



الجمهورية الجزائرية الديمقراطية الشعبية  
People's Democratic Republic of Algeria  
وزارة التعليم العالي والبحث العلمي  
Ministry of Higher Education and Scientific Research



Constantine 1 Frères Mentouri University  
Faculty of Natural and Life Sciences

جامعة قسنطينة 1 الإخوة منتوري  
كلية علوم الطبيعة والحياة

Department of Applied Biology

قسم البيولوجيا التطبيقية

Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Master

**Domain:** Natural and Life Sciences

**Field:** Biotechnology.

**Specialization:** Bioinformatics.

N° d'ordre:

N° de série:

Title:

---

# A Multimodal Framework for Explainable Chest X-ray Report Generation

---

Submitted by:

BOUGOURZI Nourhene.  
MAKHLOUF Raida Malek.  
TAIB Hadjer.

Sustained on: 25/06/2025

Board of Examiners:

**Chairperson :** Dr. I. R. AMINE KHODJA (MCB - Constantine 1 Frère Mentouri University).

**Supervisor:** Dr.H. CHEHILI (MCA - Constantine 1 Frère Mentouri University).

**Examiner :** Dr. D. Y. MEZIANI (MCB- Constantine 1 Frère Mentouri University).

Academic year

2024- 2025

## Acknowledgements

We would like to express our deepest gratitude to our thesis supervisor, ***Dr. CHEHILI Hamza***, whose mentorship went beyond the boundaries of traditional academic guidance. His valuable advice, patient encouragement, and unwavering availability throughout countless discussions were the essential foundation of this work. His belief in our potential, even during the most challenging phases of this research, gave us the strength to persevere and achieve goals we once considered out of reach.

We extend our profound appreciation to ***Dr. TAIB Abderrahmane***, the board-certified pneumologist with 14 years of clinical experience, for his patience, expertise, and invaluable contribution to the clinical validation of our model. His meticulous evaluation of the 50 generated radiological impressions and his professional insights were instrumental in establishing the clinical acceptability and diagnostic accuracy of our AI system.

We would also like to express our sincere gratitude to the members of the jury, **Dr. A BENHAMDI**, who did us the great honor of agreeing to chair the defense committee, and **Dr. Y MEZIANI**, for kindly accepting to examine this modest work.

Our thanks also go to all the teachers and administrative staff who accompanied us during our academic journey.

## Dedication

To God, the source of all wisdom, strength, and light, for guiding me through every moment of doubt and difficulty.

To the soul of my beloved parents,

Dad **Abdessalem** & Mom **BENABDERRAHMANE Samira**

Who gave me everything and asked for nothing in return. Your love lives on in every breath I take, every dream I chase, and every value I hold dear. Though you cannot witness this achievement, I always carry your hearts with me always.

"وَقُلْ رَبِّ ارْحَمُهُمَا كَمَا رَبَّيْتَنِي صَغِيرًا"

To my beloved supportive brother **Mohamed Chakib**. My mastermind, to **Yousra**, not just a sister but a soulmate, and finally to the joy of my life, my partner in crime **Kenzi** I love you so much. Thank you for always standing by my side.

To my grandfather **Djedou**, my second father, who also left this world, your wisdom and tenderness live on in me.

To my dear grandmother **Mami**, whose presence is a blessing and a source of strength.

To all my family, especially my uncles **Mohamed Riad**, **Omar**, and **Hamza**, their wives & children My aunts **Nedjoua** and **Anissa**, your support has meant more than words can say.

To my partners in this academic journey **Hadjer**, my supportive therapist, and **Malak** the pure soul whose warmth and positivity uplift everyone, you are the greatest gifts this university has given me. I am truly fortunate to have crossed paths with 'you.

To my dear cousin **Soror**, my best friends **Imene**, **Khawla**, **Djihene**, **Lina** and **Chahi**, Thank you for always loving me, caring for me, and standing by my side. Your presence has meant more to me than words can express.

And finally, to **myself**, for holding on through the pain, the silence, and the storms; for walking this journey with strength, faith, and unwavering determination; for continuing to rise, to believe, and to grow, even when it was hardest.

الحمد لله

*Bougourzi Nourhene*

## Dedication

To my amazing mom, **BENSLIMANE Ahlem** the strongest woman I've ever known. You've been my support, my best friend, my safe place. You were always there, in every situation, standing behind me with love and strength. This is for you.

