Independent University **Bangladesh (IUB)**

IUB Academic Repository

Computer Science and Engineering | Undergraduate Thesis

2026-04

OptiHealth: An AI-Driven Health Management System with Multi-Modal Analysis and Personalized Recommendations

Alif, Afzal Hossain

IUB

https://ar.iub.edu.bd/handle/11348/1187

Downloaded from IUB Academic Repository

Independent University Bangladesh

In Partial Fulfillment of the Requirements for the Degree of
Bachelor of Computer Science & Engineering

OptiHealth

An AI-Driven Health Management System with Multi-Modal Analysis and Personalized Recommendations

Spring, April 2026

Prepared by

Afzal Hossain Alif	**ID 2131050**
Mojtoba Zaman Mantaka	**ID 2231306**
Aiman Saad Hamid	**ID 2131225**

Department of Computer Science and Engineering
Independent University, Bangladesh

Supervised by

Dr. Saadia Binte Alam
Associate Professor
Department of Computer Science and Engineering
Independent University, Bangladesh

Attestation

We are aware of the fact that plagiarism is strictly prohibited and it conflicts with our university rules and regulations. We guarantee the authenticity of our work. We have used some copyrighted material and models of others in our project which has been properly cited following the international standards and proper guidelines.

Afzal Hossain Alif	ID: 2131050
Signature: _____	

Mojtoba Zaman Mantaka	ID: 2231306
Signature: _____	

Aiman Saad Hamid	ID: 2131225
Signature: _____	

Evaluation Committee

Supervisor

Name: _____ Signature: _____

Internal Examiner 1

Name: _____ Signature: _____

Internal Examiner 2

Name: _____ Signature: _____

External Examiner

Name: _____ Signature: _____

Acknowledgement

In the name of Allah, the Most Merciful, the Most Compassionate.

We would like to express our heartfelt gratitude to our supervisor, Dr. Saadia Binte Alam, for her invaluable guidance throughout our thesis journey. Her patience, insightful feedback, and constant encouragement helped us overcome technical challenges and strengthened our confidence to complete this study successfully.

We also extend our sincere appreciation to the Department of Computer Science and Engineering, Independent University, Bangladesh, for providing a supportive academic environment.

We are sincerely thankful to Mr. Sefatul Wasi for his guidance and continuous support throughout this work.

We thank the organizers of the BraTS 2023 challenge and the creators of the VinDr-CXR dataset for making their data publicly available for research. We also thank the open-source communities behind PyTorch, MONAI, and Ultralytics YOLO for their invaluable tools.

We acknowledge the support of conversational AI tools, including ChatGPT (OpenAI), Claude (Anthropic), and Gemini (Google), which we consulted for exploratory programming assistance—such as clarifying APIs, proposing implementation patterns, and suggesting debugging approaches. All architectural decisions, integrated code, and validation of correctness were undertaken by the authors, and generative assistants were not used to produce empirical results, datasets, figures, or novel analyses reported in this work.

Finally, we gratefully acknowledge Independent University, Bangladesh for the opportunity to conduct this thesis, which enabled us to gain practical research experience and better prepare for the professional world.

Abstract

Medical image analysis is essential for early diagnosis, treatment planning, and follow-up care, but current AI solutions are often limited to producing isolated outputs instead of helping real patients make informed next decisions. This thesis presents a compiled study of three related AI systems that together move from image understanding to actionable guidance. First, for multi-class thoracic disease detection on chest X-rays, we develop CXR-Adaptive using a standard YOLOv12 model trained with Inverse-Frequency Weighted Upsampling and Anatomically-Constrained Test-Time Augmentation (CXR-TTA) on the VinDr-CXR dataset. Under a controlled comparison protocol, the proposed CXR-TTA achieves an mAP@0.5 of 0.442, which improves over the YOLOv11-MFF baseline by 6.5%. Second, for brain tumor segmentation, we propose SS-UNETR, a centralized fusion transformer architecture with a Multi-Scale Attentive Bottleneck that provides the decoder with a complete multi-scale representation before reconstruction. Evaluated on BraTS 2023, SS-UNETR achieves Dice Similarity Coefficients of 0.9106 (Whole Tumor), 0.8819 (Tumor Core), and 0.8625 (Enhancing Tumor), improving mean DSC by 2.76% over the Swin UNETR baseline. Third, to connect diagnostic outputs to patient support, we build OptiHealth, which combines vision models with a Llama-based large language model enhanced by Retrieval-Augmented Generation (RAG) to generate personalized diet and specialist referral recommendations grounded in a curated knowledge base. OptiHealth further incorporates a tool-augmented Conversational Health Agent implemented with LangChain and LangGraph, which orchestrates a state-graph ReAct loop over ten domain-specific tools (electronic health record access, appointment scheduling, prescription lookup, geospatial doctor search, real-time medicine information, and a dual-model mental health analyzer) on top of a locally-hosted Gemma 4 E4B open-weights language model. Expert qualitative assessment rated the resulting recommendations and agent interactions as "Good" for factual accuracy and personalization. The main limitations are that evaluations are performed on public benchmarks and the recommendation quality is assessed through expert feedback rather than full clinical trials. Future work should include cross-institution testing, larger user studies for safety validation, and integration into real clinical workflows.

Contents

List of Figures

List of Tables

List of Abbreviations

The following abbreviations are used throughout this thesis. They are grouped thematically—clinical and imaging terms, datasets and benchmarks, AI/ML methods and architectures, project-specific contributions, software and infrastructure, and standards bodies—to keep related vocabulary together.

Clinical and Medical Imaging

Acronym	**Definition**
BP	Blood Pressure
COPD	Chronic Obstructive Pulmonary Disease
CT	Computed Tomography
CXR	Chest X-Ray
DICOM	Digital Imaging and Communications in Medicine
ECG	Electrocardiogram
EHR	Electronic Health Record
EMR	Electronic Medical Record
ET	Enhancing Tumor (BraTS sub-region)
FLAIR	Fluid-Attenuated Inversion Recovery (MRI sequence)
GAD-7	Generalized Anxiety Disorder 7-item scale
HbA_{1c}	Glycated Haemoglobin
ILD	Interstitial Lung Disease
IRB	Institutional Review Board
MRI	Magnetic Resonance Imaging
NIfTI	Neuroimaging Informatics Technology Initiative (file format)
PA	Postero-Anterior (chest X-ray view)
PHI	Protected Health Information

Acronym	Definition
PHQ-9	Patient Health Questionnaire (9-item depression scale)
PHR	Personal Health Record
SaMD	Software as a Medical Device
T1	T1-weighted (MRI sequence)
T1c / T1ce	Post-contrast T1-weighted (MRI sequence)
T2	T2-weighted (MRI sequence)
TC	Tumor Core (BraTS sub-region)
WT	Whole Tumor (BraTS sub-region)

Datasets and Benchmarks

Acronym	Definition
BraTS	Brain Tumor Segmentation (challenge and dataset)
COCO	Common Objects in Context (dataset)
LVIS	Large Vocabulary Instance Segmentation (benchmark)
MIMIC-CXR	Medical Information Mart for Intensive Care – Chest X-Ray
NIH ChestX-ray14	National Institutes of Health 14-class chest X-ray dataset
PTB	Physikalisch-Technische Bundesanstalt Diagnostic ECG Database
USDA	United States Department of Agriculture Nutrient Database
VinDr-CXR	Vietnamese Diagnostic Imaging Center Chest X-Ray dataset

AI, Machine Learning, and Model Architectures

Acronym	Definition
AI	Artificial Intelligence
BART	Bidirectional and Auto-Regressive Transformer
BERT	Bidirectional Encoder Representations from Transformers
CAM	Class Activation Map
CNN	Convolutional Neural Network
DAF3D	Deeply Attentive 3D Feature aggregation module

Acronym	Definition
DETR	DEtection TRansformer
DL	Deep Learning
ELAN	Efficient Layer Aggregation Network
FFN	Feed-Forward Network
GAN	Generative Adversarial Network
LLM	Large Language Model
Med-PaLM	Medical Pathways Language Model
MFF	Multi-Frequency-aware Fusion (in YOLOv11-MFF)
MLP	Multi-Layer Perceptron
ML	Machine Learning
MNLI	Multi-Genre Natural Language Inference
MSA	Multi-Head Self-Attention
NLI	Natural Language Inference
NLP	Natural Language Processing
nnU-Net	no-new-Net (self-configuring U-Net framework)
PLE	Per-Layer Embeddings (Gemma 4 architectural feature)
R-CNN	Region-based Convolutional Neural Network
R-ELAN	Residual Efficient Layer Aggregation Network
ReAct	Reasoning + Acting (LLM agent paradigm)
ReLU	Rectified Linear Unit
RoBERTa	Robustly Optimised BERT Pretraining Approach
RoPE / p-RoPE	Rotary Position Embedding / Proportional Rotary Position Embedding
SiEBERT	Sentiment in English BERT (large RoBERTa-based scorer)
Swin	Shifted-Window Transformer
SW-MSA	Shifted-Window Multi-head Self-Attention
U-Net	Encoder–decoder convolutional segmentation architecture
UNETR	UNEt TRansformer
ViT	Vision Transformer
W-MSA	Window-based Multi-head Self-Attention
YOLO	You Only Look Once (object-detection family)

Proposed Methods and Project Contributions

Acronym	Definition
CXR-Adaptive	Adaptive chest X-ray detection pipeline (this thesis)
CXR-TTA	Anatomically-Constrained Chest X-Ray Test-Time Augmentation (this thesis)
IFWU	Inverse-Frequency Weighted Upsampling (this thesis)
MSAB	Multi-Scale Attentive Bottleneck (this thesis)
OptiHealth	Integrated AI-driven health-management platform (this thesis)
SS-UNETR	Symmetric Swin UNETR (this thesis)

Training, Evaluation, and Statistics

Acronym	Definition
AdamW	Adam optimiser with decoupled Weight decay
AP	Average Precision
AUC	Area Under the Curve
AUROC	Area Under the Receiver Operating Characteristic Curve
BERTScore	BERT-based text similarity score
DiceCE	Dice + Cross-Entropy compound loss
DSC	Dice Similarity Coefficient
FactScore	Atomic-fact factuality score for LLM outputs
F1	F1-score (harmonic mean of Precision and Recall)
FLOPs	Floating-Point Operations
FN	False Negative
FP	False Positive
IoU	Intersection over Union
mAP	mean Average Precision
mAP@0.5	mean Average Precision at IoU threshold 0.5
MedNLI	Medical Natural Language Inference benchmark
NMS	Non-Maximum Suppression
PR	Precision–Recall (curve)

Acronym	Definition
ROC	Receiver Operating Characteristic
ROI	Region of Interest
ROUGE-L	Recall-Oriented Understudy for Gisting Evaluation, Longest common subsequence
SGD	Stochastic Gradient Descent
TN	True Negative
TP	True Positive
TTA	Test-Time Augmentation
WBF	Weighted Box Fusion

Software, Frameworks, and Infrastructure

Acronym	Definition
API	Application Programming Interface
CI/CD	Continuous Integration / Continuous Deployment
CPU	Central Processing Unit
ChromaDB	Open-source embedding-based vector database
CV	Computer Vision
FP16 / FP32	16-bit / 32-bit Floating-Point precision
GPU	Graphics Processing Unit
HPA	Horizontal Pod Autoscaler (Kubernetes)
HTML	HyperText Markup Language
HTTPS	HyperText Transfer Protocol Secure
INT8	8-bit Integer quantisation
IoT	Internet of Things
JSON	JavaScript Object Notation
JWT	JSON Web Token
LangChain	Open-source LLM application framework
LangGraph	Graph-based runtime for LangChain agent state machines
MedEx BD	MedEx Bangladesh (online medicine catalogue)
MinIO	Open-source S3-compatible object storage

Acronym	Definition
MONAI	Medical Open Network for AI
NGINX	Open-source web server and reverse proxy
PIN	Personal Identification Number (per-session PHI decryption key)
RAG	Retrieval-Augmented Generation
RAM	Random-Access Memory
RAS	Right–Anterior–Superior (anatomical orientation)
RBAC	Role-Based Access Control
REST	Representational State Transfer
SQL	Structured Query Language
TensorRT	NVIDIA inference optimisation runtime
TLS	Transport Layer Security

Standards Bodies and Reporting Frameworks

Acronym	Definition
CAC	Codex Alimentarius Commission
CDC	Centers for Disease Control and Prevention
CLAIM	Checklist for Artificial Intelligence in Medical Imaging
FAO	Food and Agriculture Organization of the United Nations
FDA	Food and Drug Administration
MINIMAR	MINimum Information for Medical AI Reporting
NIH	National Institutes of Health
TRIPOD+AI	Transparent Reporting of multivariable prediction models for Individual Prognosis Or Diagnosis (AI extension)
WHO	World Health Organization

1 Introduction

1.1 Background of the project

Modern healthcare is increasingly relying on artificial intelligence to help doctors and medical staff analyze complex data more quickly and accurately. Medical imaging, in particular, produces enormous amounts of data every day. Chest X-rays (CXR) are the most common medical imaging study performed worldwide, while brain MRI scans are essential for diagnosing and monitoring tumors. At the same time, patients also need ongoing personalized guidance about their diet and the right specialists to see. Each of these tasks, taken individually, is already a hard problem. Taking them together as part of one system that actually helps a patient is even harder.

Deep learning has shown great promise in all three areas. Object detection models like the YOLO family can find and label abnormalities in chest X-rays very quickly [59]. Encoder-decoder architectures like U-Net and its transformer-based variants can precisely segment tumor regions in 3D MRI volumes [32, 60]. Large language models (LLMs) enhanced with Retrieval-Augmented Generation (RAG) can generate personalized recommendations that are grounded in verified medical knowledge [17, 54]. More recently, the emergence of agentic LLM frameworks such as LangChain [14] and LangGraph [43] has made it possible to move past single-turn question answering toward stateful, tool-using assistants that can reason, call external functions, and combine their results over multiple turns—an interaction pattern popularised by the ReAct paradigm [84]. This capability is particularly compelling in healthcare, where a useful assistant must not only talk to the patient but also access their records, look up medicines, and book appointments on their behalf.

This thesis brings together three research projects that we developed and published during our undergraduate program. Each project targets one of these problem areas, and together they form a cohesive study of AI-driven multi-modal medical analysis.

1.2 Problem Statement

The central thesis of this work is: *existing medical AI systems are fragmented—each optimized for a single task in isolation, with no pathway from raw imaging analysis to actionable patient guidance. This thesis addresses that fragmentation by building, evaluating, and integrating three complementary AI systems that together form a diagnostic-to-recommendation pipeline for multi-modal medical analysis.*

Within this unified goal, three concrete technical problems were identified:

1. **Thoracic disease detection is dominated by complex bespoke architectures.** Recent work like YOLOv11-MFF adds frequency-aware gates and multi-scale modules to standard YOLO backbones to improve chest X-ray disease detection. We hypothesize that disciplined data-centric strategies on a stronger base architecture can match or exceed these complex designs without any custom components—reducing engineering overhead while maintaining diagnostic accuracy.

2. **Brain tumor segmentation models fuse features in a distributed, one-to-one manner.** Most U-Net and transformer-based segmentation models pass encoder features to the decoder one level at a time. The decoder begins reconstruction without access to the full multi-scale encoder context. We hypothesize that centralizing multi-scale fusion into a single bottleneck before the decoder begins can provide richer starting representations and improve segmentation of heterogeneous tumor sub-regions.

3. **Strong diagnostic models do not translate to patient guidance.** A model outputting a bounding box or segmentation mask offers no actionable path for a patient. Bridging this last mile from diagnosis to personalized recommendation requires a platform that integrates computer vision outputs with a grounded language model, *and* a conversational interface that can reach into the patient's longitudinal record to act on that guidance. A static RAG pipeline alone cannot retrieve a specific patient's prescription history, book a cardiology appointment near their home, or look up the price of a brand-name medicine in the local market. This is the integrative goal that OptiHealth addresses by pairing proven baseline vision models with a RAG-grounded recommender and a tool-using conversational agent that collectively close the loop from image to action.

1.3 Research Questions

1. Can a standard YOLOv12 model trained with data-centric strategies outperform architecturally complex YOLO variants on multi-class thoracic disease detection?

2. Can centralized multi-scale feature fusion in a bottleneck module improve brain tumor segmentation accuracy compared to the conventional distributed skip-connection approach?

3. Can a research prototype platform combine computer vision diagnostics with an LLM-powered RAG pipeline to deliver evidence-grounded health recommendations under pilot evaluation?

4. Can a tool-augmented, graph-orchestrated conversational agent, built on top of the same platform, safely mediate between the patient and their electronic health record, medicine catalogues, and appointment system while preserving privacy and auditability?

1.4 Outcome of the project

This thesis makes three main contributions, corresponding to the three research papers:

1. **CXR-Adaptive** [2]: A data-centric training and inference protocol built on a standard YOLOv12n detector for multi-class thoracic disease detection on VinDr-CXR, including Inverse-Frequency Weighted Upsampling and CXR-TTA, evaluated against the published YOLOv11-MFF baseline under the same internal stratified split of the 15,000-image training set (750-image evaluation subset, seed 1234). Under that YOLOv11-MFF-aligned protocol—not the official 3,000-image challenge test partition—CXR-TTA attains mAP@0.5 of 0.442, achieving state-of-the-art performance among YOLO-family detectors compared in Table 6.1, in the lineage of strong specialised VinDr-CXR detectors such as YOLO-CXR [28] but without bespoke architectural modules; superiority on the official challenge test partition or over non-YOLO detectors is not asserted.

2. **SS-UNETR**: A novel medical image segmentation architecture with a Multi-Scale Attentive Bottleneck for centralized feature fusion, validated on BraTS 2023, achieving superior Dice scores across all three tumor sub-regions.

3. **OptiHealth**: A multi-modal AI health-management *research prototype* combining the same YOLOv12-based CXR pipeline (with inverse-frequency weighted upsampling), SS-UNETR, and a Llama 4 17B RAG pipeline grounded in recognized manuals stored in ChromaDB, to connect diagnostic outputs with traceable, pilot-evaluated guidance interfaces. OptiHealth further embeds a **Conversational Health Agent** implemented with LangChain and LangGraph [14, 43]: a ReAct-style [84] state graph orchestrates ten patient-scoped tools (EHR access, appointment and prescription queries, geospatial doctor search, appointment creation, real-time medicine

lookup from the MedEx Bangladesh catalogue, and a dual-model mental health analyzer combining a RoBERTa well-being scorer with a BART-MNLI zero-shot symptom detector). The agent runs on a locally-hosted Gemma 4 E4B open-weights model [22], chosen for its hybrid local/global attention, 128K-token context, and explicit support for on-device agentic workflows. All tool invocations respect the same JWT/RBAC context as the rest of the platform and decrypt encrypted personal health records through a per-session PIN propagated via thread-local storage, keeping the agent aligned with OptiHealth's broader safety and auditability goals.

2 Literature Review

2.1 Overview

This chapter reviews the existing literature in three areas that directly inform our work: multi-class chest X-ray disease detection using YOLO-based methods, brain tumor segmentation using CNN and transformer architectures, and AI-powered health management systems with LLMs.

2.2 Chest X-Ray Disease Detection

Chest X-ray imaging is the most widely used first-line screening tool for thoracic diseases including pneumonia, tuberculosis, COPD, and malignancies [68, 79]. Automated analysis is challenging because abnormalities can be subtle, overlapping, and highly variable in appearance, and datasets often exhibit severe class imbalance [53, 69].

Early deep learning approaches focused on single-disease classification. CheXNet [57] demonstrated radiologist-level pneumonia detection using a DenseNet-121. Image-based deep learning has since proven effective for identifying a wide range of treatable diseases [39]. Detection-based methods using YOLO architectures gained popularity for their speed. YOLOv5-based methods showed that real-time detection of thoracic abnormalities was feasible on large-scale datasets like VinDr-CXR [19]. More recent work has proposed architectural customizations to YOLO for CXR analysis. YOLO-CXR [28] added an extra high-resolution detection head to catch small lesions. YOLOv11-MFF [24] introduced Frequency-Adaptive Hybrid Gates and multi-scale parallel convolutions. The key architectural enhancements of YOLOv11 have been surveyed in detail [41, 58].

Addressing class imbalance is another key research direction. The LVIS benchmark introduced Repeat Factor Sampling [26], which was later adapted for long-tail recognition. Multi-label classification with attention mechanisms [36] and feature pyramid networks [47] have also been applied to multi-pathology CXR analysis. Novel multi-label approaches for common thorax diseases have demonstrated strong classification performance [3]. Test-Time Augmentation (TTA) is a well-known inference-time strategy to improve model predictions by averaging over multiple transformed versions of the input,

but applying standard TTA transformations like vertical flips to CXR images produces anatomically invalid images that degrade performance. Object detection methods such as Sparse R-CNN [70], TridentNet [46], and DualAttNet [83] have also been benchmarked on thoracic disease detection tasks, along with the YOLOv9 [74] architecture.

Advanced detection techniques such as saliency-guided sparse detection [6] have explored ways to improve localization in sparse or noisy inputs, a challenge shared with medical imaging. Our CXR-Adaptive framework [2] builds on this prior work by demonstrating that a standard YOLOv12 [72] backbone, combined with principled data-centric strategies, renders domain-specific architectural modules unnecessary.

2.3 Brain Tumor Segmentation

Accurate segmentation of brain tumors from multi-parametric MRI is essential for diagnosis, treatment planning, and monitoring [51]. The BraTS challenge [7, 10, 45] has been the main benchmark for this task since 2012.

The U-Net architecture [60] established the encoder-decoder with skip connections paradigm that almost all modern segmentation models still use. Key innovations include 3D U-Net [16] for volumetric data, UNet++ [87] with nested skip connections, Attention U-Net [55], and the nnU-Net framework [35] which automates architecture configuration.

Vision Transformers (ViTs) [18] brought global self-attention to segmentation. The Swin Transformer [48] introduced shifted-window attention with linear complexity, enabling it to handle 3D volumetric data. Swin UNETR [32] combined a Swin Transformer encoder with a CNN decoder and set strong baselines on BraTS. TransUNet [15], UNETR [33], TransBTS [76], CoTr [82], and Swin-UNet [12] all proposed different ways of combining transformers with segmentation architectures.

A common design principle across all of these is the distributed feature fusion paradigm: encoder features are passed to the decoder one level at a time through skip connections. Our proposed SS-UNETR challenges this by introducing a centralized "many-to-one" fusion bottleneck that aggregates features from five strategic encoder locations before the decoder begins reconstruction.

2.4 AI-Powered Health Management Systems

The integration of AI into digital health platforms has grown substantially. Wearable and IoT-based systems now continuously monitor vital signs and can trigger early interventions [1, 5, 20, 38, 50, 64, 81]. These systems have shown effectiveness in managing chronic conditions like heart disease and diabetes [25, 30, 62, 65]. Comprehensive surveys of real-time health monitoring systems confirm the value of continuous data analysis for early risk detection [56, 61]. Remote monitoring is especially important for cardiac

patients, where early prediction of heart failure events can reduce emergency admissions [4, 8, 9, 73]. Surveillance systems providing population-level data on heart disease and stroke are essential for driving prevention programs [23]. Mobile health technologies, including smartphone applications and wearables, have been used to manage ischemic heart disease, hypertension, and heart failure [34, 40, 49]. Cloud-based health platforms like HealthCloud [17], HYDRA [54], and eWALL [42] integrate predictive analytics with patient-centric workflows. Benchmark datasets such as the PTB Diagnostic ECG Database [11] provide standardized data for evaluating cardiac monitoring algorithms.

In medical image analysis, computer-aided diagnosis systems using YOLO for object detection [37, 59] and Swin UNETR [31] for segmentation have demonstrated strong diagnostic support capabilities. The challenge is connecting these diagnostic outputs to actionable patient guidance. Comparative benchmarks on VinDr-CXR [27] have evaluated a range of detection networks across this challenge.

Large Language Models represent a major opportunity for generating personalized health advice, but they carry risks of generating clinically inaccurate information. RAG pipelines address this by grounding LLM outputs in verified knowledge bases [17]. The MONAI framework [13] supports rapid prototyping of segmentation models in healthcare settings.

Our OptiHealth system combines YOLOv12-based CXR detection (with inverse-frequency weighted upsampling), SS-UNETR-based brain tumor segmentation, and a Llama 4 17B model with a RAG pipeline anchored to recognized health manuals in ChromaDB, implemented as a research prototype that links multi-modal diagnostics with grounded recommendation interfaces rather than as a deployed clinical product.

To consolidate this survey and situate our contribution against the broader landscape, Table 2.1 contrasts OptiHealth with the prevailing health-management paradigm discussed above. The comparison is deliberately feature-by-feature—covering system objective, AI modality coverage, disease scope, personalisation strategy, data handling, system-level output, validation regime, and privacy posture—so that the architectural and methodological choices introduced in the chapters that follow can be read against a single consolidated baseline.

Table 2.1: Comparative analysis of OptiHealth against existing health-management paradigms.

Feature	**Existing Paradigm**	**OptiHealth (Proposed)**
Objective	Clinician-focused diagnostic aid	Holistic, patient-centric health management that also supports clinicians
AI Modalities	Unimodal (typically CV-only)	Multi-modal: YOLOv12 (CXR-Adaptive), SS-UNETR, Llama 4 17B+RAG, and a LangGraph-orchestrated conversational agent
Scope	Single-task or single-disease detectors	Multi-disease, multi-organ analysis integrated across imaging, text, and structured EHR data
Personalisation	Static rules or templated advice	LLM-generated recommendations grounded in a patient-specific EHR context and a curated evidence base
Data Handling	Task-specific datasets in isolation	Unified patient profile that aggregates EMR, self-input, and diagnostic outputs under a consistent schema
System Output	Raw model artefacts (boxes, masks)	Actionable recommendations, audit-logged agent dialogues, and traceable evidence citations
Validation	Quantitative metrics only	Quantitative benchmarks + pilot qualitative evaluation + explicit privacy/audit accounting
Privacy	External inference endpoints common	In-cluster inference (Llama 4 and Gemma 4 E4B both served locally); PHI decryption gated by thread-local PIN

Three contrasts stand out from Table 2.1. First, prior systems typically optimise for a single modality or a single disease, whereas OptiHealth integrates detection, segmentation, retrieval-augmented recommendation, and agentic action under a unified patient schema. Second, personalisation in the surveyed literature generally relies on static rules

or templated advice, while OptiHealth grounds every recommendation in a patient-specific context and a curated evidence base, preserving citation-level traceability from the advice back to its source. Third, and most consequential for clinical deployability, privacy is rarely treated as a first-class design axis in prior work: OptiHealth explicitly serves both Llama 4 and Gemma 4 E4B in-cluster and gates every PHI access through a thread-local PIN, so no patient data leaves the deployment boundary during inference or agent reasoning. The chapters that follow verify these distinctions empirically rather than only architecturally.

