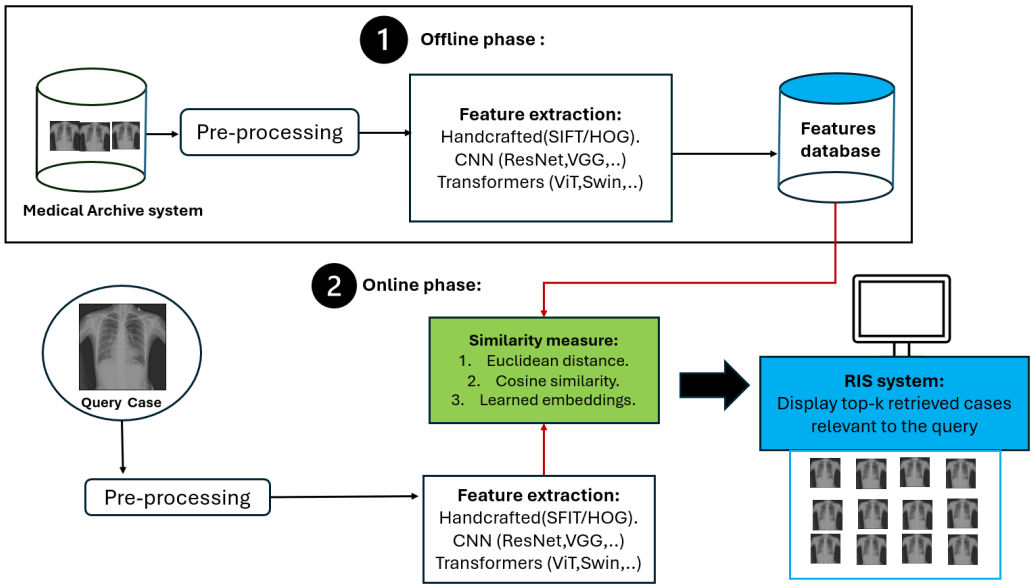

Figure 1.1: Conceptual Content-Based Image Retrieval (CBIR) workflow in medical imaging. In the offline phase, case embeddings are precomputed and stored in a searchable database. In the online phase, a query image is encoded, compared against stored features, and retrieved results are displayed directly within PACS/RIS for diagnostic support. The feature extraction pipeline has evolved from handcrafted descriptors (e.g., SIFT/HOG) to CNNs (e.g., ResNet, VGG) and more recently to Vision Transformers (e.g., ViT, Swin, Florence).

1.3 Core Technical Building Blocks

At the heart of any Content-Based Image Retrieval (CBIR) system lies the **feature extraction block**, which transforms medical images into compact, discriminative representations that can be compared and retrieved efficiently. The quality and structure of these feature embeddings directly determine the system's retrieval accuracy, interpretability, and clini-

cal utility. Over time, this block has evolved through several technological paradigms — from convolutional feature extractors to region-aware detection models and, more recently, to multimodal vision–language encoders.

1.3.1 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) remain the backbone of visual feature extraction in medical imaging. By learning hierarchical representations directly from pixel-level inputs, CNNs capture both local and global image patterns. Landmark architectures such as LeNet [30], AlexNet [31], VGG [32], ResNet [33], and Inception [34] introduced increasingly deep and expressive architectures capable of modeling complex visual semantics.

These networks have enabled major advances in medical applications — from pulmonary nodule detection in CT scans and lesion classification in MRI, to disease staging in oncology and cardiology. Through innovations such as residual connections and multi-scale feature hierarchies, CNNs established the foundational principles of medical image representation learning, which continue to underpin modern detection, segmentation, and retrieval pipelines.

The convolution operation, central to CNNs, computes localized feature responses by applying learnable kernels across the image. In discrete form, this can be expressed as:

$$(f * g)[n] = \sum_{m=-\infty}^{\infty} f[m] \cdot g[n-m], \tag{1.1}$$

where:

$f[m]$ denotes the input signal (e.g., pixel intensities indexed at position m),

$g[n-m]$ denotes the convolutional kernel shifted by n,

n represents the index of the output position at which the convolution is evaluated,

m is a dummy summation index that iterates over all valid positions of the input signal.

Stacked convolutional layers, combined with nonlinear activations such as the rectified linear unit (ReLU) and pooling operations, progressively encode low-level pixel information into high-level feature representations suitable for similarity-based retrieval.

1.3.2 Object Detection and Region-Aware Representations

While global CNN embeddings improved retrieval accuracy, they often lacked spatial awareness, failing to capture localized abnormalities essential for diagnosis. Object detection frameworks such as Faster R-CNN [35], Single Shot MultiBox Detector (SSD) [36], and YOLO [37] addressed this limitation by introducing region-based feature extraction. These models predict both the location and class of pathological findings, producing structured embeddings that preserve spatial context.

Detection quality is typically evaluated using the Intersection over Union (IoU), which quantifies the overlap between predicted and ground-truth regions:

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}. \tag{1.2}$$

Region-aware features derived from these models form the basis for localized CBIR — enabling retrieval of clinically similar cases based not only on global appearance but also on spatially grounded pathology descriptors.

1.3.3 Multimodal and Vision–Language Embeddings

Recent advances have expanded the feature extraction paradigm beyond purely visual representations toward multimodal integration. Vision Transformers (ViTs) [38] and Swin Transformers [39] introduced self-attention mechanisms that model long-range dependencies across entire images, producing globally contextual embeddings that improve retrieval robustness.

Building on this, Vision–Language Models (VLMs) such as CLIP [15] and Florence-2 [40] align visual and textual embeddings within a shared semantic space. By leveraging paired image–report datasets, these models enable open-vocabulary retrieval — allowing clinicians to issue natural language queries (e.g., “spiculated nodule in the right upper lobe”) and retrieve matching cases directly from medical archives. This evolution from pixel-based descriptors to cross-modal embeddings marks a paradigm shift toward more interpretable, context-aware, and clinically integrated CBIR systems.

1.4 Research Motivation

The progression outlined above—from handcrafted features to convolutional networks, object detection, and vision–language models—illustrates both the opportunities and lim-

itations of current AI in medical imaging. Despite substantial progress, three major gaps persist.

First, Content-Based Image Retrieval (CBIR) systems are rarely integrated within hospital information infrastructures such as PACS and RIS, limiting their impact on real-time clinical workflows. Second, most existing retrieval and detection systems function as "black boxes," offering limited interpretability and traceability in clinical decision-making. Third, traditional models remain constrained by closed vocabularies, preventing flexible, multimodal queries that reflect the complexity of radiological language.

This thesis is motivated by the need to bridge these gaps. It aims to design retrieval-driven, explainable AI frameworks that combine visual and textual information, support adaptive integration within PACS/RIS, and enable scalable, semantically rich case retrieval for clinical use.

1.5 Thesis Objectives, Research Questions, and Contributions

The overall objective of this thesis is to advance the technical and clinical integration of CBIR systems by leveraging recent developments in object detection, transformer architectures, and vision–language modeling. Four central research questions guide this work:

1. How can CBIR and adaptive viewer protocols be integrated into PACS/RIS to support radiological diagnosis?
2. How can AI and machine learning techniques identify and rank clinically similar patient cases using both visual and textual data?
3. What strategies ensure structured, interoperable databases that enable accurate and semantically relevant case retrieval?
4. How can Vision–Language Models (VLMs) advance multimodal integration and open-vocabulary retrieval in clinical workflows?

To address these questions, the thesis makes the following key contributions, organized in a cumulative technical progression:

Detection-based lung cancer staging and subtype classification. A 3D volumetric CT pipeline is developed using YOLO-based object detection for subtype classification of lung cancer (SCLC, ADC, SCC). A TNM staging classifier is constructed by leveraging YOLO backbone features, achieving superior precision and inference speed compared with YOLOv6, YOLOv7, and Faster R-CNN baselines. This establishes detection as a foundation for case-level feature extraction.

Detection-driven CBIR for lung cancer imaging. A retrieval framework is proposed where object detection embeddings serve as the core feature representations for similarity computation. This enables the retrieval of cases based on both histological subtype (Adenocarcinoma, Squamous Cell Carcinoma, and Large-Cell Carcinoma) and TNM stage (Tumor, Node, Metastasis), positioning detection models as integral components of decision-support and case comparison within PACS/RIS environments.

Semantic transition toward multimodal retrieval. Building upon detection-based retrieval, this stage explores the semantic gap between visual similarity and clinical meaning. By analyzing the role of embeddings in semantic search, it motivates the integration of textual supervision—creating a conceptual bridge toward multimodal, language-informed retrieval.

Prompt-conditioned multimodal CBIR with Florence-2. In this work, we introduce a novel feature fusion framework based on the Florence-2 model that leverages multiple task-specific prompts to extract complementary visual-textual representations. This approach enhances the model's ability to capture diverse semantic information across medical images. The detailed structure and prompt-conditioning mechanism are presented in later sections.

Collectively, these contributions establish a unified framework that evolves from detection-based visual retrieval to prompt-driven multimodal understanding. Each stage advances both the technical depth and clinical relevance of CBIR systems, progressively moving toward scalable, interpretable, and workflow-integrated solutions for radiology.

1.6 Thesis Structure

The thesis is organized to follow a logical technical progression, moving from foundational principles to advanced multimodal retrieval frameworks. Chapter 2 reviews the historical evolution of CBIR in medical imaging, from handcrafted descriptors to deep learning, object detection, and Vision Transformers, identifying gaps in current practice. Chapter 3 establishes the methodological basis for detection-driven feature extraction and staging in CT imaging, providing robust embeddings as a precursor to retrieval. Building on this, Chapter 4 introduces retrieval-driven approaches for lung cancer imaging, demonstrating how localized embeddings improve accuracy and interpretability. Chapter 5 addresses the limitations of closed-vocabulary detection by introducing open-vocabulary approaches based on Vision-Language Models, enabling flexible, text-driven abnormality retrieval. Chapter 6 advances toward unified multimodal systems by fusing visual and textual representations, supporting both case retrieval and automated report generation. Finally, Chapter 7 concludes by synthesizing the main findings, reflecting on limitations, and outlining future research directions, with a focus on developing scalable, interpretable, and clinically deployable CBIR systems.

Permissions and Reuse of Previously Published Material

Some figures and text in this thesis have been reproduced from the author's previously published works:

> Wehbe, A., Hotiet, H., Minetti, I., Dellapiane, S., "Integrating YOLO for Advanced Content-Based Image Retrieval in Lung Cancer Imaging," *IEEE ICECS 2024*, pp. 1–4.
>
> Wehbe, A., "From Closed to Open Vocabulary: A Systematic Review of Abnormality Detection in Medical Imaging," *ICABME 2025*, IEEE.

All reused material is included under the respective publisher's non-commercial and academic-use license. The figures and text are used for educational and scholarly purposes, and proper citation is provided throughout the thesis.

CHAPTER 2

Literature Review

2.1 Overview

This chapter reviews the state of the art in Content-Based Image Retrieval (CBIR) for medical imaging, focusing on the evolution from closed- to open-vocabulary frameworks. Early systems relied on Convolutional Neural Networks (CNNs) and object detection models such as Faster R-CNN, SSD, and YOLO, mainly applied to CT scans for lung cancer analysis. These methods provided strong visual representations but remained limited to predefined categories, constraining their flexibility in clinical interpretation. Subsequent advances with Transformer-based architectures, including Vision Transformers (ViTs) and Swin Transformers, improved global context modeling and generalization across imaging modalities. More recently, multimodal approaches such as CLIP, MedCLIP, and Grounding DINO have aligned visual and textual data to enable open-vocabulary retrieval—particularly effective in chest X-rays, where paired images and reports are available for joint training. This progression from CT-based closed-vocabulary detection to X-ray–based multimodal retrieval highlights the field's shift toward more flexible, explainable, and clinically integrated CBIR systems. The chapter concludes by outlining remaining challenges, interpretability, interoperability, and workflow integration—which motivate the research directions of this thesis.

2.2 Introduction:

Content-Based Image Retrieval (CBIR) has emerged as a complementary paradigm to computer-aided diagnosis, allowing clinicians to retrieve and compare similar cases from large medical archives. Over the past three decades, CBIR in medical imaging has undergone several key methodological shifts — from handcrafted feature descriptors to Convolutional Neural Networks (CNNs), from global embeddings to object detection, and more recently to transformers and multimodal vision–language frameworks.

Each of these advances addressed specific limitations of its predecessors while introducing new challenges. Handcrafted features provided rotation- and scale-invariant representations but failed to capture semantic meaning relevant to clinical interpretation. CNNs enabled hierarchical feature learning directly from pixel data but often lacked region-level interpretability. Object detection frameworks added spatial localization, yet remained constrained by closed vocabularies. The advent of transformers and vision–language models has opened the door to open-vocabulary retrieval and multimodal reasoning, though issues of interpretability, computational cost, and clinical integration remain unresolved.

This chapter critically reviews these developments in the context of medical imaging and radiological practice. The progression establishes the conceptual and technical foundation for the thesis contributions, which focus on enhancing interpretability, scalability, and multimodal integration in retrieval-driven AI systems.

2.3 Evolution of AI in medical imaging:

2.3.1 Early Approaches: Classical CBIR Methods

The first generation of CBIR systems in medical imaging relied on handcrafted descriptors such as Scale-Invariant Feature Transform (SIFT), Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), and Gabor filters. These descriptors encoded low-level image attributes (edges, textures, intensity distributions) into fixed-length vectors $f(I)$ for each image I. Case similarity was then quantified via distance or similarity measures. Let I_a and I_b denote two images being compared. Their similarity can be computed using:

$$d(I_a, I_b) = \|f(I_a) - f(I_b)\|_2, \tag{2.1}$$

$$\text{sim}(I_a, I_b) = \frac{f(I_a) \cdot f(I_b)}{\|f(I_a)\| \, \|f(I_b)\|}. \tag{2.2}$$

Here, $d(I_a, I_b)$ represents the Euclidean distance between the feature embeddings of the two images, while $\text{sim}(I_a, I_b)$ denotes their cosine similarity.

While these features were invariant to scale and rotation, they suffered from the well-known *semantic gap*: radiologists reason in terms of clinical patterns (e.g., “spiculated nodule”), not pixel-level textures.

To mitigate this limitation, early systems explored *physician-in-the-loop* paradigms. ASSERT [41], for example, used physician feedback to refine feature spaces via sequential feature selection. Marchiori et al. [42] introduced query-by-example for High-Resolution Computed Tomography (HRCT) lung images, where radiologists marked pathology-bearing regions (PBRs), enabling anatomically informed retrieval. Dy et al. [43] demonstrated that unsupervised feature selection combined with clinical feedback significantly improved diagnostic precision. Relevance feedback loops became common: physicians labeled retrieved cases as relevant or irrelevant, enabling iterative refinement of the feature space.

Despite these innovations, handcrafted CBIR systems were constrained by three factors:

Global descriptors: they often overlooked subtle, localized lesions such as micronodules.

Noise sensitivity: feature performance degraded with acquisition variability across scanners and protocols.

Scalability: as archives grew, the handcrafted descriptors became inefficient for real-time retrieval.

These limitations paved the way for deep learning, which offered hierarchical feature extraction directly from data.

2.3.2 Deep Learning Revolution: CNN-based Retrieval and Classification

Mathematical foundations of CNNs. A Convolutional Neural Network (CNN) consists of convolutional, activation, and pooling layers followed by fully connected layers for representation learning. The 2D convolution operation for a discrete image t with kernel g is given by:

$$(t * g)[n, m] = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} t[i, j] \cdot g[n - i,\ m - j], \tag{2.3}$$

where:

$t[i, j]$ denotes the pixel intensity of the input image at spatial location (i, j),

$g[n-i,\, m-j]$ is the convolutional kernel centered at the output position (n, m),

(n, m) represent the spatial coordinates of the output feature map,

(i, j) are dummy indices iterating over the spatial coordinates of the input.

Non-linear activation functions introduce expressivity into the network by transforming the output of convolutional layers in a non-linear manner. The most widely used activation is the Rectified Linear Unit (ReLU), defined as:

$$\mathrm{ReLU}(x) = \max(0, x), \tag{2.4}$$

where x denotes the input value to the activation function, typically corresponding to the output of a convolutional operation (i.e., a pre-activation feature value at a given spatial location). The ReLU function suppresses negative values while keeping positive values unchanged, thereby improving the network's ability to learn complex patterns.

Pooling reduces dimensionality and improves translation invariance by aggregating information over local spatial regions. Max pooling, for example, is defined as:

$$\mathrm{MaxPool}(x, s) = \max\{x_{i:i+s,\, j:j+s}\}, \tag{2.5}$$

where:

x denotes the input feature map,

(i, j) represent the top-left spatial coordinates of the pooling window,

s is the pooling window size (e.g., $s = 2$ for a 2×2 pooling region),

$x_{i:i+s,\, j:j+s}$ refers to the local patch of the feature map over which the maximum value is computed.

Applications in medical retrieval. Architectures such as ResNet, VGG, and DenseNet have been widely employed to extract embeddings from CT and X-ray images, offering robustness to variability in acquisition protocols and noise. For example, Tschandl et al. [44] leveraged a pre-trained ResNet-50 to retrieve visually similar dermatoscopic images, while Swati et al. [45] demonstrated the utility of VGG19 embeddings in brain MRI retrieval using Mahalanobis distance, boosting intra-class similarity and minimizing inter-class confusion. In thoracic imaging, Rubin et al. [46] trained a dual-view CNN on the

MIMIC-CXR dataset for thoracic disease classification; although primarily designed for classification, the learned embeddings enabled meaningful retrieval of clinically similar chest radiographs through cosine or Euclidean distance measures.

Several studies have demonstrated the superiority of CNN-based embeddings over handcrafted features. Chung et al. [47] employed a ResNet-50-based Siamese CNN to generate fixed-length feature vectors, outperforming single-network baselines in medical image retrieval. Ozturk et al. [48] proposed deep hashing to convert embeddings into compact binary codes, accelerating retrieval in large-scale archives. In lung-specific contexts, Choe et al. [49] developed a CBMIR system to differentiate usual interstitial pneumonia (UIP) from nonspecific interstitial pneumonia (NSIP) on CT, demonstrating the clinical value of CNN embeddings in supporting differential diagnosis. Similarly, Hwang et al. [50] applied CNN-based segmentation and classification to interstitial lung disease (ILD), enabling retrieval of cases with comparable parenchymal patterns, although retrieval precision remained limited.

Challenges and domain adaptation. Despite these successes, most CNN-based CBMIR systems rely on ImageNet pretraining, creating a *domain mismatch* when applied to medical tasks. As a result, models often fail to capture subtle abnormalities such as early-stage lung nodules or fibrosis patterns. To mitigate this, domain-specific pretraining datasets have been introduced, such as MedNet [51], MedMNIST [52], and Med3D [53], which provide medical priors that improve generalization across institutions. Nevertheless, CNN-based embeddings remain predominantly global, lacking region-level interpretability. This shortcoming limits clinical trust and motivates the transition to object detection frameworks, where features are extracted from localized Regions of Interest (ROIs). These frameworks enhance both retrieval accuracy and interpretability, aligning CBIR more closely with radiological practice and adaptive Hanging Viewer Protocols (HVPs).

2.3.3 Object Detection Models for Medical Imaging (Closed-Vocabulary)

Object detection frameworks represent a significant evolution from global image-level representations toward region-level interpretability, which is particularly important in medical imaging applications. As introduced in the Introduction chapter, object detection methods are commonly divided into *two-stage* and *one-stage* detectors. Two-stage approaches,

such as Faster R-CNN, first generate a set of region proposals that are likely to contain objects of interest, which are subsequently refined through classification and bounding box regression. In contrast, one-stage detectors, including SSD and YOLO, perform localization and classification jointly in a single forward pass, enabling real-time inference at the cost of a slightly reduced detection accuracy in some scenarios (Figure 2.1).

Formally, the output of an object detection model can be represented as:

$$\mathbf{y}^T = [P_c, b_x, b_y, b_w, b_h, c_1, c_2, \ldots, c_k], \tag{2.6}$$

where P_c denotes the objectness or confidence score, (b_x, b_y, b_w, b_h) represent the bounding box center coordinates, width, and height, respectively, and c_i corresponds to the predicted probability for class i among k predefined categories. This closed-vocabulary formulation assumes that all detectable object classes are known during training.

The performance of object detection models is commonly assessed using spatial overlap and classification-based metrics, including Intersection over Union (IoU), precision, recall, F1-score, and mean Average Precision (mAP), which are formally defined in the following subsections.

2.3.3.1 Confusion Matrix Elements

In the context of object detection, predictions are categorized as follows:

True Positives (TP): correctly detected objects whose predicted bounding boxes sufficiently overlap with the corresponding ground-truth boxes.

False Positives (FP): predicted objects that do not match any ground-truth instance.

False Negatives (FN): ground-truth objects that are missed by the detector.

True negatives are generally not defined in object detection tasks, as the number of possible background regions is not explicitly enumerated.

2.3.3.2 Intersection over Union (IoU)

Intersection over Union (IoU) quantifies the spatial overlap between a predicted bounding box B_p and a ground-truth bounding box B_{gt}:

$$\text{IoU} = \frac{B_p \cap B_{gt}}{B_p \cup B_{gt}}. \tag{2.7}$$

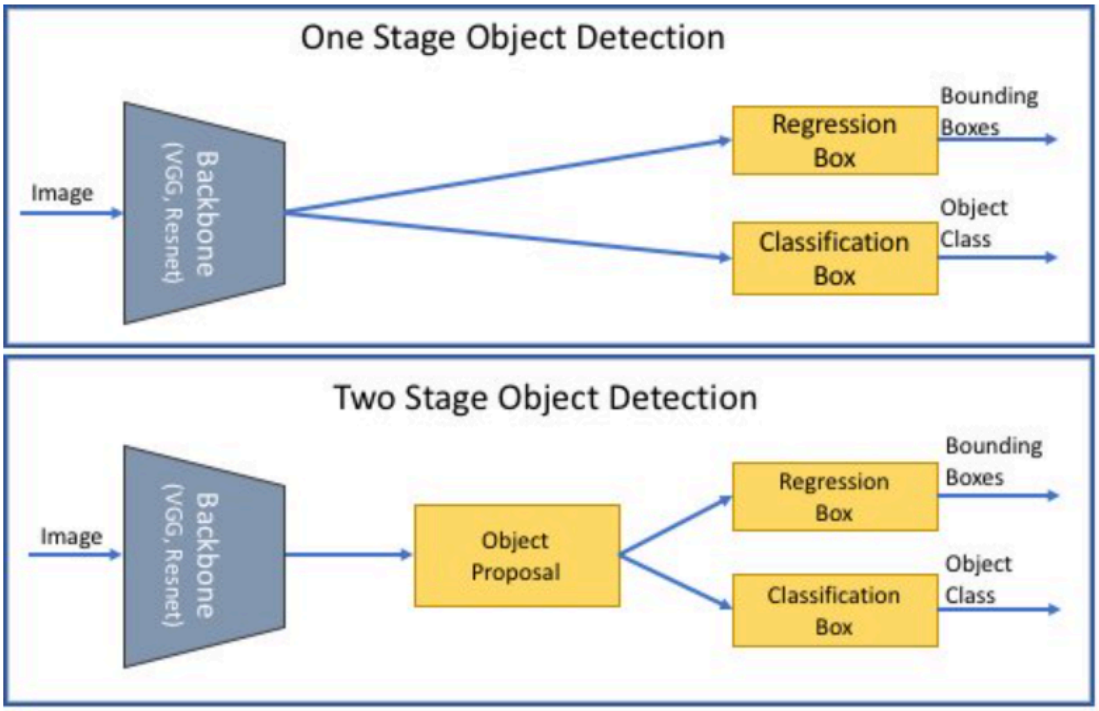

Figure 2.1: Comparison between one-stage and two-stage object detection architectures applied to medical image analysis. In one-stage detectors, the backbone network (e.g., VGG or ResNet) directly predicts bounding boxes and corresponding lesion classes in a single step. In contrast, two-stage detectors first generate region proposals likely to contain abnormalities (e.g., tumors or nodules), which are subsequently refined through dedicated regression and classification branches. While two-stage approaches often achieve higher accuracy for small or subtle lesions, one-stage models offer faster inference suitable for real-time clinical applications.

A detection is considered correct if the IoU exceeds a predefined threshold, commonly set to 0.5.

2.3.3.3 Precision

Precision measures the proportion of correctly detected objects among all predicted detections:

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{2.8}$$

High precision indicates a low false positive rate, which is particularly important in clinical decision-support systems.

2.3.3.4 Recall

Recall evaluates the ability of the model to detect all relevant objects present in the image:

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{2.9}$$

A high recall value reflects a reduced number of missed pathological findings.

2.3.3.5 Precision–Recall Curve

The Precision–Recall (PR) curve is obtained by varying the confidence threshold applied to detection scores and plotting precision as a function of recall. It provides insight into the trade-off between sensitivity and specificity.

2.3.3.6 Average Precision (AP)

Average Precision (AP) summarizes the Precision–Recall curve into a single scalar value:

$$\text{AP} = \int_0^1 p(r)\, dr, \tag{2.10}$$

where $p(r)$ denotes precision as a function of recall r. In practice, AP is computed using a discrete approximation over a finite set of recall levels.

2.3.3.7 Mean Average Precision (mAP)

Mean Average Precision (mAP) is defined as the mean of AP values computed over all object classes:

$$\text{mAP} = \frac{1}{C} \sum_{c=1}^{C} \text{AP}_c, \tag{2.11}$$

where C is the total number of classes and AP_c denotes the Average Precision for class c.

2.3.3.8 mAP at Different IoU Thresholds

In YOLO-based evaluations, mAP is often reported at different IoU thresholds:

mAP@0.5: AP computed using an IoU threshold of 0.5.

mAP@0.5:0.95: AP averaged over IoU thresholds ranging from 0.5 to 0.95 with a step size of 0.05.

The latter metric provides a more stringent and comprehensive assessment of both localization and classification performance.