To my dad, **Abdelhak**, who showed me what hard work really means and taught me how to take responsibility. Your words and actions shaped me more than you know.

To my aunt, **BENSLIMANE Amal** my inspiration and my twin soul. You're the reason I never stop hoping. I see myself in you, and that makes me proud.

To my siblings, **Racha**, **Wassim**, and **Fares** thank you for your love, your jokes, your support, and for always being my home.

To my dear friends, **Sarah** and **Nihel** thank you for being by my side in every high and low. We cried together, laughed together, and walked this path together.

And to my incredible teammates and friends, **Nourhene** and **Hadjer** your pure hearts and genuine souls made this journey unforgettable. You made everything lighter and more beautiful.

To myself, **Malak** the warrior who held on to her principles through every storm.

For always trying to leave beauty behind.

For becoming the woman, daughter, sister, and friend your younger self dreamed of.

وَمَا تَوْفِيقِي إِلَّا بِاللَّهِ

*MAKHLouF Malak Raida*

## Dedication

In a world full of fathers mine chose to be my bestfriend, to my idol leaving too soon, the loss is immeasurable, but so is the love you left behind **Moussa**.

To the one who can take the place of all others, but whose place no one can take; to my mother **Yamina BENSALEM** who gave me strength without asking for thanks and loved me without limits.

To my paternal aunt, **Houria TAIB**, who has always spoiled me with her kindness.

To the one who has my back in highs and lows; my brother **Abderrahmane TAIB** and his lovely wife **Esma BOUHILA**.

To my second mom, the one who always felt like home; to my sister **Hania Taib** and her supportive husband **Ilyes BELBEKHOUCHE**.

To the most overprotective, teasing but yet thoughtful brother **Aymene Taib** and his lovely wife **Rayene LEZZAR**.

You are the sweetest reminder that love can be tiny, loud, messy and perfect all at once, to my niece **Naya Yasmine TAIB**.

To the heartbeat I didn't know I was missing and now I can't imagine life without, to my nephew **Anis Belbekhouche**.

To the constant in every chapter of my life, the laughter in my joy, the bandage in my pain and the voice who always cheered for me; to my bestfriend **Nesrine BOUTOUHA**.

We've collected years like seashells, some chipped, some perfect but all precious. To my forever girlies **Mélissa YAICHE**, **Tayssir DAHBI**, **Inesse BENHAMOUDA**.

To the one I found in the mess of growing up while surviving life, to my unpaid therapist **Nourhene BOUGOURZI**.

Comradeship is a sacred bond forged in the crucible of purpose, tested by hardship and sealed by loyalty; to **Malak MAKHLOUF**.

To Rodolphe **Lindt**, who revolutionized the world's sweetest pleasure, you didn't just satisfy cravings, you awakened joy, evoked memory and wrapped love in foil.

To my all times favorite tv show's character, Chandler Bing, who kept me company in the loneliest nights of my life.

*TAIB Hadjer*

## **Abstract**

The growing demand for radiological services coupled with a shortage of expert radiologists has driven the development of automated report-generation systems. This thesis presents a novel explainable-AI framework for generating chest X-ray (CXR) reports by fusing visual and textual features. Using the Indiana University Chest X-ray (Open-I) dataset comprising 7,470 paired images and structured reports our approach employs a dual-branch convolutional ensemble (ResNet-50 and EfficientNet-B0) to extract complementary visual representations, alongside BERT-based embeddings of clinical “Indications” and “Findings.” These modalities are concatenated into a unified multimodal vector, which is fed into a fine-tuned Gemma-3 1B model using Low-Rank Adaptation (LoRA). Explainability is achieved via Grad-CAM heatmaps from both CNN backbones, highlighting anatomically relevant regions. Evaluated on a held-out validation subset ( $n = 300$ ), our method outperforms state-of-the-art baselines across BLEU-1 to BLEU-4, ROUGE, and METEOR metrics, while preserving clinical terminology with high precision. Qualitative analysis confirms that generated impressions align closely with radiological standards. This work demonstrates that integrating documented findings as input rather than output reduces hallucinations and enhances interpretability, paving the way for deployable, trustworthy AI-assisted radiology reporting.

### **Keywords:**

Explainable AI, Multimodal Learning, Chest X-Ray Report Generation, generative model.