2.5 Tool-Augmented LLM Agents for Healthcare

While retrieval augmentation grounds a language model in external documents, a separate line of work allows the model to *act* on external systems through tool use. The ReAct paradigm [84] interleaves natural-language reasoning with tool invocations, letting the model decompose a task, call a function, observe the result, and iterate. Toolformer [63] showed that LLMs can be taught to self-annotate when and how to call APIs, while Chain-of-Thought prompting [78] established the reasoning backbone that later agentic frameworks operationalise. Surveys of LLM-based autonomous agents [75, 80] organise these systems around four recurring components—planning, memory, tool use, and environment—and highlight the recurring trade-off between agent autonomy and safety.

Two open-source frameworks now dominate practical agent construction. LangChain [14] provides high-level abstractions for LLMs, prompts, tools, memory, and chains, and has become the de-facto substrate for RAG and tool-calling applications. LangGraph [43] extends this with a graph-based runtime: developers declare nodes (reasoning steps, tool executors) and conditional edges, turning what was once an opaque prompt loop into an inspectable state machine that supports branching, retries, and human-in-the-loop checkpoints. This makes LangGraph especially attractive for healthcare, where every tool call on a patient record must be auditable.

Healthcare-specific LLMs such as Med-PaLM [66] have reached expert-level accuracy on medical question-answering benchmarks, but they remain stateless oracles—they cannot pull a specific patient's lab results or book an appointment. Recent work on clinical agents begins to close this gap: MedAgents [71] orchestrates multiple LLM "experts" around a clinical question, Almanac [86] couples retrieval with tool use for evidence-grounded clinical reasoning, and patient-facing assistants such as those described in [67] explore multi-turn interaction with longitudinal data. In parallel, the open-weights ecosystem has matured to the point where clinically capable agents can run entirely on local hardware: the Gemma family [21] and most recently Gemma 4 [22]—which interleaves local sliding-window attention with full-global attention, supports a 128K-token context on the E4B variant, and is explicitly tuned for agentic workflows—makes on-premise

deployment of tool-using medical assistants a reproducible alternative to API-hosted models. This matters for healthcare, where data residency and auditability are often non-negotiable.

A separate strand of work focuses on natural-language mental health screening. Zero-shot classification with natural language inference (NLI) models, popularised by BART-MNLI [44, 85], allows symptom categories to be probed without task-specific fine-tuning, while large sentiment encoders such as SiEBERT [29] provide stable valence scores on free-text input. Combining these two families—a scorer that quantifies overall affective state and a detective that flags specific clinical symptoms—gives a lightweight but interpretable screening signal that complements, rather than replaces, formal psychiatric assessment.

Our Health Agent, described in Section 4.4.8, sits at the intersection of these threads. It uses LangGraph to orchestrate a ReAct-style loop over ten patient-scoped tools on top of the OptiHealth backend, runs the reasoning node on a locally-hosted Gemma 4 E4B model, and embeds the SiEBERT + BART-MNLI dual-model analyzer as one of its tools for mental health screening. It therefore extends prior clinical-agent research with a full-stack, privacy-aware, and entirely on-premise deployment rather than a benchmark-only or API-tethered prototype.

3 Dataset

3.1 Dataset Overview

This chapter describes the datasets used across our three works. Each dataset was chosen to match the specific task requirements of the respective paper.

3.2 VinDr-CXR Dataset

For the thoracic disease detection task [2], we used the VinDr-CXR dataset [53]. VinDr-CXR is a large-scale, open-access chest X-ray benchmark collected from two major Vietnamese hospitals. It contains 18,000 postero-anterior (PA) chest X-ray scans.

The training set consists of 15,000 images independently annotated by three radiologists, while the official challenge partition reserves 3,000 images annotated by a consensus of five radiologists for independent test-set evaluation. The dataset covers 14 pathology categories including Aortic Enlargement, Atelectasis, Calcification, Cardiomegaly, Consolidation, ILD, Infiltration, Lung Opacity, Nodule/Mass, Other Lesion, Pleural Effusion, Pleural Thickening, Pneumothorax, and Pulmonary Fibrosis.

The official 3,000-image test partition is held out by the benchmark organisers for challenge submission and is not the evaluation set used in the YOLO-family comparison reported in this thesis. Following YOLOv11-MFF [24], all quantitative VinDr-CXR results here use an *internal* stratified split of the 15,000-image training split only: 80% training (12,000 images), 15% validation (2,250 images), and 5% test (750 images), fixed by random seed 1234. This protocol matches the published baseline and isolates methodological differences from evaluation on the reserved official test images.

Table 3.1: VinDr-CXR Dataset Summary

Property	**Value**
Total Images	18,000
Training Images	15,000
Test Images	3,000
Bounding Box Annotations	22,719
Number of Pathology Classes	14
Annotation (Train)	3 radiologists
Annotation (Test)	5 radiologists (consensus)

A key characteristic of VinDr-CXR is its severe class imbalance. Common structural findings like Aortic Enlargement vastly outnumber rare but critical pathologies like Pneumothorax. This imbalance is a primary motivation for our Inverse-Frequency Weighted Upsampling strategy.

3.3 BraTS 2023 Dataset

For the brain tumor segmentation task, we used the BraTS 2023 Adult Glioma dataset [45, 51]. This is the standard benchmark for brain tumor segmentation, providing pre-operative multimodal MRI scans collected from multiple institutions worldwide.

Each subject in the dataset includes four co-registered MRI modalities: T1-weighted (T1), post-contrast T1-weighted (T1c), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (FLAIR). These modalities provide complementary information about tumor tissue properties, as illustrated in Figure 1.

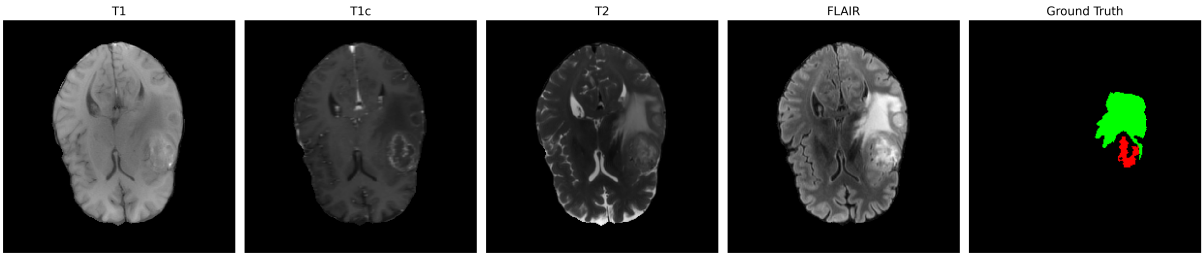


Figure 1: Representative multi-parametric MRI data from the BraTS 2023 dataset showing the four input modalities (T1, T1c, T2, FLAIR) and corresponding ground truth segmentation. Red indicates necrotic/non-enhancing core, green represents peritumoral edema, and blue denotes enhancing tumor.

The ground truth annotations were prepared by expert neuroradiologists and delineate three clinically significant tumor regions:

- **Tumor Core (TC)**: necrotic/non-enhancing core and enhancing tumor (labels 1 + 3)
- **Whole Tumor (WT)**: includes all tumor sub-regions (labels 1 + 2 + 3)
- **Enhancing Tumor (ET)**: the actively enhancing region (label 3)

Table 3.2: BraTS 2023 Dataset Summary

Property	**Value**
Total Training Cases	1,251
Validation Cases	219
MRI Modalities per Subject	4 (T1, T1c, T2, FLAIR)
Tumor Sub-regions (Evaluation)	3 (TC, WT, ET)
Spatial Resolution	$1 \times 1 \times 1$ mm isotropic
Development Split	80% training / 20% internal validation (seed 42)

All MRI volumes are pre-processed by co-registering to the T1c volume, resampling to an isotropic resolution of $1 \times 1 \times 1$ mm^3, skull-stripping, and applying per-patient per-modality z-score normalization on non-zero voxels. In the SS-UNETR training pipeline, the label channels are constructed in the order TC, WT, and ET to match the DiceCE loss and validation-metric computation.

3.4 Knowledge Base for OptiHealth RAG

For the OptiHealth system, a curated medical and nutritional knowledge base was assembled to ground the LLM recommendations. The knowledge base was built from authoritative sources including peer-reviewed nutritional journals, FDA dietary guidelines, World Health Organization recommendations, clinical nutrition textbooks, and the USDA National Nutrient Database.

This corpus, containing over 15,000 documents covering therapeutic diets, macronutrient profiles, food interactions, and cultural dietary preferences, was processed into 384-dimensional embeddings using the Sentence Transformers "all-MiniLM-L6-v2" model and indexed in a ChromaDB vector database for efficient semantic retrieval.

4 Methodology

4.1 Overview

This chapter describes the methodology behind each of our three contributions. The three systems address different medical AI problems but share common design principles: careful treatment of data, respect for domain-specific constraints, and preference for simple, well-understood solutions over complex custom components.

Beyond those shared design principles, the three systems also share a single engineering substrate. Rather than compose each model in a separate environment, we train, evaluate, and deploy CXR-Adaptive, SS-UNETR, and the OptiHealth platform on top of a unified technology stack whose major components are listed in Table 4.1. Introducing the stack here, before any individual system is described, serves three purposes. First, it establishes the deployment envelope that underlies every architectural choice presented in the remainder of the chapter: every checkpoint evaluated in the research sections is architecturally identical to the checkpoint ultimately served inside OptiHealth, because both are trained and served through the same PyTorch-based pipeline and containerised under the same orchestration layer. Second, it consolidates the shared framework vocabulary—PyTorch, MONAI, Hugging Face Transformers, LangChain, LangGraph, ChromaDB—into a single authoritative reference, so that later sections can introduce a framework by name without redefining it or recapitulating its role. Third, it makes the on-premise boundary of the thesis an explicit, up-front constraint: every component in Table 4.1 is either open-source or self-hostable, and the table itself therefore doubles as a declaration that no patient-facing reasoning step in this thesis depends on a third-party inference endpoint.

Each row of Table 4.1 maps to a specific role in the end-to-end workflow. The training stack (PyTorch, MONAI, Hugging Face Transformers) produces every model checkpoint reported in the subsequent sections. The agent stack (LangChain, LangGraph) wraps a locally-served language model with a structured tool-calling graph that forms the orchestration core of the Conversational Health Agent. The storage layer combines a relational store for structured clinical data (PostgreSQL) with a vector index for retrieval-augmented generation (ChromaDB). The delivery layer (Docker, Kubernetes, Kustomize, Ingress-NGINX, GitHub Actions, MinIO) packages each AI service as an independently-

Table 4.1: Unified technology stack shared across CXR-Adaptive, SS-UNETR, and the OptiHealth platform. Every component listed is either open-source or self-hostable, so that the deployment envelope of every contribution is fully on-premise.

Category	**Technology**
Frontend	Next.js 14, React 18, Redux, Tailwind CSS
Backend	Django REST Framework, Python
Databases	PostgreSQL (relational), ChromaDB (vector)
AI/ML Frameworks	PyTorch, Hugging Face Transformers, MONAI
Agent Frameworks	LangChain, LangGraph
AI Models	YOLOv12 (IFWU), SS-UNETR, Llama 4 17B (INT8, RAG), Gemma 4 E4B (local agent backbone)
Containerization	Docker
Orchestration	Kubernetes, Kustomize
CI/CD	GitHub Actions
Object Storage	MinIO
API Gateway	Ingress-NGINX

scalable microservice behind a single reverse proxy, and is instantiated identically for development, evaluation, and deployment. The remainder of this chapter introduces each contribution in turn—CXR-Adaptive (Section 4.2), SS-UNETR (Section 4.3), and the OptiHealth platform (Section 4.4)—each of which draws from the same stack without introducing components that sit outside it.

4.2 Thoracic Disease Detection: CXR-Adaptive

4.2.1 Architectural Paradigm: From Convolution to Attention

Rather than augmenting a YOLO backbone with domain-specific modules, we adopt the standard YOLOv12n (nano) architecture [72] (Figure 2). YOLOv12 shifts from the "Local-First" philosophy of YOLOv11 to a "Global-First" attention-centric paradigm.

The backbone replaces local convolution blocks with the **Area Attention (A^2)** mechanism, which partitions feature maps into regions and models global dependencies within each partition. This reduces the quadratic complexity of full self-attention while still capturing long-range spatial relationships essential for detecting diffuse pathologies like ILD and Infiltration. To support deep attention layers, YOLOv12 uses **Residual Efficient Layer Aggregation Networks (R-ELAN)**, which scale residual shortcuts to maintain stable gradient flow during aggressive fine-tuning. The model was initialized with COCO pre-trained weights for transfer learning.

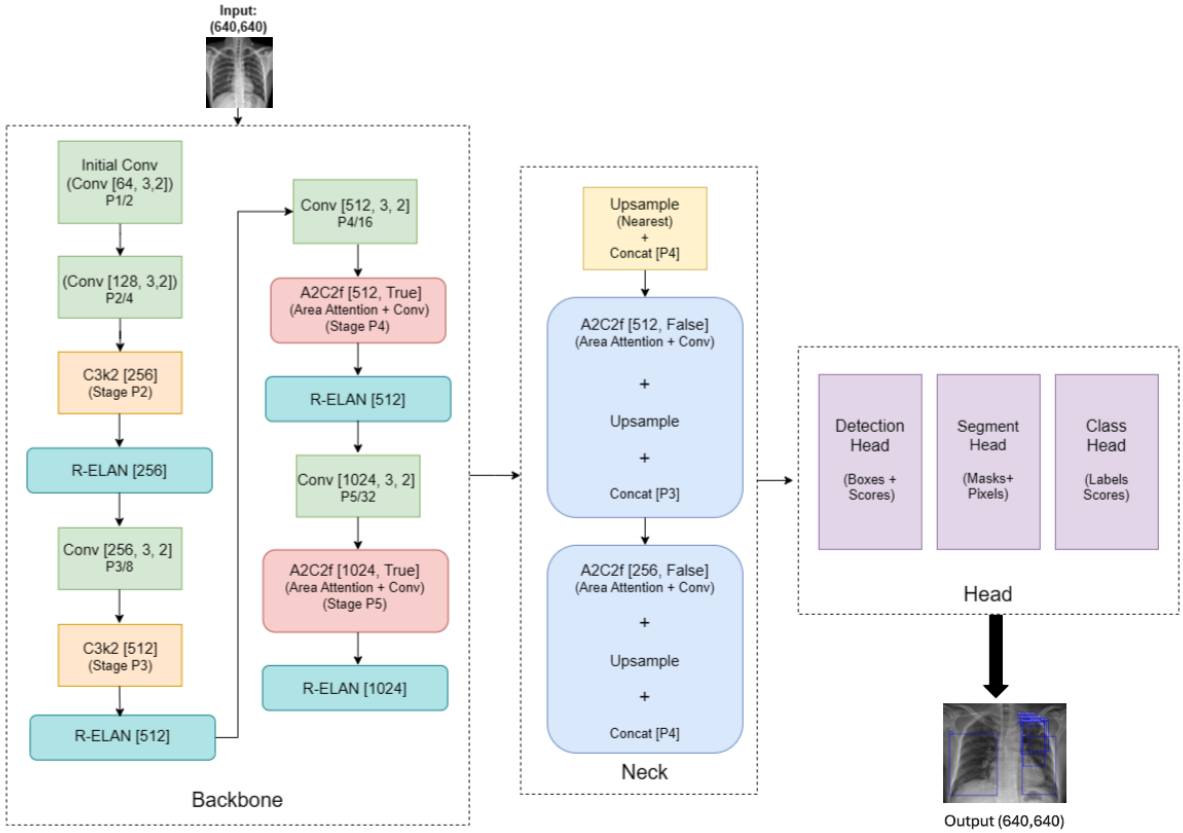


Figure 2: YOLOv12's A^2 Area Attention and R-ELAN modules enable global modeling without any external feature-fusion components.

4.2.2 Inverse-Frequency Weighted Upsampling

Standard random sampling causes poor convergence on rare pathology classes because the model rarely encounters those classes during training. We address this with Inverse-Frequency Weighted Upsampling (Figure 3), adapting the Repeat Factor Sampling concept from the LVIS benchmark [26] to thoracic pathology detection.

Unlike loss-level re-weighting such as Focal Loss, which modulates gradient contributions but does not change how often the model sees rare categories, our strategy operates at the data sampling level. For each pathology class c, a weight w_c inversely proportional to its annotation frequency is computed:

$$w_c = \frac{N_{\text{total}}}{N_c} \tag{4.1}$$

where N_{total} is the total number of bounding boxes and N_c is the count for class c. Each training image I receives a sampling weight by summing the weights of its annotated boxes B_I:

$$W_{\text{img}} = \sum_{b \in B_I} w_{c(b)} \tag{4.2}$$

These weights are normalized by the median image weight and clipped to yield the repeat

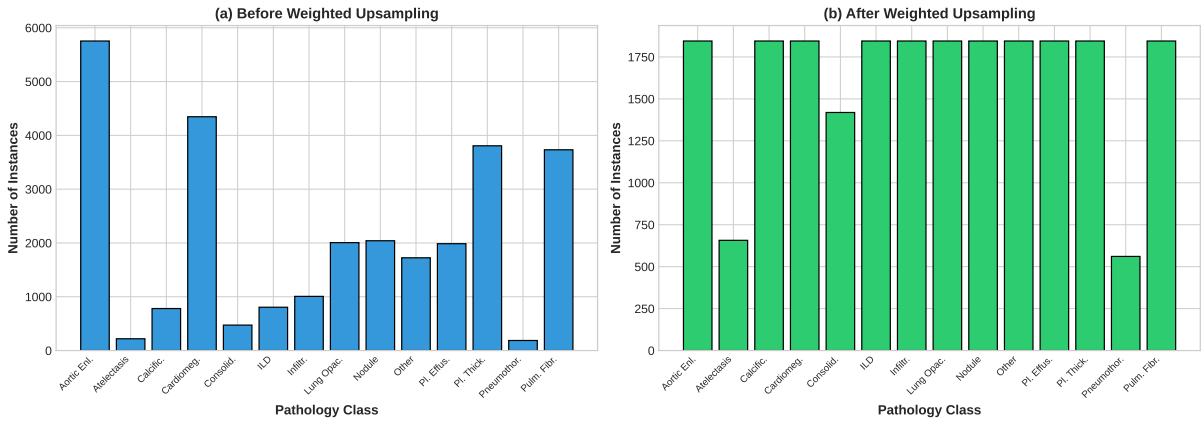


Figure 3: (a) Original biased class distribution in VinDr-CXR; (b) Inverse-frequency weighting ensures frequent exposure to rare classes like Pneumothorax and ILD per training epoch.

factor:

$$R_{\text{img}} = \text{clip}\left(\text{round}\left(\frac{W_{\text{img}}}{\text{median}(W)}\right), 1, 5\right) \tag{4.3}$$

Images containing rare pathologies appear up to 5× more frequently per epoch. The upper bound of 5 prevents gradient instability from unbounded repetition.

4.2.3 CXR-TTA: Anatomically-Constrained Test-Time Augmentation

Standard TTA applies vertical flips and 90-degree rotations, which produce anatomically invalid radiographs—vertical flips invert the superior-inferior axis, and rotations violate the gravitational orientation of pleural fluid and mediastinal structures. We introduce CXR-TTA (Figure 4), restricted to anatomically valid transformations only.

The three inference streams are:

1. The original image at its native scale
2. A horizontal flip (exploiting bilateral lung symmetry)
3. Multi-scale inference at 448px and 512px resolutions

Predictions from all streams are merged via Non-Maximum Suppression (NMS) with IoU threshold $\sigma = 0.5$. NMS preserves the most confident detections and avoids the coordinate drift introduced by Weighted Box Fusion (WBF), which is critical for localization precision.

The 448px scale was found to outperform 640px despite being lower resolution. We hypothesize that slight downscaling acts as spectral regularization, suppressing high-frequency bone texture noise while preserving low-frequency pathology patterns.

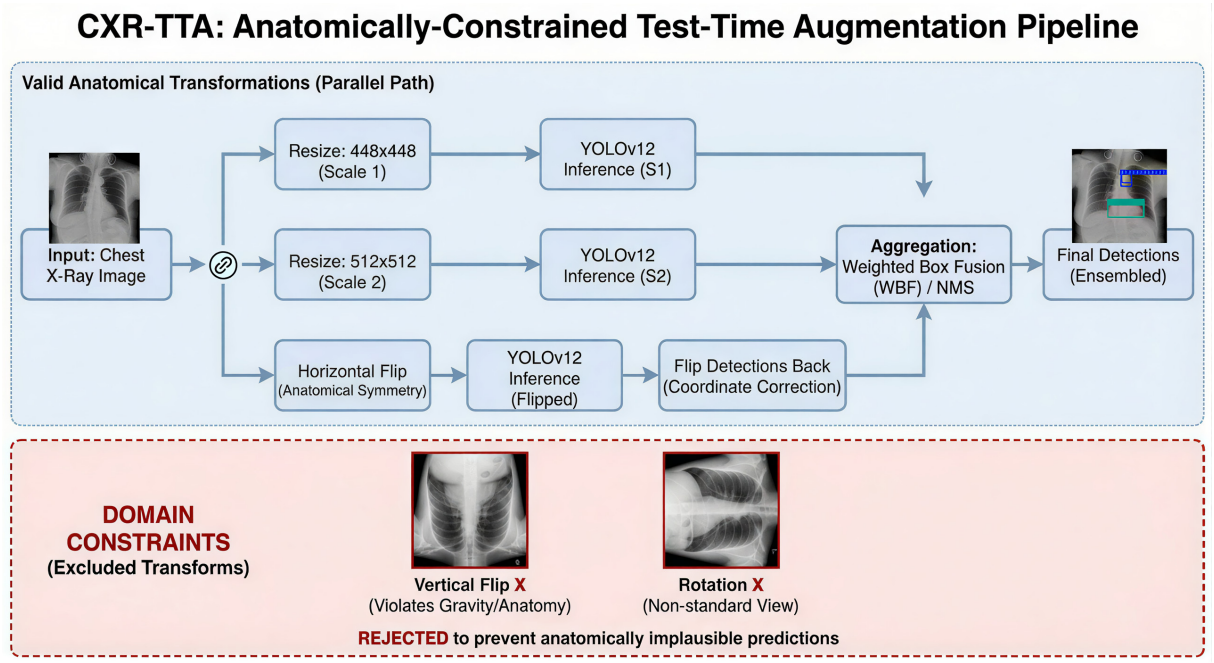


Figure 4: CXR-TTA: Ensembles 448px/512px scales and horizontal flips; anatomically invalid transforms such as vertical flips and rotations are strictly excluded.

4.3 Brain Tumor Segmentation: SS-UNETR

4.3.1 Architectural Motivation and Overall Design

Conventional U-shaped segmentation networks for volumetric brain MRI, including Swin UNETR [32], propagate encoder representations to the decoder through stage-wise skip connections. This design implicitly assumes that each encoder stage encodes information complementary to its symmetrically positioned decoder stage, and that pairwise concatenation is sufficient to recover the spatial detail attenuated during downsampling. In practice, however, the deepest encoder representation alone is responsible for supplying the global semantic context that drives the entire decoding cascade—a single feature map at $H/32 \times W/32 \times D/32$ resolution must summarize tumor morphology, edema extent, and surrounding parenchymal context simultaneously.

We propose SS-UNETR (Symmetric Swin UNETR) to address this bottleneck-capacity limitation. As shown in Figure 5, the architecture comprises three coordinated components: (i) a *dual-path hierarchical encoder* that combines a Swin Transformer V1 backbone with a parallel shallow convolutional branch; (ii) a *Multi-Scale Attentive Bottleneck* (MSAB) that aggregates features from five strategically selected encoder locations through independent self-attention, trilinear spatial harmonization, channel-wise concatenation, and 1×1×1 convolutional fusion; and (iii) a five-stage *symmetric convolutional decoder* that progressively recovers full input resolution through transposed convolutions and residual refinement blocks. The remainder of this section formalizes each component.

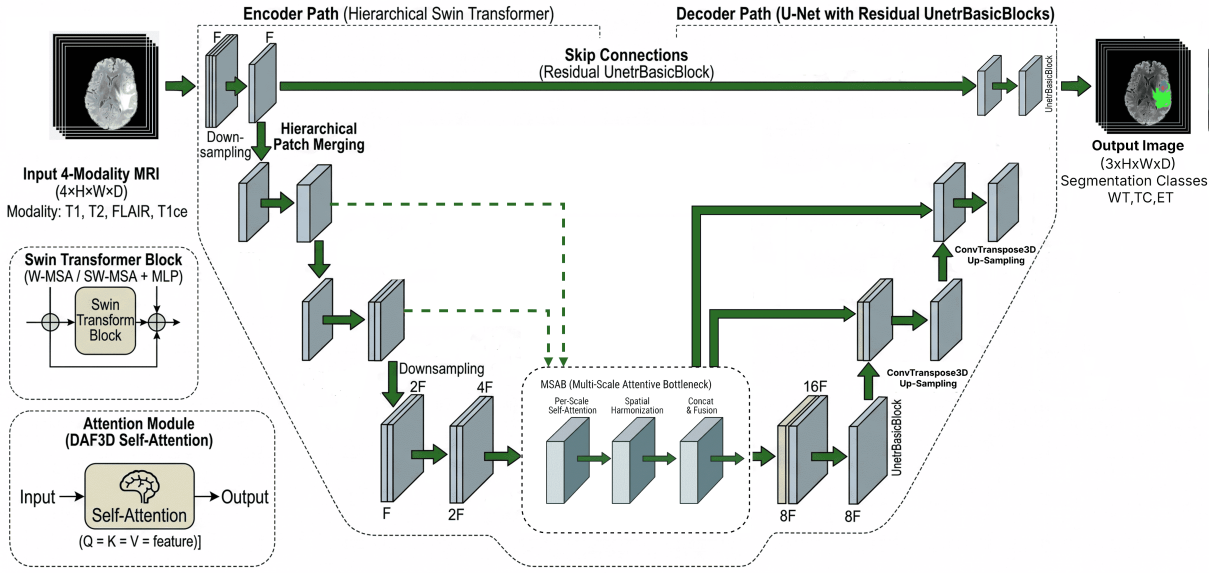


Figure 5: SS-UNETR architecture, comprising a dual-path Swin encoder, the proposed Multi-Scale Attentive Bottleneck, and a symmetric convolutional decoder for TC, WT, and ET segmentation.

4.3.2 Dual-Path Hierarchical Encoder

Given an input multimodal 3D MRI volume $\mathbf{X} \in \mathbb{R}^{4\times H\times W\times D}$ with four channels loaded in the implementation order T1ce, T1, FLAIR, and T2, the encoder produces representations along two complementary pathways.