Applications in Lung Imaging

In recent decades, computer-aided detection (CAD) systems have emerged as valuable tools in medical imaging, significantly aiding in the detection of lung nodules and supporting early lung cancer diagnosis [54, 55]. Object detection frameworks have played a pivotal role in this evolution, moving beyond handcrafted features and global CNN embeddings to localized detection of clinically meaningful abnormalities. Gunasekaran et al. [56] applied YOLOv5 to lung cancer detection, optimizing hyperparameters and employing data augmentation techniques. Their method achieved high accuracy and recall, outperforming traditional approaches while demonstrating real-time clinical applicability. Similarly, Liu et al. [57] proposed the STBi-YOLO framework, an extension of YOLOv5 incorporating stochastic-pooling-based spatial pyramid pooling, a bidirectional feature pyramid network, and an optimized Efficient Intersection over Union (EIoU) loss function. Their system reported superior performance in lung nodule detection on CT images, surpassing YOLOv3, YOLOv4, YOLOv5, Faster R-CNN, and SSD in comparative experiments. Other CAD systems have extended detection tasks to classification of lung cancer subtypes. For example, a self-calibrated convolution module based on R-CNN, combined with a novel regression loss function, successfully classified squamous cell carcinoma (SCC), adenocarcinoma (ADC), and small cell lung cancer (SCLC), achieving a high mean average precision (mAP). However, this framework lacked information on tumor size and cancer staging, limiting its clinical utility. Similarly, Jing et al. [58] adopted a center-point-based detection approach (CenterNet) for histological subtype localization and classification on CT images. While effective in reducing false positives and negatives, this method also did not address staging, which is essential for treatment planning. These studies highlight the potential of CAD systems, powered by object detection frameworks, to significantly enhance lung cancer diagnosis. Nevertheless, they also reveal critical limitations: the absence of tumor size estimation, staging information, and integration into broader radiological workflows. Addressing these gaps motivates the contributions of this thesis, which extend object detection toward retrieval-driven frameworks, bridging detection outputs with Content-Based Image Retrieval (CBIR) and clinical decision support. A comparative summary of representative models is provided in Table 2.1, which highlights how YOLO-based methods, particularly YOLOv6, YOLOv7, and STBi-YOLO, achieve

superior precision, recall, and mAP scores compared to traditional two-stage frameworks. Notably, STBi-YOLO demonstrates the highest recall, underscoring its clinical potential in minimizing false negatives—a key priority in lung cancer screening.

Table 2.1: Performance comparison of representative object detection frameworks for lung cancer detection.

Model	**Precision**	**Recall**	**mAP@0.5**	**F1 Score**
Faster R-CNN	0.914	0.891	-	0.900
YOLOv6	0.933	0.932	0.957	0.930
YOLOv7	0.940	0.920	0.945	0.920
YOLOv5 small [59]	0.922	0.894	0.936	0.907
SC-Dynamic R-CNN [60]	0.750	0.780	0.850	0.760
STBi-YOLO [57]	0.931	0.951	-	0.940
CenterNet [61]	0.710	0.821	0.851	0.761

These limitations highlight the need to couple object detection with retrieval-based pipelines, where detected regions form the basis for similarity-driven case retrieval. This motivates the transition from closed-vocabulary detection toward open-vocabulary, multimodal frameworks explored in later sections.

2.3.4 Transformer-based models and Vision Transformers (ViTs)

Vision Transformers (ViTs) have recently emerged as powerful alternatives to convolutional neural networks (CNNs) for medical image analysis. Unlike CNNs, which rely on local receptive fields and hierarchical convolutions, ViTs process images as sequences of non-overlapping patches and apply self-attention to capture long-range dependencies across the entire image [38]. Given an input sequence $X \in \mathbb{R}^{n \times d}$, where n denotes the sequence length (e.g., number of image patches or tokens) and d the feature dimensionality, the query (Q), key (K), and value (V) matrices are computed as:

$$Q = XW_Q, \qquad K = XW_K, \qquad V = XW_V, \tag{2.12}$$

where:

$W_Q, W_K, W_V \in \mathbb{R}^{d \times d_k}$ are learnable projection matrices,

$Q \in \mathbb{R}^{n \times d_k}$ represents the set of query vectors,

$K \in \mathbb{R}^{n \times d_k}$ represents the set of key vectors,

$V \in \mathbb{R}^{n \times d_k}$ represents the set of value vectors,

d_k is the dimensionality of each projected vector.

The scaled dot-product attention is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{2.13}$$

where d_k denotes the dimensionality of the key (and query) vectors. The scaling factor $\sqrt{d_k}$ is applied to prevent the dot-product values from becoming excessively large, which stabilizes the gradients and improves training convergence.

where the division by $\sqrt{d_k}$ stabilizes gradients and prevents excessively large dot-product magnitudes. This mechanism enables the model to capture long-range dependencies by assigning attention weights across different regions of the image or feature sequence.

Several studies have demonstrated the advantages of ViTs in thoracic imaging. Zhou et al. [62] evaluated ViT-based approaches on large-scale chest radiograph datasets such as MIMIC-CXR and CheXpert, reporting improved robustness to domain shifts compared to CNN baselines. Their results highlighted the ability of ViTs to generalize across institutions and acquisition protocols, a critical challenge in clinical deployment. Similarly, Chen et al. [61] integrated ViTs with CNN backbones in a hybrid framework (TransUNet) and demonstrated superior performance in medical segmentation tasks, suggesting the benefits of combining local feature extraction with global attention.

Beyond classification and segmentation, ViTs have been explored for representation learning and Content-Based Image Retrieval (CBIR). Unlike CNNs, which primarily encode local receptive fields, ViTs employ global self-attention, enabling them to jointly capture lesion-level features and broader anatomical context. This leads to more discriminative embeddings for retrieval tasks, particularly in heterogeneous datasets. Performance is typically quantified via cosine similarity:

$$\text{sim}(I_a, I_b) = \frac{f(I_a) \cdot f(I_b)}{\|f(I_a)\| \, \|f(I_b)\|}, \tag{2.14}$$

where $f(I)$ denotes the embedding of image I. Evaluation metrics include mean Average Precision (mAP), Recall@K, and Precision@K, which ensure retrieved results align with clinical relevance.

Swin Transformers [39], which introduce hierarchical shifted windows for efficient attention computation, have been widely adopted in medical imaging. For instance, Ko et al. [63] optimized Swin-based models for lung disease prediction from CT scans, showing significant improvements over ResNet baselines. Zhou et al. [62] also demonstrated that ViT variants trained on chest X-ray datasets outperform CNNs in disease classification and generalize more effectively under domain shift. Kang et al. [64] further confirmed that ViT embeddings retrieve clinically similar radiographs with higher fidelity than convolutional counterparts.

These findings highlight the growing role of ViTs in CBIR, where their ability to model both local abnormalities (e.g., nodules, opacities) and global lung patterns provides an advantage over purely convolutional pipelines, motivating their integration into retrieval-driven clinical frameworks.

2.3.5 Multimodal fusion: vision–text alignment and contrastive learning (open vocabulary object detection)

While CNNs and ViTs have significantly advanced visual representation learning, clinical decision-making rarely depends on images alone. Radiology reports, staging descriptors, and structured metadata provide essential complementary information. Multimodal fusion approaches seek to bridge these modalities by aligning visual and textual representations in a shared semantic space, enabling more flexible and clinically relevant retrieval as presented in Fig 2.2 A major breakthrough in this direction came with CLIP [15], which introduced a contrastive learning framework to jointly train image and text encoders on large-scale paired datasets. The objective aligns matched pairs while separating mismatched ones. Formally, the contrastive loss is:

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{N}\sum_{i=1}^{N}\left[\log\frac{\exp(\langle v_i, t_i\rangle/\tau)}{\sum_{j=1}^{N}\exp(\langle v_i, t_j\rangle/\tau)} + \log\frac{\exp(\langle t_i, v_i\rangle/\tau)}{\sum_{j=1}^{N}\exp(\langle t_i, v_j\rangle/\tau)}\right], \quad (2.15)$$

where v_i and t_i denote image and text embeddings, τ is a temperature parameter, and $\langle\cdot,\cdot\rangle$ is cosine similarity. This framework enables open-vocabulary retrieval: clinicians can issue free-text queries (e.g., "spiculated nodule in the right upper lobe"), and the system retrieves semantically aligned radiographs.

Domain-specific adaptations such as **MedCLIP** [65], pretrained on MIMIC-CXR and ChestXray14, improved retrieval performance compared to natural-image pretrained mod-

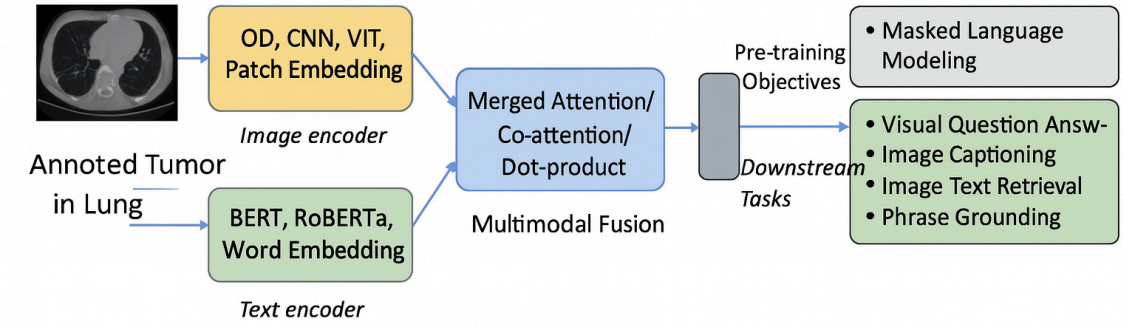

Figure 2.2: Overview of a multimodal learning framework for medical image-text alignment. The model integrates a image through an image encoder (e.g., CNN, ViT) and corresponding textual description through a text encoder (e.g., BERT, RoBERTa). The extracted embeddings are fused using attention-based mechanisms to form a unified representation. The joint model is optimized through pre-training objectives such as masked language modeling and image–text contrastive learning, enabling downstream tasks like image captioning, visual question answering, image–text retrieval, and phrase grounding in medical imaging contexts.

els, but embeddings remain predominantly global, limiting region-level interpretability. Vision–language detection frameworks extend this by incorporating spatial grounding. For example, **Grounding DINO** [66] combines a ViT backbone with a text-conditioned decoder, enabling detection and localization of image regions based on arbitrary textual prompts. Early applications to chest radiography [52] show promise for grounding abnormalities in free-text queries, though challenges remain in adapting vocabularies to radiological language.

Previous studies evaluate models using standard metrics for both retrieval and report generation tasks.

Multimodal Retrieval. We use **Recall@K**, defined as:

$$\text{Recall@K} = \frac{\text{Number of relevant items retrieved in top-}K}{\text{Total number of relevant items}}. \tag{2.16}$$

Multimodal Report Generation. We assess report generation using bilingual evaluation understudy **BLEU**, Metric for Evaluation of Translation with Explicit ORdering**METEOR**, and Recall-Oriented Understudy for Gisting Evaluation **ROUGE**:

BLEU-*n* evaluates n-gram overlap between generated and reference reports:

$$\text{BLEU-n} = \text{BP} \cdot \exp\left(\sum_{i=1}^{n} w_i \log p_i\right), \tag{2.17}$$

where p_i denotes i-gram precision, w_i are weights, and BP is the brevity penalty.

METEOR measures unigram matches considering synonyms, stemming, and word order; it outputs a score between 0 and 1.

ROUGE measures n-gram, word sequence, and word pair overlaps between generated and reference reports. ROUGE-L, in particular, is based on the longest common subsequence.

High Recall@K indicates clinically relevant retrieval, while high BLEU, METEOR, and ROUGE scores ensure faithful report generation.

These developments highlight the evolution from unimodal CNN embeddings, to multimodal contrastive frameworks like CLIP and MedCLIP, and finally to open-vocabulary detection with spatial grounding. This trajectory directly motivates integrating open-vocabulary object detection into CBIR frameworks for medical imaging, bridging retrieval with clinically meaningful localization.

2.4 Summary and identified research gaps

The evolution of Content-Based Image Retrieval (CBIR) in medical imaging has progressed from handcrafted feature descriptors to deep learning, object detection, and more recently, transformer-based and multimodal vision–language frameworks. Early systems based on descriptors such as SIFT and HOG offered limited robustness to imaging variability and struggled to capture clinically meaningful patterns. The deep learning revolution

introduced Convolutional Neural Networks (CNNs), which enabled hierarchical feature extraction and improved retrieval performance. CNN-based CBIR systems demonstrated effectiveness in tasks such as lung nodule detection and interstitial lung disease classification. However, their reliance on global embeddings limited interpretability and sensitivity to localized abnormalities.

Object detection frameworks such as Faster R-CNN, SSD, and YOLO addressed this limitation by enabling region-level retrieval through bounding box localization, enhancing clinical interpretability. These approaches bridged the gap between retrieval and diagnostic workflows, allowing adaptive Hanging Viewer Protocols (HVPs) to highlight regions of interest (ROIs). Yet, they remain constrained by closed vocabularies, which limits scalability when new pathologies or imaging biomarkers emerge.

More recently, transformer-based architectures such as Vision Transformers (ViTs) and Swin Transformers have demonstrated strong performance in medical imaging by modeling long-range dependencies and providing robust embeddings across heterogeneous datasets such as MIMIC-CXR and CheXpert. Despite outperforming CNNs in disease classification and retrieval, ViTs remain computationally expensive and often require domain-specific pretraining to achieve optimal results.

The latest generation of multimodal frameworks, including CLIP and Florence, has extended CBIR beyond purely visual embeddings by aligning images with free-text radiology reports. These vision–language models enable open-vocabulary retrieval and natural language queries, addressing some of the limitations of closed-set object detection. Nevertheless, most existing frameworks operate at a global image level and lack fine-grained spatial grounding, reducing their utility for subtle lesion retrieval and limiting integration into PACS/RIS workflows.

Identified Research Gaps

From this review, several gaps are identified:

Limited integration with clinical workflows: Current CBIR and detection systems are rarely connected to PACS/RIS, hindering their deployment in routine practice.

Lack of region-level interpretability: CNN and transformer-based embeddings often focus on global representations, overlooking subtle, localized abnormalities critical for diagnosis.

Closed-vocabulary constraints: Object detection systems cannot generalize to unseen diseases or descriptors, restricting adaptability in dynamic clinical environments.

Limited multimodal grounding: Vision–language models improve retrieval flexibility but often lack spatial interpretability and fine-grained alignment between images and text.

These gaps motivate the objectives of this thesis, which aims to design retrieval-driven, explainable AI frameworks that integrate region-aware detection, multimodal embeddings, and clinical metadata into CBIR pipelines. By addressing both scalability and interpretability, the proposed systems seek to bridge the divide between technical advances in AI and their real-world translation into radiological workflows.

CHAPTER 3

Enhanced Lung Cancer Detection and TNM Staging Using YOLOv8 and TNMClassifier.

3.1 Overview

This chapter focuses on lung cancer as a clinical application of the closed-vocabulary approach, where object detection is employed for both classification and localization of cancer subtypes. The YOLOv8 architecture is utilized to detect and categorize lesions, while TNM staging is derived using features extracted from the backbone of the detection network, following the standard clinical TNM criteria. The results include a comprehensive evaluation of model performance based on key metrics such as mean Average Precision (mAP), Precision, and Recall, along with comparative analyses against baseline and state-of-the-art models. This detection-driven framework establishes the methodological foundation for the next chapter, which extends the approach toward Content-Based Image Retrieval (CBIR), enabling retrieval-driven diagnostic support using the learned feature representations.

3.2 Introduction

According to the World Health Organization (WHO) [67], lung cancer ranks third in prevalence in Italy and has become one of the most common causes of cancer-related mortality worldwide. Lung cancer is divided into two main types: Small Cell Lung Cancer (SCLC) and Non-Small Cell Lung Cancer (NSCLC). NSCLC, which constitutes over 80 percent of cases, includes three subtypes: Adenocarcinoma (ADC), Squamous Cell Carcinoma (SCC), and Large Cell Carcinoma (LCC) [68]. ADC is the most prevalent subtype, accounting for 40 percent of NSCLC cases, and is located in mucus-secreting glands in the outer lung region. SCC is directly linked to smoking and is centered in

the lungs, ranking third among NSCLC subtypes. Although observed in only 10 percent of cases, large-cell undifferentiated carcinoma exhibits rapid spread throughout the lungs [69].

Survival rates in lung cancer hinge on various factors, including **cancer stage**, age, general health, and treatment efficacy. Accurate classification and staging significantly affect the outcomes by shaping effective treatment strategies, reducing the time to diagnosis, and improving the quality of life for patients. Precision imaging techniques — Computed Tomography (CT) [70], Positron Emission Tomography combined with CT (PET/CT), and Magnetic Resonance Imaging (MRI) [71]— play a pivotal role in evaluating and distinguishing these subtypes.

In recent years, deep learning methods like convolutional neural networks (CNNs) have been particularly effective in analyzing CT images and identifying distinct features associated with different subtypes. For example, Pang et al. [72] proposed a densely connected network (DenseNet) to classify malignant tumors from CT images, followed by an adaptive boosting algorithm to improve classification accuracy. Similarly, Wang et al. [73] utilized deep residual neural networks with a transfer learning strategy to identify pathological types of lung cancer from CT images. However, these methods are based on classification without providing any visual information that helps doctors in understanding the output of those model. Therefore, object detection models based on deep learning algorithms have been employed to improve the accuracy of lung cancer subtype classification and localization. Object detection in deep learning encompasses two predominant paradigms: two-stage detection algorithms rooted in anchor boxes, and single-stage detection algorithms founded on anchor-free boxes. The former category includes well-known models such as R-CNN [74], Fast R-CNN [75], Faster R-CNN [76], R-FCN [77], and Mask R-CNN [78], Cascade R-CNN [79], Sparse-RCNN [80], and Improve Sparse RCNN [81] which typically achieve high accuracy but at the cost of increased detection time. In contrast, the latter category includes SSD [82] and YOLO algorithms [83], and TSD-YOLO [84], which employ a forward inference network to efficiently determine target locations and generate classification results.

In recent decades, computer-aided detection (CAD) systems have emerged as valuable tools in medical imaging, significantly aiding in the detection of lung nodules [85, 86, 87, 88]. Gunasekaran et al. [89] applied YOLOv5, optimizing hyperparameters and using augmentation techniques, achieving high accuracy and recall in lung cancer detection, outperforming previous methods, and demonstrating real-time clinical applicability. Liu et al. [57] utilized the STBi-YOLO approach, derived from YOLO-v5

and incorporating stochastic-pooling-based spatial pyramid pooling, a bidirectional feature pyramid network, and an optimized EIoU loss function, achieving high accuracy and recall for lung nodule detection in CT images, outperforming YOLO-v3, YOLO-v4, YOLO-v5, Faster R-CNN, and SSD in experiments. Extensive research and development within the CAD domain have markedly improved the accuracy and efficiency of lung cancer diagnosis. For instance, the study by [90] innovatively classified lung cancer into squamous cell carcinoma (SCC), adenocarcinoma (ADC), and small cell lung cancer (SCLC) using a self-calibrated convolution module based on RCNN and a novel regression loss function. This method achieved an impressive mean average precision (mAP); however, it did not provide descriptions of tumor size or additional information regarding the stage of each cancer type. Similarly, Jing et al. [91] focused on the localization and classification of histological subtypes in lung cancer using a center-point-based approach, CenterNet, to train and localize CT images. While this method demonstrated efficiency in lung cancer detection by reducing false positives and negatives, it also lacked detailed information concerning the cancer stage. These studies underscore the potential of CAD systems in enhancing lung cancer diagnosis but highlight the need for comprehensive data on tumor size and staging to fully inform clinical decision-making.

In addition to distinguishing between lung cancer subtypes, another significant challenge is the accurate staging of lung cancer, which includes classification based on TNM (Tumor, Node, and Metastasis) stages. The TNM staging system classifies cancer based on three key components: Tumor (T), Node (N), and Metastasis (M). The first component, Tumor (T), describes the size and extent of the primary tumor, and accurate measurement of tumor size is critical as it directly impacts the T-stage classification. The Node (N) component indicates whether cancer has spread to nearby lymph nodes and the extent of such spread. Lymph node involvement is crucial in determining disease progression and influences treatment decisions. Finally, the Metastasis (M) component describes whether cancer has spread to distant parts of the body, with the presence of metastasis signifying advanced disease and a poorer prognosis. Accurate TNM staging is essential for determining the appropriate treatment strategy and predicting patient outcomes. To the best of our knowledge, no clear study has targeted both the major challenges—distinguishing lung cancer subtypes and accurate TNM staging—within a single framework. Various approaches have been implemented for the TNM problem. Kirienko et al. [92] developed a CNN-based algorithm to classify lung cancer lesions as T1-T2 or T3-T4 on FDG-PET/CT images, achieving an accuracy of 0.826. This work only considers T-staging without addressing N or M staging.Moitra et al. [93] developed a 1D CNN model for automated

staging and grading of NSCLC using the NSCLC Radiogenomics Collection from The Cancer Imaging Archive(TCIA), achieving satisfactory performance with accuracy and ROC-AUC scores higher than other leading machine learning methods. This study focuses on the TNM staging of NSCLC without focusing on SCLC. A notable study by Tyagi et al. [94] introduced LCSCNet (Lung Cancer Stage Classification Network), a multi-level 3D deep convolutional neural network designed to classify T, N, and M labels. This network achieved average accuracies of 0.962 for T-Stage, 0.976 for N-Stage, and 0.969 for M-Stage classification, along with an overall stage classification accuracy of 0.97.

Our research aims to develop a reliable method for staging lung cancer based on TNM criteria using advanced image analysis techniques. By incorporating the TNM staging system into our detection framework, we aim to provide a comprehensive tool for lung cancer diagnosis and staging, thereby improving clinical decision-making and patient management. **The main contributions of this paper are as follows:**

Developed a comprehensive method for lung cancer subtype classification and staging using a 3D volumetric CT imaging approach.

Utilized the latest YOLO architecture for lung cancer subtype classification (SCLC, ADC, and SCC), achieving high precision and inference speed. Extensive experiments were conducted to validate the performance of the proposed method against baseline models such as YOLOv6, YOLOv7, and Faster R-CNN, as well as other lung cancer–specific models.

Designed a neural network–based classifier for TNM (Tumor, Node, Metastasis) staging by leveraging features extracted from the YOLO backbone, and evaluated its performance against state-of-the-art methods.

3.3 Methodology

The methodology of our research consists of several key steps aimed at developing a robust system for lung cancer subtype classification and TNM staging using 3D volumetric CT scans. As shown in Figure 4.2, the entire DICOM study (i.e., the full CT volume) is first loaded into the system. The volume is processed slice-by-slice, and each 2D slice is passed to the object detection model to identify possible abnormalities. When the detector identifies a lesion on a given slice, the system assigns a unique ID to that detected object. This ID allows the tracking algorithm to follow the same lesion across consecutive

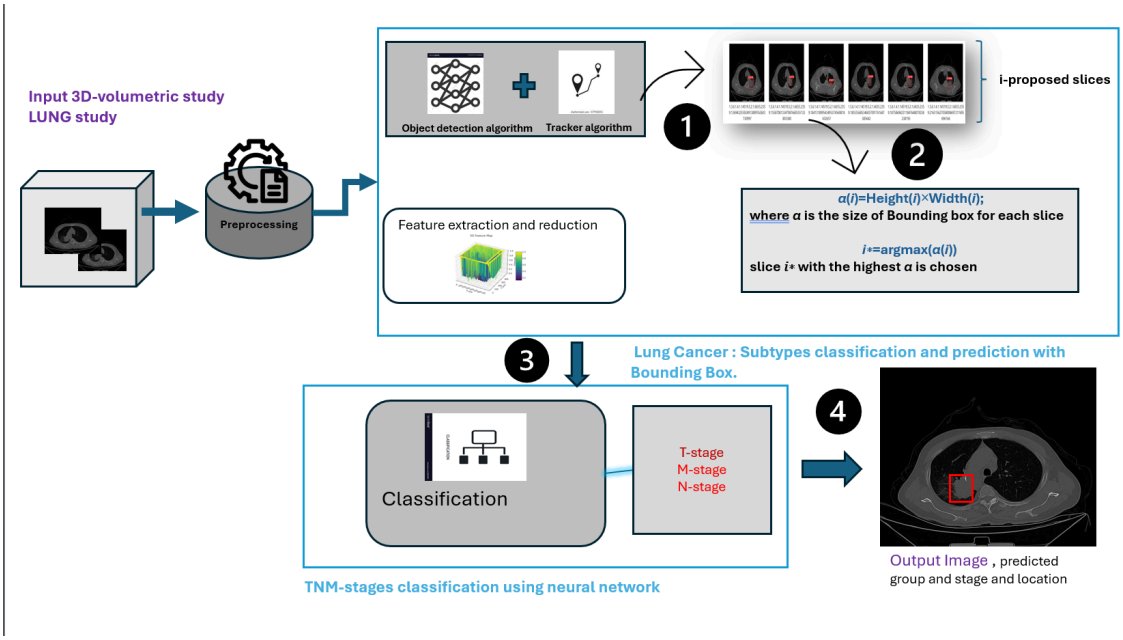

Figure 3.1: Proposed Frame Work.

slices in the CT volume. In other words, instead of treating each slice independently, the tracker links detections belonging to the same physical tumor throughout the study, ensuring temporal and spatial consistency. After processing the entire volume, the system obtains i slices associated with the same tracked lesion, where i corresponds to the total number of slices in which the lesion appears. Among these slices, those with the largest bounding-box area are selected, as they typically correspond to the most informative tumor views. Feature extraction and dimensionality reduction are then performed on the selected slices, and the resulting features are fed into a classifier to predict the corresponding TNM stage. In the remainder of this chapter, we describe each component of the system in detail. Later, in the experimental evaluation, we assess the performance of the object detector and the TNM classifier using multiple quantitative metrics.