## Résumé

La demande croissante en services radiologiques, conjuguée à une pénurie de radiologues experts, a motivé le développement de systèmes automatisés de génération de rapports. Cette thèse propose un cadre d'IA explicable pour la rédaction de rapports de radiographie thoracique (CXR) en fusionnant caractéristiques visuelles et textuelles. À l'aide du jeu de données Indiana University Chest X-ray (Open-I) comprenant 7 470 paires image et rapport structuré, notre démarche utilise un ensemble convolutionnel dual (ResNet-50 et EfficientNet-B0) pour extraire des représentations visuelles complémentaires, ainsi que des embeddings BERT des « Indications » et des « Conclusions » cliniques. Ces modalités sont concaténées en un vecteur multimodal unifié, envoyé dans un modèle Gemma-3 1B affiné via Low-Rank Adaptation (LoRA). L'explicabilité est assurée par des cartes de chaleur Grad-CAM issues des deux réseaux CNN, mettant en évidence les régions anatomiquement pertinentes. Évaluée sur un sous-ensemble de validation ( $n = 300$ ), notre méthode surpasse l'état de l'art sur les métriques BLEU-1 à BLEU-4, ROUGE et METEOR, tout en préservant la terminologie clinique avec une grande précision. L'analyse qualitative confirme que les impressions générées respectent les normes radiologiques. Ce travail montre qu'intégrer les constats cliniques en entrée plutôt qu'en sortie réduit les hallucinations et améliore l'interprétabilité, ouvrant la voie à des rapports radiologiques assistés par IA fiables et déployables.

### Mots clés:

IA explicable Apprentissage multimodal, Génération de rapports de radiographie thoracique, modèle générative

## ملخص

يزداد الضغط على خدمات الأشعة نتيجة ارتفاع أعداد الفحوصات مع نقص الأطباء المتخصصين في تفسير أشعة الصدر، مما يدفع إلى اعتماد تقنيات الذكاء الاصطناعي للمساعدة. تقدم هذه الرسالة إطار عمل يمكن تفسيره لإنشاء تقارير أشعة الصدر عبر دمج الميزات البصرية والنصية. نستخدم مجموعة بيانات جامعة إنديانا لأشعة الصدر (Open-I) التي تحتوي على 7,470 صورة مع تقرير مصاحب، حيث يستخرج نموذج إنسامبل مزدوج (ResNet-50 و EfficientNet-B0) التمثيلات البصرية، بينما تنتج طبقة BERT تضمينات نصية من قسمي "الدواعي" و "النتائج" السريرية. تُدمج هذه التمثيلات في متجه موحد يُمرر إلى نموذج Gemma-3 1B معدل باستخدام تقنية LoRA التكيف منخفض الرتبة. (ولتحقيق قابلية التفسير، نستخدم-Grad CAM لإظهار الخرائط الحرارية لكل نموذج تلافيفي، مما يسلط الضوء على المناطق الصدرية ذات الأهمية التشخيصية. عند تقييم المنهجية على 300 حالة تحقق، تجاوزت النتائج الأساليب الحالية في مقاييس BLEU-1 إلى BLEU-4 و ROUGE و METEOR مع دقة عالية في المصطلحات الطبية. وتوضح التحليلات النوعية توافق الانطباعات المؤلدة مع المعايير السريرية. يوضح هذا العمل أن استخدام المعلومات السريرية كنقطة انطلاق بدلاً من إخراج يقلل من الأخطاء اللغوية ويعزز الشفافية، مما يهيئ طريقاً لنشر تقارير أشعة موثوقة بدعم الذكاء الاصطناعي.

### الكلمات المفتاحية:

الذكاء الاصطناعي القابل للتفسير، التعلم متعدد الوسائط، توليد تقارير أشعة الصدر، التكيف منخفض الرتبة.