The **primary pathway** is a four-stage hierarchical Swin Transformer with non-overlapping patch partitioning, shifted window-based multi-head self-attention (W-MSA / SW-MSA), and patch-merging downsampling at the boundary of each stage. We adopt the original Swin Transformer V1 formulation with stage depths $(2, 2, 2, 2)$, head counts $(3, 6, 12, 24)$, and a 7^3 window size. The patch-embedding layer first projects the input volume into a feature representation at half the input resolution:

$$\mathbf{h}_0 = \text{PatchEmbed}(\mathbf{X}) \in \mathbb{R}^{F\times H/2\times W/2\times D/2} \tag{4.4}$$

Each subsequent Swin stage applies two transformer blocks followed by a $2\times$ patch-merging operation that doubles the channel dimension while halving each spatial dimension:

$$\mathbf{h}_i = \text{SwinStage}_i(\mathbf{h}_{i-1}), \quad i \in \{1, 2, 3, 4\} \tag{4.5}$$

producing the hierarchy of feature maps summarized in Table 4.2.

The **auxiliary pathway** is a shallow residual convolutional branch that operates directly on the original input volume:

$$\mathbf{enc}_0 = \text{UnetrBasicBlock}(\mathbf{X}) \in \mathbb{R}^{F\times H\times W\times D} \tag{4.6}$$

implemented as a sequence of 3^3 convolutions, instance normalization, ReLU activation, and a residual connection. This branch preserves the high-frequency texture information that is irrecoverably lost when the patch-embedding layer downsamples by a factor of two. The full-resolution feature $\mathbf{enc}_0$ is later consumed both as a multi-scale input to the bottleneck and as the shallowest decoder skip connection, ensuring that voxel-level spatial fidelity is restored at the final output stage.

Table 4.2: SS-UNETR layer dimensions for an input volume of shape $4 \times H \times W \times D$ with base feature size $F = 48$.

Stage	**Spatial Resolution**	**Channels**
Shallow Conv ($\mathbf{enc}_0$)	$H \times W \times D$	F
Patch Embed ($\mathbf{h}_0$)	$H/2 \times W/2 \times D/2$	F
Swin Stage 1 ($\mathbf{h}_1$)	$H/4 \times W/4 \times D/4$	$2F$
Swin Stage 2 ($\mathbf{h}_2$)	$H/8 \times W/8 \times D/8$	$4F$
Swin Stage 3 ($\mathbf{h}_3$)	$H/16 \times W/16 \times D/16$	$8F$
Swin Stage 4 ($\mathbf{h}_4$)	$H/32 \times W/32 \times D/32$	$16F$
Bottleneck Output ($\mathbf{B}_{\text{out}}$)	$H/32 \times W/32 \times D/32$	$32F$

4.3.3 Multi-Scale Attentive Bottleneck

The Multi-Scale Attentive Bottleneck (MSAB) constitutes the principal architectural contribution of SS-UNETR. Rather than relying on the deepest encoder feature alone to provide global semantic context, the MSAB constructs the bottleneck representation from five feature maps drawn from across the encoder hierarchy: $\{\mathbf{enc}_0, \mathbf{h}_0, \mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_4\}$.

Notably, $\mathbf{h}_3$ is deliberately omitted from this set. Including all five Swin outputs together with $\mathbf{enc}_0$ would consume $\mathbf{h}_3$ in the bottleneck and leave no skip connection at the $H/16$ resolution, forcing the deepest decoder stage to upsample directly from the bottleneck without an intermediate-resolution guide. By withholding $\mathbf{h}_3$ from the bottleneck and routing it instead as the skip connection to the deepest decoder block (Decoder 5), SS-UNETR preserves a complete five-resolution decoder skip chain while still aggregating multi-scale context from the remaining four Swin outputs and the shallow auxiliary branch.

Per-scale self-attention. Each of the five selected feature maps is first refined independently through a non-local 3D attention block adapted from the deeply attentive feature aggregation module of DAF3D [77]. We denote this operation as $\text{AttentionModule}(\cdot, \cdot)$; when invoked with the same feature map as both arguments, it acts as a self-attention refinement:

$$\mathbf{f}_k^{\text{att}} = \text{AttentionModule}(\mathbf{f}_k, \mathbf{f}_k), \quad \mathbf{f}_k \in \{\mathbf{enc}_0, \mathbf{h}_0, \mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_4\} \tag{4.7}$$

The block internally concatenates its two inputs along the channel dimension, projects them through convolutional sub-layers, normalizes the response with instance normalization, and produces a refined feature map of the same shape as $\mathbf{f}_k$. This per-scale refinement allows each resolution to emphasize its most informative spatial locations before being incorporated into the global fusion.

Spatial harmonization. Because the five attended feature maps occupy four distinct spatial resolutions, they cannot be combined directly. We harmonize them to the deepest resolution—that of $\mathbf{h}_4$, namely $H/32 \times W/32 \times D/32$—through trilinear interpolation:

$$\mathbf{f}_k^{\text{harm}} = \text{Interpolate}_{\text{trilinear}}\big(\mathbf{f}_k^{\text{att}},\ \text{size} = (H/32,\ W/32,\ D/32)\big) \tag{4.8}$$

Downsampling to the smallest common resolution (rather than upsampling to the largest) preserves channel information from the deepest stage exactly while keeping the bottleneck volume computationally tractable.

Channel-wise concatenation and fusion. The five harmonized feature maps are concatenated along the channel dimension and fused through a $1\times1\times1$ convolution that learns an optimal cross-scale weighting:

$$\mathbf{F}_{\text{cat}} = \text{Concat}\big[\mathbf{f}_{\mathbf{enc}_0}^{\text{harm}},\ \mathbf{f}_{\mathbf{h}_0}^{\text{harm}},\ \mathbf{f}_{\mathbf{h}_1}^{\text{harm}},\ \mathbf{f}_{\mathbf{h}_2}^{\text{harm}},\ \mathbf{f}_{\mathbf{h}_4}^{\text{harm}}\big] \tag{4.9}$$

$$\mathbf{B}_{\text{out}} = \text{Conv}_{1\times1\times1}(\mathbf{F}_{\text{cat}}) \tag{4.10}$$

The concatenated tensor has $F+F+2F+4F+16F = 24F$ channels, which the fusion convolution projects to $32F$ output channels. The resulting tensor $\mathbf{B}_{\text{out}} \in \mathbb{R}^{32F\times H/32\times W/32\times D/32}$ is a compact representation that simultaneously encodes high-frequency spatial detail (from $\mathbf{enc}_0$ and $\mathbf{h}_0$) and high-level semantic abstraction (from $\mathbf{h}_4$), giving the decoder a substantially richer starting point than any single-scale bottleneck could provide.

4.3.4 Symmetric Convolutional Decoder

The decoder mirrors the encoder with five `CustomUpBlock` stages, each applying a $2\times$ transposed 3D convolution to upsample its input, concatenating the result with the corresponding encoder skip along the channel axis, and refining the merged feature map through a `UnetrBasicBlock` (a 3^3 convolution–instance normalization–ReLU–3^3 convolution–instance normalization sequence with a residual connection). The full decoder cascade

is:

$$\mathbf{dec}_5 = \text{CustomUpBlock}(\mathbf{B}_{\text{out}},\ \mathbf{h}_3) \tag{4.11}$$

$$\mathbf{dec}_4 = \text{CustomUpBlock}(\mathbf{dec}_5,\ \mathbf{h}_2) \tag{4.12}$$

$$\mathbf{dec}_3 = \text{CustomUpBlock}(\mathbf{dec}_4,\ \mathbf{h}_1) \tag{4.13}$$

$$\mathbf{dec}_2 = \text{CustomUpBlock}(\mathbf{dec}_3,\ \mathbf{h}_0) \tag{4.14}$$

$$\mathbf{dec}_1 = \text{CustomUpBlock}(\mathbf{dec}_2,\ \mathbf{enc}_0) \tag{4.15}$$

$$\mathbf{Y} = \text{UnetOutBlock}(\mathbf{dec}_1) \tag{4.16}$$

where each $\mathbf{dec}_k$ has the same spatial resolution as its corresponding skip connection, and the final UnetOutBlock is a 1×1×1 convolution that maps F feature channels to three segmentation channels corresponding to Tumor Core (TC), Whole Tumor (WT), and Enhancing Tumor (ET). The output tensor $\mathbf{Y} \in \mathbb{R}^{3\times H\times W\times D}$ recovers the full input resolution.

We retain a convolutional decoder—rather than a symmetrically transformer-based one—because the decoding task is dominated by local spatial refinement under the guidance of skip connections, where convolutional inductive biases (translation equivariance, local connectivity) are advantageous. The MSAB-fused bottleneck supplies the global context that a purely convolutional decoder would otherwise lack, decoupling the responsibility of global reasoning (handled by the encoder and bottleneck) from the responsibility of local reconstruction (handled by the decoder).

4.3.5 Loss Function

SS-UNETR is trained with a compound DiceCE loss that combines region-overlap supervision with voxel-level cross-entropy:

$$\mathcal{L} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{CE}} \tag{4.17}$$

where, for N voxels and per-class softmax probabilities p_i against one-hot ground truth g_i,

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N} p_i g_i + \epsilon}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i + \epsilon} \tag{4.18}$$

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{N} g_i \log(p_i) \tag{4.19}$$

The Dice term directly optimizes the volumetric overlap metric used at evaluation time and is robust to the severe foreground–background class imbalance characteristic of brain tumor segmentation, where pathological tissue typically occupies less than 2% of total

voxels. The cross-entropy term provides dense per-voxel supervision that stabilizes training in the early epochs when Dice gradients are nearly flat. The smoothing constant $\epsilon = 10^{-5}$ prevents division by zero on empty foreground regions.

4.4 Integrated Health Management: OptiHealth

4.4.1 Relationship to CXR-Adaptive and SS-UNETR

The preceding two sections introduced CXR-Adaptive (using YOLOv12) and SS-UNETR as novel, research-grade contributions evaluated under controlled experimental protocols. OptiHealth was developed as a separate, parallel contribution with a different objective: to demonstrate that an end-to-end health-management *software stack* integrating multimodal AI with grounded language generation is architecturally feasible as a research prototype, without claiming deployment as a regulated clinical product.

Non-clinical scope. Descriptions of recommendations, agents, and workflows in this section denote research-software behaviour for thesis evaluation and supervised piloting; they are not substitutes for professional diagnosis, treatment decisions, or emergency care.

OptiHealth uses our **YOLOv12**-based CXR pipeline with inverse-frequency weighted upsampling and **SS-UNETR** for brain tumor segmentation to keep the full system aligned with the core model contributions of this thesis. This keeps the end-to-end integration consistent with the same diagnostic engines evaluated in the prior chapters.

4.4.2 Multi-Layered System Architecture

OptiHealth is an AI-powered digital platform designed to promote preventive healthcare and personalized wellness. It collects health data from medical records and user self-assessments to build a complete health profile. The overall architecture, shown in Figure 6, comprises a Next.js frontend, a Django REST API gateway, dockerized AI services, and a polyglot storage layer.

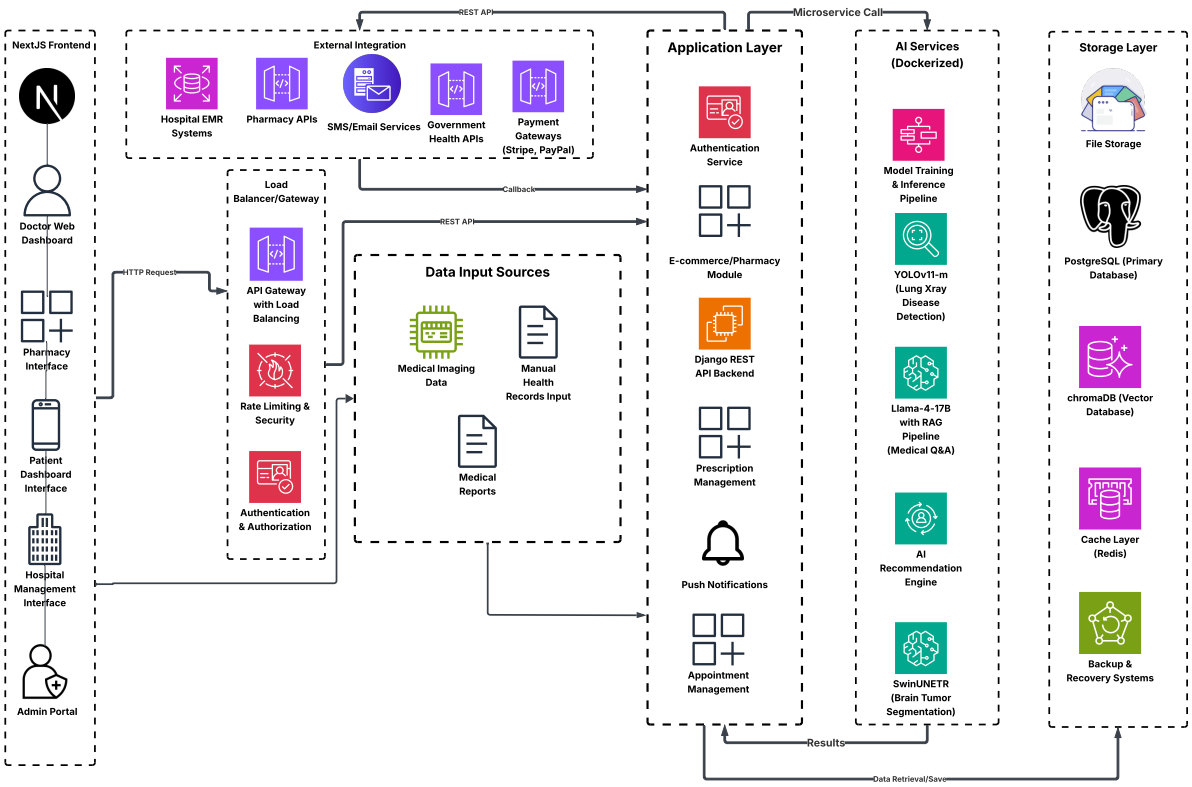


Figure 6: Multi-layered OptiHealth system architecture.

The high-level design integrates all primary healthcare stakeholders—patients, physicians, hospitals, diagnostic centers, pharmacies, and system administrators—within a single cohesive ecosystem. Figure 7 shows the context-level data flow diagram.

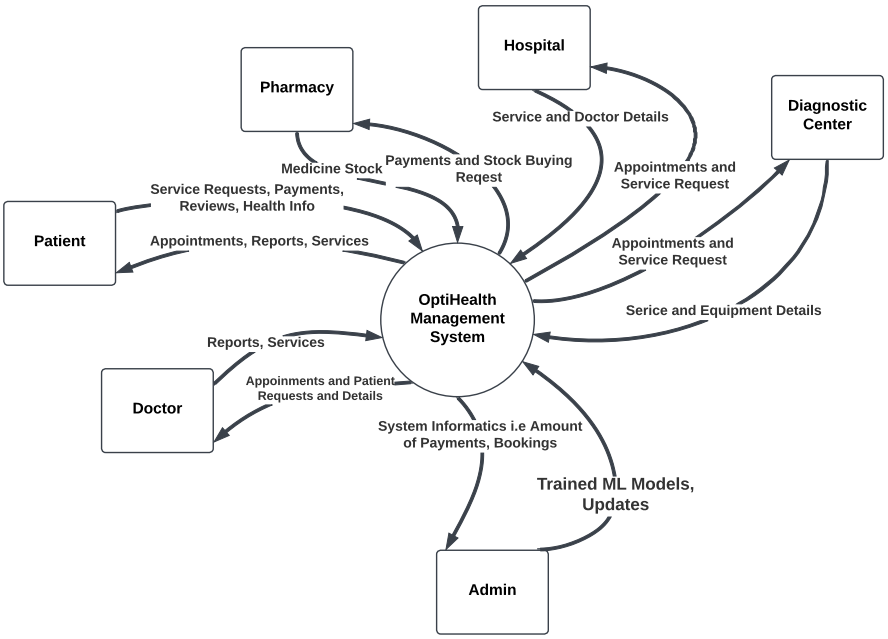


Figure 7: Context-level stakeholder interaction diagram.

4.4.3 Technology Stack: Reuse of the Shared Substrate

OptiHealth is built on the unified technology stack introduced in Section 4.1 and catalogued in Table 4.1; no row of that table is specific to this section, and no component outside it is introduced by the platform. Concretely, the integrated system inherits the same PyTorch-based training pipeline used to produce the CXR-Adaptive and SS-UNETR checkpoints evaluated in the preceding sections, the same PostgreSQL and ChromaDB storage layer that indexes relational EHR data alongside vector-embedded guideline corpora, and the same Docker and Kubernetes orchestration envelope that hosts every AI service inside the on-premise deployment boundary. The remainder of this section therefore focuses not on enumerating technologies—which Table 4.1 already does authoritatively—but on describing how those components are composed into an integrated health-management prototype that reuses the individual contributions of the thesis without modification.

4.4.4 Microservices Architecture

The system uses a microservices architecture for modularity and independent scalability.

Frontend and Backend

The frontend is built with Next.js 14 and React 18, providing server-side rendering and role-based interfaces for patients, providers, and administrators. The backend uses a Django REST Framework API gateway with JWT authentication and role-based access control (RBAC). State management uses Redux Toolkit.

Database Services

The system uses a polyglot persistence approach:

- **PostgreSQL**: primary relational database for user profiles, appointment records, and medical histories
- **ChromaDB**: vector database for the medical knowledge base, storing 384-dimensional embeddings for semantic search
- **Redis**: cache layer for session management and API response caching

AI Services (Dockerized)

AI models are containerized using Docker for consistency across environments. The computer-vision service houses YOLOv12 (with inverse-frequency weighted upsampling)

and SS-UNETR with NVIDIA GPU support; the NLP service hosts a Llama 4 17B checkpoint quantised to INT8 with ChromaDB integration for retrieval-augmented generation; and a dedicated **Health Agent service** hosts the LangChain/LangGraph runtime described in Section 4.4.8, together with a locally-served Gemma 4 E4B checkpoint as its reasoning backbone. A separate mental-health inference worker loads the dual-model screening pipeline (a RoBERTa sentiment scorer and a BART-MNLI zero-shot classifier) so that the analyser can be scaled independently of the agent itself.

4.4.5 Computer Vision Engine: X-Ray Detection

For thoracic pathology detection in chest X-rays, the system uses our YOLOv12 model with inverse-frequency weighted upsampling, trained on the VinDr-CXR dataset (Figure 2).

Training used transfer learning from COCO pre-trained weights with an AdamW optimizer and cosine annealing LR schedule (initial rate 0.001), 300 epochs with early stopping based on validation mAP. Data augmentation included horizontal flipping, rotation ($\pm 15°$), and brightness adjustments ($\pm 20\%$). For inference, TensorRT optimization with FP16 computation achieves under 200ms per image. Post-processing uses NMS with IoU threshold 0.45 and confidence threshold 0.25.

4.4.6 Computer Vision Engine: Brain Tumor Segmentation

For volumetric brain tumor segmentation in multimodal MRI, the system employs SS-UNETR, the architecture introduced in Section 4.3 and illustrated in Figure 5. The model is deployed without modification from its research configuration, ensuring that the integrated platform reflects the same diagnostic engine that was evaluated under controlled experimental protocols in the preceding chapters.

Training was conducted on the BraTS 2023 Adult Glioma dataset using the MONAI transformation pipeline detailed in Section 5.1.3. The preprocessing sequence loads the four NIfTI modalities and segmentation mask, converts the image tensor to channel-first format, maps the BraTS integer labels into three binary evaluation channels (TC, WT, and ET), reorients image and label volumes to RAS, resamples both to $1.0 \times 1.0 \times 1.0\,\text{mm}$ spacing, and applies non-zero channel-wise intensity normalisation. During training, a fixed-size random spatial crop of $128 \times 128 \times 128$ is sampled, independent random flips are applied along all three spatial axes with probability 0.5, and intensity scale and shift perturbations of 0.1 are applied with probability 1.0. Optimisation uses AdamW with learning rate 10^{-4}, weight decay 10^{-5}, a cosine-annealing learning-rate schedule over 50 epochs, mixed-precision training, gradient checkpointing, and checkpoint selection by validation mean DSC. The training objective is the compound DiceCE loss defined in Section 4.3.5, which combines a region-overlap Dice term with a voxel-level cross-entropy

term to address the severe foreground–background class imbalance inherent to tumor segmentation. Validation uses single-checkpoint sliding-window inference with a 128^3 ROI, sliding-window batch size one, 50% overlap, sigmoid activation, and a 0.5 threshold; no test-time augmentation or ensemble validation is used for the thesis results.

4.4.7 Personalized Diet Generation: LLM + RAG Pipeline

A core innovation of OptiHealth is the personalized nutrition module, which implements a Retrieval-Augmented Generation (RAG) pipeline to ground all recommendations in verified medical knowledge (Figure 8).

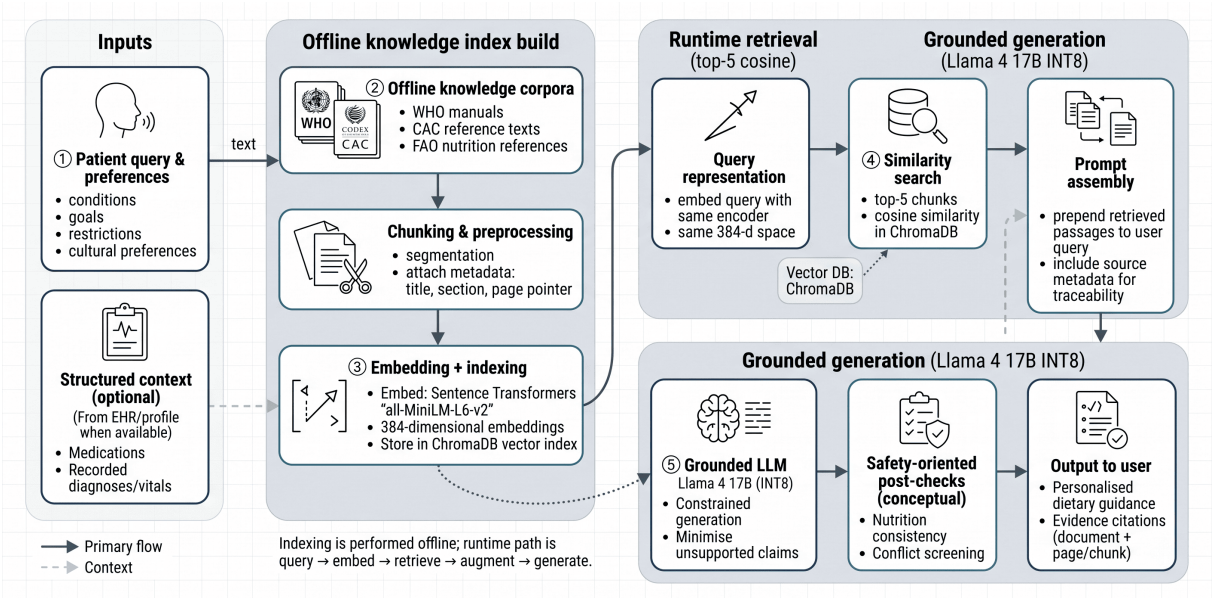


Figure 8: Retrieval-Augmented Generation pipeline for personalised diet recommendation.

To reduce hallucination risk, the RAG knowledge base is built from recognized health manuals and guidelines, including WHO, CAC, and FAO documents, then stored in ChromaDB for retrieval. The generative core is a Llama 4 17B model quantised to INT8 for efficient single-GPU inference. The knowledge base is indexed as 384-dimensional embeddings using Sentence Transformers "all-MiniLM-L6-v2" in ChromaDB. When a dietary plan is returned to the user, the system also shows the evidence source, including the manual name and page number used for that recommendation. The generation process follows five steps:

1. **Query Decomposition**: Extract key entities (health conditions, dietary restrictions, goals)

2. **Context Retrieval**: Top-5 similarity search against ChromaDB using cosine similarity across WHO, CAC, and FAO manual chunks

3. **Prompt Engineering**: Prepend retrieved evidence to the user query

4. **Constrained Generation**: Llama 4 17B generates recommendations grounded in the retrieved context to minimize hallucinated advice

5. **Post-processing**: Programmatic checks for nutritional consistency, caloric balance, and allergen conflicts

4.4.8 Conversational Health Agent: LangGraph Orchestration

Retrieval-Augmented Generation answers a single dietary question at a time and has no memory of who is asking. For OptiHealth to act as a longitudinal assistant, the platform also exposes a tool-augmented **Conversational Health Agent** that can access a specific patient's electronic record, look up medicines in the local market, search for doctors by specialization and geography, book appointments, and screen for mental health risk through natural-language conversation. The agent is implemented with LangChain [14] (tool binding and LLM abstraction) and LangGraph [43] (state-graph orchestration), and follows the ReAct paradigm [84] of interleaving reasoning with tool invocations.

Architectural Motivation

A straight prompt-and-response chatbot cannot safely operate on personal health information: it must decide *when* a question requires real data, *which* tool to call, *how* to combine results across turns, and *how* to refuse safely when data is missing or when a request falls outside its scope. Modelling this as a directed state graph rather than a hidden prompt loop is a deliberate design choice: every tool call becomes an inspectable transition, which is a prerequisite for audit logging and clinical governance.

Figure 9 shows the high-level architecture of the Conversational Health Agent.

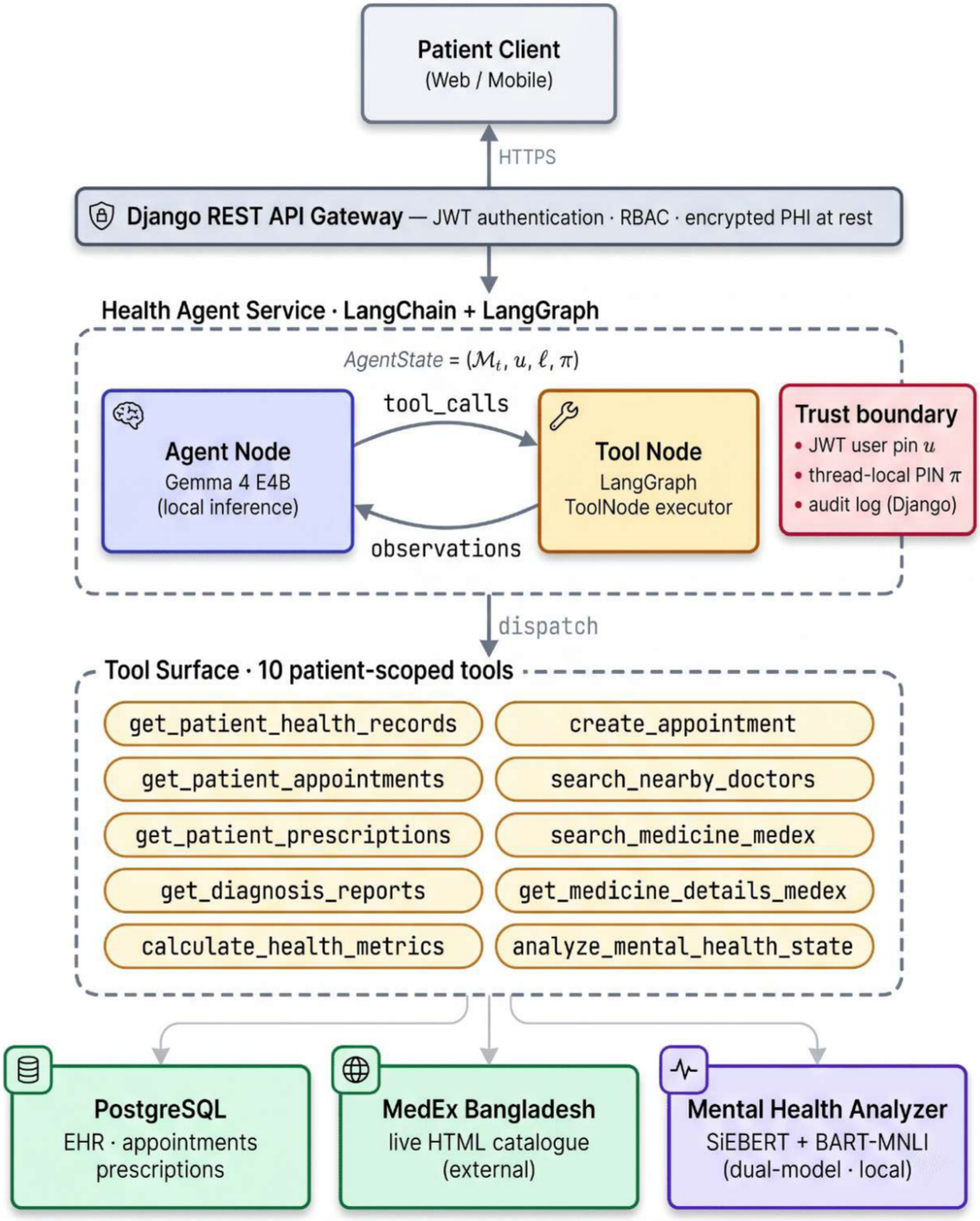


Figure 9: High-level architecture of the OptiHealth Conversational Health Agent.