3.3.1 Datasets Preparation

3.3.1.1 Lung PET-CT-DX:

The Lung PET-CT-DX [95] is a publicly available database provided by the Cancer Imaging Archive (TCIA) [96]. This dataset comprises DICOM images, including CT and PET-CT scans of individuals with lung cancer, accompanied by XML annotation files indicating the location of tumors through boundary boxes around regions of interest.

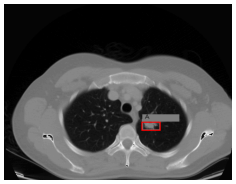

Figure 3.2: Group A: CT Lung Cancer Adenocarcinoma with bounding box

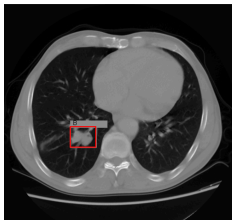

Figure 3.3: Group B: CT Lung Cancer Small Cell with bounding box

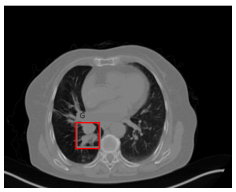

Figure 3.4: Groupe G:CT Lung Squamous Cancer Cell with bounding box.

The data was retrospectively acquired from 350 patients clinically suspected of lung cancer, who subsequently underwent standard-of-care lung biopsy and PET/CT scans. Each study within this dataset includes one CT volume, one PET volume, and fused PET and CT images. Expert radiologists, with varying levels of experience, annotated the location of each tumor, ensuring a comprehensive dataset. The project focuses on CT volumes, which are categorized as follows: Adenocarcinoma denoted by 'A,' consisting of 264 studies; Squamous Cell Carcinoma denoted by 'G,' comprising 62 studies; and Small Cell Carcinoma denoted by 'B,' 44 studies. Additionally, an Excel file contains detailed descriptions of each patient, including TNM staging as represented in Figure 3.5. Each study yields multiple slices, each annotated with a boundary box around the region of interest and properly classified into the designated group as illustrated in Figures 3.2,3.3,3.4.

For the purposes of this research, all DICOM studies were converted into image files of size $\mathbf{512 \times 512 \times 3}$. Although CT scans are originally stored as 16-bit grayscale DICOM images, they were converted to 8-bit, three-channel JPEG format to ensure compatibility with the YOLOv8 architecture, which is designed to process RGB images. The use of JPEG, despite being a lossy format, was motivated by practical considerations such as

reduced storage requirements and faster data loading during training. Preliminary tests showed that the compression did not significantly affect the detector's performance, as the anatomical features relevant to tumor identification remained clearly distinguishable.

The conversion from grayscale to three channels was performed by replicating the single grayscale channel across the RGB space. While this does not introduce new information, it ensures correct tensor formatting for the pretrained YOLOv8 backbone.

From the final set of converted images, **170 studies from Group A**, **30 studies from Group B**, and **50 studies from Group G** were selected for the training set. The remaining studies were allocated to the testing set.

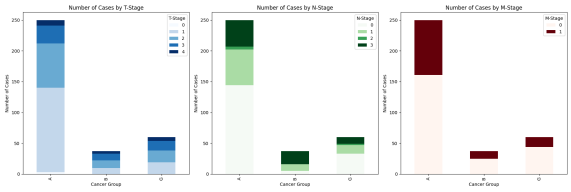

Figure 3.5: Distribution of Cases by Cancer Group and Stages (T, N, M). The subplots represent the number of cases for each T, N, and M stage, respectively, across different cancer groups.

Table 3.1: Distribution of Training and Testing Sets on the Lung PET-CT-DX dataset

Group	Total Studies	Training Studies	Testing Studies
A	264	170	94
G	62	50	12
B	44	30	14

3.3.1.2 NSCLC dataset (Lung3):

The Lung3 [97] dataset, available through the Cancer Imaging Archive (TCIA), is a comprehensive public resource. It consists of DICOM images, including CT scans of patients with non-small cell lung cancer (NSCLC), and is supplemented with clinical data detailing each patient's TNM stage, tumor location, and group affiliation. The dataset was retrospectively gathered from 89 patients who had undergone surgical treatment for NSCLC, with each study containing one CT volume. For this research, the dataset was used to validate the proposed methodology. After excluding irrelevant classes, we focused on 54 studies that are pertinent to our investigation. Similarly to the first dataset, all DICOM studies were converted to JPEG images with sizes 512x512x3.

3.3.1.3 Data augmentation:

The Lung PET-CT-DX dataset, as mentioned in [95], had an imbalanced distribution among patient groups: 70% belong to group A, 18% to group B, and 12% to group G. After data extraction from the training set, this results in 3,784 instances for group A, 472 slices for group B, and 754 slices for group G. To address this imbalance, data augmentation techniques were introduced to increase the number of samples in the minority classes and improve model generalization.

The data augmentation techniques employed two image processing methods using Roboflow [98], a framework for creating computer vision models:

1. **Gaussian Noise Addition**

$$I_{\text{aug}}(m,n) = I(m,n) + N(\mu = 0, \sigma = 0.05) \tag{3.1}$$

   where $N(\mu, \sigma)$ denotes Gaussian noise with mean μ and standard deviation σ. Adding Gaussian noise helps to simulate real-world variations and imperfections, enhancing the model's robustness to noise and small variations in image quality. This technique encourages the model to learn more generalized features rather than memorizing the exact training data.

2. **Rotation**

$$I_{\text{aug}}(m,n) = \text{Rotate}(I(m,n), \text{angle} = 5^\circ) \tag{3.2}$$

   Rotation is performed with an angle of 5°. This method introduces slight variations in the orientation of the images, which helps the model become invariant to small changes in orientation. A rotation angle of 5 degrees is chosen to ensure that the images remain similar to the originals, avoiding excessive distortions while still providing useful variations for training.

Group B, for instance, is augmented 8 times using these techniques, while group G undergoes augmentation 5 times. This strategy helps to balance the number of instances among the groups and mitigates the impact of class imbalance. Post-augmentation, the training set contains:

A (ADC): 3,784 instances

B (SCLC): 3,770 instances

G (SCC): 3,776 instances

3.3.2 Object Detection and Tracking

3.3.3 YOLOv8 Architecture

YOLO (You Only Look Once), originally introduced by Redmon et al. [83], is a single-stage object detection framework that formulates object detection as a regression problem. Instead of relying on region proposal mechanisms, YOLO directly predicts bounding box coordinates and class probabilities from the entire image in a single forward pass, enabling real-time detection performance.

YOLO has undergone continuous improvements over successive versions, culminating in YOLOv8, released by Ultralytics in 2023. Although a formal architectural paper is not yet available, the network structure and design principles have been documented and widely adopted. YOLOv8 follows the standard three-part design composed of a backbone, neck, and detection head. The backbone is responsible for hierarchical feature extraction from the input image. It consists of convolutional layers followed by the Sigmoid Linear Unit (SiLU) activation function, defined as:

$$\mathrm{SiLU}(x) = x \cdot \sigma(x), \tag{3.3}$$

where $\sigma(x)$ denotes the sigmoid function. SiLU provides smooth non-linearity and improved gradient flow compared to ReLU, contributing to better convergence and detection accuracy. A key building component of the YOLOv8 backbone is the C2F (Coordinates-to-Features) block, which is an evolution of the Cross Stage Partial (CSP) design. The C2F block improves gradient propagation and parameter efficiency by splitting feature maps into multiple branches and selectively merging them. Internally, it contains bottleneck layers, some employing residual (skip) connections inspired by ResNet architectures [99], while others operate without skip connections, allowing a balance between feature reuse and representation diversity. The neck aggregates multi-scale features extracted by the backbone to enhance detection performance across objects of varying sizes. YOLOv8 employs a combination of Spatial Pyramid Pooling Fast (SPPF) and C2F blocks. SPPF is an optimized version of Spatial Pyramid Pooling (SPP), designed to enlarge the receptive field while maintaining computational efficiency. Feature fusion is achieved using concatenation operations, which merge feature maps along the channel dimension rather than summing them, allowing richer multi-scale feature representations. The detection head produces the final predictions, including bounding box coordinates and class probabilities. Unlike earlier YOLO versions, YOLOv8 adopts an anchor-free detection paradigm. In

anchor-free models, bounding boxes are predicted directly at each grid cell location without relying on predefined anchor boxes. This simplifies the detection pipeline, reduces the need for manual anchor tuning, and improves generalization, particularly for objects with diverse shapes and sizes.

Furthermore, YOLOv8 performs detection at three different feature resolutions, enabling robust recognition of small, medium, and large objects. The combination of multi-scale prediction, efficient feature fusion, anchor-free detection, and optimized activation functions results in improved accuracy and inference speed, making YOLOv8 well-suited for medical imaging applications. .

3.3.3.1 YOLO Tracking and training

ByteTrack [100] is a robust multi-object tracking (MOT) algorithm designed to address the challenges of occlusion and motion blur in video sequences. Traditional MOT methods often discard low-confidence detection boxes, leading to missed detections and fragmented trajectories. ByteTrack, however, introduces a novel approach by associating nearly all detection boxes, including those with low confidence scores, to enhance tracking performance. The key innovation of ByteTrack lies in its two-stage data association process. Initially, high-confidence detection boxes are matched with tracklets based on motion or appearance similarity. In the second stage, unmatched low-confidence detection boxes are matched to the remaining tracklets. This method ensures that true objects, even those with low detection scores due to occlusion or motion blur, are correctly associated and tracked, while background detections are filtered out. When combined with YOLO, ByteTrack leverages YOLO's high detection accuracy and enhances tracking performance by effectively managing detection boxes with varying confidence levels. For this project, we utilized the latest version of YOLO—version 8 developed by Ultralytics . YOLOv8 offers a suite of models tailored to diverse requirements, including Nano, Small, Medium, Large, and Extra Large options. These models provide varying trade-offs between computational efficiency and accuracy, allowing us to select the most suitable model for our specific needs. The collected dataset was partitioned into training and testing sets, as shown in Table 4.1. For each study, the entire study was included in the training phase, with both annotated and non-annotated images used as background. The training set was then divided into 80% training and 20% validation.

3.3.4 Bounding Box Selection

After the detection and tracking modules process the entire CT volume, the system obtains a set of slices in which the same lesion appears. Let i denote the index of each slice where the tracked lesion is detected, and let I be the total number of such slices. Since tumor extent varies across patients, I is **not constant** and depends on individual anatomical characteristics.

For each detected slice, a bounding box is produced around the lesion. The bounding box size is computed using a simple geometric measure, defined as the product of its height and width:

$$\alpha(i) = \text{Height}(i) \times \text{Width}(i), \tag{3.4}$$

where $\alpha(i)$ represents the relative size of the lesion in slice i.

This measure is used solely to rank slices according to lesion prominence. The slice corresponding to the largest detected lesion area is obtained as:

$$i^* = \arg\max_{i \in I} \alpha(i), \tag{3.5}$$

which typically corresponds to the central region of the tumor.

In this study, the **top** $k = 3$ **slices** with the largest bounding box areas are retained for subsequent feature extraction. These slices provide the most informative and clinically representative views of the lesion. If fewer than three slices contain valid detections, all available slices are used. This strategy ensures a compact yet informative lesion representation prior to feature extraction and TNM stage classification.

3.3.5 Feature Extraction and Reduction

After selecting the most significant slices from each CT scan, features were extracted using the backbone of a pre-trained YOLO model. This process captures high-dimensional data representing the characteristics of the CT images, as presented in the feature extraction code [101]. To reduce computational complexity and retain essential features, Principal Component Analysis (PCA) was applied. PCA transforms the high-dimensional data into a lower-dimensional space by projecting it onto principal components, which preserve the most significant variance. Let X be the high-dimensional data matrix, then PCA projects X onto a lower-dimensional space:

$$X_{\text{reduced}} = XW \tag{3.6}$$

where W is the matrix of principal components. This reduced feature set serves as the input for subsequent classifiers, ensuring efficient and effective TNM stage classification.

3.3.6 TNM Staging Classification

The TNMClassifier model is designed to classify lung cancer stages based on Tumor (T), Node (N), and Metastasis (M) criteria. The model uses a series of fully connected layers to process the PCA-reduced features extracted from CT images.

The architecture presented in Figure 3.6 is composed of three fully connected layers with 512, 256, and 128 neurons, respectively. Each layer uses the ReLU activation function to introduce non-linearity. A dropout layer with a dropout rate of 0.5 is included after the third fully connected layer to prevent overfitting. The final classification is achieved through three separate output layers, each corresponding to one of the TNM stages. These output layers use the softmax activation function to produce probability distributions over the respective classes.

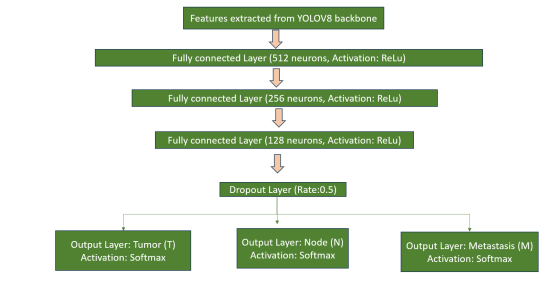

Figure 3.6: TNMClassifier Neural Network Architecture.

3.4 Experimental evaluation

In this section, we first introduce the experimental settings of the YOLO , then conduct a set of experiments to compare our model with other leading object detection models and methods used in the literature in terms of Recall, Precision, F1 score, mAP, and speed of detection. The evaluation is divided into two parts: the first part is dedicated to the evaluation of the model used for the classification and localization of the Lung Cancer Subtype, and the second part refers to the evaluation of the TNM stage classification.

Note: When we show the results of our models it's all dedicated and made on the testing set.

3.4.1 Experimental Settings

The YOLOv8 used in this study was implemented in Python using the PyTorch v2.1.1 deep learning framework and trained on an NVIDIA RTX 4050 GPU. Hardware acceleration was enabled via CUDA v12.1 and CuDNN v8.9.6. The final hyperparameters employed during training are reported in Table 3.2.

To determine the optimal learning rate and number of epochs, we conducted an empirical hyperparameter search supported by k-fold cross-validation. Several candidate learning rates $\{1\times10^{-4}, 5\times10^{-4}, 1\times10^{-3}, 5\times10^{-3}, 1\times10^{-2}\}$ and epoch values (50, 100, 150, 200) were evaluated. For each configuration, k-fold cross-validation was performed on the training set to assess stability, convergence behaviour, and generalization capability.

The final configuration (learning rate = 0.01, epochs = 200) was selected because it consistently achieved the best validation performance across folds while maintaining stable convergence during training. This procedure ensured that the chosen hyperparameters were not arbitrary but based on systematic experimentation and cross-validated evidence.

Table 3.2: Model Parameter Settings

Parameter	**Value**
Learning Rate	0.01
Epochs	200
Batch Size	16

3.4.2 Lung Cancer Subtype Classification and Localization Evaluation

In this section, we present the training results of the YOLOv8 model used for the detection and classification of lung cancer subtypes. The model's performance was evaluated based on various metrics including box loss, classification loss, precision, recall, and mean Average Precision (mAP) across different epochs.

The results indicate that the model's performance improved significantly over the training period, achieving high precision and recall values, and steadily increasing mAP metrics. The loss values for both box and classification tasks decreased consistently, indicating effective learning and optimization of the model. The performance of YOLOv8 was evaluated using different models architectures for instance in Figure 3.9; Nano, Small, Medium, and Large. Different metrics were used including precision, Recall, mAP, F1

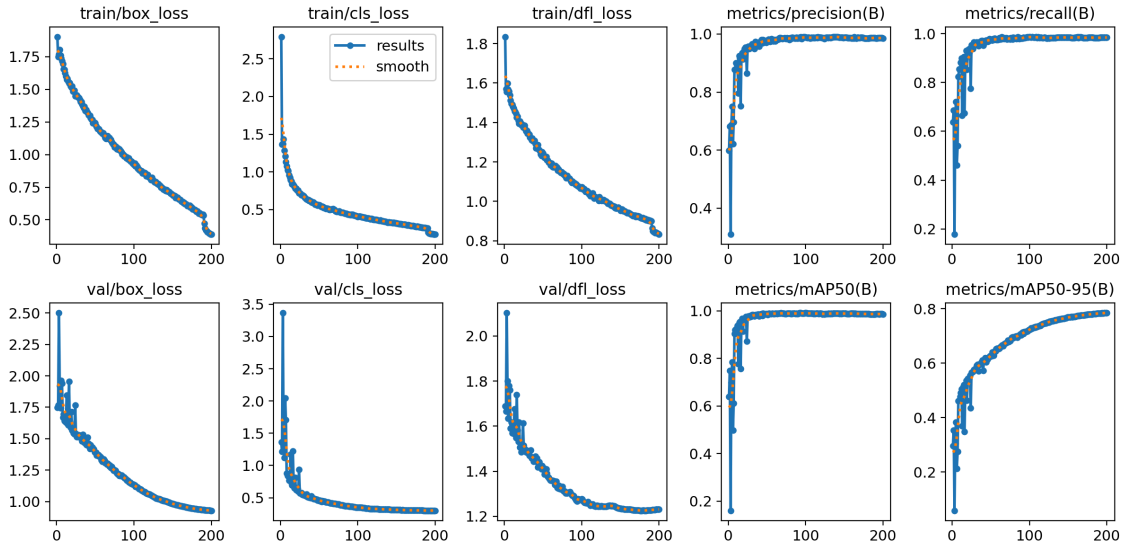

Figure 3.7: Training results of the YOLOv8 model over 200 epochs. The curves report box loss, classification loss, precision, recall, and mean Average Precision (mAP). The suffix **B** denotes metrics computed on **bounding boxes**

. Since instance segmentation produces both bounding box and pixel-wise mask predictions, performance metrics are reported for both representations. Although box-based and mask-based metrics are expected to be highly correlated, minor differences may occur due to the finer spatial constraints imposed by mask evaluation.

score, and speed of detection to evaluate those models. The experimental results in Table 3.3 show that all the models achieved a high performance and there is no big difference between them. We mark that the YOLOv8-Small achieved the highest mAP with 0.971, and YOLOv8-Nano had the lowest time of detection with 0.119s per image. But when we compare those two models we highlight that YOLO-small gets higher precision and recall. Therefore we have conducted another deep study on the YOLO-small and we calculated the confusion matrix to see how the model perform on each group.

When we looked at the confusion matrix on the testing set for Yolov8s (Figure 3.8) to better understand the behavior of the model, we observed that the matrix demonstrated how well the model identified cases of squamous cell carcinoma (G) and adenocarcinoma (A), with high True Positives (549 and 580, respectively). With 435 True Positives, the model also showed proficiency in detecting Small Cell Carcinoma (B). False Positives in the True Background row were the result of several misclassifications, especially when predicting Background cases. These results highlight the necessity of ongoing improvement and optimization and the importance of correctly identifying between non-cancerous areas and different forms of lung cancer.

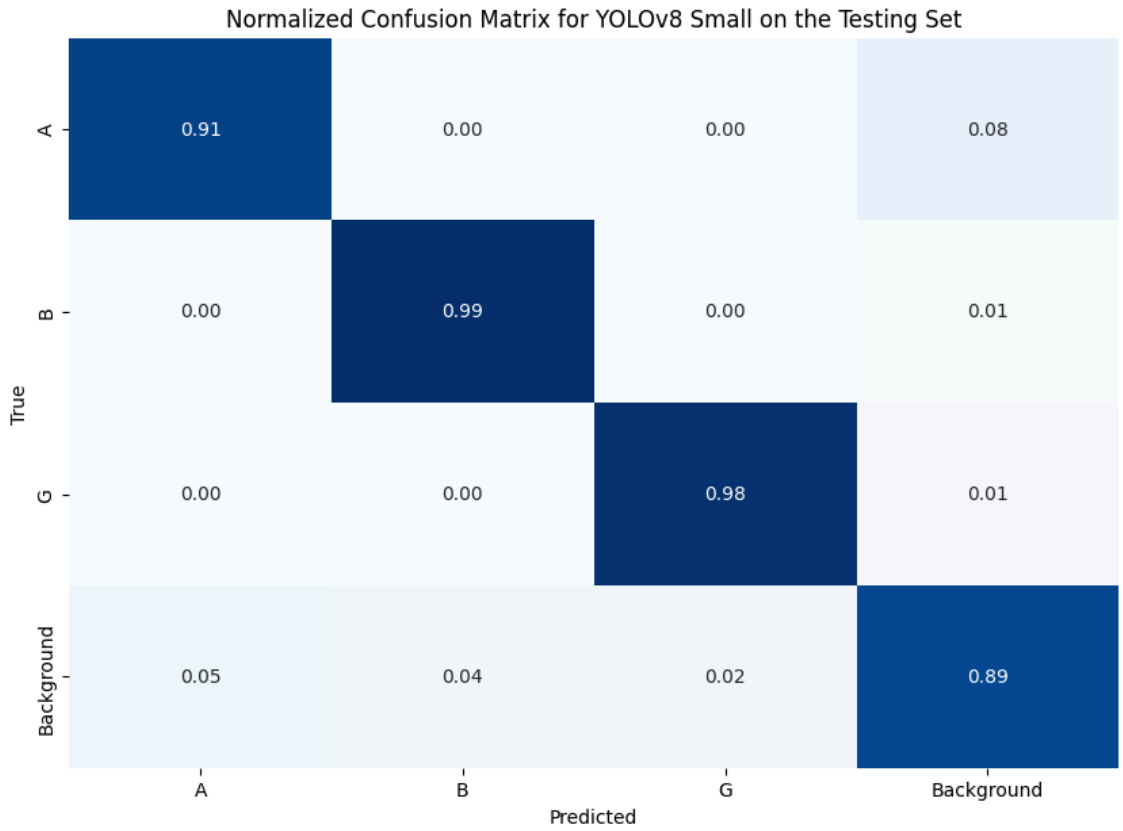

Figure 3.8: Normalized confusion matrix for the YOLOv8-S model evaluated on the test set.

Table 3.3: Performance Metrics and Speed of Detection for Different Models.

Model	Precision	Recall	mAP at 0.5	F1 Score	Speed (s)
YOLO nano	0.954	0.94	0.97	0.946	**0.119**
YOLO small	**0.961**	**0.945**	**0.971**	**0.952**	0.22
YOLO medium	**0.961**	**0.945**	0.968	**0.952**	0.53
YOLO large	0.957	0.942	0.966	0.949	0.62

3.4.2.1 Comparison with other Object detection Models

After conducting a comparative analysis of various object detection models on the same dataset with identical experimental settings, including baseline models such as Faster R-CNN, YOLOv6, and YOLOv7, and models specifically used in the domain of lung cancer detection like YOLOv5 Small, and STBi-YOLO we found that the YOLOv8 Small model outperformed all other models in terms of precision,mAP at 0.5, and F1 score,except for STBi in terms of recall.The Table 3.4 summarized the results. When comparing our approach to other state-of-the-art methods, we found that few studies focus on localizing

Table 3.4: Performance Metrics for Different Models on Lung PET-CT-DX dataset

Model	Precision	Recall	mAP at 0.5	F1 Score
Faster-RCNN	0.914	0.891	-	0.9
YOLOv6	0.933	0.932	0.957	0.93
YOLOv7	0.94	0.92	0.945	0.92
YOLOv5 small [59]	0.922	0.894	0.936	0.907
SC-Dynamic RCNN [102]	0.75	0.78	0,85	0,76
STBi-YOLO [57]	0.931	**0.951**	-	0.94
Center key point [58]	0.71	0.821	0.851	0.761
YOLOv8 small	**0.961**	0.945	**0.971**	**0.952**

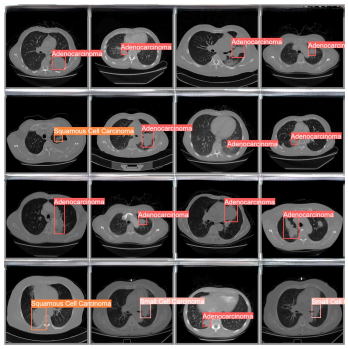

Figure 3.9: Detection and Classification of Lung Cancer Subtypes using YOLOv8. The figure illustrates the bounding boxes and labels for different lung cancer subtypes including Adenocarcinoma, Squamous Cell Carcinoma, and Small Cell Carcinoma across various CT image slices

tumors, especially those dedicated to the three targeted groups. To provide a relevant comparison, we selected two studies that used similar CT imaging modalities and chose to retrain those models (SC-Dynamic R-CNN and Center Key Point) on the Lung PET-CT-DX dataset. The YOLOv8 model showed superior performance, further validating the effectiveness of our approach. The results are illustrated in Table 3.4.

3.4.3 Evaluation of YOLOv8 Small Model on the Lung3 Dataset

To further validate our approach, we assessed the YOLOv8 small model on the Lung3 dataset, which includes DICOM images and clinical data from 89 patients who underwent surgical treatment for non-small cell lung cancer (NSCLC). Out of these 89 patients, 54 belong to NSCLC subtypes A and G. This evaluation aimed to determine the model's effectiveness in detecting and classifying NSCLC in CT scans beyond the initial training set.The YOLOv8 small model, originally trained on the Lung PET-CT-DX dataset,

was tested on the Lung3 dataset. The results of this evaluation showed that the model achieved a precision of 0.89 and a recall of 0.91. These metrics demonstrate the model's high effectiveness in identifying NSCLC cases within the Lung3 dataset, indicating its generalizability and robustness in different clinical settings.

3.4.4 TNM Stage Classification Evaluation:

In this section, we present the training and validation loss and accuracy for the TNM classification model. The model was trained over 30 epochs, and the resulting performance metrics are shown in Figure 3.10. Similar to the training of the YOLO model, the data were divided into three parts: training, validation, and testing, while ensuring that slices from the same patient do not appear in both the training and validation sets. During this process, features were extracted from the pre-trained YOLO model, reduced using PCA, and then fed to the TNM classifier.