**List of Tables**

**TABLE 1:** SUMMARY OF XAI METHODS ..... 15

**TABLE 2:** KEY MEDICAL IMAGE CAPTIONING DATASETS ..... 16

**TABLE 5:** BASIC INFORMATION ON THE HIGH PROCESSING CENTER (HPC) USED FOR DATA  
PREPROCESSING AND MODEL TRAINING. .... 23

**TABLE 6:** TECHNICAL SPECIFICATIONS OF THE COMPUTING STATION USED FOR DATA  
PROCESSING AND MODEL TRAINING ..... 23

**TABLE 7:** THE VERSIONS OF THE USED PACKAGES ..... 25

**TABLE 8:** CORE ARCHITECTURAL PARAMETERS OF GEMMA-3 1B ..... 32

**TABLE 9:** FINE-TUNING CONFIGURATION FOR GEMMA-3 1B WITH LoRA ..... 33

**TABLE 10:** EVALUATION METRICS COMPARISON BEFORE AND AFTER FINE-TUNING ..... 47

**TABLE 11:** EXPERT RADIOLOGIST EVALUATION RESULTS - CLINICAL ACCURACY ASSESSMENT  
(N=50) ..... 50

**TABLE 12:** COMPARATIVE PERFORMANCE ANALYSIS ON IU X-RAY DATASET ..... 52

## List of Figures

<b>FIGURE 1 :</b> ARTIFICIAL INTELLIGENCE’S SUBSETS OF MACHINE LEARNING AND DEEP LEARNING (TIWARI ET AL., 2018).....	8
<b>FIGURE 2 :</b> A BROADER OVERVIEW OF LLMs, DIVIDING LLMs INTO SEVEN MAIN BRANCHES: PRE-TRAINING, FINE-TUNING, EFFICIENT METHODS, INFERENCE, EVALUATION, APPLICATIONS, AND CHALLENGES (NAVEED ET AL., 2024).....	11
<b>FIGURE 3 :</b> PERFORMANCE COMPARISON OF GEMMA-3 MODELS AGAINST STATE-OF-THE-ART LANGUAGE MODELS ON BENCHMARK EVALUATIONS (GOOGLE DEEPMIND, 2025) .....	21
<b>FIGURE 4 :</b> EXAMPLE OF A CHEST X-RAY IMAGE AND CORRESPONDING STRUCTURED RADIOLOGY REPORT FROM THE INDIANA UNIVERSITY CHEST X-RAY COLLECTION .....	22
<b>FIGURE 5 :</b> MODEL PIPELINE ARCHITECTURE.....	26
<b>FIGURE 6 :</b> DATASET PARTITIONING INTO TRAINING AND VALIDATION SETS .....	27
<b>FIGURE 7 :</b> BAR CHART SHOWING THE DISTRIBUTION OF THE MOST FREQUENTLY DIAGNOSED PATHOLOGIES IN THE OPEN-I DATASET .....	28
<b>FIGURE 8 :</b> GENERAL ARCHITECTURE OF A CNN, SHOWING THE PROGRESSION FROM RAW IMAGE INPUT TO FEATURE EXTRACTION AND FINAL OUTPUT THROUGH CONVOLUTIONAL, POOLING, AND FULLY CONNECTED LAYERS (SUGANYADEVI ET AL., 2021).....	30
<b>FIGURE 9 :</b> GRAD-CAM VISUALIZATION FOR RESNET50 AND EFFICIENTNETB0 .....	34
<b>FIGURE 10:</b> HOME PAGE INTERFACE OF MEDVIS. TECH WEB APPLICATION .....	36
<b>FIGURE 11 :</b> MULTI-INPUT CLINICAL DATA ENTRY INTERFACE - COMPREHENSIVE CLINICAL INPUT PORTAL FOR AI ANALYSIS .....	38
<b>FIGURE 12 :</b> AI RADIOLOGIST REPORT INTERFACE - AI-GENERATED IMPRESSION ANALYSIS DASHBOARD.....	39
<b>FIGURE 13:</b> EXPLAINABLE AI VISUALIZATION INTERFACE - MULTI-MODAL GRADCAM ATTENTION ANALYSIS SYSTEM.....	40
<b>FIGURE 14:</b> GENERAL COMPARISON CHART SHOWING BLEU-1 (0.437), ROUGE-L-F (0.549), BERTSCORE-F1 .....	45
<b>FIGURE 15 :</b> BERT-SCORE BREAKDOWN SHOWING [ PRECISION (0.917), RECALL (0.919), F1-SCORE (0.918)] .....	46
<b>FIGURE 16 :</b> BLEU METRICS PROGRESSION FROM BLEU-1 (0.437) TO BLEU-4 (0.279) .....	46