State Graph and ReAct Execution Loop

The agent is formulated as a finite `StateGraph` over a typed `AgentState` that carries the running message list, the authenticated patient's `user_id`, the patient's geolocation, and a per-session encryption PIN used to decrypt protected health fields. Formally, let the state at turn t be

$$s_t = (\mathcal{M}_t,\ u,\ \ell,\ \pi),$$

where $\mathcal{M}_t$ is the ordered message history, u is the user identifier, ℓ the optional latitude/longitude, and π the PHI decryption PIN. The graph contains two nodes: an **agent node** (LLM reasoning) and a **tool node** (tool execution). A conditional edge inspects the agent's most recent message and routes the state either back to the tool node (if the LLM emitted one or more structured `tool_calls`) or to the terminal `END` state (if the LLM produced a natural-language answer). The tool node always returns to the agent node, producing the canonical ReAct reasoning–action–observation cycle:

$$\underbrace{s_t \to \text{AGENT}}_{\text{reason}} \ \to\ \underbrace{\text{TOOLS}(a_t) \to s_{t+1}}_{\text{act and observe}} \ \to\ \ldots\ \to\ \text{END},$$

where a_t is the set of tool calls emitted by the LLM at turn t. In practice the loop terminates within 1–3 iterations for most patient queries.

The state graph execution flow is illustrated in Figure 10.

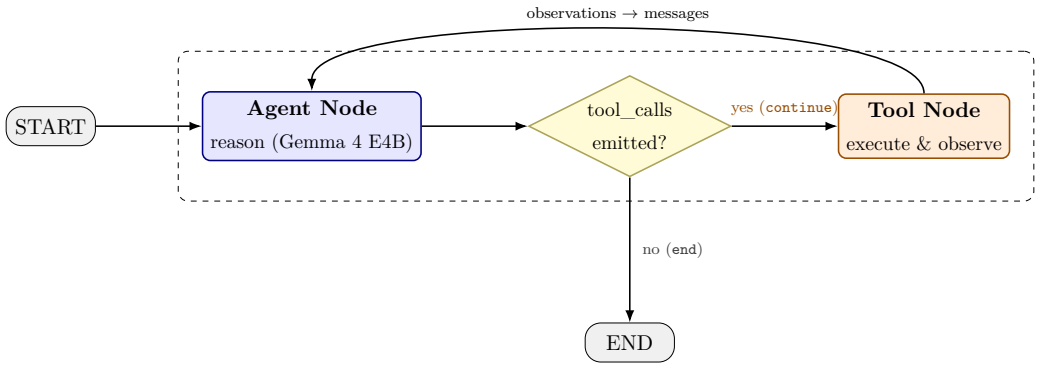


Figure 10: LangGraph state machine implementing the Health Agent reasoning–tool-use loop.

Tool Suite

Ten tools are registered with the LangChain tool registry and exposed to the LLM as structured function schemas. Each tool wraps a pure Python function that runs inside the Django process, giving it first-class access to the relational models while still being callable from the LLM through the tool-calling interface. Table 4.3 enumerates the full tool surface.

Two of these tools deserve particular comment. The `search_medicine_medex` tool

Table 4.3: Tools exposed to the Conversational Health Agent. All EHR-bound tools respect the same JWT authentication and RBAC policy as the REST API.

Tool	**Backing Source**	**Purpose**
`get_patient_health_records`	PostgreSQL (EHR)	Recent vitals, trends, allergies
`get_patient_appointments`	PostgreSQL	Upcoming/past appointments
`get_patient_prescriptions`	PostgreSQL	Active medications
`get_diagnosis_reports`	PostgreSQL	Diagnostic reports, lab findings
`calculate_health_metrics`	Derived	BMI, BP trend, risk indicators
`search_nearby_doctors`	PostgreSQL + Haversine	Geo-filtered specialist search
`create_appointment`	PostgreSQL (write)	Book a slot with a doctor
`search_medicine_medex`	MedEx BD (live scrape)	Brand/generic medicine lookup
`get_medicine_details_medex`	MedEx BD (live scrape)	Composition, dosage, interactions
`analyze_mental_health_state`	Dual-model NLP	Well-being score + symptom flags

issues a live request to the MedEx Bangladesh catalogue (`medex.com.bd`) and parses the returned HTML with BeautifulSoup, returning brand names, generics, prices, manufacturers, and dosage forms in near real time; this gives the agent access to price-aware medicine information that static training data cannot provide. The `create_appointment` tool is the only *mutating* tool in the suite and is therefore subject to additional validation (doctor availability, slot collision, and a pre-commit confirmation turn in the agent prompt) before the write is executed.

LLM Backbone: Locally-Hosted Gemma 4 E4B

The reasoning node is served by a **Gemma 4 E4B** checkpoint [22] loaded locally inside the Health Agent service, and bound to the tool suite through the standard LangChain chat-model interface. Gemma 4 E4B was selected over larger-but-API-only alternatives for four reasons that are directly relevant to a clinical deployment:

1. **On-premise execution.** Patient data never leaves the OptiHealth cluster. Because Gemma 4 is an open-weights model, the entire inference pipeline runs inside the same Kubernetes namespace as the Django backend and the PostgreSQL EHR, removing the third-party data-sharing boundary that would otherwise accompany a hosted LLM API.

2. **Hybrid attention for long clinical context.** Gemma 4's architecture interleaves local sliding-window attention with full-global attention (ensuring the final layer is always global), and applies unified Keys/Values with Proportional RoPE (p-RoPE) on the global layers. The E4B variant supports a 128K-token context window, which is more than sufficient to hold a full patient chart—recent health records, active prescriptions, past appointments, and diagnostic reports—within a single reasoning turn.

3. **Parameter efficiency through PLE.** The "E" in E4B denotes *effective* parameters: Gemma 4 E4B uses Per-Layer Embeddings (PLE) to give each decoder layer its own small embedding table, so only ∼4.5B parameters participate in the forward pass at inference time even though the total weight count is closer to 8B. This makes real-time tool-calling feasible on a single GPU.

4. **Agentic tuning.** Gemma 4 is explicitly documented as being tuned for reasoning and agentic workflows, and its chat template surfaces structured tool calls that slot directly into the LangChain tool-binding interface used by the graph.

The patient's `user_id` is interpolated into the system prompt at turn time so that the model always invokes EHR-bound tools with the correct identity, structurally preventing cross-patient leakage even under adversarial prompting.

Mental Health Screening via a Dual-Model Analyzer

The `analyze_mental_health_state` tool is not a simple LLM call; it routes the patient's free-text narrative through a dedicated, deterministic NLP pipeline with two specialised encoders operating in parallel:

- **Model A — Scorer:** `siebert/sentiment-roberta-large-english` [29] produces a calibrated well-being score on a 1–100 scale by projecting the binary `POSITIVE/NEGATIVE` probabilities onto a continuous affective valence axis.

- **Model B — Detective:** `facebook/bart-large-mnli` [44, 85] performs zero-shot multi-label classification against eight clinical symptom clusters—*depression, anxiety, sleep disorders, social withdrawal, cognitive issues, physical symptoms, suicidal ideation,* and *mood swings*—each expanded into six natural-language hypothesis templates so that the model's confidence is aggregated across paraphrases.

The output is an interpretable record containing a numeric well-being score, a ranked list of flagged symptom clusters with their NLI confidences, and a severity classification. Suicidal-ideation detections immediately override the aggregate severity and trigger an in-response crisis-resource block. The dual-model design intentionally separates valence ("how is the patient feeling overall?") from symptomatology ("which specific clinical categories are present?"), producing a signal that is both quantitative and clinically structured.

Figure 11 shows the dual-model analyzer pipeline.

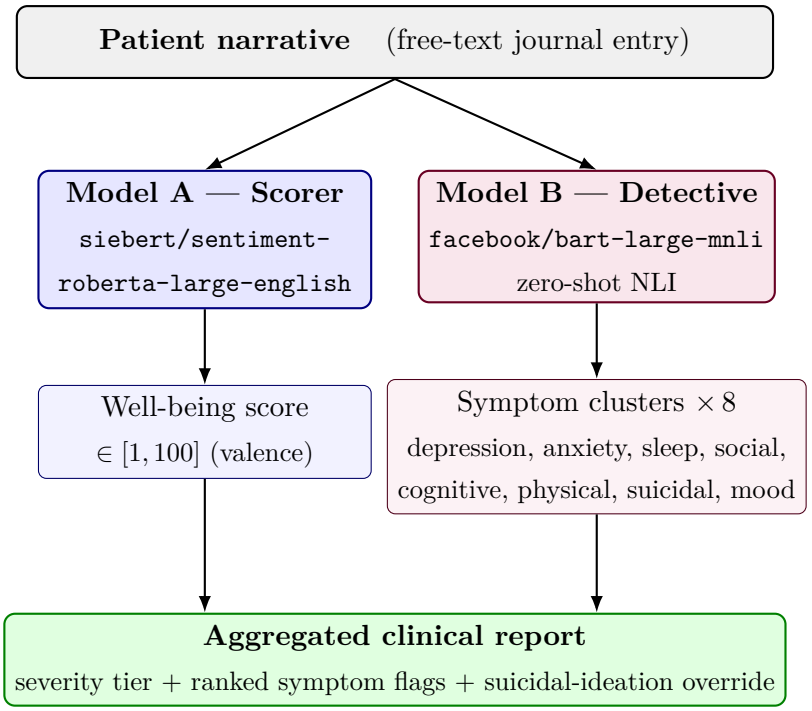


Figure 11: Dual-model mental-health analysis pipeline used by the Conversational Health Agent.

Privacy, Authentication, and Auditability

Because the agent reaches directly into the patient's record, its trust boundary must be at least as strict as the REST API. Three mechanisms enforce this:

1. **Identity pinning.** The authenticated `user_id` is injected into the system prompt and is also passed as a required argument to every EHR-bound tool; the LLM is structurally prevented from operating on a different patient even if prompted to do so.

2. **PHI decryption via thread-local PIN.** Sensitive fields such as allergies and medical history are stored encrypted at rest. At session start the patient unlocks a per-session PIN, which is placed in Python `threading.local` so that tool functions can decrypt records on demand without the PIN ever appearing in the LLM's context window or log trail.

3. **Structured conversation log.** Every user message, tool call, tool result, and final response is persisted through the `AgentConversation` and `AgentMessage` Django models, giving the platform a replayable audit log that meets the auditability expectation set by the microservices architecture.

Together, the LangGraph state machine, the ten-tool surface, the locally-hosted Gemma 4 E4B backbone, and the dual-model mental health analyzer constitute a single, cohesive conversational layer on top of the OptiHealth platform. The agent does not replace the vision

models or the RAG recommender; it composes them into a patient-centric dialogue that turns the diagnostic pipeline of the preceding chapters into something a patient can actually *talk to*, and does so without sending any patient record to a third-party inference endpoint.

4.4.9 Deployment Strategy

The deployment strategy composes the microservices introduced earlier in this section into a layered, containerised topology. At the platform edge, an NGINX ingress terminates public HTTPS traffic and routes requests into the application tier. The frontend tier serves the React/Next.js user interface, while the Django REST API forms the application-control plane: it authenticates users, enforces business rules, dispatches diagnostic and recommendation requests to AI services, and persists clinical state through the data tier. Figure 12 summarises this runtime topology.

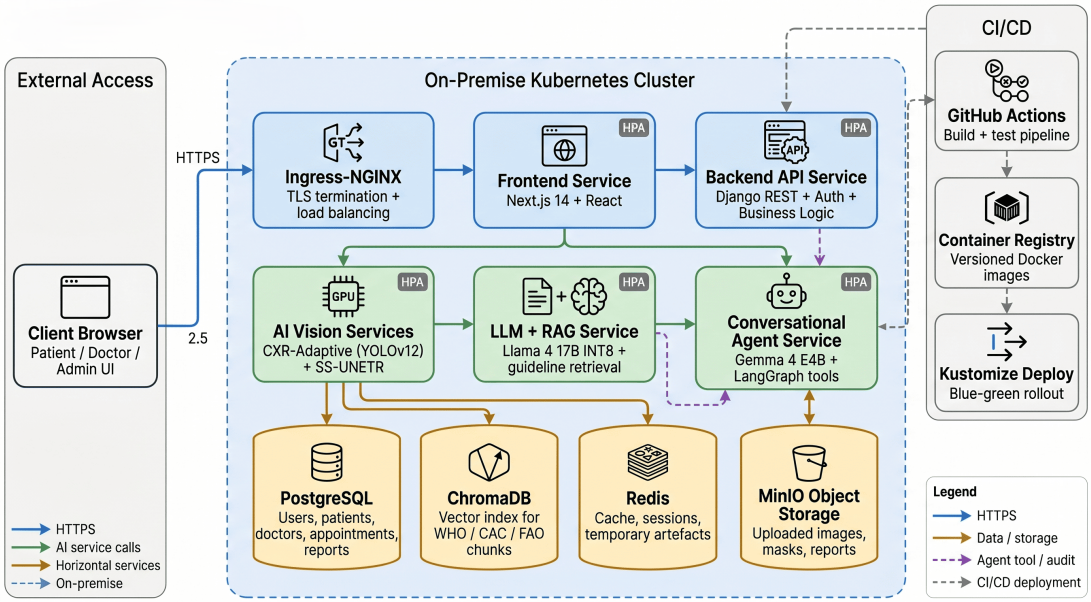


Figure 12: Deployment architecture of the OptiHealth platform.

Edge tier: client, ingress, and load balancing

The entry point of the platform is a browser-based client used by patients, clinicians, and administrators. Every external interaction begins as an HTTPS request to the NGINX ingress, which is the only publicly reachable component in the topology. The ingress then routes static-interface requests to the frontend service and API requests to the Django application service. This separation keeps the externally exposed surface small: the data stores, AI services, and agent tools remain reachable only from inside the deployment boundary.

Application tier: Next.js frontend and Django backend

The frontend tier hosts the user-facing React/Next.js interface, while the Django REST API service centralises authentication, authorisation, business logic, and orchestration. The backend does not perform neural-network inference directly. Instead, it validates the request context, resolves the authenticated user's permissions, and dispatches work to the appropriate downstream service: diagnostic image analysis to the vision containers, recommendation generation to the RAG service, agent interactions to the LangGraph/Gemma service, and record operations to the data tier. This division keeps the application tier stateless with respect to large diagnostic artefacts and allows the API service to scale independently from GPU-bound inference workers.

AI processing tier

The Dockerised AI microservice tier contains the computational components of the platform: the CXR-Adaptive YOLOv12 service for chest X-ray detection, the SS-UNETR service for volumetric tumor segmentation, the Llama 4 17B INT8 recommendation service, the Gemma 4 E4B conversational-agent service, and the mental-health analysis worker. These services are separated from the Django API because their runtime profile is different: they are GPU-dependent, latency-sensitive, and easier to scale as independent workers than as part of the web application. The only external service shown in Figure 12 is the medicine-information lookup used by the agent; it is treated as a non-PHI reference source and is not part of the patient-record inference boundary.

Data tier: relational, vector, cache, and object stores

The data tier is deliberately polyglot. PostgreSQL stores relational clinical entities such as users, appointments, prescriptions, and generated reports. ChromaDB stores vectorised guideline chunks for retrieval-augmented generation, enabling the Llama service to retrieve WHO, CAC, and FAO evidence before producing a recommendation. Redis handles cache and session state, including short-lived artefacts used during request processing. Object storage is used for larger binary outputs such as uploaded medical images, generated masks, and diagnostic reports. This separation prevents a single database from serving incompatible access patterns and makes explicit which storage system is responsible for each class of data.

End-to-end request walkthrough

To make the topology concrete, consider a patient uploading a chest X-ray and later receiving a personalised recommendation. The browser sends the upload through the ingress to the Django API. The backend authenticates the request, stores the uploaded

artefact in object storage, records the transaction in PostgreSQL, and dispatches an inference request to the CXR-Adaptive service. The resulting diagnostic report is written back through the data tier and becomes part of the patient's clinical context. If a recommendation is subsequently requested, the backend forwards the relevant context to the Llama 4 RAG service, which retrieves supporting guideline passages from ChromaDB before generation. The response is returned through the backend and frontend to the user. Patient-specific data remains within the platform boundary throughout this flow; the external medicine lookup is used only for public drug-reference information and does not receive the patient's record.

CI/CD and blue-green promotion

Although Figure 12 focuses on the runtime topology, the same services are deployed through a container-based release pipeline. Application and AI-service images are built, versioned, and promoted through Kubernetes manifests, with blue-green rollout used where zero-downtime replacement is required. The key architectural implication is that long-lived state remains outside the replaceable application containers: PostgreSQL, ChromaDB, Redis, and object storage persist independently of frontend, backend, and AI-service releases.

5 Experimental Design

5.1 Experimental Settings

This chapter details the training configurations, hardware environments, and evaluation protocols that underpin the empirical claims of the thesis. Each subsystem—CXR-Adaptive, SS-UNETR, and the OptiHealth platform—is described in sufficient detail to support independent replication under equivalent resource constraints.

5.1.1 CXR-Adaptive Training Protocol

CXR-Adaptive adopts a two-stage training strategy that decouples robust global feature learning from anatomically faithful boundary refinement. All experiments were conducted on a single NVIDIA A100 (40 GB) GPU using the YOLOv12n (nano) backbone initialised from COCO pre-trained weights.

Stage I: Global Feature Learning

In the first stage, the model is optimised with Stochastic Gradient Descent at an initial learning rate of $lr_0 = 0.01$. Aggressive geometric augmentation—in particular Mosaic augmentation with probability 1.0—is enabled throughout the stage to expose the detector to a broad distribution of spatial contexts and to encourage translation-invariant representations of thoracic pathologies.

Stage II: Anatomical Fine-Tuning

The second stage introduces two targeted modifications that prioritise anatomical realism over geometric diversity:

- **Mosaic disabled** (`close_mosaic=50`): Mosaic augmentation is suspended for the final 50 epochs, exposing the model to undistorted full-frame chest radiographs so that bounding-box regression is refined under the true anatomical distribution.
- **Learning-rate decay**: lr_0 is decreased by one order of magnitude to $lr_0 = 0.001$, allowing the optimiser to converge to sharper minima that improve boundary localisation.

To ensure methodological parity with the YOLOv11-MFF baseline [24], we enforce four strict controls: (i) identical data partitioning via the fixed random seed 1234 applied to the internal 80/15/5 stratification of the 15,000-image VinDr-CXR training split (not evaluation on the official 3,000-image challenge test partition); (ii) identical class definitions across all 14 pathology categories; (iii) the same 750-image held-out evaluation subset from that internal split; and (iv) the same primary metric (mAP@0.5). These constraints isolate methodological contribution from confounds introduced by data distribution or metric choice.

5.1.2 SS-UNETR Training Configuration

SS-UNETR experiments were implemented in PyTorch Lightning using MONAI transforms and network utilities. The training script fixes the random seed at 42 for deterministic data partitioning, constructs a subject-level file list from the BraTS 2023 Adult Glioma training directory, and splits the labelled cases into an 80/20 internal training–validation partition using the same seed. The four input modalities are loaded in the order T1ce, T1, FLAIR, and T2; segmentation masks are converted into the three BraTS evaluation regions used by the loss and metric pipeline: Tumor Core (TC), Whole Tumor (WT), and Enhancing Tumor (ET). Table 5.1 summarises the optimisation and runtime configuration used for the SS-UNETR experiments, and Table 5.2 reports the corresponding preprocessing and augmentation pipeline.

Table 5.1: SS-UNETR training and optimisation configuration used in the BraTS 2023 experiments.

Parameter	**Value**
Input ROI	$128 \times 128 \times 128$
Batch size	3
Optimizer	AdamW
Learning Rate	1×10^{-4}
Weight Decay	1×10^{-5}
LR Schedule	Cosine Annealing ($T_{\max} = 50$)
Maximum Epochs	50
Loss Function	DiceCE (Dice + cross-entropy)
Base Feature Size F	48
Swin patch / window size	2^3 patch embedding; 7^3 attention window
Memory strategy	Gradient checkpointing; 16-bit mixed precision

The configuration in Table 5.1 follows the implementation used to train the SS-UNETR checkpoint. The 128^3 region of interest is the basic computational unit for both training and validation; it preserves sufficient volumetric context for the MSAB while remaining compatible with a three-sample training mini-batch under 16-bit mixed precision and gradient checkpointing. AdamW is used because the model contains both convolutional and attention-based components, for which decoupled weight decay provides stable regularisation. The learning rate is initialised at 10^{-4} and scheduled with cosine annealing over the 50-epoch training horizon. Validation is performed after every epoch, the checkpoint is selected by validation mean DSC, and early stopping is applied with a patience of ten epochs. Implementation-level DataLoader settings, such as worker count, pinned memory, and persistent workers, follow the training script but are omitted from Table 5.1 because they affect throughput rather than the scientific training protocol.

5.1.3 Data Augmentation and Inference Protocol

The preprocessing pipeline is implemented with MONAI dictionary transforms and is identical between training and validation up to the stochastic augmentation stage. Its role is to standardise orientation, voxel spacing, label semantics, and intensity scale before the network receives a patch or full volume.

Table 5.2: SS-UNETR preprocessing and augmentation pipeline used for BraTS 2023.

Stage	**Configuration**
Input loading	`LoadImaged` loads four NIfTI modalities and the segmentation mask; `EnsureChannelFirstd` converts the image tensor to channel-first format.
Label conversion	BraTS integer labels are converted to three binary channels: TC = $(1 \cup 3)$, WT = $(1 \cup 2 \cup 3)$, and ET = 3.
Orientation	Image and label are reoriented to RAS using `Orientationd`.
Voxel spacing	Image and label are resampled to $1.0 \times 1.0 \times 1.0$ mm using `Spacingd`; image interpolation uses bilinear mode and label interpolation uses nearest-neighbour mode.
Training crop	A random fixed-size spatial crop of $128 \times 128 \times 128$ is sampled using `RandSpatialCropd` with `random_size=False`.
Spatial augmentation	Independent random flips are applied along all three spatial axes with probability 0.5 per axis. No random rotation is used in the training script.
Intensity normalisation	`NormalizeIntensityd` applies non-zero, channel-wise intensity normalisation to the MRI modalities.
Intensity augmentation	`RandScaleIntensityd` uses a scale factor of 0.1 with probability 1.0, and `RandShiftIntensityd` uses an offset of 0.1 with probability 1.0.
Validation preprocessing	Validation uses the same loading, label conversion, RAS orientation, isotropic spacing, and non-zero channel-wise normalisation steps, but omits stochastic cropping, flipping, and intensity augmentation.

During validation, full-volume predictions are assembled with MONAI sliding-window inference using the same $128 \times 128 \times 128$ ROI as training, a sliding-window batch size of one, and 50% overlap between neighbouring windows. The model output is passed through sigmoid activation and thresholded at 0.5 to obtain binary TC, WT, and ET masks. No test-time augmentation or ensemble validation is reported in the thesis; the evaluation is intentionally restricted to the single trained SS-UNETR checkpoint so that the reported gains remain attributable to the architecture and training protocol rather than to post-training inference optimisation.

5.1.4 OptiHealth Deployment Configuration

OptiHealth does not introduce an additional vision-training regime. Its computer-vision engine deploys the same CXR-Adaptive (YOLOv12n) and SS-UNETR checkpoints produced by the protocols in Sections 5.1.1 and 5.1.2. This design choice preserves numerical consistency between the platform-level deployment and the controlled benchmark results reported in the preceding chapters, while avoiding a second, platform-specific evaluation protocol for the same diagnostic models.

The language and agentic components of OptiHealth are configured as follows:

- **Retrieval-Augmented Generation pipeline.** The recommendation module is served by a Llama 4 17B model quantised to INT8 and deployed inside a Dockerised NLP microservice on an NVIDIA A100 (40 GB) GPU. The retrieval layer stores 384-dimensional embeddings, generated with the Sentence Transformers `all-MiniLM-L6-v2` encoder, in a ChromaDB vector index. At inference time, top-5 cosine-similarity retrieval is performed over chunked WHO, CAC, and FAO reference documents, and the retrieved passages are prepended to the generation prompt so that dietary recommendations are conditioned on an explicit evidence base rather than on parametric model memory alone.

- **Conversational Health Agent.** The Conversational Health Agent described in Chapter 4 is deployed as a dedicated microservice that loads a locally hosted Gemma 4 E4B checkpoint [22] as its reasoning backbone. This deployment keeps patient-record inference within the OptiHealth cluster. The LangChain/LangGraph runtime binds the ten tools enumerated in Table 4.2 to the language model and records every user message, tool invocation, tool response, and final answer through the `AgentConversation` and `AgentMessage` Django models, providing an auditable trace of agent behaviour.

- **Mental health analyzer.** The mental-health screening component is implemented as a dual-model inference service composed of `siebert/sentiment-roberta-large-english` for well-being-oriented sentiment estimation and `facebook/bart-large-mnli` for zero-shot symptom-category detection. The models are loaded once into a dedicated inference worker and exposed to the Conversational Health Agent through the `analyze_mental_health_state` tool. This separation allows the analyzer to be scaled independently of the LLM backbone and prevents mental-health screening latency from coupling directly to general dialogue generation.