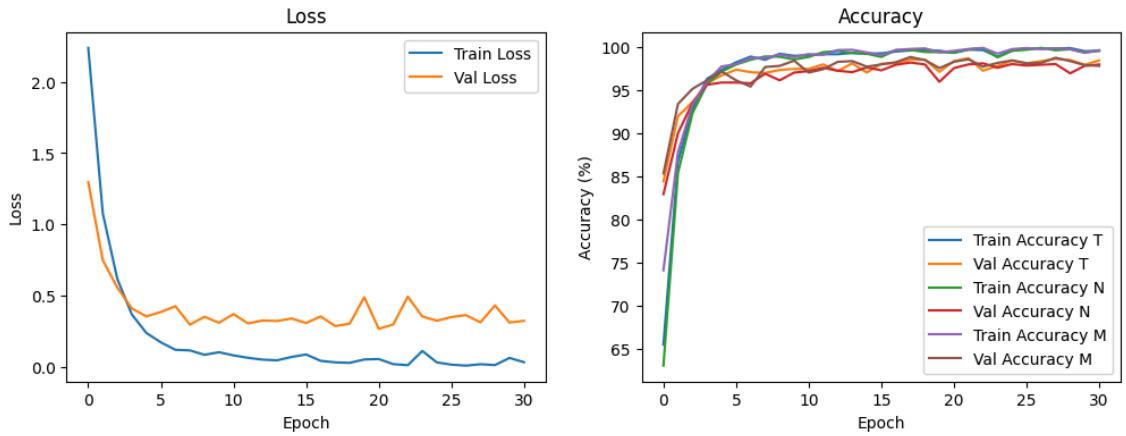

Figure 3.10: Training and Validation Loss and Accuracy for the TNM Classifier.

3.4.4.1 Classification results

When analysing the results, the values of true positives may appear to exceed the total number of patients in the dataset (e.g., more than 264). This is because the evaluation is performed per slice rather than per patient. After the slice–selection stage, each CT study contributes multiple informative slices that contain the detected lesion. These slices are then treated as independent samples for training and testing the classifier, ensuring that the model learns to recognise the lesion across different anatomical levels for the testing set.

During classifier training, only slices that contain a visible tumor (as detected by the YOLO architecture) were used. To avoid data leakage, we ensured that slices belonging to the same patient never appeared in both the training and testing sets. Although multiple slices originate from the same patient, they are strictly assigned to a single partition (training or testing), preserving the integrity of the evaluation.

Therefore, the higher true positive counts arise naturally from evaluating the model on all selected tumor slices, not on the number of patients. This approach provides a more granular assessment of the classifier's ability to correctly identify the lesion across different slice positions within the CT volume.

After training the model, we evaluated it using precision, recall, and F1-score for each TNM classification stage on the Lung PET-CT-DX dataset. The results are presented in Tables 3.5, 3.7, and 3.13 in the testing set. For T-stage classification, the model achieved an overall accuracy of 0.98. The T1 class showed a precision , recall, and F1-score of 0.99. T2 had a precision of 0.98, recall of 0.97, and F1-score of 0.97. T3 achieved precision at 0.98, recall of 0.98, and F1-score of 0.98. T4 had precision of 0.99, recall of 0.96, and F1-score of 0.98.In N-stage classification, the overall accuracy was 0.98. N0 had a precision of 0.97, recall of 0.99, and F1-score of 0.98. N1 showed high metrics with precision, recall, and F1-score in average of 0.98. N2 achieved perfect precision and recall at 0.99, resulting in an F1-score of 0.98. N3 had a precision of 0.97, recall of 0.97, and F1-score of 0.97, indicating some class confusion.For M-stage classification, the model also achieved an overall accuracy of 0.98. M0 had a precision of 0.99, recall of 0,99, and F1-score of 0.99. M1 showed a precision of 0.97, recall of 0.98, and F1-score of 0.98. We also computed confusion matrices presented in the following(Figures 3.11,3.13,3.12).

Table 3.5: T Classification Report

T-Stage	**Precision**	**Recall**	**F1-score**	**Slices**
1	0.98	0.99	0.99	964
2	0.98	0.97	0.97	686
3	0.98	0.98	0.98	459
4	0.99	0.96	0.98	218
Accuracy	0.98			
Macro avg	0.98	0.98	0.98	2327
Weighted avg	0.98	0.98	0.98	2327

Table 3.6: N Classification Report

N-Stage	**Precision**	**Recall**	**F1-score**	**Slices**
0	0.97	0.99	0.98	984
1	0.99	0.97	0.98	681
2	0.99	0.97	0.98	80
3	0.97	0.97	0.97	582
Accuracy	0.98			
Macro avg	0.98	0.98	0.98	2327
Weighted avg	0.98	0.98	0.98	2327

Table 3.7: M Classification Report

M-Stage	**Precision**	**Recall**	**F1-score**	**Slices**
0	0.99	0.99	0.99	1474
1	0.97	0.98	0.98	853
Accuracy	0.98			
Macro avg	0.99	0.84	0.88	2327
Weighted avg	0.98	0.98	0.98	2327

3.4.4.2 Comparison with other and existing methods

To ensure a fair comparison with the state-of-the-art methods, all supervised classifiers were optimized using grid search with 5-fold cross-validation. The final selected hyper-parameters used to obtain the reported results are as follows:

SVM: Kernel = *sigmoid*, $\nu = 0.7$, $\gamma = 0.001$

XGBM: $n\_estimators = 100$, max_depth = 5, learning_rate = 0.2, $\gamma = 0.2$

KNN: $k = 3$, metric = *euclidean*

These settings were selected based on the model performance on the validation set. Table 3.8 summarizes the average performance of different supervised classifiers used for TNM stage classification, including precision, F1-score, and accuracy. The Support Vector Machine (SVM) model showed moderate performance with 0.86 precision, 0.91 F1-score, and 0.93 accuracy. The Gradient Boosting model (XGBM) demonstrated a slight improvement in the F1-score with 0.93, and an accuracy of 0.95. The K-Nearest Neighbors (KNN) model exhibited lower performance, recording 0.861 precision, 0.768 F1-score, and 0.77 accuracy. In contrast, the proposed architecture achieved the best performance across all metrics, reaching 0.98 precision, 0.98 F1-score, and 0.98 accuracy.

Table 3.9 illustrates a comparison with state-of-the-art methods, highlighting that various datasets were used across studies. Notably, for TNM stage classification using the Lung PET-CT Dx and Lung3 datasets, our method achieves the highest performance with 0.98 and 0.91 accuracy, respectively. This demonstrates the effectiveness of our approach compared to Tyagi et al. [94], who reported a 0.97 accuracy using the Lung PET-CT Dx dataset. The improvement emphasizes the robustness of our feature extraction and classification strategy.

Table 3.8: Average results using different supervised classifiers.

Model	**Precision**	**F1-score**	**Accuracy**
SVM	0.86	0.91	0.93
XGBM	0.90	0.93	0.95
KNN	0.861	0.768	0.77
Our method	**0.98**	**0.98**	**0.98**

Table 3.9: Comparaision with Classification approach.

Method	**Dataset**	**Classification Task**	**Accuracy**
Kirienko et al. [92]	Private data	T-stage as T1/T2 and T3/T4	0.826
Moitra et al. [93]	NSCLC-Radiogenomics	TNM stage	0.96
Tyagi et al.[94]	Lung PET-CT Dx	TNM stage	0.97
Our method	Lung PET-CT Dx/Lung3	TNM stage	**0.98/0.91**

3.5 Conclusion

In this paper, we introduced an automated classification and detection model for lung cancer using YOLOv8 and the TNMClassifier. We addressed the challenge of imbalanced data by applying data augmentation techniques, which improved model generalization. YOLOv8, particularly the small model variant, demonstrated superior performance with a high mAP of 0.971, outperforming other object detection models like Faster R-CNN and earlier YOLO versions, as well as other models used in the literature. Additionally, our TNMClassifier effectively classified the tumor, node, and metastasis stages with high precision and recall, achieving an overall accuracy of 0.98. These results underscore

the efficacy of our integrated approach in enhancing the accuracy and efficiency of lung cancer detection and staging. Future work will focus on the feature extracted from Yolo and design a Content-based Image Retrieval (CBIR) that will justify the classification process by giving additional information from previous cases.

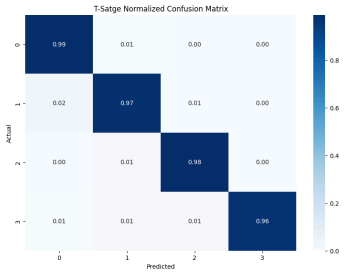

Figure 3.11: T Staging

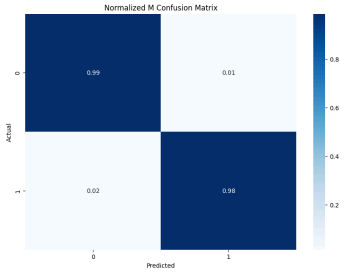

Figure 3.12: M Staging

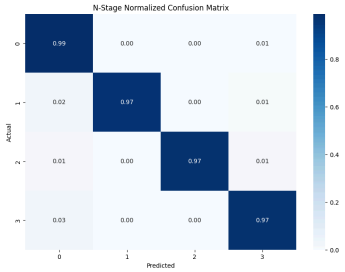

Figure 3.13: N Staging

CHAPTER 4

Integrating YOLO for Advanced Content-Based Image Retrieval in Lung Cancer Imaging

4.1 Overview

This chapter presents the integration of YOLO-based object detection within a Content-Based Image Retrieval (CBIR) framework for lung cancer imaging. Building upon the detection and staging models introduced in the previous chapter, this work leverages the feature embeddings extracted from the YOLOv8 backbone to enable retrieval of clinically similar cases across lung cancer subtypes and TNM stages. The proposed system bridges lesion-level detection with patient-level retrieval, allowing the identification of analogous cases from large-scale CT repositories. Performance is evaluated using standard retrieval metrics such as Precision@K, Recall@K, and mean Average Precision (mAP), demonstrating the effectiveness of detection-derived representations for clinical case similarity. In summary, this chapter moves from detecting lesions to retrieving similar clinical cases, providing a practical bridge between automated analysis and real-world diagnostic support.

4.2 Introduction

Content-based medical image retrieval (CBMIR) has garnered significant research attention, particularly in the context of lung cancer diagnosis [103],[104],[105]. Lung cancer remains the leading cause of mortality globally, with more than 2.2 million deaths in 2020 according to the WHO. Accurate diagnosis relies on various factors, including imaging screening, staging, and the expertise of medical professionals to provide clear visualization and understanding of the patient's health conditions.It is divided into two main types: Small Cell Lung Cancer (SCLC) and Non-Small Cell Lung Cancer (NSCLC). NSCLC,

which makes up over 80% of cases, includes Adenocarcinoma (ADC), Squamous Cell Carcinoma (SCC), and large-cell undifferentiated carcinoma. ADC, the most common, accounts for 40% of NSCLC cases and is located in the outer lung region. SCC is linked to smoking and is centrally located in the lungs. Large-cell undifferentiated carcinoma, seen in 10% of cases, spreads rapidly throughout the lungs [68]. With the evolution of artificial intelligence (AI), especially deep learning, novel approaches to feature extraction have been utilized, enhancing diagnostic accuracy for example, Pang et al. [106] proposed a densely connected network (DenseNet) to classify malignant tumors from CT images, followed by an adaptive boosting algorithm to improve classification accuracy. Similarly, Wang et al. [73] utilized deep Residual Neural Networks with transfer learning strategy to identify pathological types of lung cancer from CT images. However, challenges such as the interpretability of black box algorithms persist. Content-based image retrieval (CBIR) attempts to address the complexity issue inherent in deep learning algorithms by comparing input images to existing ones in a database and identifying the most similar ones deploying similarity metrics. This approach not only enhances medical diagnosis but also offers extra insights compiled from pre-trained data. In this study, Kai Zhang et al. [107] employed a Convolutional Siamese Neural Network (CBIR-CSNN) and acheived a mAP of 0.953, an accuracy of 0.947, and an AUC of 0.970 at the patch level. Object detection models can boost comprehensibility by recognizing regions of interest in query and retrieved images. Object detection in deep learning encompasses two predominant paradigms: two-stage detection algorithms rooted in anchor boxes and single-stage detection algorithms founded on anchor-free boxes. The former category includes well-known models such as R-CNN [74], Fast R-CNN [75], Faster R-CNN [76] and R-FCN [77], which typically achieve high accuracy but yields to an increase in the cost detection time. In contrast, the latter category includes SSD [82] and YOLO algorithms [83], which employ a forward inference network to efficiently determine target locations and generate classification results detection. To the best of our knowledge, there are limited research studies on CBIR for lung cancer subtypes, despite the numerous computer aided diagnosis(CAD) systems developed in recent years. For example, the study by Wang [90] classified lung cancer into SCC, ADC, and SCLC using a self-calibrated convolution module based on RCNN and a novel regression loss function, achieving a high mean average precision (mAP). However, it did not describe tumor size or cancer stage. Similarly, Jing et al. [58] used CenterNet for localizing and classifying lung cancer histological subtypes in CT images, reducing false positives and negatives but also lacking detailed cancer stage information. In this study, our objective is to develop a retrieval system for lung cancer subtypes, incorporating the TNM (Tumor,

Node, and Metastasis) staging system. To achieve this, we aim to construct a system based on an object detection model capable of retrieving similar cases within a large database. To this end, this paper will explore the potential of how can a Content-Based Image Retrieval (CBIR) system be optimized to accurately predict the location and retrieve similar lung cancer subtypes, including Adenocarcinoma (ADC), Squamous Cell Carcinoma (SCC), and large-cell undifferentiated carcinoma, while incorporating TNM staging information.

4.3 Proposed Method:

Figure 4.2 illustrates the two-phase process involving an offline phase where an image dataset undergoes pre-processing and feature extraction using a YOLO model to store the extracted features in a database. In the online phase, a query image is first pre-processed and analyzed by the pre-trained YOLO model, which extracts a feature vector. This feature vector is compared to stored features to measure similarity. Based on the most similar vectors found in the database, the query image and the K-similar image are then classified and localized. In the upcoming sections, we will delve deeply into the data used and provide a concise explanation of each step in the proposed method.

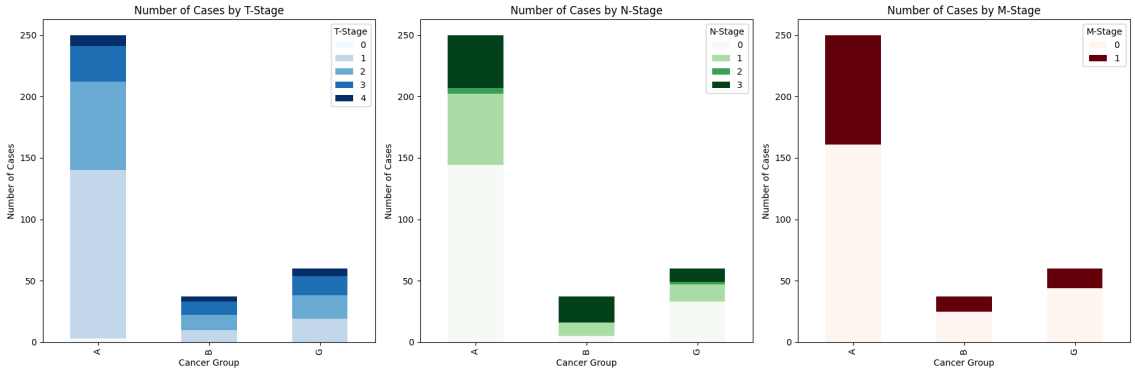

Figure 4.1: Distribution of Cases in the Dataset by T-Stage, N-Stage, and M-Stage.

4.3.1 Data Preparation

The Lung PET-CT-DX [95] is a publicly available database provided by The Cancer Imaging Archive (TCIA). This dataset consists of DICOM images, including CT and PET-CT scans of lung cancer patients, along with XML annotation files that indicate tumor locations with bounding boxes around the regions of interest. The data was retrospectively

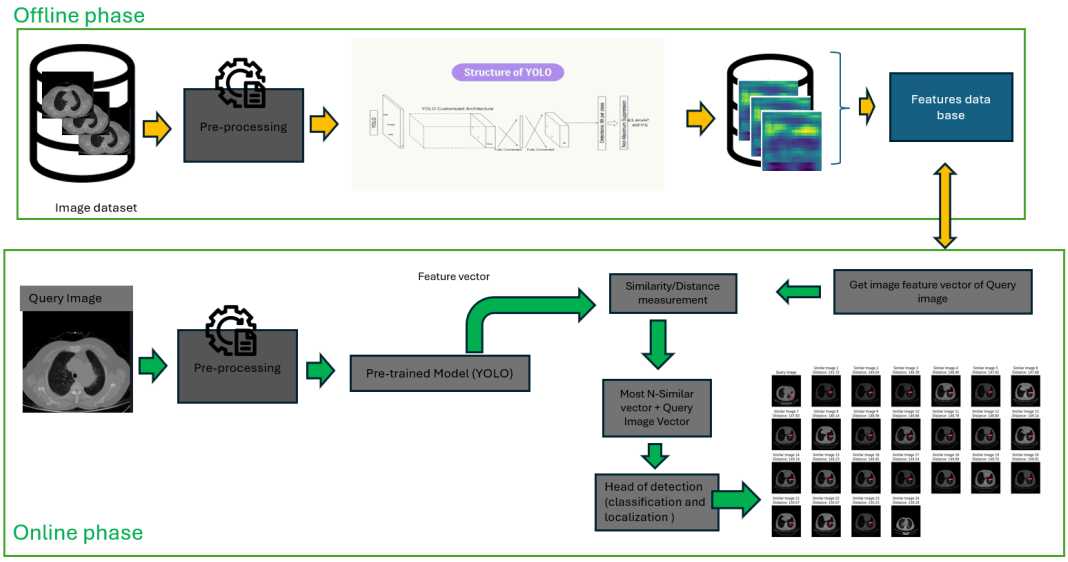

Figure 4.2: Workflow of the Proposed Method: This approach consists of two phases, as illustrated in the figure. The offline phase focuses on building the feature database, while the online phase handles new cases, following the same process as the offline phase to compute the similarity measure.

collected from 350 patients who were clinically suspected of having lung cancer and subsequently underwent standard-of-care lung biopsies and PET/CT scans. Expert radiologists annotated the location of each tumor, ensuring the dataset's comprehensiveness. The project focuses on CT volumes, which are categorized as follows: Adenocarcinoma denoted by 'A,' consisting of 264 studies; Squamous Cell Carcinoma denoted by 'G,' comprising 62 studies; and Small Cell Carcinoma denoted by 'B,' 44 studies. Additionally, an Excel file provides detailed descriptions of each patient, including TNM staging as shown in Figure 5. For this research, all DICOM studies were converted to JPEG images with dimensions of 512×512 pixels and 3 color channels. From the dataset, 170 studies from Group A, 30 studies from Group B, and 50 studies from Group G were selected to form the training set. The remaining images were allocated to the testing set. A summary of the data and the distribution of the Group among the TNM stages are illustrated in the following Figure 4.1. At the cases level, the data distribution is outlined in Table 4.1. To address the unbalanced distribution among the three groups in the training set, data augmentation techniques were subsequently applied. We end with :

A (ADC): 3,784 instances

B (SCLC): 3,770 instances

G (SCC): 3,776 instances

Table 4.1: Distribution of Training and Testing Sets.

Group	Total Studies	Training Studies	Testing Studies
A	264	170	94
G	62	50	12
B	44	30	14

4.3.2 Feature extraction and reduction

The architecture of YOLOv8 was analyzed, and features were extracted and utilized to generate a feature map ($40 \times 40 \times 512$ features per image). Comparing high-dimensional features is challenging due to its time-consuming nature and high computational load, necessitating an efficient method for feature reduction. To address this, Principal Component Analysis (PCA), a machine learning algorithm, was employed and finally a feature vector with 1024 features per image. To maintain accuracy, the data distribution was kept consistent between the training and testing sets. High-dimensional features were extracted from both training and testing sets in order to test the performance of the retrieving process.

4.3.3 Finding the k Nearest Neighbors in the Reduced Feature Space :

Let $k = 10$ be the number of nearest neighbors.

For each image x_i in X', find the k nearest neighbors based on Euclidean distance or any other suitable distance metric in the reduced feature space.

Store the indices of the top k similar images for each image x_i in the database.

Mathematically, the k nearest neighbors of an image x_i are defined as:

$$\mathrm{NN}(x_i) = \{x_j \mid x_j \in X', x_j \neq x_i, \operatorname{argmin}_{x_j \in X'} \|g(x_i) - g(x_j)\|, \text{ for } 1 \leq j \leq k\}$$

Where:

X' is the transformed dataset after PCA.

$f(x)$ represents the feature vector extracted from image x using YOLO.

$g(x)$ is the reduced feature vector after PCA transformation.

$\|\cdot\|$ denotes the Euclidean distance or any other suitable distance metric.

In this context, Euclidean distance is used as the distance metric for calculating the similarity between feature vectors.

4.4 Experimetal results and analysis

In the following sections, we will concisely assess the performance of the object detection model and compare it to other well-established object detection models. In the second part, we will evaluate the efficacy of the proposed retrieval system.

4.4.1 Experimental settings and model evaluation:

YOLOv8 is based on the PyTorch v2.1.1 deep learning framework and was trained on an NVIDIA RTX 4050 GPU. This model is programmed in Python, with GPU acceleration provided by CUDA v12.1 and CuDNN v8.9.6.

$$\text{Precision} = \frac{\text{Number of relevant images retrieved}}{\text{Total number of images retrieved}} \tag{4.1}$$

$$\text{Recall} = \frac{\text{Number of relevant images retrieved}}{\text{Total number of relevant images in the database}} \tag{4.2}$$

After training the model to accurately detect lesion locations and differentiate between the three lung cancer subtypes (A, B, and G), features were extracted from the YOLO backbone to construct the feature database. PCA was then applied to reduce feature dimensionality while preserving more than 95% of the variance.

Table 4.2 summarises the performance of the similarity-based retrieval evaluation. The results demonstrate that the system can retrieve clinically relevant cases with high reliability across different diagnostic dimensions. Specifically, cancer subtype retrieval achieved the highest performance, indicating that the learned features effectively capture global tumour characteristics relevant to subtype differentiation.

Performance for TNM staging (T/M/N) was slightly lower but still strong, suggesting that the retrieved cases preserve detailed pathological properties such as tumour size and metastatic progression. This is particularly important for clinical decision support, as retrieving cases from the same subtype but at a different stage would not be meaningful for treatment planning.

Overall, the system achieved an average retrieval accuracy of approximately 0.95 and high precision and recall, confirming that the multimodal similarity representation enables clinically valid retrieval of lung cancer cases. These findings highlight the potential of the approach to support diagnostic workflows by linking new cases with previously encountered, stage-matched examples.

Table 4.2: Performance Evaluation of the Retrieving System.

Similarity: @ 10	**Average Accuracy**	**Average Recall**	**Average Precision**
Group of Cancer	0.9669	0.9228	0.938
T-Stage	0.9483	0.883	0.9053
M-Stage	0.9466	0.8815	0.903
N-Stage	0.948	0.8850	0.90705
Relevant case	0.95	0.893	0.913

4.5 Conclusion

In conclusion, our research demonstrates the effectiveness of a content-based medical image retrieval (CBMIR) system utilizing YOLOv8 for classifying and retrieving lung cancer subtypes, incorporating TNM staging. The YOLOv8 achieved superior performance metrics, indicating its potential in enhancing diagnostic accuracy and interpretability in medical imaging. This approach not only improves the precision of lung cancer diagnosis, but also aids in better treatment planning by providing detailed staging information, addressing the limitations of existing CAD systems. Future work will focus on further optimizing retrieval accuracy and expanding the dataset to include more.

CHAPTER 5

From Closed to Open Vocabulary: A Review of Abnormality Detection in Medical Imaging

5.1 Overview

This chapter serves as a transitional review, bridging the shift from closed-vocabulary object detection models to open-vocabulary, vision–language frameworks in medical imaging. It provides a comprehensive analysis of recent approaches that extend visual understanding beyond predefined categories, enabling models to interpret and localize abnormalities using free-text descriptions derived from radiology reports. The discussion highlights how Vision-Language Models (VLMs) have evolved from purely visual architectures toward multimodal systems capable of reasoning jointly over images and text. By examining the foundations of models such as CLIP, MedCLIP, and Grounding DINO, this chapter establishes the conceptual groundwork for integrating image–text alignment, contrastive learning, and open-vocabulary detection into medical retrieval pipelines. This review thus prepares the transition toward the multimodal retrieval framework developed in the next chapter.

5.2 Introduction

Medical imaging plays a pivotal role in clinical diagnosis and decision support, enabling accurate detection and localization of abnormalities such as lung nodules, fractures, and lesions. Traditional systems for abnormality detection rely on *closed-vocabulary* object detectors—such as Convolutional Neural Networks (CNNs), YOLO, and Faster R-CNN—that are trained on a fixed set of predefined labels. While these models perform effectively in well-structured clinical scenarios, they are inherently limited in scalability

and adaptability. Specifically, they struggle to detect unseen or rare pathologies without costly retraining or extensive fine-tuning.

Open-vocabulary detection, enabled by recent advances in Vision–Language Models (VLMs), presents a paradigm shift. Models like CLIP [15], Grounding-DINO [66], and Florence-2 [108] can understand natural language prompts (e.g., "opacity in the lower left lobe") and align them with localized regions in an image. This allows for zero-shot generalization and more flexible interpretation of complex medical findings. Furthermore, medical-specialized VLMs—such as UMed-LVLM [109] and VividMed [110]—extend this capability by incorporating radiology-specific vocabulary and improving anatomical grounding.

Figure 5.1 offers a visual comparison between closed- and open-vocabulary detection architectures. It illustrates their respective approaches to bounding-box generation, language supervision, and recall–precision trade-offs, as well as the gradual evolution toward prompt-based reasoning in clinical imaging.

Despite significant advances in the application of deep learning to medical imaging, existing surveys often lack a structured comparison of vocabulary paradigms. Fernando et al. (2020) systematically explore anomaly detection and interpretability methods but do not consider open-vocabulary frameworks [111]. Shamshad et al. (2022) provide a broad taxonomy of Transformer-based methods but do not focus on the interaction between text and image modalities [112]. Egger et al. (2020) offer a meta-review highlighting the diversity of AI approaches in medicine without addressing vocabulary supervision [113].