**FIGURE 17 : A REPRESENTATIVE EXAMPLE OF IMPRESSION GENERATION QUALITY**

IMPROVEMENT THROUGH LORA FINE-TUNING ..... 48

**FIGURE 18 : MULTI-ARCHITECTURAL GRAD-CAM ANALYSIS FOR IMAGE CXR3688\_IM-1839-**

0001-0001.PNG ..... 49

## **ACRONYMS**

**ACR:** American College of Radiology

**AI:** Artificial Intelligence.

**AP:** Anteroposterior

**API:** Application Programming Interface

**ARDS:** Acute Respiratory Distress Syndrome

**BERT:** Bidirectional Encoder Representations from Transformers

**BLEU:** Bilingual Evaluation Understudy

**CheXpert:** Chest X-ray Expert

**CNN:** Convolutional Neural Network

**CPU:** Central Processing Unit

**CSV:** Comma-Separated Values.

**CT:** Computed Tomography

**CXR:** Chest X-Ray

**DL:** Deep Learning.

**DRF:** Django REST Framework

**EfficientNet:** Efficient Neural Network

**EHR:** RElectronic Health Record

**F1:** F1-Score.

**GPU:** Graphics Processing Unit

**Grad-CAM:** Gradient-weighted Class Activation Mapping

**HPC:** High Performance Computing

**IT:** Instruction-Tuned (referring to Gemma-3 1B-IT)

**IU:** Indiana University

**JSON:** JavaScript Object Notation

**LIME:** Local Interpretable Model-agnostic Explanations

**LLMs:** Large Language Models.

**LoRA:** Low-Rank Adaptation

**METOR:** Metric for Evaluation of Translation with Explicit ORderin

**MIMIC-CXR:** Medical Information Mart for Intensive Care - Chest X-Ray

**ML:** Machine Learning.

**MLP:** Multi-Layer Perceptron (MLP).

**MRI:** Magnetic Resonance Imaging

**MVC:** Model-View-Controller

**NLP:** Natural Language Processing.

**NLTK:** Natural Language Toolkit

**NumPy:** Numerical Python

**OpenCV:** Open Source Computer Vision Library

**PA:** Posteroanterior

**PEFT:** Parameter-Efficient Fine-Tuning.

**PIL:** Python Imaging Library

**PyTorch:** Python-based deep learning framework

**RAM:** Random Access Memory

**ResNet:** Residual Network

**RGB:** Red Green Blue

**ROCO:** Radiology Objects in Context

**ROUGE:** Recall-Oriented Understudy for Gisting Evaluation

**SHAP:** SHapley Additive exPlanations

**URL:** Uniform Resource Locator

**XAI:** Explainable Artificial Intelligence



## Table of Contents

**Acknowledgements**

**Dedication**

**Abstract**

**Résumé**

**الملخص**

**List of Tables**

**List of Figures**

**ACRONYMS**

General Introduction .....	2
CHAPTER 1: Bibliographic Synthesis .....	1
Introduction.....	6
1. Medical imaging and radiology context .....	6
1.1 Chest X-Ray imaging in clinical practice .....	6
1.2 Radiological report structure and standards .....	6
1.3 Current challenges in radiology reporting .....	7
1.4 The need for automated report generation.....	7
2. Artificial Intelligence (AI) .....	7
3. Deep learning in medical imaging .....	8
3.1 Convolutional Neural Networks for medical image analysis .....	8
3.2 Ensemble learning and pre-trained models .....	9
3.2.1 ResNet architecture and applications.....	9
3.3.2 EfficientNet for medical imaging .....	9
4. Natural Language Processing in healthcare .....	9

4.1	BERT and Transformer models in medical NLP .....	10
4.2	Large Language Models for medical text generation .....	10
5.	Evaluation metrics for medical text generation .....	12
6.	Multimodal learning approaches in healthcare .....	12
7.	Explainable Artificial Intelligence (XAI) in medical imaging .....	13
7.1	Terminology .....	13
7.2	The need for XAI, motivations .....	13
7.3	Common explainability methods.....	14
7.4	Real-world applications and case studies .....	15
8.	Rated works .....	16
	Conclusion . .....	18
CHAPTER 2: Materials and Methodology .....		19
Introduction.....		20
1.	Materials .....	20
1.1	Large Language Model Gemma-3 .....	20
1.2	Dataset description and characteristics .....	21
1.3	High Performance Computing (HPC).....	22
1.4	Computing station (Desktop) .....	23
1.5	Python .....	24
1.6	Packages.....	24
1.6.1	Core deep learning & transformers.....	24
1.6.2	Computer vision & feature extraction.....	24
1.6.3	Natural Language Processing .....	24
1.6.4	Data handling & utilities.....	24
1.6.5	Visualization tools .....	24
1.7	Web framework .....	24
2.	Methodology .....	25