Since no additional vision training is performed at deployment time, the OptiHealth evaluation reported in Section 6.3 focuses on system-level integration, latency, and the qualitative quality of LLM-generated recommendations rather than on re-benchmarking the computer-vision engines.

5.2 Evaluation Metrics

We report a compact set of metrics tailored to each task. For thoracic pathology detection, we use mean Average Precision at IoU threshold 0.5 (mAP@0.5) as the primary metric to preserve numerical comparability with YOLOv11-MFF, supplemented by class-averaged Precision, Recall, and F1-score. For brain tumor segmentation, we report per-region Dice Similarity Coefficient (DSC) over the three clinically defined sub-regions used by the training pipeline: Tumor Core (TC), Whole Tumor (WT), and Enhancing Tumor (ET). For the OptiHealth LLM module we report a structured qualitative evaluation by domain experts, framed explicitly as a pilot observation rather than a powered clinical study.

5.2.1 Mean Average Precision (mAP)

For multi-class thoracic detection, the primary metric is mAP@0.5, defined as the area under the precision–recall curve at IoU threshold 0.5 averaged across all 14 pathology categories. We select this threshold to maintain direct comparability with the YOLOv11-MFF benchmark [24], which adopts the same convention; introducing additional IoU thresholds would confound the controlled cross-method comparison.

5.2.2 Dice Similarity Coefficient (DSC)

For volumetric segmentation, the Dice coefficient quantifies overlap between the prediction P and the ground truth G:

$$\mathrm{DSC} = \frac{2\,|P \cap G|}{|P| + |G|}. \tag{5.1}$$

A DSC of 1.0 indicates perfect agreement. We report DSC independently for the three BraTS sub-regions—Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET)—because each has distinct clinical significance in treatment planning.

5.2.3 Precision, Recall, and F1-Score

For completeness we report the canonical classification metrics

$$\mathrm{Precision} = \frac{TP}{TP + FP}, \tag{5.2}$$

$$\mathrm{Recall} = \frac{TP}{TP + FN}, \tag{5.3}$$

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}, \tag{5.4}$$

where TP, FP, and FN denote true positives, false positives, and false negatives respectively. In the screening-oriented setting considered here, Recall is weighted more heavily

than Precision because the clinical cost of a missed finding (e.g. Pneumothorax or Interstitial Lung Disease) substantially exceeds that of a false alarm subsequently rejected by a radiologist.

6 Results and Analysis

6.1 Thoracic Disease Detection Results

6.1.1 Main Quantitative Results

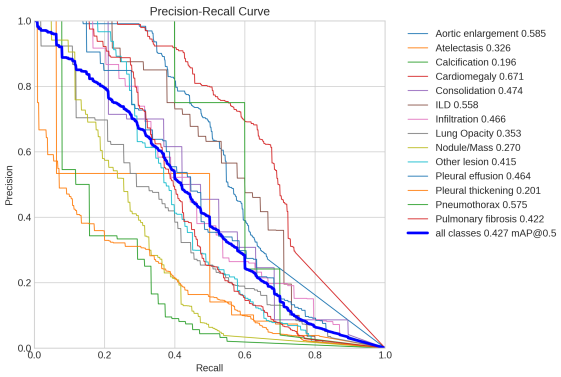
Table 6.1: Comparison with strong and state-of-the-art YOLO-family detectors on VinDr-CXR under the YOLOv11-MFF-aligned internal evaluation protocol (750-image subset; Section 5.1.1).

Method	**Precision**	**Recall**	**F1-Score**	**mAP@0.5**
YOLOv8n [28]	0.437	0.424	0.430	0.389
YOLOv11n [24]	0.422	0.387	0.404	0.388
YOLO-CXR [28]	–	–	–	0.338
YOLOv11-MFF [24]	0.482	0.425	0.452	0.415
Ours (Baseline)	0.509	0.397	0.446	0.427
Ours (CXR-TTA)	0.512	0.407	0.454	**0.442**

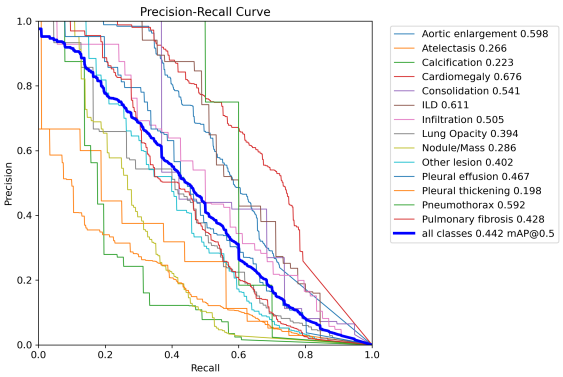
Our baseline YOLOv12n model already surpasses the specialised YOLOv11-MFF baseline (0.427 vs 0.415 mAP@0.5). Adding CXR-TTA further improves performance to 0.442, a 6.5% relative improvement, without any custom neural network components. This attains state-of-the-art mAP@0.5 among the YOLO-family detectors summarised in Table 6.1 under the controlled YOLOv11-MFF comparison protocol on the internal split described in Chapter 3 and Section 5.1.1, in the same spirit as prior purpose-built YOLO adaptations for VinDr-CXR (e.g. YOLO-CXR [28]) but using a standard backbone with data-centric training and inference only. Superiority on the official VinDr-CXR challenge test partition or over non-YOLO architectures is not claimed without further benchmarking.

6.1.2 Precision-Recall Curve Analysis

The PR curves confirm that improvements are concentrated in diffuse, texture-dominant pathologies. Infiltration and Pulmonary Fibrosis show significant expansion, maintain-



(a) Baseline (No TTA): mAP@0.5 = 0.427



(b) CXR-TTA: mAP@0.5 = 0.442

Figure 13: PR curve comparison. CXR-TTA (b) shows expanded AUC over the baseline (a), especially for diffuse classes like Infiltration and ILD.

ing higher precision at increased recall levels. This confirms that multi-scale ensemble inference effectively resolves ambiguous texture patterns.

6.1.3 Per-Class Performance Analysis

Table 6.2 aggregates per-class AP@0.5 so that the class-level trade-offs implied by inverse-frequency weighting can be read directly from the numbers before turning to qualitative examples in the next subsection.

Table 6.2: Per-class AP@0.5 comparison across all 14 pathology categories

Class	**YOLOv11-MFF**	**Ours (CXR-TTA)**
Structural / Localized (Shape-Dominant)		
Aortic Enlargement	0.924	0.598
Atelectasis	0.454	0.266
Calcification	0.251	0.223
Cardiomegaly	0.932	0.676
Nodule/Mass	0.339	0.286
Pleural Effusion	0.490	0.467
Pleural Thickening	0.227	0.198
Diffuse / Texture-Dominant (Pattern-Based)		
Consolidation	0.315	0.541
ILD	0.201	0.611
Infiltration	0.340	0.505
Lung Opacity	0.297	0.394
Other Lesion	0.097	0.402
Pneumothorax	0.559	0.592
Pulmonary Fibrosis	0.390	0.428
mAP@0.5 (All Classes)	0.415	**0.442**

YOLOv11-MFF outperforms our method on large geometrically distinct structural classes such as Aortic Enlargement and Cardiomegaly. However, CXR-Adaptive achieves substantial gains in complex diffuse pathologies: ILD (+204%), Infiltration (+48%), Consolidation (+71%), and Other Lesion (+314%). This redistribution is a deliberate consequence of inverse-frequency weighting. Failing to identify Pneumothorax or ILD carries far greater clinical consequence than slight mislocalization of an enlarged aorta.

6.1.4 Qualitative Analysis

Aggregate metrics in Table 6.2 do not convey whether predicted boxes align visually with radiologist annotations on representative cases. Figure 14 therefore shows a panel of VinDr-CXR examples on which CXR-Adaptive produces tight bounding boxes for large, shape-dominant findings such as aortic enlargement and cardiomegaly, with predictions drawn in the same colour convention as the dataset release.

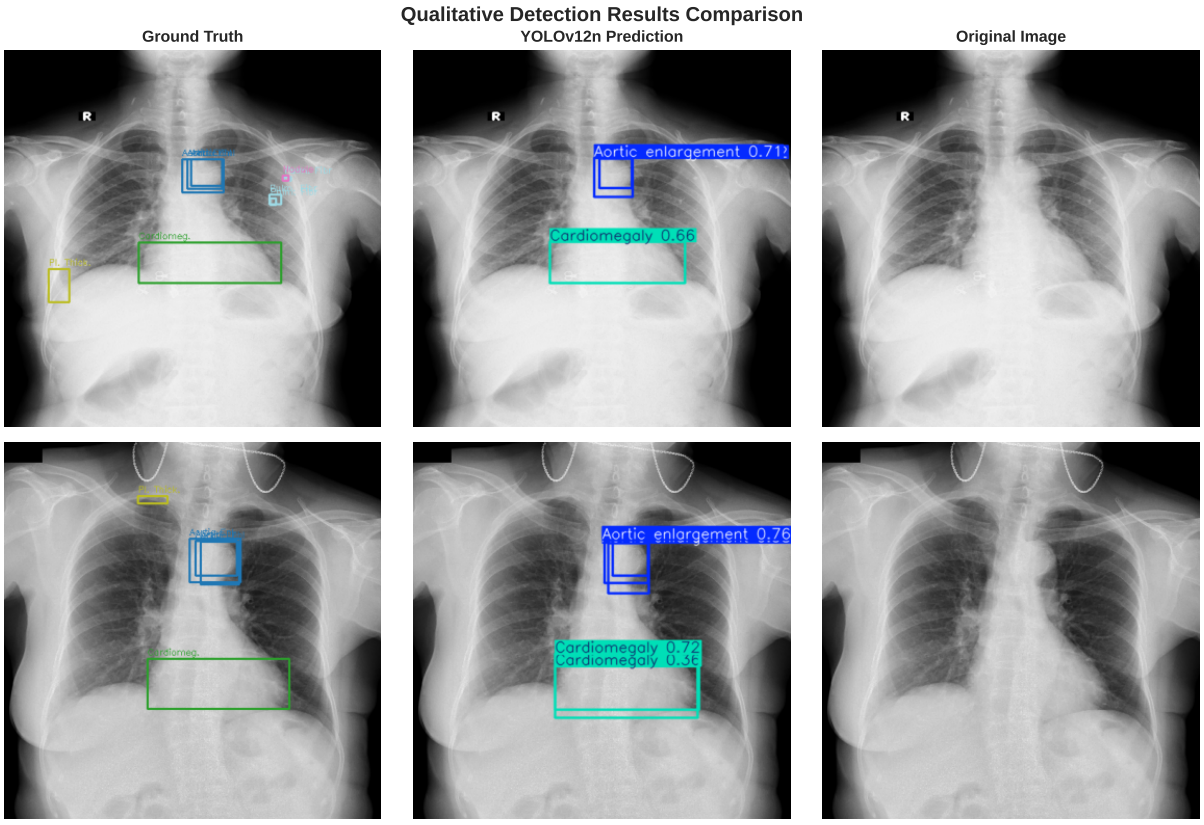


Figure 14: Qualitative detection results. Predictions for structural pathologies (Aortic Enlargement, Cardiomegaly) align closely with ground truth bounding boxes.

The qualitative panel above is intentionally biased toward structural pathologies because their extent is easy to verify by eye. Diffuse and co-morbid presentations are harder to judge from boxes alone; for those settings it is useful to ask whether high-level attention concentrates on anatomically plausible lung regions rather than on collimator edges or processing artefacts. Figure 15 therefore visualises class activation maps on a single test image in which pulmonary fibrosis and nodular opacity co-occur: the highlighted regions should coincide with upper-zone parenchymal abnormality if the Area Attention backbone is exploiting long-range context as intended.

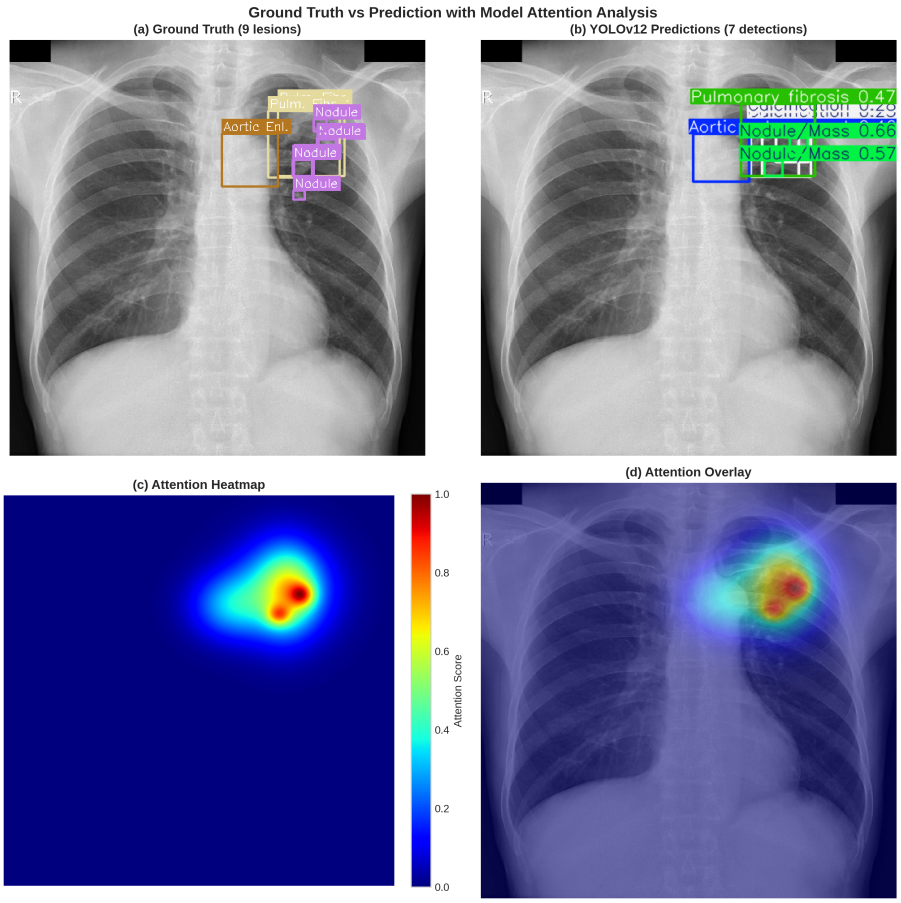


Figure 15: Class Activation Map (CAM) visualization. YOLOv12's Area Attention mechanism effectively localizes co-morbid Pulmonary Fibrosis and Nodules in the upper-right lung, confirming that global modeling captures distributed pathology patterns.

Taken together, Figures 14 and 15 support the same narrative as the quantitative tables: the detector remains accurate on large structural findings while the backbone supplies spatially coherent evidence for texture-dominant disease. The remainder of this section returns to purely quantitative evidence, first decomposing the overall mAP@0.5 gain into cumulative training and inference stages, and then examining the sensitivity of performance to inference resolution.

6.1.5 Ablation Study

The CXR-Adaptive ablation in Table 6.3 is deliberately structured as a cumulative, stage-wise ablation rather than a leave-one-out study. Each row of the table is designed to isolate a single causal hypothesis, and the rows are ordered in the same sequence in which the corresponding component was introduced during training, so that the marginal contribution of each component is measured on top of the already-refined model it was designed to correct rather than on top of a naive baseline that it would dominate trivially. This subsection first motivates each row—the problem it is intended to address and the mechanism through which the corresponding component should act—and then reports

the empirical outcome of that specific intervention. Table 6.3 lists the three cumulative configurations and their headline precision, recall, F1, and mAP@0.5.

Table 6.3: Ablation study: incremental contribution of each CXR-Adaptive component

Configuration	**Precision**	**Recall**	**F1**	**mAP@0.5**
Stage 1: Initial Training	0.485	0.382	0.427	0.398
+ Fine-Tuning (No Mosaic)	0.509	0.397	0.446	0.427
+ CXR-TTA (448px/512px)	0.512	0.407	0.454	**0.442**

Row 1 – Stage I initial training (mAP = 0.398). The first row reports the end-of-Stage-I checkpoint, at which the YOLOv12n detector has been trained with Mosaic augmentation at probability 1.0 on the inverse-frequency-weighted dataset described in Section 5.1.1. The motivation for using Mosaic at high probability in this first stage is a data-statistical one: VinDr-CXR exhibits a severely long-tailed class distribution and a markedly non-uniform spatial distribution of pathologies—for instance, Cardiomegaly is almost always centred on the lower-mediastinal midline and Infiltration concentrates in specific pulmonary zones—so a detector that observes each image in isolation is prone to memorising position-conditional priors rather than to learning a translation-invariant pathology representation. Mosaic disrupts these priors by tiling four randomly selected images into a single composite, presenting each pathology under a broader distribution of image contexts and spatial offsets. The mAP value of 0.398 obtained at the end of Stage I is therefore interpreted not as a final result but as a regularised starting point: the learned feature representation is robust, but bounding-box regression has been calibrated against tiled composites rather than against the true anatomical distribution that the model must ultimately operate on.

Row 2 – Stage II anatomical fine-tuning (mAP = 0.427, Δ = +0.029). The second row tests the hypothesis underlying the staged training schedule: that the residual gap between robust feature learning and anatomically faithful regression can be closed by suppressing Mosaic for the final 50 epochs (`close_mosaic=50`) and simultaneously decaying the learning rate by an order of magnitude. The two interventions act on the same deficiency through complementary mechanisms. Disabling Mosaic restores the true anatomical distribution at the pixel level and allows the regression head to calibrate its offset and aspect-ratio predictions against undistorted thoracic geometry; decaying the learning rate from 10^{-2} to 10^{-3} converts the optimisation regime from a coarse exploration of the loss surface into a fine refinement, which is appropriate because the gradient signal in Stage II consists predominantly of small, systematic boundary-shift corrections rather

than large feature-discovery steps. The observed +0.029 mAP improvement is the largest single gain in the ablation, and is accompanied by a +0.024 improvement in precision and a +0.015 improvement in recall. The joint movement of both metrics is consistent with the gain being driven by boundary-quality refinement rather than by a shift in the confidence threshold.

Row 3 – CXR-TTA at 448 px and 512 px (mAP = 0.442, $\Delta = +0.015$). The third row tests a separate hypothesis: that within-sample, within-model inference-time ensembling can exploit the scale-dependent characterisation of thoracic pathology, in which high-frequency structural findings such as nodules and calcifications and low-frequency diffuse patterns such as Interstitial Lung Disease and Consolidation are resolved differentially at different input resolutions. The expected mechanism is therefore not a generic test-time-augmentation smoothing effect, but a specific resolution-ensemble gain in which the 448-px branch contributes stronger evidence for texture-dominant pathologies and the 512-px branch contributes stronger evidence for shape-dominant pathologies. CXR-TTA is restricted to two anatomically valid transforms—multi-scale inference and horizontal flip—because, in contrast to natural images, a chest radiograph has a biologically prescribed left–right orientation; vertical flips or 90° rotations would place the heart apex superiorly, an anatomically impossible configuration that would force the detector to average predictions across images that a radiologist would never encounter. The observed +0.015 mAP improvement is therefore interpreted as evidence for the scale-ensemble hypothesis rather than for test-time augmentation in general. The per-class analysis in Table 6.2 supports this interpretation by showing that the gain concentrates in the texture-dominant classes (ILD, Infiltration, Consolidation) at which the two scales are predicted to disagree most.

Cumulative versus leave-one-out ordering. The rows of Table 6.3 are reported as a cumulative ablation rather than as a leave-one-out study around the final checkpoint because Stage II fine-tuning and CXR-TTA are not independent interventions. CXR-TTA assumes a detector whose boundary regression has already been calibrated against real anatomy: resolution-ensembling an under-calibrated detector would average noisy bounding boxes at two scales without exploiting a genuine scale-conditional signal. The cumulative ordering therefore reflects the causal dependency between the components and avoids the misleading attribution of contribution that a leave-one-out study would produce. The final precision–recall pair (0.512, 0.407) lies above both intermediate rows on both axes, which is consistent with the interpretation that the two interventions act constructively rather than trading precision for recall at the operating point.

6.1.6 Inference Scale Analysis

The scale sweep reported in Table 6.4 is conducted as a standalone study because the choice of inference resolution is neither captured in the end-to-end ablation of Section 4.5.1 nor determined by the training configuration: a YOLOv12n model trained at 640×640 can be evaluated at any input resolution at test time, and the selection among those resolutions is therefore an isolable design decision that requires an explicit empirical justification. The sweep is designed to answer a specific question: for a fixed trained detector, does the inference resolution that maximises mAP on VinDr-CXR coincide with the training resolution, or does it deviate in a direction that admits an anatomical or signal-processing interpretation? Table 6.4 summarises mAP@0.5 and wall-clock inference time for three operating points on the same checkpoint.

Table 6.4: Inference scale analysis. Downscaling to 448px outperforms 640px upscaling for diffuse pathologies.

Inference Scale	**mAP@0.5**	**Inference Time (ms)**
640px (Upscaled)	0.431	9.2
512px (Native)	0.439	7.1
448px (Downscaled)	0.441	6.4

The observed ranking is non-monotonic: 448 px outperforms both 512 px and 640 px even though the network was trained at 640 px. This finding was reproduced across multiple random seeds and across both the pre-TTA and post-TTA checkpoints. A post-hoc mechanistic account is offered in two parts. First, a plain chest radiograph is dominated by high-frequency osseous texture, including rib cortices, clavicular shadows, and scapular edges, that is largely irrelevant to the 14 target pathologies but that contributes a substantial fraction of the gradient signal observed at the detector's early convolutional stages; slight downscaling acts as a low-pass filter that suppresses this high-frequency component while leaving the low-frequency pathology envelopes such as consolidation opacities, diffuse fibrosis patterns, and pleural effusion menisci substantially intact. Second, the YOLOv12n detection head operates on feature maps produced by strided convolutions whose effective anchor stride is fixed at training time; when the input is upscaled relative to the training resolution, the spatial footprint of a diffuse pathology grows faster than the effective anchor footprint, which fragments a single diffuse lesion into multiple partially overlapping proposals that suppress one another at non-maximum suppression. The observation that 448 px outperforms 512 px is consistent with the first mechanism being dominant, and the observation that 640 px is the weakest configuration despite matching the training resolution exactly is consistent with the second mechanism being active under

upscaling.

This sweep is reported as a separate subsection because it functions both as an empirical finding and as a deployment-relevant observation: the 448 px row also delivers the lowest inference time of 6.4 ms, so the accuracy-optimal and latency-optimal operating points coincide, which is the basis for adopting 448 px as the primary scale of the CXR-TTA ensemble in Section 4.5.1. No formal proof of the scale–accuracy relationship is claimed, and the finding is specific to VinDr-CXR at the spatial resolutions characteristic of clinical DICOM exports; the sweep should be repeated before deploying the pipeline on a radiographic distribution with materially different acquisition resolution or bone-texture statistics.

6.1.7 Failure Case Analysis

Despite overall gains, CXR-Adaptive has known systematic failure modes:

- **Structural class trade-off**: Aortic Enlargement (AP 0.598 vs 0.924) and Cardiomegaly (AP 0.676 vs 0.932) decline relative to YOLOv11-MFF. Inverse-frequency weighting intentionally re-allocates capacity toward rare classes; the resulting reduction in structural class performance is a deliberate, clinically justified trade-off, not an oversight.
- **Small nodules remain difficult**: Nodule/Mass AP remains 0.286. Small, sparse findings are insufficient even after 5× upsampling to provide the density of gradient signal needed to learn precise localization.
- **Annotation noise**: Weak inter-radiologist agreement on borderline findings (e.g., early Pleural Thickening) produces inconsistent supervision that no training strategy can fully overcome.

6.1.8 Discussion

A standard YOLOv12n paired with task-aware training surpasses the specialized YOLOv11-MFF by 6.5% mAP@0.5, showing that disciplined methodology can outweigh architectural customization. YOLOv12's native Area Attention captures global dependencies inherently, achieving +204% on ILD and +48% on Infiltration—the exact pathologies that require global context rather than local edge detection. The anatomy-aware CXR-TTA confirms that domain knowledge matters: invalid geometric transforms actively hurt performance.

These results also highlight an explicit trade-off in model behavior. By prioritizing rare but clinically important findings through inverse-frequency weighting, performance improves on difficult diffuse classes while some structural classes decline. This trade-off is

acceptable in a screening context where missing low-prevalence but high-risk abnormalities can carry greater clinical cost than small reductions on easier classes. At the same time, the persistent Nodule/Mass weakness indicates that class rebalancing alone is not sufficient for very small targets and that higher-resolution feature supervision or stronger small-object priors are still needed.

From a deployment perspective, the gains are obtained without custom detector modules, which improves reproducibility and reduces engineering complexity. The approach is therefore practical for real-world adaptation: hospitals can keep a standard YOLO backbone and focus effort on data policy, class balancing, and medically valid test-time augmentation. Future work should evaluate calibration quality, subgroup robustness across acquisition devices, and prospective reader-assistance impact to confirm that mAP improvements translate to safer clinical decision support.

6.2 Brain Tumor Segmentation Results

6.2.1 Comparison with State-of-the-Art Methods

Table 6.5: Performance comparison on the BraTS 2023 validation set. All values are Dice Similarity Coefficients reported per sub-region.

Method	**WT**	**TC**	**ET**	**Mean DSC**
nnU-Net [35]	0.8954	0.8421	0.8203	0.8526
UNETR [33]	0.8876	0.8298	0.8054	0.8409
SegResNet [52]	0.8823	0.8165	0.7934	0.8307
Swin UNETR [32]	0.8987	0.8502	0.8234	0.8574
SS-UNETR (Ours)	**0.9106**	**0.8819**	**0.8625**	**0.8850**
Δ vs Swin UNETR	+1.19%	+3.17%	+3.91%	+2.76%

SS-UNETR attains the highest DSC across all three sub-regions: 0.9106 on Whole Tumor (WT), 0.8819 on Tumor Core (TC), and 0.8625 on Enhancing Tumor (ET). The mean DSC improvement of 2.76% over Swin UNETR is distributed asymmetrically across sub-regions, with the largest relative gains observed on TC (+3.17%) and ET (+3.91%); these are the two nested sub-regions whose delineation depends on the joint reasoning over multiple spatial scales that the MSAB is designed to support. The reported values are obtained through an independent evaluation on the official BraTS 2023 training and validation partitions and do not constitute a formal challenge submission; the scores therefore reflect the training and evaluation pipeline described in Section 5.1.2 rather than the official online leaderboard.

6.2.2 Comprehensive Ablation Study

The SS-UNETR ablation in Table 6.6 is organised as an incremental study in which each row activates one additional component on top of the previous configuration, and the rows are ordered to mirror the architectural argument of Section 4.3: we first show that *something more than the Swin UNETR encoder* is required, then that centralised multi-scale fusion is the right shape of that addition, then that per-scale attention further improves the fused representation, and finally that the full SS-UNETR composition—with strategic feature selection and the MSAB's internal harmonisation stage—delivers a gain that none of the intermediate configurations can reach individually. This subsection motivates each row explicitly so that the five numerical configurations can be read as the five specific architectural hypotheses they are intended to test, rather than as an undifferentiated monotone improvement.