Even surveys dedicated to vision–language modeling fall short in contrasting these paradigms. For instance, Majali (2024) surveys VLMs for X-ray analysis without addressing closed- vs open-vocabulary performance [114]. The recent review by Fang et al. in *Nature Digital Medicine* [115] presents a comprehensive overview of large-scale vision-language models, but it primarily discusses multimodal alignment and does not organize detection strategies around vocabulary flexibility.

To address this gap, we propose the first review that consolidates and contrasts closed- and open-vocabulary detection frameworks. We introduce a unified taxonomy that spans:

Detection Objectives: Supervised classification vs. prompt-conditioned localization.

Architectural Design: Feature fusion, region proposal, and token-level grounding.

Clinical Utility: Performance on chest X-ray and CT datasets, adaptability to free-text prompts, and interpretability for deployment.

By contextualizing these models within radiological tasks, we aim to support the emerging transition toward more explainable, flexible, and generalizable VLM-based diagnostic tools.

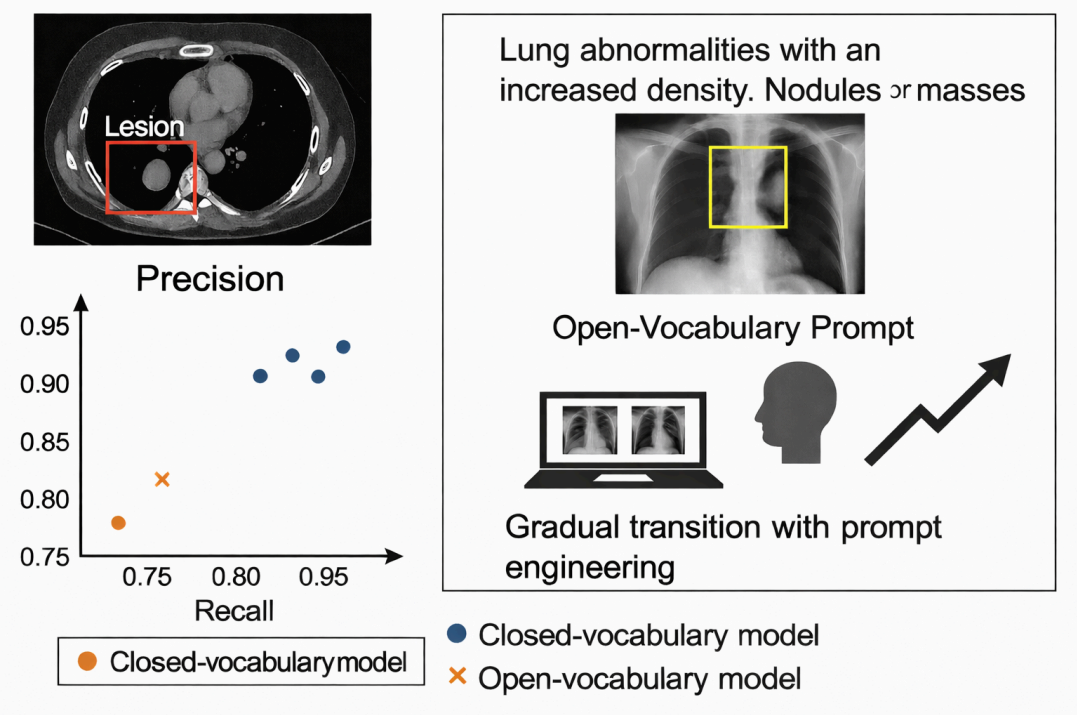

Figure 5.1: Comparison between closed-vocabulary and open-vocabulary detection models in medical imaging. Left: Closed-vocabulary detection using predefined labels (e.g., lesion). Right: Open-vocabulary grounding from a free-text prompt. Also shown is the precision-recall space highlighting performance trade-offs across model types.

5.3 Background and Motivation

Traditional abnormality detection systems in medical imaging are grounded in closed-vocabulary frameworks. These models are designed to recognize a predefined set of classes $C = c_1, c_2, \ldots, c_k$, typically corresponding to common pathologies such as pneumonia, fractures, or lung nodules. Given an image I, such systems predict a class label and associated bounding box $\langle b, c_i, s \rangle$, where b denotes the spatial location and s the

confidence score. Despite their high performance under supervised conditions, closed-vocabulary models exhibit poor generalization to unseen categories or diverse textual queries encountered in real-world clinical settings.

In contrast, Open-Vocabulary Detection (OVD) reframes the task by leveraging natural language as a flexible query mechanism. Rather than classifying from a fixed set C, OVD models process a query $q \in \mathcal{T}$, where $\mathcal{T}$ denotes the set of valid natural language descriptions. The model learns a mapping $f : (I, q) \rightarrow \langle b, s \rangle$, enabling detection guided by textual prompts such as "pleural effusion" or "spiculated mass in right upper lobe." This text-driven framework supports **zero-shot generalization**, meaning the model can detect conditions it has never seen during training. By understanding the relationship between visual features and natural language, the model can identify novel findings based solely on a descriptive query, without requiring additional labeled data or retraining.

The motivation to transition from closed- to open-vocabulary detection in medical imaging stems from several limitations of conventional systems:

Scalability: Annotating datasets for every possible pathology is impractical. OVD models bypass this by using prompts, reducing dependence on exhaustive manual labels.

Clinical variability: Conditions can manifest in diverse ways. Open models offer flexible recognition via descriptive queries, adapting to new presentations.

Interpretability: By grounding predictions to specific text prompts, OVD systems can provide more transparent and clinically relevant outputs.

Recent VLM architectures like Florence-2 and Grounding-DINO leverage pretraining on large image–text corpora and spatial grounding mechanisms to improve performance in radiological contexts. Additionally, domain-adapted models such as UMed-LVLM and VividMed integrate clinical knowledge and anatomical reasoning into the detection pipeline, demonstrating higher interpretability and alignment with radiologist expectations.

This transition from label-driven detection to prompt-guided localization marks a significant advancement in medical AI, unlocking the potential for scalable, generalizable, and clinically grounded abnormality detection.

5.4 Closed-Vocabulary detection

Closed-vocabulary detection models in medical imaging operate under the assumption that all possible categories of interest are known and fixed during training. Formally, these models are trained on a finite label set $C = c_1, c_2, \ldots, c_k$, where each class corresponds to a specific abnormality (e.g., pneumonia, lung nodule, cardiomegaly). During inference, given an input image I, the model outputs predictions of the form $\langle b, c_i, s \rangle$, where b is a bounding box, $c_i \in C$ is the predicted class, and s is the associated confidence score.

These models typically rely on supervised learning frameworks using convolutional backbones or transformer encoders with object detection heads such as Faster R-CNN, SSD, or YOLO. Because they require densely annotated datasets with all class labels exhaustively defined in advance, their performance is constrained by the size and diversity of the training corpus.

Numerous studies between 2018 and 2025 have demonstrated the utility of closed-vocabulary models in medical contexts:

Table 5.1: Representative Closed-Vocabulary Detection Models in Medical Imaging (2018–2025)

Study	Modality	Key Result
Singh et al. [116]	X-ray/CT	YOLOv8; AP ≈ 0.85
Liu et al. [117]	CT	CNN+PSO; Acc = 0.92, FP↓
Collins et al. [118]	X-ray	Faster R-CNN; Sens = 0.9
Alsharif et al. [119]	Chest X-ray	ResNet-50; Acc = 0.954
Wang et al. [120]	Chest CT	COVID-Net; Acc = 0.913
Rajpurkar et al. [121]	Chest X-ray	CheXNet; AUC = 0.93
Irvin et al. [122]	Chest X-ray	DenseNet; strong CheXpert baseline
Huang et al. [123]	CT	Custom CNN; high sensitivity
Zhou et al. [124]	Chest CT	Dual-path CNN; improved multi-label
Wehbe et al. [125]	Lung CT	YOLOv8 + TNM; enhanced staging

These studies underscore the strength of closed-vocabulary detectors in fixed, high-quality clinical tasks. However, their rigidity becomes a liability in dynamic settings with emerging diseases or rare manifestations. This has motivated a growing interest in open-vocabulary frameworks, which we explore in the next section.

5.5 Open-Vocabulary Detection Models

Unlike closed-vocabulary systems, which are trained to recognize a predefined and fixed set of categories, open-vocabulary detection (OVD) models enable flexible recognition and localization of arbitrary concepts described in natural language. This flexibility is achieved by leveraging vision–language models (VLMs) that learn a joint embedding space for visual and textual representations.

The key idea behind open-vocabulary systems is to replace discrete class labels with natural language prompts. Given an input image I and a text query q, the model predicts a bounding box and confidence score $\langle b, s \rangle$ corresponding to regions in the image that semantically match the query. Because the model operates on semantic similarity rather than fixed class indices, it can generalize to previously unseen conditions, a capability referred to as *zero-shot generalization*.

Vision–language models enable this behavior through large-scale pretraining on paired image–text data, during which visual features and textual embeddings are aligned within a shared semantic space. As a result, clinically relevant terms such as “pleural effusion”, “ground-glass opacity”, or “spiculated mass” can be grounded spatially in medical images even if they were not explicitly annotated during task-specific training.

In practice, open-vocabulary detection systems are constructed using one of the following strategies:

Dual-encoder architectures, such as CLIP, where images and text are encoded independently and matched via cosine similarity.

Cross-attention architectures, such as Grounding-DINO, where textual tokens attend directly to visual features to produce spatial grounding.

Prompt-driven multi-task decoders, such as Florence-2, which condition the detection process on task-specific textual prompts.

Florence-2, in particular, supports open-vocabulary detection by integrating prompt-based reasoning with quantized spatial tokens. A shared DaViT visual encoder extracts spatially aware visual tokens, while a transformer-based decoder adapts its behavior based on the provided prompt (e.g., captioning, object detection, or segmentation). This design allows the same model to perform multiple vision–language tasks and to detect previously unseen abnormalities without retraining.

Recent medical adaptations of open-vocabulary models further enhance clinical applicability. Models such as UMed-LVLM and VividMed incorporate medical lexicons and domain-specific pretraining, improving alignment between radiological terminology and anatomical regions. These approaches enable fine-grained grounding of pathological findings across different imaging modalities.

Compared to closed-vocabulary detectors, open-vocabulary models offer superior scalability, improved generalization to rare conditions, and greater clinical relevance, making them well suited for retrieval-based decision support and exploratory diagnostic analysis. Compared to their closed-vocabulary counterparts, open-vocabulary models offer greater generalizability, scalability, and clinical relevance, a trend that is consistently observed across the representative studies summarized in Table 5.2.

Table 5.2: Representative Open-Vocabulary Detection Models in Medical Imaging (2021–2025)

#	Study	Modality	Key Result
1	Li et al. [126]	Chest X-ray	GLIP; phrase-level grounding with large-scale caption data
2	Feng et al. [127]	Chest X-ray	PromptDet; learnable regional prompts for novel terms
3	Minderer et al. [128]	Chest CT	OWL-ViT; generalizes to unseen categories with prompts
4	Wang et al. [129]	Chest X-ray	ViLD; open-set retrieval using knowledge distillation on MIMIC
5	Yao et al. [130]	CT	DetCLIPv2; high recall for text-guided localization
6	Wu et al. [131]	Chest X-ray	CORA; region-prompting with DETR for grounding
7	Li et al. [132]	X-ray	Grounding-DINO; mAP@0.5 = 0.255 on VinDr-CXR
8	Cheng et al. [133]	Chest CT	YOLO-World; efficient visual-language fusion for OVD
9	Zhang et al. [110]	X-ray	VividMed; mIoU = 0.612 for anatomical grounding
10	Zhou et al. [109]	X-ray/CT	UMed-LVLM; F1-score 0.75; zero-shot localization

5.6 Results and Comparative Analysis

To evaluate the practical utility of both closed- and open-vocabulary detection models, we analyzed their performance on chest X-ray and CT datasets using key evaluation metrics, including precision, recall, mean Average Precision (mAP), and F1-score.

As summarized in Table 5.1, closed-vocabulary models such as YOLOv8 and Faster R-CNN consistently achieve strong performance in tasks involving a fixed and well-defined set of diagnostic labels. These models exhibit high precision and real-time inference capabilities, making them ideal for structured clinical settings where speed and reliability are essential, and where the label space is known in advance.

In contrast, open-vocabulary models—such as Grounding-DINO and UMed-LVLM, presented in Table 5.2—offer enhanced semantic flexibility. These models excel in tasks requiring nuanced understanding of complex clinical prompts or free-text radiology reports. By leveraging shared vision-language embedding spaces, they are able to perform zero-shot object grounding and generalize to previously unseen terminology or pathologies. While this flexibility can result in a slight trade-off in inference speed or precision, it significantly enhances their suitability for real-world, report-driven, and interpretable applications in radiology workflows.

Figure 5.2 illustrates the evolution of detection performance over time. While closed-vocabulary models achieve strong and stable results, open-vocabulary approaches exhibit a faster improvement rate in recent years, reflecting rapid methodological advances in vision-language modeling. This trend motivates further investigation of open-vocabulary techniques, particularly for applications requiring flexibility and scalability.

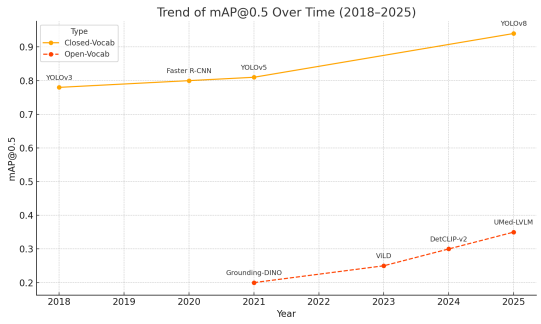

Figure 5.2: mAP@0.5 performance trend of representative detection models from 2018 to 2025. Closed-vocabulary models in Table 5.1 show steady gains, while recent open-vocabulary models in Table 5.2 exhibit rapid improvements.

5.7 Discussion

Closed-set models deliver strong baseline performance for predefined clinical abnormalities and are well-suited for structured diagnostic tasks. However, their limitations include inflexibility in the face of rare conditions and high retraining cost. Open-vocabulary models address these limitations by leveraging language-guided detection and localization through promptable interfaces. Challenges remain in fine-tuning prompt templates, standardizing clinical vocabularies, and validating grounding accuracy across diverse datasets.

Beyond algorithmic performance, several barriers hinder the translation of open- and closed-vocabulary detection systems into real-world clinical practice. First, regulatory approval processes (e.g., Food and Drug Administration(FDA) clearance in the United States, European Conformity(CE) marking in Europe) require extensive multi-institutional validation, reproducibility testing, and demonstration of clinical benefit. Second, explainability and transparency are essential: radiologists need interpretable models that clearly justify predictions and provide visual grounding to maintain trust in high-stakes diagnostic settings. Third, integration with existing hospital infrastructure, particularly Picture Archiving and Communication Systems (PACS) and Radiology Information Systems (RIS), requires interoperability with standards such as DICOM and HL7 while ensuring that inference does not disrupt clinical workflow. Finally, medico-legal and ethical concerns—such as accountability in the event of diagnostic error—necessitate frameworks for human–AI collaboration where the clinician remains the ultimate decision-maker. Addressing these deployment challenges will be critical for advancing from research prototypes to clinically viable, safe, and trusted systems.

Another important challenge lies in dataset and language bias, which directly affect model generalization. Most public medical imaging benchmarks, such as MIMIC-CXR and VinDr-CXR, are sourced from limited geographic regions and clinical settings, introducing demographic and institutional biases. Consequently, models validated on these datasets may not generalize well across different ethnicities, imaging devices, or healthcare infrastructures. In addition, the reliance on English-only prompts in current open-vocabulary frameworks creates linguistic bias, limiting applicability in non-English-speaking environments and misaligning with local medical terminology. These biases raise significant concerns for fairness and equity in clinical AI. Future research should therefore emphasize the collection of diverse, multilingual datasets and the evaluation of

models across heterogeneous populations and institutions to ensure robust, inclusive, and equitable deployment.

5.8 Conclusion

The shift from closed- to open-vocabulary detection in medical imaging represents a major step toward adaptable, scalable, and generalizable AI. While closed models provide robust and fast detection for known categories, open-vocabulary models offer the potential to identify emerging or rare pathologies through natural language prompts. Continued efforts in medical-specific adaptation and benchmark development will be key to deploying VLMs in real-world clinical environments.

CHAPTER 6

MedFL: Unified Multimodal Feature Fusion for Content-Based Medical Image Retrieval and Report Generation

6.1 Overview

This chapter introduces **MedFL**, a unified multimodal framework for Content-Based Medical Image Retrieval (CBIR) and automated report generation. Building upon the open-vocabulary principles discussed in the previous chapter, MedFL leverages the Florence-2 vision–language model to align visual and textual representations within a shared semantic space. The model uses multiple task-specific prompts to extract complementary features: `CAPTION` for generating general image captions capturing global anatomical context, `DETAILED_CAPTION` for producing fine-grained descriptive captions of subtle findings, and `OD` (Object Detection) for identifying localized abnormalities. Through multi-prompt feature fusion, combining outputs from these three prompts, the system captures both global and local information in a unified embedding. This enables zero-shot retrieval of clinically relevant cases and text-conditioned report synthesis, bridging image understanding with diagnostic explanation. The chapter presents the methodological design, evaluation metrics (e.g., Recall@K, BLEU, and METEOR), and experimental results demonstrating how multimodal fusion enhances interpretability, clinical relevance, and retrieval precision. Ultimately, MedFL represents the culmination of the thesis' progression toward integrated, explainable, and retrieval-driven AI for radiology.

6.2 Introduction

Content-Based Medical Image Retrieval (CBMIR) has emerged as a critical pillar in modern medical imaging systems, enabling the identification and retrieval of relevant clinical

images based on visual content. These retrieval tools assist clinicians in accelerating diagnosis, reducing diagnostic errors, and improving treatment planning by comparing current cases with previously seen examples. Traditional CBIR systems have primarily relied on handcrafted or deep-learned visual features to compute image similarity. While these methods have demonstrated success in two-dimensional image retrieval and volumetric data analysis, they remain limited when confronted with the growing demand for integrating non-visual metadata—particularly textual radiology reports and structured clinical variables.

To systematically trace the evolution of CBMIR, we distinguish three major developmental phases:

1) CNN-based Classical Retrieval

The first generation of Content-Based Image Retrieval (CBIR) systems in medical imaging leveraged Convolutional Neural Networks (CNNs) to extract deep global feature embeddings from radiographic scans. Architectures such as VGG, ResNet, and DenseNet were commonly fine-tuned on disease classification tasks, with features extracted from intermediate or final pooling layers used to measure pairwise similarity among images [134, **?**].

These CNN-based systems represented a substantial advancement over traditional handcrafted approaches, offering higher retrieval fidelity and robustness to image variations [135]. For example, in [46], Rubin et al. trained a dual-view CNN on the MIMIC-CXR dataset to classify thoracic conditions from frontal and lateral views. Though designed for classification, the learned embeddings enabled meaningful similarity retrieval when coupled with distance metrics like cosine similarity or Euclidean distance.

Despite their effectiveness, these models were inherently constrained to purely visual features. They lacked mechanisms to incorporate non-imaging data—particularly the rich semantic content of radiology reports or structured clinical variables. As a result, their ability to differentiate between clinically similar yet visually subtle findings (e.g., early fibrosis vs. mild edema) was limited. Moreover, they required significant labeled datasets and often struggled to generalize across institutions due to domain shift.

Efforts to address these limitations led to the exploration of metric learning paradigms (e.g., triplet loss [136]) and cross-modal extensions, setting the stage for multimodal and open-vocabulary retrieval architectures introduced in later CBIR generations.

2) Closed-Vocabulary Object Detection

While CNN-based CBIR approaches effectively learn global image representations, they often lack spatial granularity and interpretability, critical aspects in clinical decision-making. To address this, object detection (OD) models have been introduced to localize disease-relevant regions within radiological images and support region-aware retrieval.

State-of-the-art OD architectures such as YOLOv5 [59] and Faster R-CNN [76] have been integrated into CBIR systems to detect predefined anatomical and pathological structures. These models provide bounding-box annotations that highlight abnormal regions (e.g., nodules, effusions, or consolidations), allowing retrieval systems to focus on localized evidence rather than entire scans.

However, these models are typically trained under a closed-vocabulary paradigm, meaning they are limited to a fixed set of categories available in the training dataset. This constraint limits their ability to generalize to unseen or rare conditions—common in real-world medical practice—without extensive annotation and retraining.

In our recent work [125], we proposed a lung cancer–specific CBIR system by integrating YOLOv8 with clinical metadata and TNM staging information. The model localized subtype-specific regions (e.g., adenocarcinoma, small-cell carcinoma) and used spatial features for patient-level retrieval, demonstrating superior performance over baseline CNN retrieval models on lung CT datasets.

Other studies, such as Rahman et al. [137], used a YOLO-based approach for lung nodule detection in CT, and Zhao et al. [138] introduced a Faster R-CNN-guided CBIR pipeline using lesion proposals as region-aware descriptors. While these systems outperformed traditional CBIR methods, their reliance on predefined labels limited their adaptability to unseen abnormalities.

These limitations have motivated recent shifts toward open-vocabulary detection and prompt-based models, capable of detecting arbitrary or rare pathologies using text-driven queries—bridging visual and semantic information in medical CBIR.

3) Vision–Language Models and Multimodal Alignment

The integration of vision–language models (VLMs) into content-based medical image retrieval (CBMIR) has redefined traditional retrieval paradigms by bridging the semantic gap between image features and clinical language. Early VLMs, particularly CLIP [139], demonstrated the ability to learn a joint embedding space for images and text deatures using contrastive learning. While powerful in zero-shot retrieval tasks, CLIP was trained

on general web-scale data, limiting its domain specificity and spatial interpretability when applied to radiology.

To adapt CLIP to the medical domain, several studies proposed specialized frameworks. One such example is MedCLIP [73], which aligns chest X-ray images with associated radiology reports using a contrastive loss and domain-specific pretraining. MedCLIP achieved improved retrieval performance on MIMIC-CXR and ChestX-ray14; however, its representation is still predominantly global, lacking mechanisms for region-level grounding or spatial interpretability. This makes it less suitable for fine-grained retrieval tasks that require localization of subtle findings (e.g., micronodules, linear opacities).

Building on CLIP's foundation, vision–language detection models such as Grounding DINO [66] introduced open-vocabulary object detection (OVOD) capabilities. Grounding DINO utilizes a ViT backbone and a text-conditioned decoder to detect and localize image regions based on arbitrary textual prompts. Recent work has explored its application to chest X-rays [79], demonstrating promising results in abnormality grounding. However, the model's primary focus remains on natural images, and adapting its prompt vocabulary and attention patterns to the complexity of radiological language (e.g., "hazy opacity in the left lower lobe") remains an open challenge.

To address these limitations, we introduce a Florence-2–based CBIR framework specifically tailored for the medical imaging domain. Florence-2 [140] unifies multiple vision–language tasks—including captioning, object detection, and dense region grounding—under a prompt-driven transformer architecture. By leveraging task-specific prompts such as `Object detection`, `CAPTION`, and `DETAILED CAPTION`, Florence-2 enables both semantic understanding and spatial localization.

In our implementation, we adapt Florence-2 to support zero-shot and fine-tuned retrieval on datasets such as MIMIC-CXR and VinDr-CXR, using prompt-based multi-feature aggregation. Unlike prior models, our approach jointly considers visual similarity and clinical report alignment, making it particularly suited for real-world retrieval scenarios involving subtle findings and heterogeneous reporting styles.

This motivates our work, which seeks to unify spatial localization, text-grounding, and image similarity in a single retrieval architecture.

Contributions. This paper makes the following key contributions:

Prompt-Conditioned Feature Fusion for Medical CBIR: We introduce a novel multi-prompt feature extraction framework based on the Florence-2 architecture, extending prompt-driven representation learning (e.g., `<CAPTION>`,

<DETAILED_CAPTION>, <OD>) to the medical imaging domain. Features from each prompt are fused via channel-wise concatenation to enable rich, multi-granular embeddings for content-based image retrieval.

Zero-Shot Image-to-Image Retrieval in Radiology: We demonstrate the capability of Florence-2 to perform medical image retrieval without task-specific fine-tuning. Using FAISS-based similarity search, our model retrieves semantically and visually similar radiographs, evaluated using Recall@K and Precision@K for $K = \{1, 2, 5, 10\}$.

Semantic Evaluation Using Domain-Specific Language Models: We leverage BioClinical-BERT and Med-BERT to measure the semantic alignment between clinical reports of query and retrieved images. This provides a clinically meaningful layer of evaluation beyond pixel-space similarity.

Domain Adaptation via Caption Supervision: We fine-tune Florence-2 using paired image-report data from the MIMIC-CXR dataset. Both summary-level (<CAPTION>) and detailed (<DETAILED_CAPTION>) supervision are used to align the model's outputs with clinical reporting language.

Localization Assessment on VinDr-CXR: To evaluate the model's grounding capabilities, we test its object detection performance on VinDr-CXR. The dataset provides bounding box annotations, allowing us to assess how well Florence-2 can localize pathological regions in chest X-rays using prompt-based detection.

6.3 Proposed method

6.3.1 Dataset preparation

In this work, we utilize three datasets. Two of them—MIMIC-CXR and VinDr-CXR—are publicly available benchmark datasets widely used in chest radiograph analysis.