2.1 Overall pipeline architecture .....	25
2.2 Data preparation and preprocessing .....	26
2.2.1 Data quality assessment and filtering .....	26
2.2.2 Dataset partitioning .....	26
2.2.3 Data exploration and visualization .....	27
2.2.4 Image preprocessing pipeline .....	28
2.2.5 Text preprocessing and tokenization .....	29
2.3 Visual feature extraction module .....	29
2.3.1 ResNet50 implementation and configuration .....	30
2.3.2 EfficientNetB0 integration .....	30
2.3.3 Ensemble learning implementation .....	30
2.3.4 Feature map processing and dimensionality reduction .....	31
2.4 Textual feature extraction .....	31
2.4.1 Report parsing & cleaning .....	31
2.4.2 Tokenization & embedding extraction .....	31
2.4.3 Embedding normalization & concatenation .....	31
2.4.4 Text feature output .....	31
2.5 Ensemble fusion .....	32
2.6 Language model fine-tuning .....	32
2.6.1 LoRA implementation and configuration .....	32
2.6.2 Training strategy and hyperparameters .....	33
2.6.3 Report generation process .....	33
2.7 Explainability integration .....	34
2.7.1 Heatmap generation and processing .....	34
2.7.2 Comparative activation analysis .....	35
2.8 System integration and deployment .....	35
2.8.1 Django web application development .....	36

2.8.2 Model serving and lazy loading .....	37
Conclusion .....	41
Chapter 3: Results and discussion .....	38
Introduction.....	43
1.Results.....	43
1.1Model performance evaluation .....	43
1.1.1 Detailed metric analysis.....	44
1.1.2 Comprehensive performance visualization.....	44
1.2    Fine-tuning impact: comparative analysis .....	47
1.2.1 Quantitative results - performance metrics transformation .....	47
1.2.2 Qualitative analysis of impression generation .....	47
1.3 Interpretability analysis: Grad-CAM and visual attention model.....	49
1.4 Clinical validation: expert radiologist evaluation .....	50
1.4.1 Expert assessment methodology .....	50
1.4.2 Clinical accuracy results .....	50
2. Discussion.....	51
2.1 Architectural innovation: findings-as-input approach .....	51
2.2 Comparative analysis: state-of-the-art performance benchmarking .....	51
2.3Clinical acceptability analysis .....	53
General conclusion .....	55
Reference .....	57
Appendices.....	68
A Multimodal Framework for Explainable Chest X-ray Report Generation .....	82



# **GENERAL INTRODUCTION**

## General Introduction

In recent years, medical imaging has become increasingly central to the diagnosis and management of a wide range of health conditions. Among all imaging modalities, chest X-rays (CXR) are the most widely performed due to their accessibility, speed, and relatively low cost. They are routinely used to detect infections, tumors, cardiovascular issues, and other thoracic abnormalities. However, reading and interpreting CXRs is a highly specialized task. It requires both deep clinical experience and precise pattern recognition. Unfortunately, the number of qualified radiologists is not keeping pace with the growing demand for imaging services, resulting in delays, missed diagnoses, and overburdened healthcare systems.

In this context, artificial intelligence (AI) has emerged as a promising tool to assist radiologists by automating certain aspects of the diagnostic process. One area where AI can play a transformative role is the generation of radiology reports. These reports are the final and most critical step in the imaging workflow, as they translate visual data into clinical decisions. Automatic report generation aims to produce meaningful textual descriptions of imaging findings that can be used directly by healthcare professionals. But doing so is far from simple: it requires combining accurate visual recognition with domain-specific language generation.

Unlike typical image captioning tasks where a short and generic sentence is enough radiology report generation is far more complex. It involves producing structured, detailed, and context-aware narratives. This task must reflect clinical terminology, handle uncertainties, and support differential diagnoses. It must also avoid common AI pitfalls such as hallucinations (fabricated information) or overly generic conclusions. As a result, building such systems demands a fusion of computer vision, natural language processing (NLP), and medical knowledge.