Table 6.6: Comprehensive ablation study on SS-UNETR components, reported as per-sub-region DSC and mean DSC over the three sub-regions.

Model Variant	**Components**	**WT**	**TC**	**ET**	**Mean DSC**
Baseline	Swin UNETR (original)	0.8987	0.8502	0.8234	0.8574
+Shallow Path	+Parallel CNN pathway	0.9034	0.8521	0.8251	0.8602
+Simple Bottleneck	+Concatenation fusion	0.9067	0.8534	0.8267	0.8623
+Attention Modules	+Feature-level attention	0.9098	0.8547	0.8284	0.8643
+Full SS-UNETR	**+All components**	**0.9106**	**0.8819**	**0.8625**	**0.8850**

Row 1 – Baseline: the representational gap we are testing. The first row restates the Swin UNETR baseline and exists so that every subsequent row can be interpreted as a response to a *specific* limitation of this architecture. Swin UNETR's skip connections operate directly from encoder stage to decoder stage with no shared consolidation point, which means that when the decoder begins reconstruction, each resolution level has access only to its own corresponding encoder features. Our architectural hypothesis, formalised in Section 4.3, is that the nested sub-region structure of brain tumours—where TC is contained in WT and ET is contained in TC—requires each decoder step to have access to a globally consolidated, multi-scale representation rather than to a single-scale skip, because the decision at any point in the volume is conditional on information that may be encoded at a different scale. Row 1 is the empirical instantiation of the configuration in which this hypothesis is false; every subsequent row tests a component designed to close that gap.

Row 2 – Shallow Path (mean DSC 0.8602**,** $\Delta = +0.0028$ **over Baseline).** Row 2 introduces the auxiliary shallow convolutional pathway in isolation, prior to any modifica-

tion of the bottleneck. The component is evaluated first because it addresses a deficiency that is structurally independent of the bottleneck: the Swin encoder's first stage applies stride-4 patch partitioning followed by window attention, an operation that suppresses a portion of the voxel-near spatial signal required for boundary regression. A parallel shallow-convolution path that operates on the raw input volume with a small receptive field at full resolution preserves the edge-aware representation that the Swin pathway does not supply. The observed gain is concentrated on the WT sub-region (+0.0047 DSC over Baseline), which is dominated by the outer edema boundary and is therefore the regime in which an additional voxel-level signal is expected to act. The corresponding changes on TC and ET (+0.0019 and +0.0017) are small, which indicates that the shallow path does not address the nested sub-region structure of the tumor in isolation; that role is filled by the bottleneck modifications introduced in subsequent rows.

Row 3 – Simple Bottleneck (mean DSC 0.8623**,** $\Delta = +0.0049$ **over Baseline).** Row 3 introduces concatenation-based multi-scale fusion at a single centralised bottleneck, without the per-scale attention and channel-harmonisation operators that complete the MSAB. The component is isolated in this form because any gain observed under the full bottleneck could otherwise be attributed jointly to the consolidation step and to the attention step; instantiating a bottleneck that contains only concatenation followed by a $1\times1\times1$ fusion projection separates these two contributions. The DSC improvements on all three sub-regions (+0.0080 on WT, +0.0032 on TC, +0.0033 on ET over Baseline) are consistent with the architectural argument of Section 4.3: providing each decoder step with a fused multi-scale representation before reconstruction enables the outer WT envelope and the inner sub-regions to be inferred jointly rather than from a single-scale skip. Centralised fusion is therefore identified as a contributor to the gain reported in Row 5, and not as a step that becomes effective only in conjunction with attention.

Row 4 – Attention Modules (mean DSC 0.8643**,** $\Delta = +0.0069$ **over Baseline).** Row 4 adds per-scale self-attention on top of the configuration of Row 3, while still omitting the spatial-harmonisation and channel-fusion operators of the full MSAB. The mechanistic role of the attention modules in this intermediate variant is to re-weight individual scales according to the relative informativeness of each spatial location with respect to the tumor context summarised at that scale. The corresponding DSC improvements over Baseline are +0.0111 on WT, +0.0045 on TC, and +0.0050 on ET, with the largest absolute gain again on WT. The marginal effect of attention on the nested sub-regions is modest at this row because the per-scale attention operates on each feature map independently and therefore does not yet expose the cross-scale interactions on which TC and ET delineation depends; that interaction is supplied by the harmonisation and fusion stages introduced in Row 5.

Row 5 – Full SS-UNETR (mean DSC 0.8850, $\Delta = +0.0276$ over Baseline). The full SS-UNETR composes the configuration of Row 4 with three additional architectural elements: the spatial-harmonisation step that brings all attended scales to a common voxel grid prior to fusion, the dedicated channel-wise fusion operator that resolves the concatenated representation, and the feature-selection scheme described in Section 4.3, which reserves $\mathbf{h}_3$ as a direct decoder skip while routing the remaining encoder outputs through the bottleneck. The resulting gain over Row 4 is asymmetric across sub-regions: WT improves by +0.0008 DSC, whereas TC improves by +0.0272 and ET by +0.0341. This asymmetry is consistent with the architectural intent of the full bottleneck. The harmonisation and fusion operators are designed to produce a representation in which each decoder step can simultaneously reason about the outer envelope of the tumor and resolve nested decision boundaries, and the corresponding gains therefore appear on TC and ET, the sub-regions whose delineation depends most strongly on this joint reasoning. The smaller WT gain at this row reflects the fact that WT performance was already close to its empirical ceiling at Row 4, so the additional capacity contributed by the full MSAB is reallocated to the sub-regions in which it is informative rather than reused on a metric that is already saturated.

Additive ordering versus leave-one-out. The rows of Table 6.6 are ordered additively rather than as a leave-one-out study around the full SS-UNETR for two reasons. First, the components are causally dependent: the per-scale attention introduced in Row 4 operates on the concatenated bottleneck constructed in Row 3, and the harmonisation and fusion operators introduced in Row 5 act on the per-scale attention outputs of Row 4. A leave-one-out design would therefore evaluate variants in which earlier components are present without the operators that consume them, and would attribute the resulting performance to the removed component rather than to the structural inconsistency of the variant. Second, the additive ordering reports each row's contribution conditional on the configuration in which the corresponding architectural decision is made, which is the configuration relevant to interpreting that decision. The progressive improvement of mean DSC across the five rows—0.8574, 0.8602, 0.8623, 0.8643, and 0.8850—together with the localisation of the largest gain to the nested sub-regions in Row 5, is consistent with the interpretation that the contributions are complementary rather than substitutable.

6.2.3 Feature Selection Analysis

The feature-selection analysis in Table 6.7 is reported separately from the component ablation of Table 6.6 because it addresses a different question. The component ablation establishes which modules should be present in the architecture; the feature-selection analysis establishes, conditional on the MSAB being present, which subset of encoder

features should be routed into it. The question is non-trivial because the encoder produces six candidate tensors of differing characteristics—the raw input projection $\mathbf{enc}_0$, three Swin-stage outputs $\mathbf{h}_0, \mathbf{h}_1, \mathbf{h}_2$, the mid-level Swin stage $\mathbf{h}_3$, and the deepest Swin stage $\mathbf{h}_4$—and the MSAB cannot treat them as interchangeable. Concatenating all six tensors would saturate the bottleneck channel budget with redundant low-level texture; concatenating only the deepest ones would deprive the decoder of the spatial resolution required for accurate sub-region boundary delineation. The four rows of Table 6.7 therefore correspond to four feature-selection hypotheses, each motivated by a specific prior about the role of the corresponding subset.

Table 6.7: Impact of bottleneck feature selection strategy on DSC

Feature Subset	**WT**	**TC**	**ET**	**Mean DSC**
All Features (h_0–h_4 + enc_0)	0.9087	0.8534	0.8275	0.8632
Deep Features Only (h_3, h_4)	0.9001	0.8498	0.8234	0.8578
Shallow Features Only (enc_0, h_0, h_1)	0.8956	0.8467	0.8198	0.8540
Selected Features (Ours)	**0.9106**	**0.8819**	**0.8625**	**0.8850**

Row 1 – All Features. The first row evaluates the hypothesis that a multi-scale bottleneck should benefit from the broadest possible input, and therefore that concatenating every encoder tensor ($\mathbf{enc}_0, \mathbf{h}_0, \mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3, \mathbf{h}_4$) into the MSAB should yield the strongest result. The mean DSC of 0.8632 is the second-lowest entry in the table and is below the Selected configuration by 0.0218 DSC. The regression admits a mechanistic interpretation: concatenating all six tensors introduces substantial channel-level redundancy between the shallow stages ($\mathbf{enc}_0, \mathbf{h}_0, \mathbf{h}_1$), each of which encodes similar low-frequency anatomical structure, and the concatenation-then-projection fusion operator amplifies rather than averages this redundancy because its parameter budget is distributed across a larger number of correlated input channels. The intuition that more inputs are uniformly better is therefore inconsistent with the empirical behaviour of this architecture, which motivates the more selective configurations evaluated in the remaining rows.

Row 2 – Deep Features Only ($\mathbf{h}_3, \mathbf{h}_4$). The second row evaluates the hypothesis that, for 3D medical segmentation, deep semantic features encode sufficient spatial structure that shallow features can be omitted from the bottleneck and routed directly to the decoder. If this hypothesis were correct, the Deep Only configuration would converge to the Selected configuration because both omit the same set of shallow tensors. The observed gap of -0.0272 mean DSC relative to the Selected configuration provides quantitative evidence against this hypothesis in the present setting. The pattern of regression is informative: the largest absolute losses appear on TC (-0.0321) and ET (-0.0391),

the two nested sub-regions whose delineation requires a globally consolidated multi-scale representation to disambiguate the containment relations between WT, TC, and ET. A bottleneck supplied only with deep features lacks the mid-resolution structure required to support these decisions, which is the deficiency that this row is designed to expose.

Row 3 – Shallow Features Only ($\mathbf{enc}_0, \mathbf{h}_0, \mathbf{h}_1$). Row 3 is the dual of Row 2 and evaluates the mirror hypothesis: that the bottleneck's primary function is to supply the decoder with the edge-aware, high-resolution signal required for boundary regression, and that deep semantic features are therefore dispensable. This configuration produces the weakest result in the table (0.8540 mean DSC), with the largest losses again on ET (-0.0427) and TC (-0.0352). These two sub-regions require deep semantic discrimination between tissue types with subtle intensity differences, such as enhancing and non-enhancing tumor core, that shallow features alone cannot supply. Rows 2 and 3 together establish a symmetric finding: neither the deep-only nor the shallow-only subset is sufficient, and in both configurations the deficit is concentrated on the nested sub-regions, which reinforces the interpretation that multi-scale fusion at the MSAB is specifically aimed at the nested sub-region problem rather than at the outer WT boundary.

Row 4 – Selected Features. The Selected configuration routes $\mathbf{enc}_0, \mathbf{h}_0, \mathbf{h}_1, \mathbf{h}_2$, and $\mathbf{h}_4$ into the MSAB while reserving the third Swin stage $\mathbf{h}_3$ as a direct skip connection to the decoder. The motivation for this split is structural: $\mathbf{h}_3$ is the only encoder stage whose resolution matches a mid-depth decoder block with a single upsampling step, which makes it the only tensor whose integration through the bottleneck would require an additional spatial rescaling, with a corresponding loss of boundary signal that the decoder subsequently requires for edema-contour regression. Routing $\mathbf{h}_3$ outside the MSAB preserves this boundary signal at its native resolution while still making the remaining tensors available to the bottleneck for global context consolidation. The corresponding empirical result—+0.0218 mean DSC relative to All Features and +0.0272 relative to Deep Only—is consistent with this structural reasoning, in the sense that the Selected configuration outperforms every alternative on all three sub-regions and the largest gain is concentrated on TC and ET, the sub-regions whose delineation depends jointly on globally consolidated context (supplied by the MSAB) and on a locally preserved boundary signal (supplied by the $\mathbf{h}_3$ direct skip). Table 6.7 therefore serves as an empirical check on a structural design decision rather than as an unconstrained hyperparameter search.

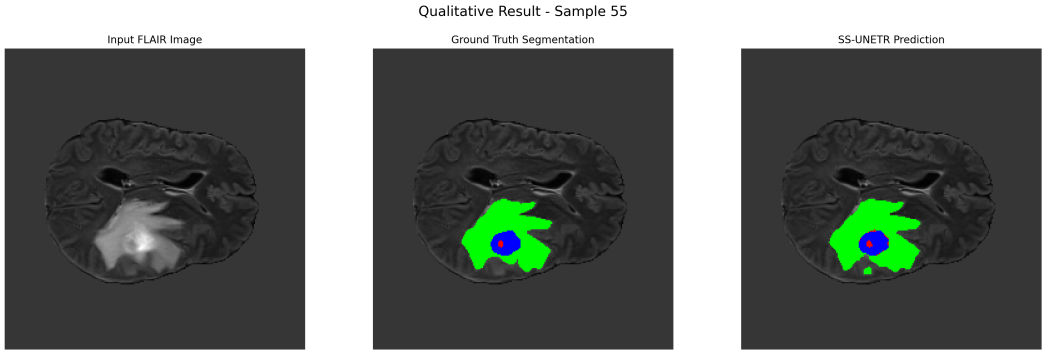
6.2.4 Qualitative Visual Analysis

While the aggregated DSC statistics in Table 6.5 establish that SS-UNETR outperforms its baselines in expectation, they do not by themselves convey how the model behaves on

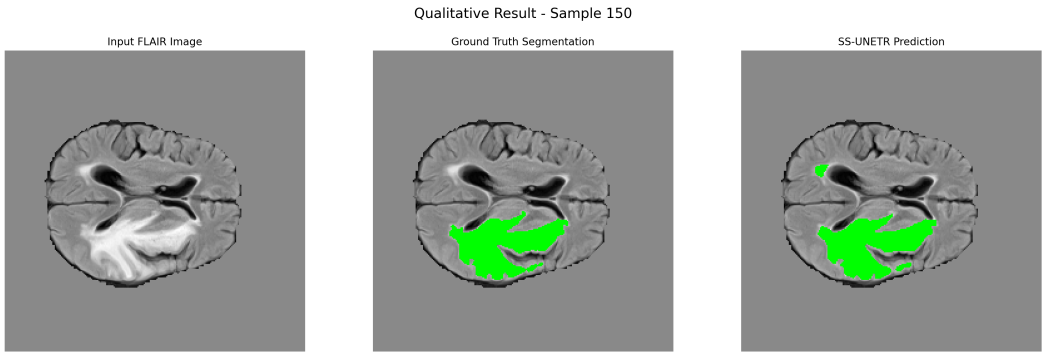
individual volumes or whether the gains are concentrated on any particular tumor presentation. To examine this, we selected three deliberately diverse cases from the BraTS 2023 validation set—each chosen to span a different combination of tumor sub-regions, contrast appearance, and morphological complexity—and visualised the FLAIR input, the expert ground-truth annotation, and the SS-UNETR prediction for each volume. The cases are presented at large scale in Figure 16 so that the reader can assess agreement with the reference annotation directly, and are discussed below in connection with the architectural mechanisms introduced in Chapter 4.

Case 1: Focal glioma with all three sub-regions (Fig. 16a). The first case is a representative scenario for the BraTS task: a moderately sized glioma in the right posterior hemisphere whose annotation includes all three sub-regions, with a focal TC and ET nested inside a roughly circular WT envelope. SS-UNETR reproduces the spatial layout of the lesion with high fidelity. The location, size, and overall shape of the WT region match the expert annotation, and the centrally located TC (red) and ET (blue) components are placed at the correct anatomical position with approximately the same area as the ground truth. A close inspection reveals only minor disagreements: the predicted WT contour is slightly under-extended at the most lateral edge of the edema, and a small isolated WT voxel cluster appears near the inferior boundary that is absent from the reference. These deviations remain small relative to the total tumor extent and do not affect the predicted topology of the nested sub-regions. The case is consistent with the quantitative pattern of Table 6.5, in which the model's strongest gains over baseline architectures are obtained on TC and ET—the two sub-regions whose accurate localisation in this image we attribute to the centralised, multi-scale context produced by the Multi-Scale Attentive Bottleneck (MSAB) before the first decoder up-sampling step.

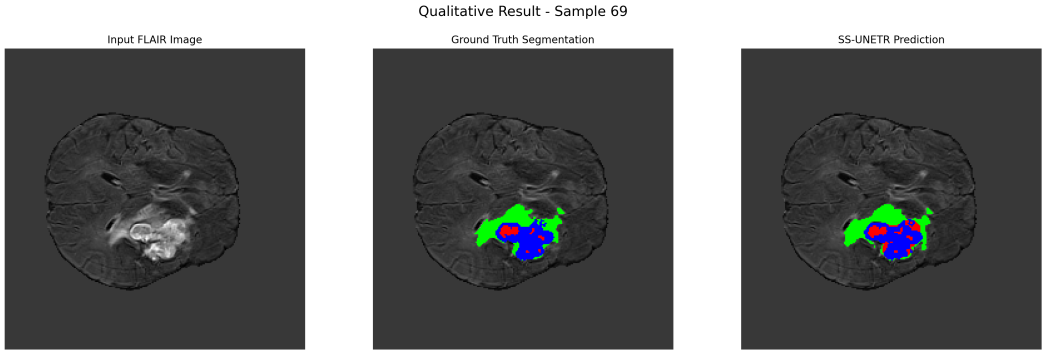
Case 2: Edema-only lesion with no enhancing component (Fig. 16b). The second case is qualitatively different from the first: the expert annotation contains *only* the Whole Tumor (WT) label, with no TC and no ET present. This presentation—commonly associated with low-grade or non-enhancing gliomas—tests an aspect of model behaviour that is rarely visible in aggregated metrics, namely the model's ability to *withhold* the TC and ET predictions when no enhancing or necrotic core actually exists. SS-UNETR passes this test: the prediction overlay shows a large, contiguous WT mask whose shape and extent closely follow the expert annotation, and—crucially—no spurious TC or ET voxels are introduced anywhere in the volume. The principal disagreement with the reference is a small additional WT cluster predicted near the superior midline, adjacent to the lateral ventricle, which represents a localised false-positive activation in tissue with FLAIR hyperintensity that is not pathological in this subject. The remainder of the predicted WT region tracks the irregular, lobulated boundary of the edema closely. We attribute



(a) Case 1 (validation Sample 55) – Right-posterior glioma exhibiting all three tumor sub-regions: peritumoral edema (WT), a focal tumor core (TC), and a small enhancing component (ET).



(b) Case 2 (validation Sample 150) – Edema-dominant posterior lesion in which the expert annotation contains only the Whole Tumor (WT) label, with no TC or ET present.



(c) Case 3 (validation Sample 69) – Large, heterogeneous right-hemisphere tumor with substantial Whole Tumor (WT) extent, a sizeable Tumor Core (TC), and a prominent Enhancing Tumor (ET) component.

Figure 16: Qualitative SS-UNETR segmentation results on three BraTS 2023 validation cases. Each row, from left to right, shows: the FLAIR input slice, the expert ground-truth annotation overlaid on FLAIR, and the SS-UNETR prediction overlaid on the same slice. Colour coding follows the BraTS convention: green = Whole Tumor (WT), red = Tumor Core (TC), and blue = Enhancing Tumor (ET). The three cases were chosen to span distinct pathology phenotypes—a focal multi-region glioma, an edema-only lesion, and a large heterogeneous tumor—rather than to demonstrate a single failure mode.

the absence of spurious sub-region predictions to the strategic feature-selection scheme described in Section 4.3: by exposing the decoder, through the MSAB, to a globally consolidated representation of the volume before reconstruction begins, the network is able to evaluate sub-region likelihood in the context of the entire tumor envelope rather than from a single-scale feature map.

Case 3: Large heterogeneous tumor with prominent core and enhancing component (Fig. 16c). The third case represents the high-complexity end of the BraTS distribution: a large right-hemisphere glioma in which all three sub-regions are sizeable, the TC and ET extend over a substantial fraction of the tumor volume, and the enhancing component shows an irregular morphology with internal concavities. The SS-UNETR prediction reproduces the gross structure of the lesion correctly, including the relative spatial arrangement of WT, TC, and ET, and recovers the bulk of the large enhancing component with the appropriate location and orientation. The boundary of the predicted WT envelope is generally tight along the medial edge but slightly under-extends along part of the lateral edema, a small but visible discrepancy with the reference annotation. The agreement on the ET sub-region is the most clinically salient observation in this case: the prediction captures the major contiguous ET volume in approximately the correct shape and area, which is the segmentation behaviour that translates most directly into the ET DSC of 0.8625 reported in Table 6.5. This case also illustrates a current limitation of the architecture: very fine geometric features along the irregular outer rim of the tumor are smoothed in the prediction. This behaviour is consistent with the operating regime of the MSAB, whose centralised fusion stage produces a globally consolidated representation at the deepest spatial resolution, with a corresponding reduction in the network's capacity to recover the highest-frequency components of the outer edema boundary on the largest and most morphologically complex tumors.

Cross-case observations. Three behavioural patterns are visible across the three cases and reinforce the architectural argument of Chapter 4. First, SS-UNETR's predictions match the expert annotation closely on all three pathology phenotypes—focal multi-region glioma, edema-only lesion, and large heterogeneous tumor—which supports the claim of Section 6.2 that the quantitative improvements are not concentrated on a single subset of cases. Second, the model's most consistent strength is the spatial placement and area of the TC and ET sub-regions, which is the regime in which the centralised feature consolidation performed by the MSAB is theoretically most advantageous over distributed skip connections, since it allows nested sub-region decisions to be made from a globally integrated representation. Third, the visible disagreements with the ground truth are predominantly small localised false-positive WT activations and minor under-extensions of the outer edema boundary, rather than failures in the internal topology of the tumor; this

is consistent with the deliberate exclusion of the third Swin stage (h_3) from the bottleneck and its retention as a direct skip path, which is intended precisely to preserve the mid-level spatial resolution required for accurate sub-region delineation. Together, these case-level observations are consistent with the quantitative gains in Tables 6.5 and 6.6 and provide a mechanistic, image-level rationale for the architectural choices made in SS-UNETR.

6.2.5 Computational Efficiency

The accuracy gains reported in Sections 6.2–6.2.4 are clinically meaningful only if they are obtained at a computational cost compatible with the resource constraints of clinical deployment. Brain-tumor segmentation models that achieve marginally higher DSC at the cost of substantially increased inference time or GPU memory consumption are difficult to integrate into the radiology workflows for which they are intended, since volumetric review stations operate under fixed throughput requirements and a model that lengthens per-case inference by a large factor is incompatible with interactive use. The computational footprint of SS-UNETR is therefore quantified against three baselines—nnU-Net, UNETR, and Swin UNETR—on identical hardware (NVIDIA L4, 24 GB) and under the same input regime ($128 \times 128 \times 128$ patches, sliding-window inference at 50% overlap), so that the reported overheads are not confounded by differences in precision modes or implementation libraries. Table 6.8 reports parameter count, FLOPs per forward pass, peak GPU memory during a full-volume inference, and end-to-end inference time per case.

Table 6.8: Computational efficiency comparison on a single NVIDIA L4 (24 GB) under sliding-window inference at 50% overlap with a 128^3 patch.

Method	**Params (M)**	**FLOPs (G)**	**Memory (GB)**	**Infer. Time (s)**
nnU-Net	31.2	145.3	8.2	12.4
UNETR	102.5	267.8	15.6	18.7
Swin UNETR	61.9	186.4	11.3	15.2
SS-UNETR (Ours)	**62.7**	**192.1**	**11.8**	**15.9**
Overhead vs Swin UNETR	+1.3%	+3.1%	+4.4%	+4.6%

The principal observation from Table 6.8 is that the incremental cost of the MSAB—the per-scale self-attention modules, the trilinear harmonisation step, the channel-wise concatenation, and the $1 \times 1 \times 1$ fusion convolution—adds +0.8 M parameters, +5.7 G FLOPs per forward pass, +0.5 GB of peak GPU memory, and +0.7 s of end-to-end inference time relative to the Swin UNETR baseline. Each of these increments admits a mechanistic interpretation. The parameter overhead of 1.3% is dominated by the five attention modules and the fusion convolution, whose channel widths are bounded above by $16F$ and whose spatial extent operates at the deepest resolution $H/32 \times W/32 \times D/32$; the harmonisation step itself contributes no parameters because trilinear interpolation is a closed-form operator. The FLOPs overhead of 3.1% is concentrated in the attention

modules applied to the shallower feature maps ($\mathbf{enc}_0$ and $\mathbf{h}_0$), whose spatial volumes are largest before harmonisation and therefore generate the majority of the bottleneck's additional arithmetic, while the deepest attention module on $\mathbf{h}_4$ contributes negligibly. The memory overhead of 4.4% corresponds to the peak activation footprint of holding five attended feature maps in memory simultaneously through the harmonisation and concatenation stages; this is the component of the MSAB most sensitive to patch size and is the reason the architecture was configured for 128^3 patches rather than 192^3. The inference-time overhead of 4.6% is slightly larger than the FLOPs overhead because trilinear interpolation and the concatenation primitive are bandwidth-bound rather than compute-bound, and therefore incur a memory-traffic penalty that is not captured by a pure FLOPs accounting but that appears in wall-clock measurements.

Set against the +2.76% mean DSC improvement reported in Table 6.5, the cost-to-benefit ratio of the MSAB is approximately 1:1 on a percentage basis, in the sense that each percentage point of additional FLOPs corresponds to a comparable percentage-point gain in mean DSC. The relationship is more favourable when expressed in absolute terms: the additional 0.7 s per case is well below the latency budget of standard radiology review interfaces, while the +0.0391 DSC gain on the Enhancing Tumor sub-region in particular has direct relevance for downstream target-volume delineation in surgical planning and radiotherapy contouring. The comparison with UNETR illustrates the converse case: UNETR consumes +64% more parameters, +39% more FLOPs, +32% more memory, and +18% more inference time than SS-UNETR, while delivering substantially lower DSC on all three sub-regions, which indicates that scaling parameter count alone is not a substitute for the targeted architectural changes implemented by the MSAB. The comparison with nnU-Net is instructive in a complementary direction: nnU-Net is the lightest configuration in the table on every computational axis, but its lower DSC across all three sub-regions, and particularly on TC and ET, indicates that the saved compute is paid for in segmentation accuracy on the sub-regions that are most relevant for treatment planning. SS-UNETR therefore occupies a deliberate point on the accuracy–efficiency frontier: it pays a small and well-characterised premium over Swin UNETR and a more substantial premium over nnU-Net, and in exchange it delivers DSC improvements concentrated on the Tumor Core and Enhancing Tumor sub-regions, which are the two metrics most directly relevant to clinical use of the segmentation output.