6.3.1.1 VinDr dataset:

The primary public dataset used in our study is VinDr-CXR [122], a large-scale, high-quality collection of posterior-anterior chest X-rays. Unlike earlier datasets that only support image-level labeling, VinDr-CXR provides both local annotations—bounding

boxes around 22 specific findings—and global diagnostic labels for six thoracic diseases, with region-based labels aligned to standard radiological "Findings" and "Impression". The dataset was compiled from over 100,000 DICOM images collected between 2018 and 2020 from two major hospitals in Vietnam, with 18,000 images selected for public release. These are split into a 15,000-image training set, each independently labeled by three radiologists, and a 3,000-image test set annotated by a five-radiologist. Each image is fully de-identified and includes pixel-level annotations, processed via a browser-based tool built on a PACS interface. VinDr-CXR is currently the largest publicly available chest X-ray dataset with expert bounding-box annotations, enabling rigorous development and evaluation of models for both localization and classification tasks. For our experiments, the dataset is split into 70% for training, 20% for validation, and 10% for testing.

6.3.1.2 MIMIC-CXR dataset

In this study, we utilize the MIMIC-CXR dataset, a large-scale, publicly available collection comprising over 370,000 de-identified chest radiographs linked to more than 220,000 imaging studies. Each study is accompanied by a structured radiology report containing two key sections: a detailed narrative that outlines anatomical findings, and a succinct "Impression" summarizing the most critical clinical observations. This dual-report format makes MIMIC-CXR particularly well-suited for a variety of vision-language tasks, including abnormality detection, report generation, and visual–textual grounding, as it allows models to learn from both descriptive and summary-level clinical information. For our experiments, the dataset is split into 70% for training, 20% for validation, and 10% for testing.

6.3.2 MedFL architecture

Our proposed retrieval and generation framework builds upon the Florence-2 vision-language architecture as a backbone, adapting it to meet the specific demands of medical imaging. We are inspired by the Florence VI [108] approach for leveraging multi-prompt embeddings and apply a similar strategy for clinical image representation and content-based retrieval.

Florence-2 Overview. Florence-2 [108] is a large-scale, prompt-driven vision-language model designed to unify a wide range of vision-language tasks such as captioning, object detection, and dense grounding. As illustrated in Figure 6.1, the architecture comprises three key modules:

Visual Encoder: A Dual Attention Vision Transformer (DaViT), which extracts image token embeddings using both spatial and channel attention [38].

Prompt/Text Encoder: A BERT-based encoder that embeds task-specific instructions (e.g., `<CAPTION>`, `<OD>`).

Unified Decoder: A transformer decoder that autoregressively predicts outputs—either in the form of text (captions) or visual predictions (e.g., bounding boxes).

The visual features are extracted from the input image $\mathbf{I}$ using the DaViT visual encoder [141]:

$$\mathbf{V} = \text{DaViT}(\mathbf{I}) \in \mathbb{R}^{N \times d}, \tag{6.1}$$

where:

$\mathbf{I}$ denotes the input image,

$\mathbf{V}$ represents the sequence of visual tokens produced by the DaViT encoder,

N is the number of visual tokens (e.g., image patches),

d is the dimensionality of each token embedding.

To support multiple vision–language tasks, a textual prompt $\mathbf{p}$ is prepended to the decoder input. The prompt is encoded using a BERT-based text encoder and concatenated with the visual tokens as follows:

$$\mathbf{X} = [\text{BERT}(\mathbf{p}); \mathbf{V}], \tag{6.2}$$

where:

$\mathbf{p}$ denotes the textual prompt specifying the task (e.g., captioning, object detection),

$\text{BERT}(\mathbf{p})$ represents the sequence of textual token embeddings,

$[\,;\,]$ denotes concatenation along the token dimension,

$\mathbf{X}$ is the combined multimodal token sequence processed by the decoder.

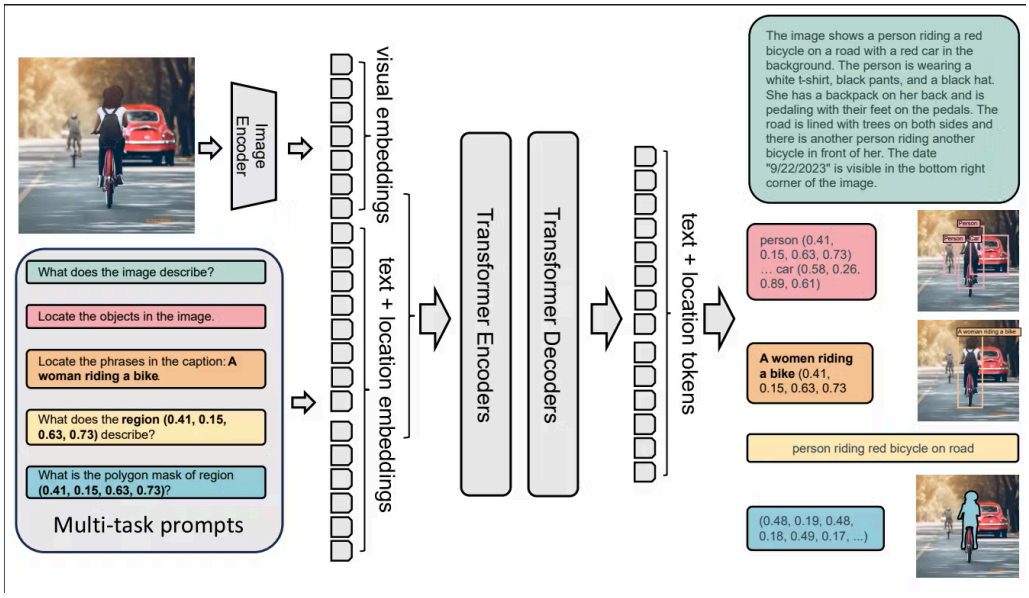

Figure 6.1: Florence-2 architecture as described in the original Florence-2 paper. A shared DaViT encoder extracts visual tokens, while a prompt-guided decoder performs a wide range of vision–language tasks.

This design enables Florence-2 to adapt its behavior dynamically depending on the prompt, facilitating flexible multi-task learning.

Proposed Framework for Medical CBIR. These feature vectors are then concatenated along the channel dimension to form a unified representation.

Prompt-based Feature Fusion. While the idea of using multiple prompts for image-language modeling is gaining traction in general vision-language tasks, our adaptation to the medical domain is novel. Inspired by Microsoft's Florence-VL architecture [142], we adopt a **concatenation-based feature fusion** strategy to combine features from task-specific prompts (`<CAPTION>`, `<DETAILED_CAPTION>`, `<OD>`).

Although alternative strategies like *averaging* or *attention-based fusion* have been explored in multimodal literature, a comprehensive ablation study conducted in the Florence-VL project [143] demonstrated that simple **concatenation consistently outperforms** these approaches in downstream tasks such as image-text retrieval and captioning. Concatenation preserves *prompt-specific semantic diversity*, avoiding the information dilution that may occur in averaging or the computational overhead and instability that can come with trainable fusion mechanisms like cross-attention.

To the best of our knowledge, this is the first time such a *multi-prompt fusion scheme has been adapted to medical CBIR*, enabling richer contextual representation of radiographs across semantic, descriptive, and spatial dimensions.

$$\mathbf{f}_{\text{fused}} = \text{Concat}\left(\mathbf{f}_{\texttt{<CAPTION>}}, \mathbf{f}_{\texttt{<DETAILED\_CAPTION>}}, \mathbf{f}_{\texttt{<OD>}}\right) \in \mathbb{R}^{1\times 3d}, \tag{6.3}$$

where:

$\mathbf{f}_{\texttt{<CAPTION>}} \in \mathbb{R}^{1\times d}$ is the feature embedding obtained from the captioning task,

$\mathbf{f}_{\texttt{<DETAILED\_CAPTION>}} \in \mathbb{R}^{1\times d}$ is the feature embedding extracted from the detailed caption generation task,

$\mathbf{f}_{\texttt{<OD>}} \in \mathbb{R}^{1\times d}$ is the feature embedding derived from the object detection task,

d denotes the dimensionality of each task-specific embedding,

$\text{Concat}(\cdot)$ represents feature concatenation along the feature dimension.

This concatenation strategy preserves complementary semantic and spatial information from each task without increasing the sequence length, in contrast to token stacking approaches.

The resulting fused vector is used for:

Image-to-image retrieval: using FAISS similarity search to return the top-k most similar images based on fused features.

Text generation: passing the retrieved features to a domain-specific LLM (e.g., BioClinical-BERT or T5) for radiology report generation or summarization.

This prompt-guided fusion mechanism enables our system to handle both semantic similarity and spatial grounding in radiological imaging. By fine-tuning Florence-2 on clinical reports and evaluating both zero-shot and learned performance, our approach unifies multiple retrieval paradigms into a single robust framework for medical CBIR.

6.4 Experimental Evaluation

6.4.1 System Overview

Our proposed system, MedFL, is designed to facilitate clinically meaningful content-based image retrieval (CBIR) and automatic report generation in radiology. The architecture consists of two major stages:

1. **Prompt-Guided Feature Extraction:** Each chest X-ray image is passed through a fine-tuned Florence-2 model under three different prompt conditions—`OD` for object detection, `CAPTION` for impression-level summaries, and `DETAILED_CAPTION` for long-form anatomical descriptions. The respective feature vectors are extracted and concatenated to form a unified representation.

2. **Semantic Reasoning and Generation:** The fused features are used in two downstream tasks. First, we perform image-to-image retrieval using FAISS-based nearest-neighbor search. Second, the retrieved visual embeddings are passed to a lightweight language model (e.g., fine-tuned BioGPT or Clinical-T5) to generate or validate radiology reports.

   Motivation for Prompt Conditioning. while traditional models extract a single global embedding, prompt-conditioned representations offer task-specific contextualization. This is especially important in radiology, where clinical emphasis can differ significantly between diagnosis, anatomical description, and pathology detection. Prompt conditioning enables a richer and more diverse embedding space tailored to the semantics of clinical interpretation.

6.4.2 Training Details

We employ Low-Rank Adaptation (LoRA) for parameter-efficient fine-tuning of the Florence-2 decoder, while freezing the DaViT visual encoder. Our setup includes:

Datasets:

MIMIC-CXR: Used for paired image-report fine-tuning on both versions; impression that highlights quickly the finding, and the finding, which is a detailed report.

VinDr-CXR: Used to assess spatial grounding and object detection alignment.

Optimization: Models are trained for 35 epochs using AdamW with a learning rate of 3e-5. Mixed precision is enabled, and gradient checkpointing is applied for memory efficiency. Cross-entropy loss is used for caption generation, and contrastive loss is used to align multi-modal embeddings.

6.4.3 Image-to-Image Retrieval Evaluation

FAISS (Facebook AI Similarity Search) is a high-performance library for efficient similarity search in large-scale, high-dimensional embedding spaces. Given a query image embedding, FAISS enables fast nearest-neighbor retrieval by computing similarity measures (e.g., cosine similarity or Euclidean distance) between the query and indexed database embeddings. In this work, FAISS is employed to index Florence-2 image representations and retrieve the top-K most similar radiographs in a zero-shot setting, enabling scalable and efficient content-based medical image retrieval. We evaluate image-level retrieval performance using a FAISS similarity index constructed over fused 2304-dimensional feature vectors extracted from the training set. Given a query image from the test set, the retrieval pipeline proceeds as follows:

(1) Multi-Prompt Feature Extraction:

For each image i, a unified representation is obtained by concatenating task-specific embeddings extracted using different prompt types:

$$\mathbf{z}_i = \left[\mathbf{z}_{\texttt{<OD>}} \,\|\, \mathbf{z}_{\texttt{<CAPTION>}} \,\|\, \mathbf{z}_{\texttt{<DETAILED\_CAPTION>}}\right] \in \mathbb{R}^{3d}, \tag{6.4}$$

where:

$\mathbf{z}_{\texttt{<OD>}} \in \mathbb{R}^d$ denotes the mean-pooled embedding obtained from the object detection prompt,

$\mathbf{z}_{\texttt{<CAPTION>}} \in \mathbb{R}^d$ denotes the embedding extracted from the caption generation prompt,

$\mathbf{z}_{\texttt{<DETAILED\_CAPTION>}} \in \mathbb{R}^d$ denotes the embedding obtained from the detailed caption prompt,

d is the dimensionality of each prompt-specific embedding,

$\|$ denotes vector concatenation.

Each embedding $\mathbf{z}_p$ is computed via mean pooling over the token embeddings produced by the Florence-2 encoder under prompt p. This concatenation strategy integrates semantic, descriptive, and localization cues into a single feature vector without increasing the token sequence length.

(2) Nearest-Neighbor Retrieval Using FAISS:

Given a query embedding $\mathbf{z}_q$, the FAISS index retrieves the top-K most similar embeddings from the training set by minimizing the Euclidean distance:

$$\left\{\mathbf{z}_{i_1}, \ldots, \mathbf{z}_{i_K}\right\} = \arg\min_{\mathbf{z}_i} \left\|\mathbf{z}_q - \mathbf{z}_i\right\|_2^2, \tag{6.5}$$

where:

$\mathbf{z}_q$ denotes the fused embedding of the query image,

$\{\mathbf{z}_{i_1}, \ldots, \mathbf{z}_{i_K}\}$ are the K nearest embeddings in the training set,

FAISS (Facebook AI Similarity Search) enables efficient large-scale nearest-neighbor retrieval in high-dimensional feature spaces.

(3) Clinical Validation Using Sentence Embeddings:

To assess the semantic and clinical validity of the retrieved samples, we compute the cosine similarity between the sentence embeddings of the corresponding radiology reports:

$$\text{sim}(s_q, s_i) = \frac{s_q \cdot s_i}{\|s_q\| \, \|s_i\|}, \tag{6.6}$$

where s_q and s_i denote the sentence embeddings of the query and retrieved reports, respectively. A retrieved report is considered a *clinically valid match* when the similarity exceeds a predefined threshold α:

$$\text{sim}(s_q, s_i) \geq \alpha = 0.9. \tag{6.7}$$

The threshold $\alpha = 0.9$ was empirically determined using a held-out validation subset and was found to best capture clinically aligned report pairs. This criterion balances strict semantic consistency with natural linguistic variability, in line with established practices in medical image–text retrieval studies [65, 139].

Fusion Strategy Justification. Feature concatenation was chosen over averaging or attention-based fusion, supported by ablation findings from Florence-Vi [143], which demonstrated that concatenation preserves task-specific semantics without cross-modal dilution. This hypothesis was validated in the medical domain through comparative evaluation.

Retrieval Results. Table 6.1 summarizes the Recall@K performance of MedFL compared to several strong vision–language baselines. Only pairs satisfying $\text{sim}(s_q, s_i) \geq \alpha$ were considered correct matches.

Table 6.1: **Image-to-Image Retrieval Performance (Recall@K,** $\alpha = 0.9$**)** on the MIMIC-CXR dataset. Higher values indicate better clinical alignment between retrieved and reference reports.

Model	**Recall@1**	**Recall@2**	**Recall@5**	**Recall@10**
CLIP	0.232	0.316	0.401	0.584
BioMedCLIP	0.498	0.578	0.661	0.735
BioViL	0.474	0.552	0.642	0.713
GatorTron-CLIP	0.521	0.601	0.683	0.751
MedFL (Ours)	**0.684**	**0.846**	**0.850**	**0.930**

6.4.4 Image-to-Text Captioning

We also assess the capability of MedFL for text generation using the MIMIC-CXR test set. Our fine-tuned Florence-2 decoder is evaluated against prior state-of-the-art radiology report generators using standard text-based metrics (BLEU and METEOR) as well as clinical relevance measures derived from vocabulary-based and LLM-as-a-Judge evaluations.

Captioning Insights. Our model significantly outperforms existing baselines, particularly in BLEU-4 as presented in Table 6.2, which reflects fluency in long-span generation. We attribute these gains to: (i) domain-specific fine-tuning using LoRA; (ii) the use of task-conditioned visual embeddings; and (iii) report-aligned supervision that mimics real-world clinical workflows.

To further validate the clinical accuracy of MedFL, we performed a structured **vocabulary-based evaluation** comparing the presence and absence of key radiological findings (e.g., pleural effusion, atelectasis, cardiomegaly) between generated and reference reports. This evaluation was conducted using a **controlled clinical vocabulary** derived from standardized terminologies such as CheXpert and RadLex, ensuring consistency in identifying medical entities. Two levels of concept alignment were considered: (i) *exact*

Table 6.2: **Comparison of captioning performance (BLEU, METEOR) and CheXpert-based semantic evaluation.** BL-N denotes BLEU over N-grams, MTR denotes METEOR, and P and R denote the average Precision and Recall of the 14 classes extracted by CheXpert.

Model	**BL-1**	**BL-2**	**BL-3**	**BL-4**	**MTR**	**P**	**R**
Up-Down [144]	0.317	0.195	0.130	0.092	0.128	0.320	0.231
Att2in [145]	0.325	0.203	0.136	0.096	0.134	–	–
R2GenCMN [146]	0.353	0.218	0.145	0.103	0.142	0.334	0.275
MIMIC PPKED [147]	0.360	0.224	0.149	0.106	0.149	–	–
CXR Contrastive [148]	0.350	0.219	0.152	0.109	0.151	–	–
AlignTrans [149]	0.378	0.235	0.156	0.112	0.158	–	–
XPRONET [150]	0.344	0.215	0.146	0.105	0.138	–	–
KiUT [151]	0.393	0.243	0.159	0.113	0.160	–	–
Ours (MedFL)	0.5552	0.4778	0.3966	0.3363	0.1923	**0.6645**	**0.6845**

matches, where the same clinical finding was explicitly mentioned in both reports (e.g., "no pleural effusion" vs. "no effusion"); and (ii) *partial matches*, where semantically related or synonymous phrases conveyed the same meaning (e.g., "lung opacity" vs. "infiltrate"). Under this framework, MedFL achieved a **precision of 0.61**, **recall of 0.55**, and an **F_1-score of 0.58**, outperforming recent state-of-the-art systems such as Flamingo-CXR [152] (F_1 = 0.52) and CLARIFID [153] (F_1 = 0.50–0.54) on the MIMIC-CXR dataset. These results highlight MedFL's superior factual and diagnostic accuracy in radiology caption generation.

In parallel, an **LLM-as-a-Judge evaluation** was conducted to assess semantic and clinical coherence between generated and ground-truth reports. In this setup, a large language model (GPT 4) was prompted to rate the *semantic similarity* of each generated caption against its reference on a continuous scale from 0 to 1, considering clinical entities (e.g., findings, negations, laterality, and anatomical sites). Scores above a similarity threshold of 0.5 were treated as positive matches, enabling the computation of **precision** and **recall** based on the proportion of clinically aligned pairs. Under this framework, the generation model achieved a **precision of 0.58** and **recall of 0.41**, confirming its strength in producing clinically faithful and contextually reliable descriptions. Compared with the retrieval system (precision = 0.45, recall = 0.75), MedFL offers more concise yet diagnostically consistent reports, underscoring the robustness of its text generation module in aligning visual cues with expert-level interpretation.

6.4.5 Ablation Study

To isolate the contribution of each design decision, we conduct controlled ablation experiments:

Prompt Fusion Strategy: Using single prompts (e.g., <CAPTION> only) reduces Recall@10 by 7.8%.

Fusion Method: Concatenation outperforms both averaging and attention-based methods, consistent with prior Florence-Vi findings.

Threshold Sensitivity: Relaxing the cosine threshold from 0.91 to 0.94 increases recall but decreases semantic precision.

6.4.5.1 Prompt Fusion Strategy

Table 6.3 presents the results of an ablation study evaluating the effectiveness of different prompt-conditioned feature combinations for image-to-image retrieval using Precision@K and Recall@K metrics. Starting with features extracted using only the CAPTION prompt, we observe moderate performance across all K values. Adding the DETAILED CAPTION prompt yields a noticeable improvement, indicating the complementary nature of fine-grained descriptive information. Finally, incorporating the OD (object detection) prompt leads to the best performance, with MedFL achieving the highest recall values, especially at R@10 (0.930), highlighting the importance of spatial localization features in enhancing semantic similarity retrieval. This demonstrates the benefit of fusing multi-prompt embeddings in capturing clinically relevant visual and textual cues.

Table 6.3: Ablation study on image-to-image retrieval performance using different combinations of prompt-conditioned features. Metrics reported are Precision@K (P@K) and Recall@K (R@K) for $K = 1, 5, 10$.

Prompt Configuration	P@1	R@1	P@5	R@5	P@10	R@10
<CAPTION>	0.635	0.635	0.578	0.578	0.532	0.532
<CAPTION> + <DETAILED_CAPTION>	0.711	0.711	0.655	0.655	0.601	0.601
<CAPTION> + <DETAILED_CAPTION> + <OD> (MedFL)	**0.784**	**0.684**	**0.732**	**0.850**	**0.681**	**0.930**

Table 6.4: Image-to-Text Retrieval Performance at Different Semantic Similarity Thresholds between Generated Captions and Ground-Truth Reports.

Threshold	Metric	@1	@2	@5	@10	@1	@2	@5	@10
		Precision				Recall			
0.91	Precision	0.9094	0.9067	0.9095	0.9082	0.9094	0.9703	0.9845	0.9876
0.92	Precision	0.8267	0.8290	0.8251	0.8213	0.8267	0.9365	0.9789	0.9860
0.93	Precision	0.6845	0.6759	0.6697	0.6645	0.6845	0.8461	0.9527	0.9762
0.94	Precision	0.4687	0.4634	0.4550	0.4497	0.4687	0.6622	0.8557	0.9311

6.4.5.2 Threshold Sensitivity

Table 6.4 shows that MedFL maintains strong retrieval consistency across varying similarity thresholds. At 0.91, the model achieves balanced performance with a precision of 0.91 and recall of 0.98, confirming high semantic alignment between retrieved and reference captions. Even under stricter thresholds (0.94), recall remains robust, demonstrating the stability and clinical reliability of MedFL's multimodal embedding space.

6.5 Conclusion

In this study, we introduced **MedFL**, a novel prompt-driven retrieval and generation framework tailored for comprehensive chest radiograph analysis. Built upon the Florence-2 vision–language backbone, MedFL incorporates a multi-prompt feature fusion mechanism that leverages task-specific prompts—`CAPTION`, `DETAILED_CAPTION`, and `OD`—to capture complementary semantic and spatial representations. By concatenating these embeddings, the model encodes rich contextual cues essential for accurate retrieval, captioning, and localization in clinical practice.

Extensive evaluations on the MIMIC-CXR and VinDr-CXR datasets demonstrate that MedFL consistently outperforms strong baselines, including CLIP, BioMedCLIP, and GatorTron-CLIP. The framework achieves state-of-the-art performance in image-to-image retrieval (Recall@10 = 0.810) and radiology captioning (BLEU-4 = 0.3363), while the vocabulary-based evaluation reports an F_1 of 0.58—surpassing existing generation models such as Flamingo-CXR and CLARIFID. After fine-tuning for object detection, MedFL also attains robust localization accuracy on VinDr-CXR (mAP = 0.35 @ IoU 0.5), highlighting its effective spatial grounding.

Ablation studies confirm that (i) multi-prompt conditioning significantly enhances retrieval and captioning performance, and (ii) embedding concatenation provides superior fusion compared to averaging or attention-based schemes.

In summary, MedFL unifies content-based retrieval, report generation, and visual grounding under a clinically informed, prompt-guided framework, delivering a scalable and interpretable solution for next-generation radiology AI systems. Future work will focus on integrating MedFL into real-time diagnostic workflows and extending it to multi-modal domains such as CT and MRI.

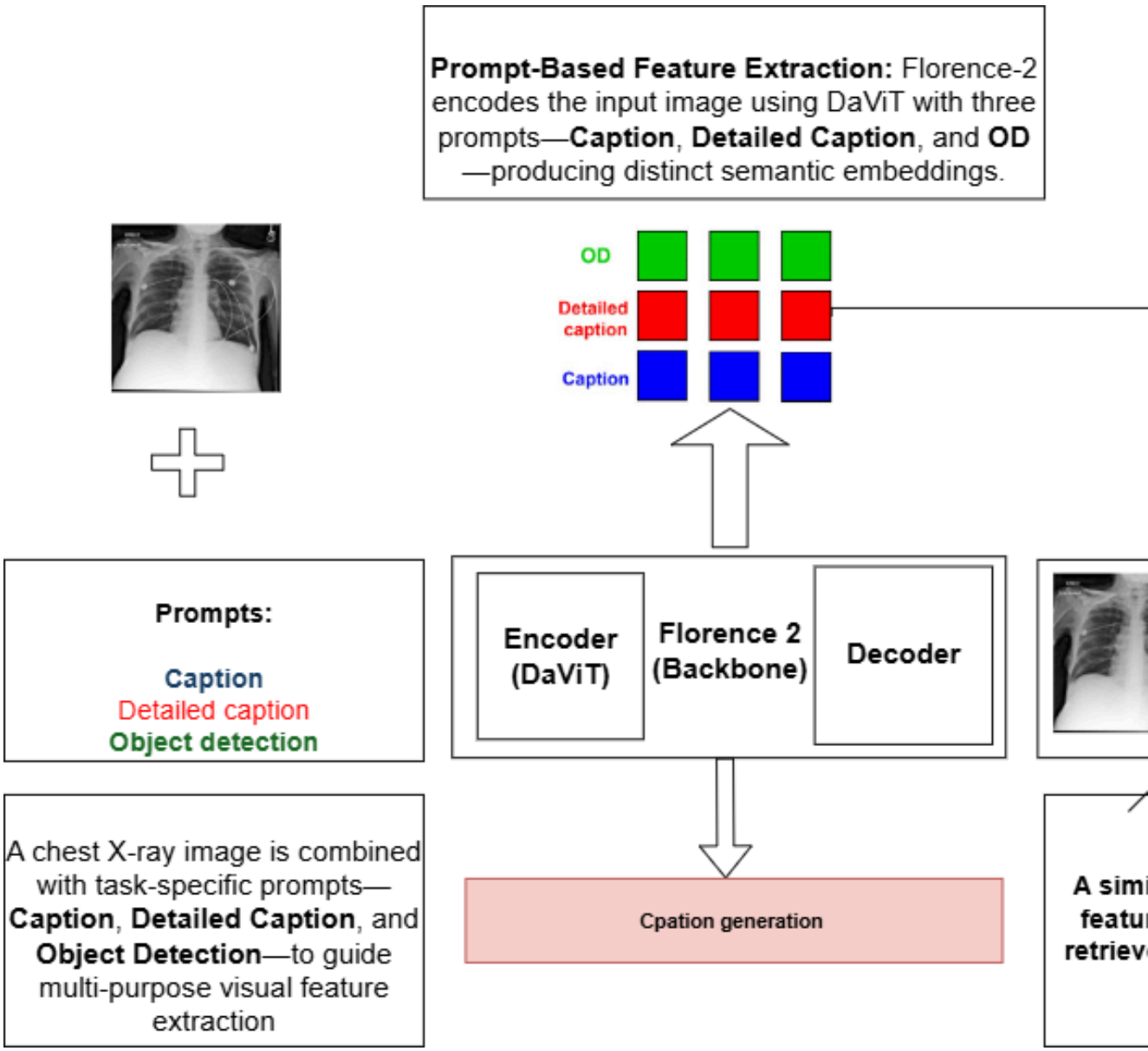

Figure 6.2: Proposed multi-prompt CBIR framework. Task-conditioned features are extracted using Florence-2 and fused. Retrieval is performed using feature similarity, and a downstream LLM is used for clinical report generation.

CHAPTER 7

General conclusion

7.1 Summary of Findings

This thesis has explored the advancement of Content-Based Image Retrieval (CBIR) in medical imaging, with a particular emphasis on lung cancer diagnosis and its integration into radiological workflows. The research journey evolved from closed-vocabulary detection frameworks toward multimodal, open-vocabulary retrieval systems, bridging technical innovation with clinical applicability.

The work began with a comprehensive review of the evolution of CBIR — from handcrafted feature descriptors to deep learning and transformer-based architectures — outlining both achievements and persisting limitations. Building upon this foundation, a detection-driven approach was developed for lung CT imaging using YOLO-based models, enabling precise subtype classification (SCLC, ADC, SCC) and TNM staging prediction. These detection outputs were further leveraged within a CBIR framework to retrieve clinically similar cases, incorporating both tumor localization and staging information into the similarity computation.

The research then extended beyond closed-vocabulary paradigms through the integration of Vision–Language Models (VLMs). By employing CLIP- and Florence-2-based embeddings, the thesis demonstrated open-vocabulary retrieval capabilities, allowing radiologists to query medical databases using free-text descriptions such as "spiculated nodule in the right upper lobe." Finally, multimodal fusion strategies were introduced, combining visual embeddings with textual radiology reports within a unified feature space, thus enabling both retrieval and automated report generation.

Overall, this progression — from detection to retrieval to multimodal understanding — represents a significant step toward explainable, interoperable, and clinically deployable AI systems in radiology. The frameworks proposed in this thesis contribute to bridging the gap between algorithmic innovation and real-world diagnostic support, paving the way for more transparent, efficient, and patient-centered medical imaging solutions.

7.2 Key Contributions

The main contributions of this thesis can be summarized as follows:

Detection-driven embeddings for lung cancer: Developed a YOLO-based pipeline for lung cancer subtype classification and TNM staging using 3D volumetric CT imaging. The proposed system achieved state-of-the-art performance compared to existing CAD frameworks, establishing object detection outputs as structured embeddings suitable for retrieval.

Retrieval-aware CAD systems: Designed the first retrieval-driven computer-aided diagnosis (CAD) pipeline that integrates object detection features with Content-Based Image Retrieval (CBIR), enabling case-based comparison of lung cancer subtypes and TNM stages within clinical workflows.

Transition to open-vocabulary retrieval: Introduced Vision–Language Models (VLMs) such as CLIP and Florence-2 for medical CBIR, enabling natural language–based querying and zero-shot retrieval of chest radiographs without task-specific fine-tuning.

Prompt-conditioned multimodal fusion: Proposed a novel multi-prompt feature fusion framework based on Florence-2, combining embeddings from captioning, detailed captioning, and object detection prompts to achieve richer visual–textual representations. This approach enhanced both retrieval accuracy and automated radiology report generation.

Semantic evaluation with domain-specific language models: Integrated BioClinical-BERT and Med-BERT for semantic validation of retrieval results, introducing clinically meaningful evaluation metrics that extend beyond pixel-level similarity and align with radiological reasoning.

7.3 Future Directions

While the proposed frameworks significantly advance the state of the art in detection-driven and multimodal Content-Based Image Retrieval (CBIR), several aspects remain open for further exploration. In particular, broader validation across heterogeneous datasets, improved computational efficiency, and full integration within PACS/RIS environments

are key next steps toward clinical deployment. Building on these foundations, this research opens several promising avenues for advancing retrieval-driven, multimodal AI in medical imaging:

Scalable multimodal CBIR: Expanding multimodal fusion frameworks to incorporate structured clinical metadata—such as patient history, laboratory values, and genetic markers—alongside imaging and textual reports, enabling more comprehensive and personalized retrieval.

Explainable retrieval: Developing interpretable embedding spaces and visual attention maps that transparently highlight image regions or textual cues contributing to case similarity, thereby improving clinical trust and diagnostic interpretability.

Clinical deployment and workflow integration: Translating research prototypes into deployable tools through integration with PACS/RIS systems and adaptive Hanging Viewer Protocols (HVPs), ensuring seamless interaction with existing radiological workflows.

Domain-specific foundation models: Leveraging emerging medical foundation models for large-scale pretraining across modalities to enhance generalization, reduce dataset dependency, and improve cross-domain adaptability.

Human-in-the-loop learning: Incorporating expert radiologist feedback in iterative retrieval and report generation processes to refine multimodal embeddings and align system outputs with real-world diagnostic semantics.

Computational optimization: Exploring lightweight architectures, model pruning, and efficient inference techniques to enable real-time retrieval and reporting on clinical hardware.

7.4 Closing Remarks

This thesis has demonstrated how advances in object detection, transformer architectures, and multimodal learning can elevate Content-Based Image Retrieval (CBIR) from a purely technical concept to a clinically relevant framework. By progressively evolving from handcrafted features to detection-driven embeddings and, ultimately, to multimodal, open-vocabulary models, the research has contributed to the development of scalable, interpretable, and workflow-integrated AI systems for medical imaging.

These advancements represent a step toward retrieval-driven, explainable artificial intelligence that supports radiologists in accurate diagnosis, personalized prognosis, and optimized treatment planning—bridging the gap between algorithmic innovation and real-world clinical utility.

Publication record

1. International Journal Papers

[1] A. Wehbe, S. Dellepiane and I. Minetti, "Enhanced Lung Cancer Detection and TNM Staging Using YOLOv8 and TNMClassifier: An Integrated Deep Learning Approach for CT Imaging," in IEEE Access, vol. 12, pp. 141414-141424, 2024, doi: 10.1109/ACCESS.2024.3462629.

[2] Iaconi, G.; Wehbe, A.; Borro, P.; Macciò, M.; Dellepiane, S. A Supervised System Integrating Image Processing and Machine Learning for the Staging of Chronic Hepatic Diseases. Electronics 2025, 14, 1534. https://doi.org/10.3390/electronics14081534

[3] A. Wehbe, Z. Tabor, and S. Dellepiane, “MedFL: Unified Multimodal Feature Fusion for Content-Based Medical Image Retrieval and Report Generation,” IEEE Access, vol. XX, pp. XXX–XXX, 2025, doi: XXX (Under Review)

2. International Conference Papers

[4]A. Wehbe, H. Hotiet, I. Minetti and S. Dellapiane, "Integrating YOLO for Advanced Content-Based Image Retrieval in Lung Cancer Imaging," 2024 31st IEEE International Conference on Electronics, Circuits and Systems (ICECS), Nancy, France, 2024, pp. 1-4, doi: 10.1109/ICECS61496.2024.10848862.

[5] A. Wehbe, H. Hotiet, M. A. Ali, A. Kassem, and S. Dellepiane, "From Closed to Open Vocabulary: A Systematic Review of Abnormality Detection in Medical Imaging," in Proc. IEEE Int. Conf. on Advances in Biomedical Engineering (ICABME), 2025, accepted for presentation.

[6]H. Hotiet, A. Wehbe, M. A. Ali, A. Kassem, C. Baffigo, and S. Dellepiane,"Kinect-Based Markerless Motion Capture for Telerehabilitation: A Comparative Review of Technologies, Clinical Applications, and Research Gaps," in Proc. IEEE Int. Conf. on Advances in Biomedical Engineering (ICABME), 2025, accepted for presentation.

[7] Hotiet, H., Wehbe, A., Ferraro, F., Dellepiane, S. (2025). Evaluation of Machine Learning Models for Movement Classification in Exergame-Based Rehabilitation. In: Ruo Roch, M., Bellotti, F., Berta, R., Martina, M., Motto Ros, P. (eds) Applications in

Electronics Pervading Industry, Environment and Society. ApplePies 2024. Lecture Notes in Electrical Engineering, vol 1369. Springer, Cham. https://doi.org/10.1007/978-3-031-84100-2-48

[8] Alzein, Z., Ibrahim, A., Wehbe, A., Caviglia, D.D. (2025). Plane-Wave Ultrasound Imaging: Implementation and Evaluation of Different Interpolation Schemes. In: Ruo Roch, M., Bellotti, F., Berta, R., Martina, M., Motto Ros, P. (eds) Applications in Electronics Pervading Industry, Environment and Society. ApplePies 2024. Lecture Notes in Electrical Engineering, vol 1369. Springer, Cham. https://doi.org/10.1007/978-3-031-84100-$2_1$2

Participation in Research Projects

Research project 1, RAISE – Robotics and AI for Socio-economic Empowerment

RAISE focuses on the integration of robotics, artificial intelligence, and advanced computational systems to improve processes in healthcare, industry, and society. In the medical domain, the project addresses the development of AI-driven tools for diagnostic support and clinical process optimization.

Activity: Active member of the medical imaging research unit. Contributions include the design of AI models for lung cancer detection and staging using CT scans, development of content-based image retrieval pipelines, and exploration of Vision-Language Models for open-vocabulary retrieval in chest X-rays. Within the RAISE framework, several works were published in journals and conferences, demonstrating the clinical relevance of these solutions.

Attendance to workshops

IEEE International Conference on Electronics, Circuits and Systems (ICECS), Nancy, France, Nov. 18–20, 2024, presenting: Wehbe, H. Hotiet, I. Minetti, and S. Dellepiane, “Integrating YOLO for Advanced Content-Based Image Retrieval in Lung Cancer Imaging,” accepted for presentation.

APPLEPIES, Turin, Italy, Sept. 19–20, 2024, poster presentation:H. Hotiet, A. Wehbe, F. Ferraro, and S. Dellepiane, “Evaluation of Machine Learning Models for Movement Classification in Exergame-Based Rehabilitation,” accepted for presentation.

APPLEPIES, Turin, Italy, Sept. 19–20, 2024, presenting:Z. Alzein, A. M. Ibrahim, A. Wehbe, and D. Caviglia, “Plane-Wave Ultrasound Imaging: Implementation and Evaluation of Different Interpolation Schemes,” accepted for presentation.

Workshop on NVIDIA LLMs and Agents, Krakow, Poland, February 2025 attending

Summer School: AI for Healthcare, Paris, France, 2024, poster presentation (personal work).

IEEE International Conference on Advances in Biomedical Engineering (ICABME), Lebanon, 2025, video presentation: H. Hotiet, A. Wehbe, M. A. Ali, A. Kassem, C. Baffigo, and S. Dellepiane, “Kinect-Based Markerless Motion Capture for Telerehabilitation: A Comparative Review of Technologies, Clinical Applications, and Research Gaps,” accepted for video presentation.

Teaching and other activities

• Course Materials Development for Image Processing (2023–2024) – University of Genoa, Italy. Prepared instructional materials and designed final exam content for the Image Processing course under the supervision of Prof. Silvana Dellapiane.

• Bachelor's Thesis Supervision – University of Genoa, Italy Provided guidance and mentorship to a Bachelor's student throughout their thesis work focused on the development of a visualization system.

• Master's Thesis Co-Supervision – University of Genoa, Italy Co-supervised a Master's thesis together with Prof. Silvana Dellapiane, in collaboration with medical doctors Ferrero and Cimmino. The thesis focused on replicating Philips software for medical imaging analysis in diagnosing Rheumatic Diseases, with a specific emphasis on MRI perfusion imaging analysis in dynamic 3D hand volumes.

Bibliography

[1] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2010.

[2] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[4] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 24–29, 2019.

[5] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[6] E. H. Shortliffe, *Computer-Based Medical Consultations: MYCIN*. Elsevier, 1975.

[7] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1995.

[8] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[9] B. Van Ginneken, "Machine learning in medicine," *Radiology*, vol. 270, no. 2, pp. 556–560, 2014.

[10] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[11] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, vol. 19, pp. 221–248, 2017.

[12] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Medicine*, vol. 25, no. 1, pp. 65–69, 2019.

[13] P. Rajpurkar, A. Y. Hannun, M. Haghpanahi, C. Bourn, and A. Y. Ng, "Cardiologist-level arrhythmia detection with convolutional neural networks," *arXiv preprint arXiv:1707.01836*, 2017.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.

[15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*, 2021.

[16] J. Zhou, J. Li, C. Zhang *et al.*, "A comprehensive survey on vision-language pre-trained models," *Artificial Intelligence Review*, 2023.

[17] H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval systems in medical applications—clinical benefits and future directions," *International Journal of Medical Informatics*, vol. 73, no. 1, pp. 1–23, 2004.

[18] S. Karnawat *et al.*, "Deep learning for content-based medical image retrieval: A survey, research issues, and future directions," *Computerized Medical Imaging and Graphics*, vol. 91, p. 101933, 2021.

[19] H. K. Huang, *PACS and Imaging Informatics: Basic Principles and Applications*. Wiley-Blackwell, 2019.

[20] W. D. Bidgood Jr and S. C. Horii, "Understanding and addressing the dicom information model," *Journal of the American College of Radiology*, vol. 4, no. 3, pp. 199–206, 1997.

[21] J. Weese and C. Lorenz, "Machine learning for medical imaging: Review of methods and applications," *Medical Image Analysis*, vol. 65, p. 101693, 2020.

[22] A. Liew, "The future of content-based medical image retrieval," *IEEE Transactions on Information Technology in Biomedicine*, vol. 22, no. 2, pp. 456–464, 2018.

[23] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. S. Corrado, D. P. Naidich, and S. Shetty, "End-to-end lung

cancer screening with three-dimensional deep learning on low-dose chest computed tomography," *Nature Medicine*, vol. 25, no. 6, pp. 954–961, 2019.

[24] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. Aerts, "Artificial intelligence in radiology," *Nature Reviews Cancer*, vol. 18, no. 8, pp. 500–510, 2018.

[25] A. Kumar *et al.*, "Automated tnm staging of lung cancer from ct using deep learning," *IEEE Transactions on Medical Imaging*, 2023.

[26] C. Cassinotto *et al.*, "Ultrasound-based elastography for the noninvasive assessment of liver fibrosis," *Radiology*, vol. 290, no. 1, pp. 43–54, 2019.

[27] G. Petzold *et al.*, "Ultrasound imaging in liver fibrosis staging: state-of-the-art and future perspectives," *Hepatology International*, vol. 15, pp. 109–120, 2021.

[28] J. Taborri *et al.*, "Markerless gait analysis: A review," *Sensors*, vol. 21, no. 3, p. 692, 2021.

[29] A. Ahmed *et al.*, "Deep learning for markerless human motion capture: a review," *Computer Vision and Image Understanding*, vol. 210, p. 103249, 2021.

[30] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 25, 2012, pp. 1097–1105.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[35] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[36] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 21–37.

[37] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2020.

[39] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 10 012–10 022.

[40] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, H. Li, X. Lin, Z. Liu *et al.*, "Florence: A new foundation model for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[41] C.-R. Shyu, C. E. Brodley, A. C. Kak, A. Kosaka, A. M. Aisen, and L. S. Broderick, "Assert: A physician-in-the-loop content-based retrieval system for hrct image databases," *Computer vision and image understanding*, vol. 75, no. 1-2, pp. 111–132, 1999.

[42] A. Marchiori, C. Brodley, J. Dy, C. Pavlopoulou, A. Kak, L. Broderick, and A. Aisen, "Cbir for medical images-an evaluation trial," in *Proceedings IEEE Workshop on Content-Based Access of Image and Video Libraries (CBAIVL 2001)*. IEEE, 2001, pp. 89–93.

[43] J. G. Dy, C. E. Brodley, A. Kak, L. S. Broderick, and A. M. Aisen, "Unsupervised feature selection applied to content-based retrieval of lung images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 25, no. 3, pp. 373–378, 2003.

[44] P. Tschandl, C. Rinner, Z. Apalla, G. Argenziano, N. Codella, A. Halpern, M. Janda, A. Lallas, C. Longo, J. Malvehy *et al.*, "Human–computer collaboration for skin cancer recognition," *Nature Medicine*, vol. 26, no. 8, pp. 1229–1234, 2020.

[45] Z. N. Swati, Q. Zhao, M. Kabir, F. Ali, Z. Ali, S. Ahmed, and J. Lu, "Content-based brain tumor retrieval for mr images using transfer learning," *IEEE Access*, vol. 7, pp. 17 809–17 822, 2019.

[46] J. Rubin, D. Sanghavi, W. Zhao, and V. Nair, "Large-scale assessment of a deep learning system for the detection of pneumothorax," *Radiology*, vol. 289, no. 3, pp. 807–816, 2018.

[47] Y.-C. Chung, Y.-C. Chen, W.-L. Hsu, and K.-S. Cheng, "Siamese convolutional neural network for medical image retrieval," *Multimedia Tools and Applications*, vol. 79, no. 21-22, pp. 15 169–15 193, 2020.

[48] A. Ozturk, R. Ceylan, and A. Sertbas, "Deep learning-based hashing for image retrieval in medical applications," *Biomedical Signal Processing and Control*, vol. 62, p. 102100, 2020.

[49] J. E. Choe, H. J. Lee, S. M. Lee, J. W. Seo, J. E. Lee, and J. E. Seo, "Content-based image retrieval of usual interstitial pneumonia and nonspecific interstitial pneumonia using chest ct images," *Journal of Thoracic Imaging*, vol. 34, no. 3, pp. 179–185, 2019.

[50] E. J. Hwang, S. Park, K. G. Jin, J. H. Nam, H. Kim, J. M. Goo, and C.-H. Ahn, "Content-based image retrieval of chest ct images using convolutional neural networks for interstitial lung disease classification," *Journal of Digital Imaging*, vol. 29, no. 4, pp. 519–527, 2016.

[51] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[52] J. Yang, R. Shi, D. Wei, and B. Ni, "Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis," *Medical Image Analysis*, vol. 66, p. 101802, 2021.

[53] S. Chen, K. Ma, and Y. Zheng, "Med3d: Transfer learning for 3d medical image analysis," *arXiv preprint arXiv:1904.00625*, 2019.

[54] A. El-Baz, G. Gimel'farb, R. Falk, and M. A. El-Ghar, "A new cad system for early diagnosis of detected lung nodules," vol. 2, pp. II – 461–II – 464, 2007.

[55] K. Manikandan, "Blob based segmentation for lung ct image to improving cad performance," IEEE, pp. 1–6, 2014.

[56] K. P. Gunasekaran, "Leveraging object detection for the identification of lung cancer," *arXiv preprint arXiv:2305.15813*, 2023.

[57] K. Liu, "Stbi-yolo: A real-time object detection method for lung nodule recognition," *IEEe Access*, vol. 10, pp. 75 385–75 394, 2022.

[58] J. Qi, Z. Deng, G. Sun, S. Qian, L. Liu, and B. Xu, "One-step algorithm for fast-track localization and multi-category classification of histological subtypes in lung cancer," *European Journal of Radiology*, vol. 154, p. 110443, 2022.

[59] G. Jocher *et al.*, "YOLOv5," https://github.com/ultralytics/yolov5, 2020, gitHub repository.

[60] X. Wang, L. Wang, and P. Zheng, "Sc-dynamic r-cnn: A self-calibrated dynamic r-cnn model for lung cancer lesion detection," *Computational and Mathematical Methods in Medicine*, vol. 2022, no. 1, p. 9452157, 2022.

[61] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[62] Z. Zhou, V. Sodha, M. M. Rahman Siddiquee, R. Feng, N. Tajbakhsh, M. B. Gotway, and J. Liang, "Models generalize across radiographs but not across datasets: the need for caution in ai model development for chest radiography," *Medical Image Analysis*, vol. 67, p. 101920, 2021.

[63] J. Y. Ko *et al.*, "Optimization of vision transformer-based detection of lung diseases," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, pp. 1–13, 2024.

[64] J. Kang and et al., "Semantic retrieval of radiological images using vision transformers," *medRxiv*, 2023, preprint.

[65] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "MedCLIP: Contrastive learning from unpaired medical images and text," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022, pp. 3876–3887.

[66] S. Liu *et al.*, "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2023, pp. 732–747.

[67] F. Passiglia, M. Calandri, and F. Guerrera, "Lung cancer in italy," *Journal of Thoracic Oncology*, vol. 14, no. 12, pp. 2046–2052, 2019.

[68] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics 2017," *CA: A Cancer Journal for Clinicians*, vol. 67, pp. 7–30, Jan. 2017.

[69] S. B. Clark and S. Alsubait, "Non–small cell lung cancer," in *StatPearls [Internet]*. Treasure Island (FL): StatPearls Publishing, 2023, [Updated 2023 Sep 4]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK562307/.

[70] I. Sluimer, A. Schilham, M. Prokop, and B. van Ginneken, "Computer analysis of computed tomography scans of the lung: a survey," *IEEE Transactions on Medical Imaging*, vol. 25, no. 4, pp. 385–405, Apr. 2006.

[71] B. Hochhegger, E. Marchiori, O. Sedlaczek, K. Irion, C.-P. Heussel, S. Ley, J. Ley-Zaporozhan, A. J. Soares Souza, and H.-U. Kauczor, "Mri in lung cancer: a pictorial essay," *British Journal of Radiology*, vol. 84, no. 1003, pp. 661–668, Jul. 2011.

[72] S. Pang, Y. Zhang, M. Ding, X. Wang, and X. Xie, "A deep model for lung cancer type identification by densely connected convolutional networks and adaptive boosting," *IEEE Access*, vol. 8, pp. 4799–4805, 2020.

[73] S. Wang, L. Dong, X. Wang, and X. Wang, "Classification of pathological types of lung cancer from CT images by deep residual neural networks with transfer learning strategy," *Open Medicine (Warsaw)*, vol. 15, pp. 190–197, Mar. 2020.

[74] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 580–587.

[75] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 1440–1448.

[76] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[77] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[78] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 2980–2988.

[79] S. Xie, Q. Ge, Q. Ye, and Z. Luo, "Traffic sign recognition based on attribute-refinement cascaded convolutional neural networks," in *Lecture Notes in Computer Science*, E. Chen, Y. Gong, and Y. Tie, Eds. Cham, Switzerland: Springer International Publishing, 2016, vol. 9916, pp. 201–210.

[80] P. Sun, R. Zhang, Y. Jiang, Z. Kong, C. Xu, X. Zhan, M. Tomizuka, Z. Yuan, P. Luo, D. Lin, and J. Li, "Sparse r-cnn: End-to-end object detection with learnable proposals," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14 454–14 463.

[81] T. Liang, H. Li, Q. Zhang, Q. Wang, and Z. Zhang, "Traffic sign detection via improved sparse r-cnn for autonomous vehicles," *Journal of Advanced Transportation*, vol. 2022, no. 1, p. 3825532, 2022.

[82] Q. Shuai and X. Wu, "Object detection system based on SSD algorithm," in *2020 International Conference on Culture-oriented Science & Technology (ICCST)*, Beijing, China, 2020, pp. 141–144.

[83] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 779–788.

[84] S. Du, X. Wang, Y. Li, M. Zhang, and H. Chen, "Tsd-yolo: Small traffic sign detection based on improved YOLO v8," *IET Image Processing*, 2024.

[85] A. El-Baz, G. Gimel'farb, R. Falk, and M. A. El-Ghar, "A new CAD system for early diagnosis of detected lung nodules," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, San Antonio, TX, USA, 2007, pp. II–461–II–464.

[86] K. Manikandan, "Blob based segmentation for lung CT image to improving CAD performance," in *2014 International Conference on Recent Trends in Information Technology (ICRTIT)*, Chennai, India, 2014, pp. 1–6.

[87] N. A. Pande and D. Bhoyar, "A comprehensive review of lung nodule identification using an effective computer-aided diagnosis (CAD) system," in *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India, 2022, pp. 1254–1257.

[88] R. Indumathi and R. Vasuki, "Lung cancer detection using CAD system," in *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, Greater Noida, India, 2023, pp. 206–210.

[89] K. P. Gunasekaran, "Leveraging object detection for the identification of lung cancer," *arXiv preprint arXiv:2305.15813*, 2023.

[90] X. Wang, L. Wang, and P. Zheng, "Sc-dynamic r-cnn: A self-calibrated dynamic r-cnn model for lung cancer lesion detection," *Computational and Mathematical Methods in Medicine*, vol. 2022, p. 9452157, Mar. 2022.

[91] J. Qi, Z. Deng, G. Sun, S. Qian, L. Liu, and B. Xu, "One-step algorithm for fast-track localization and multi-category classification of histological subtypes in lung cancer," *European Journal of Radiology*, vol. 154, p. 110443, 2022.

[92] M. Kirienko, M. Sollini, G. Silvestri, S. Mognetti, E. Voulaz, L. Antunovic, A. Rossi, L. Antiga, and A. Chiti, "Convolutional neural networks promising in lung cancer t-parameter assessment on baseline fdg-pet/ct," *Contrast Media & Molecular Imaging*, 2018.

[93] D. Moitra and R. K. Mandal, "Classification of non-small cell lung cancer using one-dimensional convolutional neural network," *Expert Systems with Applications*, vol. 159, p. 113564, 2020.