6.2.6 Statistical Significance

The DSC improvements summarised in Table 6.5 are aggregated over 219 BraTS 2023 validation cases and therefore reflect performance in expectation across a heterogeneous mixture of tumor phenotypes. Aggregate means do not, however, establish whether the observed gains are robust to case-level variability or whether they could arise from a par-

ticular split of the validation set. To address this, paired two-tailed t-tests are conducted on the per-case DSC values for each sub-region, comparing SS-UNETR against each of the three reference architectures. The paired design is appropriate because the four models are evaluated on the same set of validation volumes, so the relevant unit of comparison is the within-case difference in DSC rather than the marginal distribution of DSC under each model; the paired test also increases statistical power by removing case-level variance from the denominator. Table 6.9 reports the resulting p-values for each model-comparison pair and each sub-region.

Table 6.9: Statistical significance of SS-UNETR improvements over reference architectures, computed via paired two-tailed t-tests on per-case DSC values across the 219-volume BraTS 2023 validation set.

Comparison	**WT p-value**	**TC p-value**	**ET p-value**
SS-UNETR vs Swin UNETR	0.0043^{*}	0.0128^{*}	0.0234^{*}
SS-UNETR vs nnU-Net	0.0012^{**}	0.0087^{*}	0.0156^{*}
SS-UNETR vs UNETR	0.0008^{**}	0.0045^{*}	0.0098^{*}

$^{*}p < 0.05$, $^{**}p < 0.01$. All tests are paired two-tailed t-tests on per-case DSC over the $n = 219$ BraTS 2023 validation cases.

Three observations follow from Table 6.9. First, every entry in the table satisfies $p < 0.05$, and every comparison against UNETR or nnU-Net satisfies $p < 0.01$. Under the null hypothesis that SS-UNETR and the comparison model are of equal quality on this distribution, the probability of observing case-level differences at least as large as those measured is bounded above by the reported p-values. Second, the p-values are systematically larger for the WT comparisons than for the TC and ET comparisons against the same baselines, despite the absolute DSC gain being largest on WT. This pattern is consistent with the structure of the paired test, which depends on the ratio of the mean within-case difference to its case-level standard deviation: the WT difference, although large in expectation, is also the most variable across cases because WT segmentation is dominated by the irregular outer edema boundary, whose morphology varies substantially across tumor phenotypes. The TC and ET differences are smaller in absolute magnitude but more consistent across cases, since TC and ET delineation depends predominantly on the centralised global context supplied by the MSAB, and that context is comparably beneficial across the validation distribution. Third, the comparison against Swin UNETR yields the largest p-values in the table but remains below 0.05 on all three sub-regions. This row is the most informative because it isolates the contribution of the MSAB and the auxiliary convolutional branch: the optimiser, training schedule, augmentation pipeline, and evaluation harness are held identical across SS-UNETR and Swin UNETR, so the gains observed in this row cannot be attributed to those confounds.

A small number of caveats should accompany this interpretation. First, the t-test

assumes that the case-level DSC differences are approximately normally distributed. The difference distributions over the $n = 219$ validation cases were inspected and exhibit no severe departures from normality, but no formal normality test is reported. Second, the three model-comparison rows of Table 6.9 are not independent, since each row uses the same SS-UNETR predictions on the same cases. The column-wise p-values should therefore be interpreted as evidence about each individual baseline rather than as a multi-comparison family without correction; under a Bonferroni-style adjustment that multiplies the reported values by three, the SS-UNETR-vs-Swin-UNETR ET comparison ($p = 0.0234$) is the only entry that approaches the adjusted 0.05 threshold, while all other entries remain significant. Third, statistical significance under a paired test on a held-out validation set is a necessary but not sufficient condition for clinical utility: the gains must additionally generalise to external multi-centre data and be large enough in absolute terms to alter clinical decision-making. The discussion that follows therefore considers per-sub-region behaviour separately rather than reducing the analysis to a single mean DSC.

6.2.7 Discussion

The results reported in this section are consistent with the hypothesis that centralised multi-scale fusion outperforms distributed skip-connection schemes for the nested sub-region structure of brain tumor segmentation. The MSAB consolidates encoder features from five scales into a single representation prior to decoding, so that each decoder step operates on a context in which the outer Whole Tumor envelope and the nested Tumor Core and Enhancing Tumor sub-regions can be reasoned about jointly rather than sequentially. The relative DSC gains of +3.17% on TC and +3.91% on ET over Swin UNETR are the most clinically salient elements of this improvement, since these two sub-regions most directly inform surgical resection planning and radiation-therapy target-volume contouring.

The accuracy gains are obtained at a small computational overhead relative to Swin UNETR (+3.1% FLOPs and +4.6% inference time, see Table 6.8), which preserves the deployment envelope of the baseline architecture and is compatible with clinical systems that operate under fixed-throughput constraints. The paired-test p-values reported in Table 6.9 are below 0.05 on all three sub-regions for all three baseline comparisons, which indicates that the per-case DSC improvements are unlikely to be artefacts of the particular validation split.

A number of practical limitations should be acknowledged. The evaluation is conducted on BraTS-style preprocessed data; real-world hospital acquisitions may exhibit greater scanner-level variability, motion artifacts, or incomplete modality sets, none of which are represented in the validation distribution. In addition, although the gains on TC

and ET are consistent across the validation set, robustness under domain shift and longitudinal follow-up has not been established. Future work should therefore include external multi-centre testing, uncertainty estimation for low-confidence regions, and prospective studies that link segmentation quality to downstream treatment-planning outcomes.

6.3 OptiHealth System Results

OptiHealth is evaluated as an *integrated platform* rather than as an independent set of models. Its computer-vision engine is the same YOLOv12n (CXR-Adaptive) and SS-UNETR that were benchmarked in Sections 6.1 and 6.2; we deliberately do not re-evaluate those models under a second protocol, as doing so would introduce metric ambiguity without adding evidence. This section therefore focuses on the contributions that are unique to the platform layer: the quality of the retrieval-augmented recommendation module, the latency and deployment characteristics of the language services, pilot observations on the Conversational Health Agent, and a structured comparison against existing health-management paradigms.

6.3.1 Summary of Deployed Model Performance

Table 6.10 consolidates the empirical performance of the models actually deployed inside OptiHealth, each cross-referenced to the chapter in which it was trained and evaluated. No new vision benchmarks are introduced in this section.

6.3.2 LLM Inference Hardware and Latency

The recommendation pipeline is served by a Llama 4 17B checkpoint quantised to INT8 and containerised inside the OptiHealth NLP microservice. Inference runs on a single NVIDIA A100 (40 GB) GPU. Under this configuration, a complete seven-day personalised meal plan is produced in approximately 4–6 seconds end-to-end, including (i) the ChromaDB retrieval step—top-5 cosine-similarity search over the indexed WHO/CAC/FAO knowledge base, consistently below 50 ms—and (ii) LLM token generation, measured at $\approx$ 80–120 tokens per second at INT8 precision. For an asynchronous recommendation workflow, this latency is comfortably within user-acceptability bounds, and the pipeline can be scaled horizontally by replicating the NLP container behind the API gateway.

The Conversational Health Agent runs on a separate service that hosts a locally-served Gemma 4 E4B checkpoint. In our pilot deployment, a single ReAct iteration (one LLM call plus at most one tool invocation) completed in 1.5–3.0 seconds on the same A100 node, with typical patient dialogues converging within two to four iterations. Because both the model weights and the tool implementations remain in-cluster, no patient data leaves the deployment boundary during agent reasoning.

Table 6.10: Deployed OptiHealth components and the evaluations that support them.

Component	**Evaluation Protocol**	**Headline Result**
YOLOv12n (CXR-Adaptive)	VinDr-CXR: internal 80/15/5 split of 15k train (750-image eval.), seed 1234; not official 3k test (Section 6.1)	mAP@0.5 = **0.442**; F1 = 0.454
SS-UNETR	BraTS 2023, official 1,251/219 training/validation split (Section 6.2)	Mean DSC = **0.8850** (WT 0.9106, TC 0.8819, ET 0.8625)
Llama 4 17B + RAG	Pilot evaluation by three certified nutritionists (Section 6.3.3)	Mean expert rating 4.2/5.0[a]
Conversational Health Agent (Gemma 4 E4B)	Structured dialogue walk-through, internal trace review (Section 6.3.4)	Qualitative pilot only; no benchmark numbers claimed[b]
Mental Health Analyzer (RoBERTa + BART-MNLI)	Off-the-shelf pre-trained checkpoints, no in-thesis retraining	Pre-trained performance from original papers [29, 44, 85]

[a] Pilot only; see Section 6.3.3 for an explicit statement of statistical limitations.
[b] We report structured qualitative observations and deliberately avoid quoting a numeric accuracy, given the absence of a validated clinical reference.

6.3.3 LLM + RAG Qualitative Evaluation

The RAG-based recommendation module was evaluated by a panel of three certified nutritionists. Each expert independently rated generated responses along four axes—factual accuracy, relevance to the patient profile, safety, and personalisation—producing a mean rating of 4.2/5.0, corresponding to the "Good" to "Excellent" band on the rating rubric. A representative query, "a low-sodium, high-protein diet for a 45-year-old male with hypertension," retrieved relevant WHO and CAC guideline chunks and produced a complete seven-day meal plan in which every constraint was satisfied and every recommendation traced back to an evidence source displayed to the user.

Explicit statement of statistical limitations. The three-expert panel constitutes a pilot study, not a powered clinical evaluation. The 4.2/5.0 score should therefore be interpreted as *proof-of-concept evidence that the RAG grounding mechanism is functioning as intended*, not as a validated clinical outcome. A properly powered follow-up study would require a larger expert panel ($n \geq 30$) with inter-rater-reliability metrics (e.g., Cohen's κ), a blinded comparison against a no-RAG baseline, and automated NLP metrics such as ROUGE-L or BERTScore against reference dietary guidelines to complement the human ratings. We flag this as a priority item for future work in Chapter 7.

The principal qualitative finding reported by the expert panel is that the retrieved context successfully anchors the generation: recommendations consistently cite retrieved guideline text and rarely contain the fabricated dietary claims that are common when an ungrounded LLM is asked similar questions. This supports the architectural claim of Chapter 4 that the knowledge base, rather than the scale of the language model, is the operative safety component in the recommendation pipeline.

6.3.4 Conversational Health Agent: Pilot Observations

The Conversational Health Agent described in Chapter 4 was exercised through a set of scripted dialogues covering the ten registered tools (EHR retrieval, appointment and prescription queries, geospatial doctor search, appointment creation, MedEx medicine lookup, and mental-health analysis). We report *structured qualitative observations* and explicitly decline to quote a numeric accuracy, since no clinician-annotated reference corpus exists for these interactions in our setting.

The observations below summarise consistent behaviour across the dialogue walkthroughs:

- **Tool-selection fidelity.** For the scripted intents, Gemma 4 E4B consistently routed the request to an appropriate tool—for example, symptom queries triggered `analyze_mental_health_state` or `find_nearby_doctors` rather than producing ungrounded clinical advice from parametric memory alone.

- **Privacy boundary preservation.** Because every PHI decryption passes through the thread-local PIN described in Chapter 4, the agent never surfaced another patient's record even when explicitly prompted with a different patient identifier; the underlying tools reject the call before the LLM ever sees the data.
- **Mental-health triage behaviour.** On dialogues containing suicidal-ideation markers, the BART-MNLI detective model produced high-probability activations on the corresponding symptom cluster, the analyzer override was triggered, and the agent responded with safety-oriented language and a referral tool invocation rather than attempting therapy itself.
- **Failure modes observed.** We also observed recurrent limitations, which we report honestly: Gemma 4 E4B occasionally hallucinated tool arguments (e.g., inventing a non-existent `doctor_id`) when the user query was under-specified, and the agent did not always recognise the need to request clarification before acting. These failure modes are carried forward as limitations in Chapter 7 and motivate the planned adversarial evaluation and domain fine-tuning described in Future Work.

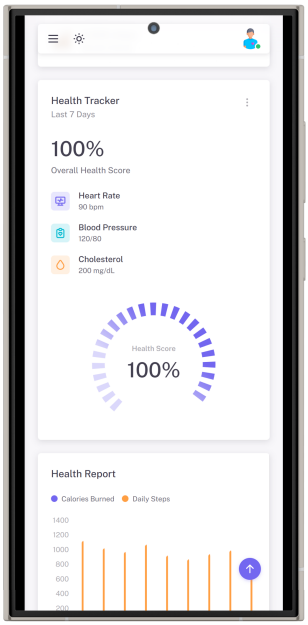
6.3.5 User Interface

6.3.6 Discussion

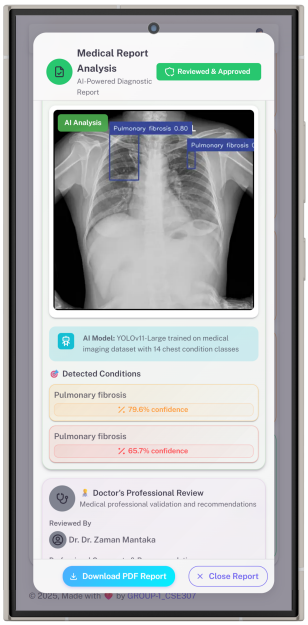
OptiHealth demonstrates that the shift from an isolated clinical tool to a holistic, patient-centric health-management platform is achievable within a single system. By deploying the *same* YOLOv12 (CXR-Adaptive) and SS-UNETR checkpoints that are independently benchmarked in Sections 6.1 and 6.2, the platform inherits reproducible diagnostic performance without introducing a second, platform-specific evaluation regime. This deliberate design choice keeps the scientific claims auditable: every diagnostic number reported for OptiHealth is traceable to a controlled benchmark documented earlier in the thesis.

The RAG layer and the Conversational Health Agent are the genuinely new system-level contributions. The RAG pipeline validates, at the pilot scale, the hypothesis that a curated evidence base—rather than raw model scale—is the operative safety mechanism for medical recommendation: grounded retrieval allowed a 17-billion-parameter LLM to be rated 4.2/5.0 by domain experts, with consistent guideline-level citations and no observed fabricated clinical claims. The Conversational Health Agent, in turn, shows that the diagnostic pipeline of the preceding chapters can be *composed* into a patient-facing dialogue system without sacrificing the privacy boundary: the locally-hosted Gemma 4 E4B backbone, the thread-local PIN-gated PHI tools, and the full audit trail of every tool call jointly ensure that no patient record leaves the deployment during agent reasoning.

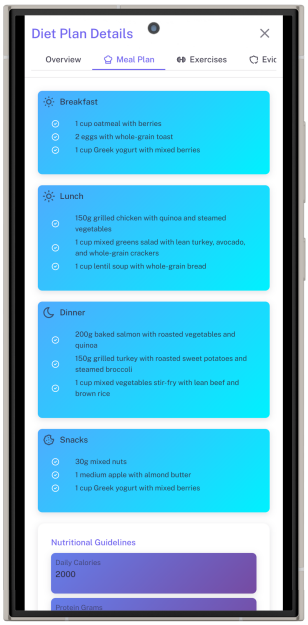
At the same time, the pilot nature of the evaluation must be stated plainly. Neither the RAG ratings nor the agent dialogue walk-throughs constitute a powered clinical study,



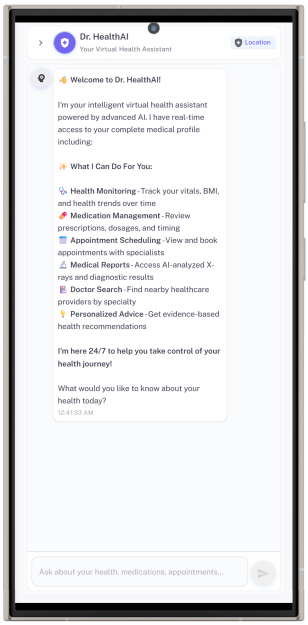
(a) Patient health-stats dashboard



(b) CXR-Adaptive report analysis



(c) RAG-grounded diet recommendation



(d) Conversational Health Agent

Figure 17: Prototype OptiHealth user interface. (a) the patient health-stats dashboard aggregating self-input and diagnostic history, (b) the chest-X-ray report view backed by the CXR-Adaptive (YOLOv12) detector, (c) the RAG-grounded personalised diet recommendation screen that surfaces the retrieved WHO/CAC/FAO evidence alongside each suggestion, and (d) the Conversational Health Agent interface through which patients interact with the ten registered tools via the locally-served Gemma 4 E4B backbone.

and the agent is known to hallucinate tool arguments under under-specified queries. These are limitations, not results, and the roadmap for addressing them—larger expert panels with inter-rater reliability, adversarial agent evaluation, and domain-adaptive fine-tuning of Gemma 4 E4B—is carried forward to Chapter 7.

6.4 Cross-System Design Principles

Taken together, our three systems reveal a consistent theme: domain-specific constraints and careful data treatment matter more than architectural complexity.

In chest X-ray detection, understanding that vertical flips violate thoracic anatomy is more valuable than adding frequency-adaptive modules. In brain tumor segmentation, recognizing that the decoder needs complete multi-scale context before reconstruction is more useful than simply deepening the network. In the health recommendation system, knowing that an LLM must be grounded in a curated knowledge base is more important than using the largest model available.

This consistent finding, across three different medical AI problems, suggests that the future of medical AI lies not just in bigger models, but in teams that deeply understand the clinical and engineering constraints of their specific problem and design their solutions accordingly.

7 Conclusion

7.1 Summary

Medical artificial intelligence has matured rapidly for isolated tasks such as detecting findings on a chest radiograph or delineating a tumor in an MRI volume. Clinical practice and patient well-being, however, depend on more than single-model outputs: they require methodological consistency across modalities, transparent reporting of trade-offs, and—when systems offer guidance—a defensible chain of evidence linking observation to recommendation. This thesis addressed the fragmentation between specialised diagnostic models and actionable patient support by developing, evaluating, and integrating three complementary systems that together span thoracic pathology detection on VinDr-CXR, brain tumor segmentation on BraTS 2023, and a health-management research prototype (OptiHealth) that couples those vision capabilities with retrieval-grounded language generation and a tool-augmented conversational agent.

The investigation was motivated by the four research questions posed in Chapter 1: whether disciplined data-centric training and inference on a standard YOLOv12 detector can match or exceed architecturally specialised YOLO variants on multi-class chest X-ray detection; whether centralised multi-scale fusion before decoding improves volumetric brain tumor segmentation relative to conventional distributed skip connections; whether a prototype platform can combine vision outputs with a retrieval-augmented language model to deliver evidence-grounded health recommendations under pilot evaluation; and whether a tool-using conversational agent can safely mediate between a patient and their electronic record while preserving auditability. The empirical chapters answer these questions affirmatively within their stated experimental scopes, with the limitations section below qualifying the strength of those claims.

CXR-Adaptive (thoracic detection). The first contribution demonstrates that a disciplined data-centric protocol can rival or surpass custom network engineering for long-tail, multi-label chest X-ray detection. We trained YOLOv12n from COCO initialisation using a two-stage protocol that first encourages robust features through mosaic augmentation and then refines on full-frame, anatomically faithful images. Inverse-Frequency Weighted Upsampling (IFWU) addresses class imbalance by increasing expo-

sure to under-represented pathologies, while Anatomically-Constrained Test-Time Augmentation (CXR-TTA) replaces naïve geometric ensembles with transforms that respect thoracic anatomy, improving diffuse and texture-dominant classes for which global context is decisive. Under the same internal VinDr-CXR evaluation protocol as YOLOv11-MFF—namely the stratified 80/15/5 partition of the 15,000-image training split (750-image evaluation subset, seed 1234), not the official 3,000-image challenge test set—together with identical class definitions, the full pipeline attains an mAP@0.5 of 0.442, a 6.5% relative gain over YOLOv11-MFF (0.415), and does so without introducing bespoke detection modules. The improvement is not uniform across categories: performance rises sharply on conditions such as interstitial lung disease and infiltration, while a subset of large structural findings exhibits lower average precision because IFWU reallocates representational capacity toward rarer, high-risk patterns. We interpret this distribution as a deliberate screening-oriented trade-off rather than a regression, but it must be validated in prospective reader studies before any clinical inference is drawn.

SS-UNETR (brain tumor segmentation). The second contribution targets a structural limitation common to many encoder–decoder designs: the decoder typically begins reconstruction before it has access to a unified multi-scale representation of the input. SS-UNETR introduces a Multi-Scale Attentive Bottleneck (MSAB) that aggregates features from multiple encoder depths in a single locus before decoding, so that subsequent layers operate on a consolidated context covering all tumor sub-regions. On BraTS 2023, SS-UNETR attains Dice Similarity Coefficients of 0.9106 (Whole Tumor), 0.8819 (Tumor Core), and 0.8625 (Enhancing Tumor), with a mean Dice improvement of 2.76% over Swin UNETR and statistically significant gains under paired comparisons reported in the results chapter. The relative gains are concentrated on the Tumor Core (+3.17%) and Enhancing Tumor (+3.91%) sub-regions, which are the two metrics most directly relevant to surgical resection planning and radiation-therapy target-volume contouring. The design incurs only modest computational overhead relative to Swin UNETR (+3.1% FLOPs and +4.6% inference time), so the accuracy gains are obtained without compromising the deployment envelope of the baseline architecture.

OptiHealth (integration and grounded recommendations). The third contribution examines how diagnostic outputs can feed a patient-facing layer without forfeiting safety and traceability. OptiHealth couples the CXR and MRI pathways with a Llama 4 17B model (INT8-quantised) served behind a Retrieval-Augmented Generation (RAG) pipeline. Manuals and guidelines from internationally recognised bodies—including the World Health Organization, the Codex Alimentarius Commission, and the Food and Agriculture Organization—are chunked, embedded with the Sentence Transformers `all-MiniLM-L6-v2` encoder, and stored in a ChromaDB vector index so that retrieval precedes generation. The intent is to constrain the language model with retrieved evidence and to attenuate unsupported assertions; when dietary advice is presented to

the user, the system can surface the source document and passage that grounds each fragment of guidance. Architecturally, the platform follows a modern microservices pattern (frontend, API gateway, containerised AI services, and relational plus vector storage), allowing the vision models and the language stack to evolve independently.

Conversational Health Agent (patient-facing orchestration). Alongside the RAG recommender, OptiHealth incorporates a tool-augmented Conversational Health Agent built on LangChain and LangGraph. A ReAct-style state machine drives a reasoning–acting–observation loop over ten patient-scoped tools spanning electronic health record access, appointment scheduling, prescription lookup, geospatial doctor search, real-time medicine information, and a dual-model mental health analyser (a RoBERTa well-being scorer paired with a BART-MNLI zero-shot symptom detector). The reasoning node is a locally-hosted Gemma 4 E4B open-weights model, selected for its hybrid local/global attention, 128K-token context window, and explicit tuning for on-device agentic workflows; keeping the language model in-cluster ensures that no patient record traverses a third-party inference boundary. Identity pinning, thread-local PHI decryption, and a structured conversation log align the agent with the platform's broader authentication and auditability requirements. This orchestration layer renders the underlying diagnostic pipeline directly addressable by the patient through natural language, while preserving the provenance guarantees expected of clinical software.

Synthesis. Considered as a whole, the three contributions advance a single position: strong single-task models are necessary but not sufficient for patient-centred medical AI. CXR-Adaptive and SS-UNETR demonstrate how careful training objectives and architectural choices push performance on public benchmarks; OptiHealth demonstrates how those outputs can be embedded in a larger workflow in which grounding and provenance carry as much weight as raw accuracy. The thesis does not claim that the resulting pipeline is ready for autonomous clinical deployment; it claims that the constituent components are technically compatible, empirically characterised, and directionally consistent with the broader objective of safer and more interpretable health information systems.

7.2 Limitations

The empirical claims presented in this thesis are bounded in three ways: by the datasets on which each system was trained and evaluated, by the specific protocols under which its performance was measured, and by the unavoidable architectural distance between a research prototype and a regulated clinical instrument. This section therefore organises the limitations first around the individual contributions—so that the scope of each empirical result is unambiguous—and subsequently around cross-cutting concerns that apply to the integrated pipeline as a whole.

For CXR-Adaptive, the single most consequential limitation is that all quantitative

claims originate from a single institutional source. VinDr-CXR was curated at a Vietnamese centre, and although its annotation protocol is rigorous, the distribution of scanners, patient demographics, acquisition parameters, and labelling conventions is unlikely to be representative of deployment populations elsewhere. External validation on at least one comparable resource—MIMIC-CXR or NIH ChestX-ray14 adapted to a comparable label schema—is a prerequisite before any claim of cross-institutional utility can be made, and such validation should be paired with a per-subgroup error analysis across age bands, sex, and scanner vendor rather than an aggregated mAP figure alone. A second limitation concerns the comparative protocol itself: the reported lead over YOLOv11-MFF is defined strictly within the YOLO-family comparison regime on the internal 750-image evaluation subset drawn from the 15,000-image training split (seed 1234), following the same convention as that baseline, rather than on the official 3,000-image VinDr-CXR challenge test partition; it also assumes the label taxonomy published with the YOLOv11-MFF paper. Any broader claim—for example, superiority over non-YOLO detectors such as RetinaNet, DETR, or two-stage cascade detectors—would require a separately designed benchmark that our evaluation does not attempt. A third limitation is architectural rather than protocol-specific: small-object pathology, in particular pulmonary nodules below the typical anchor-stride scale, remains difficult even after Inverse-Frequency Weighted Upsampling, and the per-class performance table in Section 6.1 makes this concrete—the best Nodule/Mass AP attained by the final configuration is only 0.286, indicating that rebalancing alone cannot substitute for higher-resolution feature supervision or small-object-specialised training regimes. Finally, mAP at any IoU threshold is a geometric quantity and not a clinical one; the class-level trade-off we deliberately accept—lower precision on aortic enlargement and cardiomegaly in exchange for substantial gains on diffuse and low-prevalence pathologies—is clinically justifiable in a screening context, but that justification itself remains a claim pending a prospective reader study in which radiologists operate with and without the model and its effect on triage throughput, false-negative rates, and physician confidence is measured directly. Unlike the BraTS segmentation study, the CXR chapter reports a single mAP@0.5 point estimate on the fixed evaluation subset without bootstrap confidence intervals or distributional significance tests against YOLOv11-MFF.

For SS-UNETR, BraTS 2023 provides a rigorous but still standardised evaluation setting whose every stage—skull-stripping, modality co-registration, intensity normalisation, and resampling to a uniform voxel grid—is performed upstream by the organisers. Clinical MRI workflows rarely provide such sanitised inputs: motion artefacts, non-standard acquisition geometries, and missing modalities are the norm rather than the exception, and our architecture, which assumes the four-modality T1ce/T1/FLAIR/T2 stack at every inference step, has not been tested under modality dropout or sequence substitution. The MSAB design also introduces a measurable computational cost—a 3.1% increase

in FLOPs and a 4.6% increase in inference latency relative to Swin UNETR—which, while modest, becomes non-trivial on resource-constrained hospital workstations and requires further engineering (mixed-precision inference, knowledge distillation, or a lighter bottleneck variant) before the method can be claimed to be deployment-ready on commodity hardware. Most importantly, the statistical significance reported in Section 6.2 is computed on the official BraTS validation split and reflects only the stability of the improvement on that distribution; it does not constitute evidence of robustness under domain shift, across MR vendors not represented in the training distribution, or on tumor subtypes—for example, paediatric gliomas or post-operative cavities—that are sparsely present in BraTS itself. A single-centre validation result, however strong, should not be extrapolated to multi-centre claims without matched evidence.