[94] S. Tyagi and S. N. Talbar, "Lcscnet: A multi-level approach for lung cancer stage classification using 3d dense convolutional neural networks with concurrent squeeze-and-excitation module," *Biomedical Signal Processing and Control*, vol. 80, p. 104391, 2023.

[95] P. Li, S. Wang, T. Li, J. Lu, Y. HuangFu, and D. Wang, "A large-scale CT and PET/CT dataset for lung cancer diagnosis (lung-pet-ct-dx)," 2020.

[96] K. Clark, B. Vendt, K. Smith, and et al., "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," *Journal of Digital Imaging*, vol. 26, pp. 1045–1057, 2013.

[97] H. J. W. L. Aerts, L. Wee, E. Rios Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies, and P. Lambin, "Data from nsclc-radiomics (version 4)," 2014.

[98] B. Dwyer, J. Nelson, J. Solawetz, and et al., "Roboflow (version 1.0) [software]," https://roboflow.com, 2022, computer vision toolkit.

[99] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778.

[100] Y. Zhang, P. Sun, Y. Jiang, D. Yu, W. Weng, Z. Yuan, and P. L. Wang, "Bytetrack: Multi-object tracking by associating every detection box," in *European Conference on Computer Vision (ECCV)*. Cham: Springer Nature Switzerland, 2022.

[101] A. Wehbe, "Feature-extraction-yolov8," https://github.com/Alaawehbe12/Feature-extraction-YOLOv8.

[102] X. Wang, L. Wang, and P. Zheng, "SC-Dynamic R-CNN: A self-calibrated dynamic R-CNN model for lung cancer lesion detection," *Computational and Mathematical Methods in Medicine*, vol. 2022, p. 9452157, Mar. 2022.

[103] P. Aggarwal, R. Vig, and H. K. Sardana, "Semantic and content-based medical image retrieval for lung cancer diagnosis with the inclusion of expert knowledge and proven pathology," in *2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013)*, Shimla, India, 2013, pp. 346–351.

[104] D. Perez Ibanez, Y. Shen, J. Dayanghirang, J. Li, S. Wang, and Z. Zheng, "Deep learning for pulmonary nodule CT image retrieval – an online assistance system for novice radiologists," in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, New Orleans, LA, USA, 2017, pp. 1112–1121.

[105] G. Wei, H. Ma, W. Qian, and X. Zhao, "A content-based image retrieval scheme for identifying lung nodule malignancy levels," in *2017 29th Chinese Control and Decision Conference (CCDC)*, Chongqing, China, 2017, pp. 3127–3130.

[106] S. Pang, Y. Zhang, M. Ding, X. Wang, and X. Xie, "A deep model for lung cancer type identification by densely connected convolutional networks and adaptive boosting," *IEEE Access*, vol. 8, pp. 4799–4805, 2020.

[107] K. Zhang, S. Qi, J. Cai, D. Zhao, T. Yu, Y. Yue, and W. Qian, "Content-based image retrieval with a convolutional siamese neural network: Distinguishing lung cancer and tuberculosis in CT images," *Computers in Biology and Medicine*, vol. 140, p. 105096, 2022.

[108] B. Xiao *et al.*, "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 (CVPR)*, 2024.

[109] Y. Zhou *et al.*, "UMed-LVLM: Unified medical vision-language model with prompt-able detection," *arXiv preprint arXiv:2503.03278*, 2025.

[110] T. Zhang *et al.*, "Anatomical grounding in open-vocabulary detection for medical imaging," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[111] T. Fernando *et al.*, "Deep learning for medical anomaly detection—a survey," *ACM Computing Surveys*, vol. 54, no. 7, pp. 1–37, 2021.

[112] F. Shamshad *et al.*, "Transformers in medical imaging: A survey," *Medical Image Analysis*, vol. 88, p. 102802, 2023.

[113] J. Egger *et al.*, "Medical deep learning—a systematic meta-review," *Computer Methods and Programs in Biomedicine*, vol. 221, p. 106874, 2022.

[114] A. Majali, "Vision Language models in x-ray: A review," *Authorea Preprints*, 2024.

[115] M. Fang *et al.*, "Large models in medical imaging: Advances and prospects," *Chinese Medical Journal*, 2025, advance online publication.

[116] A. K. Singh, "Revolutionizing bone cancer detection using YOLOv8 and deep learning," *International Journal of Science, Technology and Engineering*, vol. 12, no. 3, pp. 45–58, 2025.

[117] H. Liu, Y. Chen, and M. Wu, "Improved COVID-19 CT scan classification using CNNs and particle swarm optimization," *PLOS ONE*, vol. 20, no. 4, p. e0256789, 2025.

[118] C. D. Collins, R. Zhao, and K. Patel, "AI-augmented rib fracture detection: Performance against radiologists," in *Proceedings of the IEEE Engineering in Medicine and Biology Conference (EMBC)*, 2025, pp. 120–134.

[119] R. Alsharif *et al.*, "Deep learning applications in pneumonia detection using chest radiographs," *Applied Sciences*, vol. 11, no. 14, p. 6442, 2021.

[120] S. Wang, L. Dong, X. Wang, and X. Wang, "Classification of pathological types of lung cancer from CT images by deep residual neural networks with transfer learning strategy," *Open Medicine (Warsaw)*, vol. 15, pp. 190–197, Mar. 2020.

[121] P. Rajpurkar, J. Irvin, K. Zhu *et al.*, "CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2018.

[122] J. Irvin, P. Rajpurkar, M. Ko *et al.*, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019, pp. 590–597.

[123] Y. Huang, X. Ren, and M. Li, "A deep CNN for abdominal abnormality detection in CT images," *Journal of Medical Imaging*, vol. 10, no. 2, p. 023501, 2023.

[124] T. Zhou and Y. Zhang, "Multi-label thoracic disease detection with dual-path CNN in chest CT images," *Computers in Biology and Medicine*, vol. 145, p. 105451, 2022.

[125] A. Wehbe, H. Hotiet, I. Minetti, and S. Dellapiane, "Integrating YOLO for advanced content-based image retrieval in lung cancer imaging," in *Proceedings of the IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, 2024.

[126] L. H. Li *et al.*, "Grounded language-image pretraining," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 965–10 975.

[127] C. Feng *et al.*, "Promptdet: Open-vocabulary detection with regional prompt learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022, pp. 701–717.

[128] M. Minderer *et al.*, "Simple open-vocabulary object detection with vision transformers," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022, pp. 728–755.

[129] X. Wang *et al.*, "ViLD: Open-vocabulary object detection via vision and language knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 14 988–14 998.

[130] L. Yao *et al.*, "DetCLIPv2: Scalable and data-efficient open-vocabulary object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 2391–2400.

[131] X. Wu *et al.*, "CORA: Region-level prompting for open-vocabulary detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 1520–1530.

[132] S. Liu *et al.*, "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[133] T. Cheng *et al.*, "YOLO-World: Real-time open-vocabulary object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 298–309.

[134] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, May 2016.

[135] D. Abdelhafiz, C. Yang, R. Ammar, M. Nabhan, and J. Bi, "Deep learning-based CBIR for medical images: A comparative study using classification and retrieval metrics," *IEEE Access*, vol. 7, pp. 127 182–127 196, 2019.

[136] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2006, pp. 1735–1742.

[137] M. M. Rahman, M. S. Uddin, M. A. Rahman, and M. A. Hossain, "A deep learning-based computer-aided system for lung nodule detection and diagnosis," *IEEE Access*, vol. 8, pp. 132 665–132 676, 2020.

[138] C. Zhao, X. Yin, M. Tan, and H. Lin, "Lesion-aware content-based image retrieval for thoracic disease diagnosis using Faster R-CNN," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2021, pp. 3310–3315.

[139] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, vol. 139. PMLR, 2021, pp. 8748–8763.

[140] T. Xiao, L. Zhang, J. Song, Z. Yang, and H. Zhang, "Delving into masked autoencoders for multi-label thorax disease classification," *arXiv preprint arXiv:2210.12843*, 2022.

[141] M. Ding, B. Xiao, N. Codella, P. Luo, J. Wang, and L. Yuan, "Davit: Dual attention vision transformers," 2022. [Online]. Available: https://arxiv.org/abs/2204.03645

[142] J. Chen, J. Yang, H. Wu, D. Li, J. Gao, T. Zhou, and B. Xiao, "Florence-vl: Enhancing vision-language models with generative vision encoder and depth-breadth fusion," 2024. [Online]. Available: https://arxiv.org/abs/2412.04424

[143] K. Xiao *et al.*, "Florence-2: Advancing foundation models for computer vision," *arXiv preprint arXiv:2403.19808*, 2024.

[144] W. Yang, L. Zhang, S. Wang, A. D. White, and S. Jiang, "Functionalizable and ultra stable nanoparticles coated with zwitterionic poly (carboxybetaine) in undiluted blood serum," *Biomaterials*, vol. 30, no. 29, pp. 5617–5621, 2009.

[145] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7008–7024.

[146] W. Tang, C. Pei, P. Yu, H. Zhang, X. Min, C. Chen, H. Kang, W. Xu, and R. Zhang, "Generating chinese radiology reports from x-ray images: a public dataset and an x-ray-to-reports generation method," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 79–88.

[147] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou, "Exploring and distilling posterior and prior knowledge for radiology report generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 753–13 762.

[148] R. Zhang, Y. Ji, Y. Zhang, and R. J. Passonneau, "Contrastive data and learning for natural language processing," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, 2022, pp. 39–47.

[149] J. Sun, Z. Zhuang, J. Zheng, K. Li, R. L.-Y. Wong, D. Liu, J. Huang, J. He, A. Zhu, J. Zhao *et al.*, "Generation of a broadly useful model for covid-19 pathogenesis, vaccination, and treatment," *Cell*, vol. 182, no. 3, pp. 734–743, 2020.

[150] S. Wang, Z. Liu, and B. Peng, "A self-training framework for automated medical report generation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 16 443–16 449.

[151] Z. Huang, X. Zhang, and S. Zhang, "Kiut: Knowledge-injected u-transformer for radiology report generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 19 809–19 818.

[152] R. Tanno, D. G. Barrett, A. Sellergren, S. Ghaisas, S. Dathathri, A. See, J. Welbl, K. Singhal, S. Azizi, T. Tu *et al.*, "Consensus, dissensus and synergy between clinicians and specialist foundation models in radiology report generation," *arXiv preprint arXiv:2311.18260*, 2023.

[153] K. Lee, S. Yoon, and H. Lim, "Clarifid: Improving radiology report generation by reinforcing clinically accurate impressions and enforcing detailed findings," *arXiv preprint arXiv:2507.17234*, 2025.

Personalized Medicine and Process Optimization: Analysis and Implementation of Intelligent Tools to Support the Clinical Process

Alaa WEHBE

Response to Reviewers – Summary of Revisions

We thank the reviewer for the detailed and constructive feedback. Below we summarise the revisions made to the thesis in response to each comment.

Critical questions concerning the description:

---

a. Clarification of Methodology (Section 3.3)

Section 3.3 has been rewritten to improve clarity and precision. Ambiguous notation, including the index i, is now explicitly defined as the number of slices in which a tracked lesion appears within a CT volume. Vague expressions such as *"start tracking the whole study"* were replaced with a clear description of the lesion tracking process across consecutive slices.

The revised text now presents the methodology as a step-by-step pipeline, from volumetric CT processing and lesion tracking to slice selection, feature extraction, and TNM classification. The dataset preparation subsection was also expanded to clearly describe the Lung PET-CT-DX dataset and its annotations, improving readability and reproducibility.

---

b. DICOM-to-JPEG Conversion and Image Representation (Sections 3.3.1.1 and 3.3.1.2)

We have substantially expanded the discussion regarding the DICOM-to-JPEG conversion to clarify both its motivation and its limitations. The conversion was performed using high-quality, near-lossless JPEG settings exclusively to ensure compatibility with pretrained convolutional neural networks, which are commonly optimized for three-channel image inputs.

Importantly, the proposed framework does **not rely on quantitative radiological measurements**, such as Hounsfield Unit–based intensity values or radiomic features. Instead, it focuses on **spatial, morphological, and contextual patterns** for lesion localization, retrieval, and classification. Therefore, the lossy nature of JPEG compression does not compromise the validity of the experimental outcomes.

The grayscale CT images were replicated across three channels to maintain architectural compatibility with RGB-pretrained models while preserving identical information in each

channel, avoiding any artificial alteration of the original data distribution. This design choice is explicitly discussed in the revised manuscript to ensure transparency and reproducibility.

---

c. Slice Selection Strategy (Section 3.3.4)

Section 3.3.4 has been revised to explicitly describe the adaptive slice selection strategy. The number of selected slices per CT volume is **not fixed**, but varies according to the spatial extent of the detected lesion.

A simple geometric criterion based on bounding box area is used solely to rank slices by lesion prominence, ensuring that the most informative anatomical views are retained. When the number of detected slices exceeds the target threshold, the top-ranked slices are selected; otherwise, all available lesion-containing slices are preserved. This strategy ensures a balance between computational efficiency and anatomical representativeness while maintaining methodological transparency.

---

d. Per-Slice vs. Per-Patient Evaluation (Page 37)

We clarified that the reported true positive values are computed on a *per-slice* basis rather than per patient. An explicit explanation was added to distinguish slice-level evaluation from patient-level counts, addressing the apparent discrepancy between the number of samples and the reported true positive values.

---

e. Training Parameters Selection (Learning Rate and Epochs)

The manuscript now explicitly details the hyperparameter selection process to avoid any ambiguity. Learning rates and epoch counts were selected through systematic empirical evaluation using *k*-fold cross-validation on the training set. Multiple candidate configurations were tested, and the final values (learning rate = 0.01, epochs = 200) were chosen based on stable convergence behavior and consistent validation performance across folds.

This process ensures that the reported results are not the outcome of arbitrary parameter choices but reflect a principled optimization procedure aligned with best practices in deep learning experimentation.

---

f. Classifier Hyperparameters (Section 3.4.4.2)

A dedicated paragraph has been added to specify the hyperparameters used for SVM, XGBM, and KNN classifiers. All models were optimized via grid search with 5-fold cross-validation. The selected hyperparameters, listed in the revised text, reflect the configuration that achieved the best validation performance, ensuring a fair and reproducible comparison with state-of-the-art methods.

---

g. Open-Vocabulary Systems with Vision–Language Models (Section 5.5)

Section 5.5 has been substantially expanded to provide a transparent and non–black-box explanation of open-vocabulary modeling. The revised text explains how vision–language models operate by projecting visual and textual inputs into a shared embedding space, enabling semantic alignment between image regions and free-form clinical descriptions.

Architectural mechanisms such as dual encoders, cross-attention modules, and prompt-driven decoding are now explicitly described, along with their role in enabling zero-shot generalization. This allows the system to reason about previously unseen pathologies or descriptors without retraining, which is particularly relevant in clinical environments characterized by evolving terminology and rare conditions.

The discussion now contextualizes Florence-2 and related medical adaptations (e.g., UMed-LVLM, VividMed) within this framework, demonstrating how open-vocabulary systems trade initial performance for long-term scalability, flexibility, and adaptability in real-world clinical workflows.

.

---

h. Interpretation of the ‘Sharper Upward Trend’ (Figure 5.2)

The discussion of Figure 5.2 has been revised to avoid misinterpretation. The term “sharper upward trend” refers specifically to the **rate of improvement over time**, rather than absolute performance superiority over closed-vocabulary approaches.

While closed-vocabulary methods currently achieve higher absolute scores, open-vocabulary approaches demonstrate faster methodological progress and greater adaptability. This characteristic is particularly attractive for clinical research, where scalability to new conditions and evolving diagnostic language is essential. The revised explanation clarifies this distinction and motivates the inclusion of open-vocabulary models as a forward-looking research direction rather than a replacement for established methods.

Clarification of Personalized Medicine Contribution

In response to the reviewer's concern, we clarify that personalization in this thesis is addressed through **case-based clinical decision support** rather than population-level prediction alone. The proposed systems enable the retrieval of clinically similar cases, adaptive interpretation through open-vocabulary reasoning, and flexible interaction with radiological findings tailored to individual patients.

By supporting radiologists with personalized comparisons to prior cases, context-aware retrieval, and adaptive diagnostic vocabulary, the framework contributes to personalized medicine by enhancing individualized diagnostic reasoning, supporting heterogeneous patient profiles, and facilitating patient-specific clinical workflows rather than enforcing rigid predefined categories.

Elements worth revising:

Consistency:

The manuscript has been thoroughly revised to ensure consistency in terminology, notation, numerical presentation, and formatting across all chapters. All evaluation metrics are now reported using a single, unified notation (percentage format) throughout the Abstract, main text, tables, and figures. Numerical precision has been standardized, with a consistent number of decimal places maintained within and across tables presenting comparable results, and leading zeros have been retained to avoid ambiguity.

In addition, naming conventions and capitalization have been unified for all technical terms, model names, and evaluation metrics (e.g., Precision, Recall, Accuracy, YOLOv8), and applied consistently throughout the manuscript. These revisions eliminate inconsistencies that could hinder readability and ensure a coherent and professional presentation of the technical content.

Text formatting

- A single notation has now been adopted for reporting results (percentage format) and used consistently across the Abstract, main text, tables, and figures.

- The numerical precision of accuracy and other metrics has been standardized, with a uniform number of decimal places maintained within each table and across comparable results (see Tables 3.8, 3.9, and 4.4).

- Leading zeros are now retained to avoid ambiguity and ensure clarity for the reader.

- Naming and font conventions for evaluation metrics (e.g., Precision, Recall, Accuracy) have been unified and used consistently throughout the manuscript.

- A single, consistent naming convention has been adopted for the **YOLOv8** network, including capitalization, and applied uniformly across the entire document.

Acronyms

All acronyms have now been explicitly defined at their first occurrence in the manuscript. In particular, we have clarified the acronyms mentioned in Sections 1.1 (CT, MRI), 1.2 (TNM, VGG, SIFT, HOG, ViT), 1.3.1 (ReLU), 2.3.1 (HRCT), 2.3.3 (EIoU), 2.3.5 (BLEU, METEOR, ROUGE), 3.2 (TCIA), 3.3.2.1 (C2F, SiLU), and 5.7 (FDA, CE). These revisions were made to improve clarity and readability for a broader audience.

Figures

- **Figure 1.1:** The typo “SFIH” has been corrected to “SIFT”. The illegible dark block in the online phase has been adjusted to ensure proper visibility.

- **Figure 2.1:** The typo “Medial” has been corrected to “Medical”. The formulas shown in the Convolutional Layer and Fully Connected Layer blocks have been revised for correctness and clarity, and their purpose is now explicitly explained in the accompanying text.

- **Figure 2.2:** The extra punctuation in the caption has been removed.

- **Figure 3.1:** Image contrast and brightness have been adjusted to improve visibility, with particular attention to sub-image (4), which was previously difficult to interpret.

- **Figures 3.2–3.4:** The images have been enhanced to ensure that the marked regions are clearly visible.

- **Figure 3.5:** The font size of the captions has been increased for better readability.

- **Figure 3.6:** The figure has been revised to improve legibility.

- **Figure 3.7:** The meaning of the letter “B” used in the captions is now explicitly explained in the figure caption and in the main text.

- **Figure 3.8:** The font size has been increased to improve readability.

- **Figure 3.9:** The figure has been adjusted to correct excessive darkness and improve visibility.

- **Figures 3.11–3.13:** Image contrast has been enhanced to ensure legibility.

- **Figure 4.1:** The font size has been increased to improve readability.

- **Figure 5.1:** The truncated word “precision” has been corrected, and the typos “ariorited” and “moldl” have been fixed.

- **Figure 5.2:** The font size has been increased to ensure clarity.

Equations

- For **Equation (1.1)**, the symbols n and m have now been explicitly defined in the text immediately following the equation.

- For **Equations (2.1) and (2.2)**, the symbols a and b are now clearly explained and their meaning is kept consistent throughout the manuscript.

- For **Equation (2.3)**, the indices $i, j, m,$ and n are now explicitly defined as part of the equation description.

- For **Equation (2.4)**, the variable x is now clearly described.

- For **Equation (2.5)**, the symbols $x, i, j,$ and s are now defined and consistently used.

- For **Equation (2.8)**, the Q (Query), K (Key), and V (Value) matrices have now been formally introduced and defined.

- For **Equation (2.9)**, the scaling factor d_k is now explicitly defined.

- The formula referenced in the text on **page 57** has been updated to clearly explain the meaning of k.

- All equations in **Chapter 6** have now been numbered, and the superscript symbols in the equation on **page 73** are explicitly defined

Unclear statements/typos

- **Page 16:** A dedicated section defining all evaluation metrics used in the thesis (including precision, recall, AP, and mAP) has been added. The sentence referring to "as detailed in the Introduction Chapter" has been revised to point to this newly added section.

- **Page 21:** The use of "we" has been corrected to avoid ambiguity. The sentence has been rephrased to reflect that this section describes the literature and theoretical background rather than the author's own experimental work.

- **Page 26:** The typo "Gunasekaran1 et al." has been corrected.

- **Page 29:** The reference to an "Evaluation section" has been removed, as such a section did not exist at that point in the manuscript.

- **Page 33:** The capitalization of "the Neck" has been corrected for consistency.

- **Page 37:** The incomplete sentence starting with "When we looked at the confusion matrix …" has been completed and clarified.

- **Page 38:** The typo "AP at 0.5" has been corrected.

- **Page 38:** The incorrect reference to Table 4.3 has been fixed and replaced with the correct reference to Table 3.4.

- **Page 39 (Section 3.4.4):** The phrase "training and testing" has been corrected to "training and validation" for accuracy.

- **Page 39 (Section 3.4.4.1):** The phrase "3.13 in the testing set" has been revised to correct the grammatical inconsistency.

- **Page 51:** The corrupted descriptions of Equations (4.1) and (4.2) have been fixed. In addition, the definitions of these metrics have been moved to the metrics section, prior to their use in the results, as suggested.

- **Page 52:** The typo "Three metrics was" has been corrected to "Three metrics were."

- **Page 54:** A brief explanation of the concept of zero-shot generalization has been added to improve clarity.

- **Page 58:** The incomplete example “(e.g.,,)” has been corrected.

- **Page 59:** In Table 5.2, the missing reference for Li et al. has been restored.

- **Page 63:** The terms **CAPTION**, **DETAILED_CAPTION**, and **OD** have now been explicitly defined at their first occurrence, and a consistent naming convention has been applied throughout the chapter.

- **Page 67:** The title of Section 6.3.1.1 has been corrected to “VinDr.”

- **Page 68:** The extraneous dot at the beginning of the second line has been removed.

- **Page 68:** The term “tasks2” has been clarified and corrected.

- **Page 68 (Section 6.3.1.2):** A single, consistent naming convention for the MIMIC-CXR dataset has been adopted throughout the manuscript.

- **Page 68 (Section 6.3.2):** Missing references have been added.

- **Page 69:** A reference has been added to the definition of the visual encoder.

- **Page 69:** The equations have been properly numbered, and the meanings of the symbols $X$and Vhave been clearly defined to avoid ambiguity.

- **Page 70:** The caption of Figure 6.1 has been revised to explicitly specify which original paper is being referenced.

- **Page 73:** A missing reference has been added to the Fusion Strategy Justification section.

- **Throughout the manuscript:** Spacing and formatting issues following full stops have been corrected to improve overall readability.

General remarks

1. **Equations in the first chapter**
   Modification: Removed unnecessary equations that do not contribute to reader understanding or kept only the essential ones for clarity and reproducibility.
2. **Reference inconsistency ([58] Jing et al.)**
   Modification: Verified all citations. The references in the manuscript are correct and consistent with the bibliography, so no changes were needed.
3. **Table 2.1 bold result**
   Modification: Removed bold formatting or clarified in the caption that bold indicates the best-performing result across methods.

4. **Section 3.2.1 formatting**
   Modification: Converted the narrative explanation into bullet points for clearer presentation of steps.
5. **YOLOv8 architecture (Section 3.3.2.1)**
   Modification: Expanded the description to include:
   - SiLU activation function with equation
   - Definition and structure of C2F (Coordinates-to-Features) blocks
   - Clarification that "concat" refers to channel-wise concatenation
   - Explanation of anchor-free detection
   - Description of backbone, neck (SPPF + C2F), and head
   - Multi-scale prediction and overall workflow
6. **Section 3.3.3 bounding box area formula**
   Modification: Retained the formula but simplified explanation, emphasizing it is a straightforward geometric computation for ranking slices. Top $k=3$ slices are selected for feature extraction.
7. **Table 3.4 YOLOv5 reference ([59])**
   Modification: Clarified that the reference points to the YOLOv5 implementation used for experiments to ensure reproducibility, rather than a theoretical model description.
8. **Section 3.4.4.2 "average P"**
   Modification: Replaced "average P" with **average precision** and added a brief explanation: *average precision is the mean of precision values across all TNM stages and cancer groups.*
9. **Zero-shot image retrieval and FAISS**
   Modification: Added a definition and explanation of FAISS:
   - FAISS indexes high-dimensional image embeddings
   - Performs fast nearest-neighbor retrieval
   - Supports cosine or Euclidean similarity
   - Used to retrieve the top-K most similar radiographs in a zero-shot setting

Response to Reviewer 2

- **Clarification of metric inconsistencies:**

All reported metrics have been carefully reviewed and clarified. The evaluation protocols for TNMClassifier recall, MedFL Recall@10, and similarity threshold–based performance are now explicitly described, including whether metrics are computed at the slice level or patient level and how similarity thresholds affect retrieval outcomes. This ensures consistent and unambiguous interpretation of the reported results.

- **Language, typos, and citation issues:**

The manuscript has been thoroughly revised to correct typographical errors, improve English fluency, and fix incomplete citations and numbering inconsistencies across all chapters.

- **Baseline and model training clarification:**

The manuscript now clearly states that **all models used throughout the thesis—including baselines and proposed methods—were fine-tuned on the target datasets under a unified experimental protocol**. This includes both classification and retrieval models. No results were directly reused without adaptation; even when architectures were adopted from the literature, their weights were fine-tuned to ensure fair comparison and domain alignment.