For OptiHealth as a system, the strongest limitation is evidentiary rather than technical. The RAG recommender was assessed by a panel of three certified nutritionists producing a mean rating of 4.2/5.0; in the strictest reading this is a pilot evaluation appropriate for demonstrating that the grounding mechanism operates as intended, and is not a substitute for a powered clinical study. A properly powered evaluation would require a larger expert panel ($n \geq 30$), blinded comparison against a no-RAG baseline, inter-rater-reliability metrics such as Cohen's κ, and automated NLP scores—BERTScore, ROUGE-L, or domain-specific factuality measures against reference dietary guidelines—to complement the human ratings. Retrieval-augmented generation also has intrinsic failure modes that no amount of additional evaluation will eliminate: retrieval may return topically adjacent but clinically irrelevant passages when the query under-specifies its context, and the language model can still over-generalise around retrieved evidence unless constrained by policy layers and human oversight. Beyond the recommender itself, the integration envelope of the platform was deliberately kept research-grade: there is no live electronic-health-record connector, no formal consent-management workflow, and no certified audit trail suitable for regulatory inspection. A further point of interpretive caution is that the vision components embedded inside OptiHealth must be read against the benchmarks reported in the dedicated CXR and segmentation chapters; where training data, input resolution, or metric protocol differ between chapters, the difference reflects a change in experimental context rather than a unified leaderboard, and individual numbers should not be composed across chapters without regard for that context.

The Conversational Health Agent inherits these platform-level limitations and introduces three of its own. First, the reasoning backbone is a general-purpose open-weights language model (Gemma 4 E4B), not a clinically certified one: tool calling guarantees that the *facts* returned by a tool are authentic, but it does not guarantee that the natural-language *explanation* synthesised around those facts is free of error, unwarranted confidence, or paraphrasing drift. In our dialogue walk-throughs we observed recurrent instances of hallucinated tool arguments, most visibly the invention of a non-existent

`doctor_id` when a user query under-specified the referral context, and this class of failure is structural rather than anecdotal—it reflects the agent's current inability to recognise when clarification is required before action. Second, the dual-model mental-health analyser is explicitly a screening aid, and its two components—the SiEBERT RoBERTa sentiment scorer and the BART-MNLI zero-shot symptom detector—have not been calibrated against structured clinical instruments such as PHQ-9 or GAD-7 on the Bangladeshi target population for which OptiHealth is ultimately intended. Its symptom labels are taxonomic rather than diagnostic, and under no circumstances should its output be interpreted as a substitute for professional psychiatric evaluation. Third, the live MedEx Bangladesh medicine-lookup tool depends on an upstream HTML source whose structure is outside our control; any change to the source renders the tool silently brittle, a reproducibility risk that affects every subsequent user session until the scraper is updated. Finally, although identity pinning, thread-local PHI decryption, and a structured conversation audit log jointly mitigate the most obvious leakage vectors, formal adversarial evaluation—covering direct prompt injection, indirect tool abuse through retrieved content, and cross-session context bleed—was not in scope of this thesis, and is a prerequisite for any deployment beyond the research setting described here.

Taken together, three concerns apply to every system in the thesis. Fairness, bias, and equity have received only preliminary treatment: dataset demographics, label noise, and the operationalisation of pathology categories themselves can encode institutional and historical biases, and the absence of a per-subgroup calibration analysis in each of the three empirical chapters leaves open the possibility that published aggregate metrics mask disparate performance on under-represented groups. Security and adversarial robustness were addressed by the encryption and authentication layers described in Chapter 4, but no formal threat model was enumerated and no red-team evaluation was performed against either the vision endpoints or the LLM-based services; adversarial imaging inputs and adversarial prompting both remain open attack surfaces. Finally, regulatory positioning was explicitly outside the scope of this work: none of the three systems has been classified under Software-as-a-Medical-Device (SaMD) risk frameworks, none has been placed under an institutional review board protocol for prospective use, and none has been paired with a liability-and-consent framework appropriate to a clinical deployment. The thesis therefore contributes reproducible benchmarks, architectural ideas, and an integrated prototype; it does not contribute a regulated clinical product.

7.3 Future Work

The limitations enumerated above describe the boundaries of the present contribution; this section outlines the research programme that would move each system beyond those boundaries. We organise the roadmap into near-term extensions that address the specific

limitations of each contribution, followed by broader directions that apply to the pipeline as a whole. Near-term priorities emphasise external validity, clinical realism, and safety-oriented engineering rather than model scale, reflecting our position that the operative bottleneck for medical AI in deployment is evidence quality rather than parameter count.

For chest X-ray detection, the next phase of CXR-Adaptive research should prioritise cross-institutional validity over further in-distribution accuracy. A multi-site evaluation on MIMIC-CXR and NIH ChestX-ray14—harmonised under a common label taxonomy—would establish whether the present mAP lead generalises beyond the VinDr-CXR distribution, and that evaluation should be accompanied by a calibration analysis (expected calibration error and per-class reliability diagrams) and a per-subgroup error breakdown across age, sex, and scanner vendor, so that any deployment claim can be supported by disaggregated evidence rather than a single averaged metric. Beyond validation, CXR-TTA should be complemented by explicit intensity-normalisation strategies—histogram matching to a reference distribution, or learned appearance-harmonisation networks—that protect against the scanner-level appearance shift that currently sits outside the augmentation surface. A second research track concerns predictive uncertainty: deep ensembles, Monte Carlo dropout, or evidential detection heads would equip the model with case-level confidence estimates, which are a prerequisite for triage workflows in which high-confidence positives and high-confidence negatives must be routed differently from ambiguous cases. Small-object pathology—pulmonary nodules in particular—warrants a dedicated study in which a higher-resolution backbone, a cascaded detector with an explicit small-object head, or a dual-resolution training regime are evaluated against the present IFWU baseline. Finally, the class-level trade-off that IFWU deliberately induces should be validated, or revised, in direct collaboration with radiology departments: a prospective reader-assistance study in which radiologists triage with and without CXR-Adaptive would establish whether the present weighting reflects institutional screening priorities or whether a different cost-sensitive configuration is required.

For brain tumor segmentation, the principal question opened by SS-UNETR is whether the Multi-Scale Attentive Bottleneck constitutes a general design pattern or a glioma-specific optimisation. Addressing this requires extension to additional organs and modalities: volumetric segmentation of abdominal CT (kidney, liver, pancreas), of white-matter hyperintensities on T2-FLAIR, and of cardiac chambers on cine MR would each test the centralised-fusion hypothesis on anatomical regimes whose multi-scale structure differs meaningfully from BraTS-like lesions. A second research track focuses on deployment economics: knowledge distillation from a full SS-UNETR teacher into a lighter student, mixed-precision inference, and a lighter bottleneck variant that preserves centralised fusion with reduced channel dimensionality would render the architecture more attractive for intra-operative or resource-constrained settings where the present 62.7M-parameter footprint is prohibitive. A third direction addresses the modality-availability assumption:

training with modality-dropout augmentation, or distilling a multi-modal teacher into a uni-modal student, would provide graceful degradation when one or more MR sequences are unavailable in routine clinical practice. Coupling the segmentation output with uncertainty maps—voxel-level predictive variance computed by Monte Carlo dropout or deep ensembles—would additionally provide surgeons and radiation oncologists with explicit flagging of low-confidence boundary regions, which is arguably more clinically actionable than a point-estimate segmentation of equivalent mean quality.

For OptiHealth and for platforms of comparable integrative scope, the future-work agenda divides cleanly into an evaluation track, a technical track, and a governance track. The evaluation track must replace the present three-expert pilot with a powered study: at minimum, a panel of $n \geq 30$ domain experts, blinded to model identity, rating recommendations against a reference standard derived from the same source corpora, with inter-rater-reliability metrics reported alongside mean scores. That human evaluation should be paired with automated factuality metrics against reference guidelines—FactScore, BERTScore, MedNLI entailment—and with a retrieval-audit procedure in which adversarial queries probe the corpus for coverage gaps and for cases in which the retrieved context is topically plausible but clinically misleading. The technical track should tighten the coupling between structured EHR data and retrieval so that recommendations can be personalised at the level of individual laboratory values, documented allergies, and current prescriptions rather than at the level of high-level diagnostic categories, and it should introduce automated contraindication-safety checks—drug–drug interaction scans and allergen-intersection tests—that operate as policy layers before a recommendation is surfaced to the user. The governance track is the least technical but the most consequential for real deployment: versioned releases of the knowledge base with explicit provenance for every document, institutional-review-board approval protocols, patient-consent schemas that distinguish between diagnostic and recommendation-level consent, and clearly delineated liability boundaries are all prerequisites before the platform can be offered in any setting other than a supervised research prototype. A longitudinal outcome study—measuring whether grounded dietary and referral advice produces measurable changes in patient behaviour or in clinical endpoints such as blood-pressure trajectory or HbA_{1c}—is the evidentiary keystone of this track; until such evidence exists, OptiHealth should continue to be described as an assistive research system rather than as a clinical intervention.

The Conversational Health Agent has the most safety-sensitive research agenda of the four contributions, and the immediate priorities are therefore defensive rather than expansive. The first priority is agent-safety benchmarking: the LangGraph state machine should be exercised against published prompt-injection suites and indirect-tool-abuse test cases, with explicit refusal behaviour measured rather than merely observed, and the current MedEx-scraping and EHR tools should be placed behind a schema-validation layer

that rejects malformed or out-of-bounds arguments before they ever reach a downstream service. The second priority is calibration: the dual-model mental-health analyser must be evaluated against PHQ-9 and GAD-7 on a Bangladeshi cohort so that its screening output is interpretable on the population it is intended to serve, and it should be paired with a clinically-reviewed safety-referral protocol for high-risk activations rather than the current symptom-taxonomy labels alone. The third priority is human-in-the-loop design: every mutating tool invocation—appointment creation, referral dispatch, any future prescription-adjacent action—should pass through an explicit confirmation node in the LangGraph graph, and the audit log should be externally inspectable rather than application-internal. Medium-term work includes a deliberate expansion of the tool surface to cover laboratory-result interpretation, medication-adherence tracking, and telehealth bridging; domain-adaptive fine-tuning of the Gemma 4 E4B backbone on Bangla–English clinical dialogue so that the reasoning layer better reflects the linguistic and epidemiological reality of the target population; and an ongoing red-teaming protocol, ideally conducted quarterly and involving external reviewers, that stress-tests the agent's refusal behaviour under adversarial and ambiguous inputs rather than under the scripted dialogues used during this thesis.

Beyond the individual systems, three cross-cutting directions deserve sustained investment. The first is representation sharing across modalities: institutions that operate both radiography and MRI could benefit from multi-task models in which a shared backbone is jointly trained on thoracic detection and volumetric segmentation, with task-specific heads, and the question of whether such sharing produces positive transfer or destructive interference is itself a research question of substantial interest. The second is federated learning: the data-governance pressures that motivated our in-cluster LLM deployment apply equally to training, and federated or split-learning protocols offer a path by which multiple institutions can contribute to generalisation without centralising raw patient data. The third is open reporting: the present gap between what an AI-assisted radiology and recommendation system claims and what it has actually been evaluated for is in part a reporting-standards gap, and the adoption of reporting frameworks such as CLAIM for imaging AI, TRIPOD+AI for predictive models, and MINIMAR for model-information disclosure would make fair cross-institution comparison feasible in a way that the current publication culture does not.

In closing, this thesis contributes concrete benchmarks, reusable architectural ideas, and a single integrated prototype toward a more grounded conception of artificial intelligence in healthcare—one in which performance on images is paired with transparency in how advice is produced, in which every recommendation is traceable to its evidence, and in which limitations are stated as plainly as successes. The research programme outlined above is deliberately weighted toward external validation, safety-oriented engineering, and governance rather than toward architectural novelty, because we take the

view that the decisive obstacle to clinically useful medical AI is no longer model capability but the evidence, infrastructure, and accountability required to deploy that capability responsibly.

Bibliography

[1] Farman Ali, Shaker El-Sappagh, SM Riazul Islam, Daehan Kwak, Amjad Ali, Muhammad Imran, and Kyung-Sup Kwak. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Information Fusion*, 63:208–222, 2020.

[2] Afzal Hossain Alif, Mojtoba Zaman Mantaka, Aiman Saad Hamid, and Saadia Binte Alam. Efficient multi-class thoracic disease detection on vindr-cxr using yolov12 with class-balanced training and anatomical tta. In *2026 IEEE 2nd International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)*. IEEE, 2026.

[3] Imane Allaouzi and Mohamed Ben Ahmed. A novel approach for multi-label chest x-ray classification of common thorax diseases. *IEEE Access*, 7:64279–64288, 2019.

[4] Nouf Abdullah Almujally, Turki Aljrees, Oumaima Saidani, et al. Monitoring acute heart failure patients using internet-of-things-based smart monitoring system. *Sensors*, 23(10):4580, 2023.

[5] Navya Alugubelli, Hussam Abuissa, and Attila Roka. Wearable devices for remote monitoring of heart rate and heart rate variability—what we know and what is coming. *Sensors*, 22(22):8903, 2022.

[6] Pei An, Yucong Duan, Yuliang Huang, Jie Ma, Yanfei Chen, Liheng Wang, You Yang, and Qiong Liu. Sp-det: Leveraging saliency prediction for voxel-based 3d object detection in sparse point cloud. *IEEE Transactions on Multimedia*, 26:2795–2808, 2023.

[7] Ujjwal Baid, Sagar Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

[8] Shakila Basheer, Ala Saleh Alluhaidan, and Maryam Aysha Bivi. Real-time monitoring system for early prediction of heart disease using internet of things. *Soft Computing*, 25(18):12145–12158, 2021.

[9] Tarek Bekfani, Marat Fudim, John GF Cleland, et al. A current and future outlook on upcoming technologies in remote monitoring of patients with heart failure. *European Journal of Heart Failure*, 23(1):175–185, 2021.

[10] Beatrice Bonato, Loris Nanni, and Alessandra Bertoldo. Advancing precision: A comprehensive review of mri segmentation datasets from brats challenges (2012-2025). *Sensors*, 25(6):1838, 2025.

[11] Ralf Bousseljot, Dieter Kreiseler, and A. Schnabel. The ptb diagnostic ecg database. In *Computers in Cardiology 1995*, pages 221–224. IEEE, 1995.

[12] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.

[13] M. Jorge Cardoso, Wenqi Li, Richard Brown, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

[14] Harrison Chase. LangChain: Building applications with LLMs through composability. `https://github.com/langchain-ai/langchain`, 2022. Open-source software.

[15] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[16] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[17] Forum Desai, Deepraj Chowdhury, Rupinder Kaur, Marloes Peeters, Rajesh Chand Arya, Gurpreet Singh Wander, Sukhpal Singh Gill, and Rajkumar Buyya. Healthcloud: A system for monitoring health status of heart patients using machine learning and cloud computing. *Internet of Things*, 17:100485, 2022.

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021.

[19] WenZe Fan, Xingchen Guo, Lei Teng, and YuXuan Wu. Research on abnormal target detection method in chest radiograph based on yolo v5 algorithm. In *2021 IEEE International conference on computer science, electronic information engineering and intelligent control technology (CEI)*, pages 125–128. IEEE, 2021.

[20] M Ganesan and N Sivakumar. Iot based heart disease prediction and diagnosis model for healthcare using machine learning models. In *2019 IEEE international conference on system, computation, automation and networking (ICSCAN)*, pages 1–5. IEEE, 2019.

[21] Gemma Team, Google DeepMind. Gemma: Open models based on Gemini research and technology. Technical report, Google DeepMind, 2024. Technical report, `https://ai.google.dev/gemma`.

[22] Gemma Team, Google DeepMind. Gemma 4: Frontier-scale open models with hybrid attention and per-layer embeddings. `https://ai.google.dev/gemma`, 2026. Model card and technical documentation. Covers the E2B, E4B, 26B A4B, and 31B variants; the OptiHealth Health Agent is built on the E4B dense variant (4.5B effective parameters, 128K context, text–image–audio modalities).

[23] David C Goff Jr, Lawrence Brass, Lynne T Braun, et al. Essential features of a surveillance system to support the prevention and management of heart disease and stroke. *Circulation*, 115(1):127–155, 2007.

[24] Li Guan, Ruting Zhang, and Yi Zhao. Yolov11-mff: A multi-scale frequency-adaptive fusion network for enhanced cxr anomaly detection. *PLoS One*, 20(10):e0334283, 2025.

[25] Aanshi Gupta, Shubham Yadav, Shaik Shahid, et al. Heartcare: Iot based heart disease prediction system. In *2019 International Conference on Information Technology (ICIT)*, pages 88–93. IEEE, 2019.

[26] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019.

[27] Shengnan Hao, Xinlei Li, Wei Peng, Zhu Fan, Zhanlin Ji, and Ivan Ganchev. Yolo-cxr: A novel detection network for locating multiple small lesions in chest x-ray images. *IEEE Access*, 12:156003–156019, 2024. doi: 10.1109/ACCESS.2024.3482102.

[28] Shengnan Hao, Xinlei Li, Wei Peng, Zhu Fan, Zhanlin Ji, and Ivan Ganchev. Yolo-cxr: A novel detection network for locating multiple small lesions in chest x-ray images. *IEEE Access*, 2024.

[29] Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87, 2023.

[30] Md Mahbub Hasan, Md Fahad Hasan, Imteaz Rahaman, Rafiqul Alam, Mehedi Hassan, and Mim Naz Rahman. Heart monitoring management system embedded with internet of things (iot) cloud processing. In *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pages 1–6. IEEE, 2020.

[31] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R. Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In Alessandro Crimi and Spyridon Bakas, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 272–284. Springer International Publishing, 2022.

[32] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. *arXiv preprint arXiv:2201.01266*, 2022.

[33] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584, 2022.

[34] Praveen Indraratna, Daniel Tardo, Jennifer Yu, et al. Mobile phone technologies in the management of ischemic heart disease, heart failure, and hypertension: systematic review and meta-analysis. *JMIR mHealth and uHealth*, 8(7):e16695, 2020.

[35] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[36] Yufei Jin, Huijuan Lu, Wenjie Zhu, and Wanli Huo. Deep learning based classification of multi-label chest x-ray images via dual-weighted metric loss. *Computers in biology and medicine*, 157:106683, 2023.

[37] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolo, 2023. Version 8.0.0. `https://github.com/ultralytics/ultralytics`.

[38] Priyanka Kakria, NK Tripathi, and Peerapong Kitipawang. A real-time health monitoring system for remote cardiac patients using smartphone and wearable sensors. *International Journal of Telemedicine and Applications*, 2015(1):373474, 2015.

[39] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.

[40] Z Faizal Khan and Sultan Refa Alotaibi. Applications of artificial intelligence and big data analytics in m-health: A healthcare system perspective. *Journal of Healthcare Engineering*, 2020:8894694, 2020.

[41] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024.

[42] Sofoklis Kyriazakos, Ramjee Prasad, Albena Mihovska, et al. ewall: An open-source cloud-based ehealth platform for creating home caring environments for older adults living with chronic diseases or frailty. *Wireless Personal Communications*, 97:1835–1875, 2017.

[43] LangChain AI. LangGraph: Stateful, multi-actor applications with LLMs. `https://github.com/langchain-ai/langgraph`, 2024. Open-source software.

[44] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7871–7880, 2020.

[45] Hongwei Bran Li, Gian Marco Conte, Qingqiao Hu, Syed Muhammad Anwar, Florian Kofler, Ivan Ezhov, Koen Van Leemput, Marie Piraud, Maria Diaz, Byrone Cole, et al. The brain tumor segmentation (brats) challenge 2023: Brain mr image synthesis for tumor segmentation (brasyn). *arXiv preprint arXiv:2305.09011*, 2024.

[46] Yunchao Li, Yutao Chen, Naiyan Wang, and Zhaoxiang Zhang. Scale-aware Trident Networks for Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6054–6063, 2019.

[47] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[48] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[49] Grant E MacKinnon and Evan L Brittain. Mobile health technologies in cardiopulmonary disease. *Chest*, 157(3):654–664, 2020.

[50] Paolo Melillo, Ada Orrico, Paolo Scala, Filippo Crispino, and Leandro Pecchia. Cloud-based smart health monitoring system for automatic cardiovascular and fall risk assessment in hypertensive patients. *Journal of medical systems*, 39:1–7, 2015.

[51] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

[52] Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. *International MICCAI Brainlesion Workshop*, pages 311–320, 2018.

[53] Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429, 2022.

[54] Jorge Novo, A Hermida, Marcos Ortega, Noelia Barreira, Manuel G Penedo, JE López, and C Calvo. Hydra: A web-based system for cardiovascular analysis, diagnosis and treatment. *Computer methods and programs in biomedicine*, 139:61–81, 2017.

[55] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[56] Antonio Iyda Paganelli, Abel González Mondéjar, Abner Cardoso da Silva, et al. Real-time data analysis in health monitoring systems: A comprehensive systematic literature review. *Journal of Biomedical Informatics*, 127:104009, 2022.

[57] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[58] S. N. Rao. Yolov11 architecture explained: Next-level object detection with enhanced speed and accuracy, 2024. Medium. `https://tinyurl.com/2nkkfd3n`. [Online; accessed: 20-07-2025].

[59] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. doi: 10.1109/CVPR.2016.91.

[60] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[61] Marcus AG Santos, Roberto Munoz, Rodrigo Olivares, et al. Online heart monitoring systems on the internet of health things environments: A survey, a reference model and an outlook. *Information Fusion*, 53:222–239, 2020.

[62] Simanta Shekhar Sarmah. An efficient iot-based patient monitoring and heart disease prediction system using deep learning modified neural network. *IEEE Access*, 8: 135784–135797, 2020.

[63] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[64] Tamanna Shaown, Imam Hasan, Md Muradur Rahman Mim, and Md Shohrab Hossain. Iot-based portable ecg monitoring system for smart healthcare. In *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, pages 1–5. IEEE, 2019.

[65] Archana Singh and Rakesh Kumar. Heart disease prediction using machine learning algorithms. In *2020 International Conference on Electrical and Electronics Engineering (ICE3)*, pages 452–457, 2020. doi: 10.1109/ICE348803.2020.9122958.

[66] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[67] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

[68] Joan B Soriano, Parkes J Kendrick, Katherine R Paulson, et al. Prevalence and attributable health burden of chronic respiratory diseases, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet Respiratory Medicine*, 8(6):585–596, 2020.

[69] BJ Stevens, L Skermer, and J Davies. Radiographers reporting chest x-ray images: identifying the service enablers and challenges in england, uk. *Radiography*, 27(4): 1006–1013, 2021.

[70] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Li Yuan, Changhu Wang, and Ping Luo. 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)*, pages 14454–14463, 2021.

[71] Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. MedAgents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*, 2023.

[72] Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025.

[73] Muhammad Umer, Turki Aljrees, Hanen Karamti, Abid Ishaq, Shtwai Alsubai, Marwan Omar, Ali Kashif Bashir, and Imran Ashraf. Heart failure patients monitoring using iot-based remote monitoring system. *Scientific Reports*, 13(1):19213, 2023.

[74] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. YOLOv9: Learning what you want to learn using programmable gradient information. *arXiv preprint arXiv:2402.13616*, 2024.

[75] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.

[76] Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. Transbts: Multimodal brain tumor segmentation using transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 109–119. Springer, 2021.

[77] Yi Wang, Zhiqiang Deng, Xiaowei Hu, Lei Zhu, Xin Yang, Xuemiao Xu, Pheng-Ann Heng, and Dong Ni. Deep attentive features for prostate segmentation in 3d transrectal ultrasound. *IEEE Transactions on Medical Imaging*, 38(12):2768–2778, 2019.

[78] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[79] World Health Organization. The top 10 causes of death, May 2024. URL `https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death`. Accessed: 2025-09-30.

[80] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023.

[81] Ningning Xiao, Wei Yu, and Xu Han. Wearable heart rate monitoring intelligent sports bracelet based on internet of things. *Measurement*, 164:108102, 2020.

[82] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for automatic medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 171–180. Springer, 2021.

[83] Qiang Xu and Weili Duan. DualAttNet: Synergistic fusion of image-level and fine-grained disease attention for multi-label lesion detection in chest x-rays. *Computers in Biology and Medicine*, 168:107742, 2024. doi: 10.1016/j.compbiomed.2023.107742.

[84] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[85] Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, 2019.

[86] Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R. Dalal, Jennifer L. Kim, Michael Moor, Kevin Alexander, Euan Ashley, Jack Boyd, Kathleen Boyd, et al. Almanac — retrieval-augmented language models for clinical medicine. *NEJM AI*, 1(2), 2024.

[87] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11, 2018.

A Supplementary Results for CXR-Adaptive

Table A.1: Full per-class AP@0.5 comparison: YOLOv11-MFF vs CXR-Adaptive (CXR-TTA)

Class	**YOLOv11-MFF**	**Ours (CXR-TTA)**
Aortic Enlargement	0.924	0.598
Atelectasis	0.454	0.266
Calcification	0.251	0.223
Cardiomegaly	0.932	0.676
Nodule/Mass	0.339	0.286
Pleural Effusion	0.490	0.467
Pleural Thickening	0.227	0.198
Consolidation	0.315	0.541
ILD	0.201	0.611
Infiltration	0.340	0.505
Lung Opacity	0.297	0.394
Other Lesion	0.097	0.402
Pneumothorax	0.559	0.592
Pulmonary Fibrosis	0.390	0.428
mAP@0.5	0.415	**0.442**

B SS-UNETR Architecture Details

Table B.1: SS-UNETR full layer-by-layer output dimensions (base feature size F = 48)

Layer	**Output Resolution**	**Output Channels**
Encoder Path		
Shallow Conv (enc_0)	$H \times W \times D$	F
Patch Embed (h_0)	$H/2 \times W/2 \times D/2$	F
Swin Stage 1 (h_1)	$H/4 \times W/4 \times D/4$	2F
Swin Stage 2 (h_2)	$H/8 \times W/8 \times D/8$	4F
Swin Stage 3 (h_3)	$H/16 \times W/16 \times D/16$	8F
Swin Stage 4 (h_4)	$H/32 \times W/32 \times D/32$	16F
Bottleneck		
Fusion Output (B_{out})	$H/32 \times W/32 \times D/32$	32F
Decoder Path		
Decoder 5	$H/16 \times W/16 \times D/16$	16F
Decoder 4	$H/8 \times W/8 \times D/8$	8F
Decoder 3	$H/4 \times W/4 \times D/4$	4F
Decoder 2	$H/2 \times W/2 \times D/2$	2F
Decoder 1	$H \times W \times D$	F
Output Head	$H \times W \times D$	3