This thesis presents an explainable, multimodal AI framework that addresses these challenges. Our approach leverages a combination of convolutional neural networks (CNNs) for image analysis, transformer-based language models for text generation, and specialized adaptation techniques to fine-tune performance. We incorporate two key components: GEMMA (Gradient Episodic

Memory Model Adaptation), which enables continual learning, and LoRA (Low-Rank Adaptation), which allows efficient fine-tuning with minimal computational overhead. These components are integrated into a pipeline that processes both image and textual data (such as clinical indications) to generate coherent and clinically sound CXR reports.

To ensure transparency and trustworthiness, our system is designed with explainability in mind. Visual explainability is implemented using Grad-CAM, a method that generates heatmaps to highlight the parts of the X-ray image that influenced the model's decision. This helps clinicians verify and understand the AI's reasoning process. Additionally, we use expert-guided transformer modules to simulate the collaborative diagnostic reasoning found in clinical practice. This contributes to richer and more nuanced report generation, emulating how multiple radiologists might approach a complex case.

Our experiments were conducted using the publicly available Indiana University Chest X-ray (Open-I) dataset, which includes thousands of paired images and reports. We evaluate our system on multiple standard metrics such as BLEU, ROUGE and METEOR. The results show that our model performs competitively with state-of-the-art systems, while offering better interpretability and alignment with clinical expectations. We also conduct qualitative evaluations to assess how closely the generated reports match human-written ones in terms of style, clarity, and accuracy.

By bridging the gap between image understanding and medical language generation, our work demonstrates the feasibility of developing AI systems that can assist in radiological workflows. The goal is not to replace radiologists, but to support them with tools that are fast, reliable, and capable of reducing cognitive burden especially in settings where expertise is scarce. Ultimately, such systems can contribute to improved patient care by enabling faster diagnosis, more consistent reporting, and better resource allocation.

This thesis is organized into three main chapters:

- Chapter 1: Bibliographic Synthesis, This chapter provides a comprehensive overview of the medical imaging field, particularly chest X-ray interpretation, the structure and standards of radiology reports, and current challenges in diagnostic radiology. It also reviews related work in AI-based report generation and outlines the theoretical background for deep learning, NLP, and explainable AI.
- Chapter 2: Materials and Methods, here we describe the datasets used, the technical specifications of the computing infrastructure, and the software libraries employed. We explain how data was preprocessed, how the multimodal model architecture was constructed, and how the model was trained and fine-tuned. Special emphasis is placed on the integration of explainability mechanisms such as Grad-CAM and attention layers.
- Chapter 3: Results and Discussion, in this chapter, we present the performance results of the system across multiple evaluation tasks. We compare different model configurations, analyze the outputs in terms of explainability and clinical accuracy, and discuss the system's strengths and limitations. We also highlight areas for future improvement and potential real-world applications of this work.

Through these three chapters, we aim to show how a thoughtfully designed AI system can contribute to more interpretable, efficient, and scalable radiology workflows laying the groundwork for future clinical integration of AI in medical imaging.

# **CHAPTER 1: Bibliographic** **Synthesis**



## Introduction

In this chapter, we explore the pivotal role of chest X-rays in clinical practice, detailing their diagnostic value, reporting standards, and the persistent challenges radiologists face. We then introduce how artificial intelligence particularly deep learning and natural language processing has begun to reshape medical imaging workflows. By examining key architectures, evaluation metrics, and explainability methods, we lay the groundwork for our explainable, multimodal report-generation framework.

## 1. Medical imaging and radiology context

### 1.1 Chest X-Ray imaging in clinical practice

Chest X-rays (CXRs) are indispensable in diagnosing thoracic conditions due to their accessibility, speed, and low cost. Standard imaging involves posteroanterior (PA) and lateral views, offering complementary insights. Interpreting these images demands skill due to anatomical overlap and projection limitations. Tools like the ABCDEFGHI mnemonic guide structured evaluations. Common findings include consolidations, effusions, and device misplacements. Proper identification is vital for patient safety. Limitations like poor positioning or exposure can lead to diagnostic errors. Misinterpretations, especially among trainees, highlight the need for structured training and decision support tools (Radiopaedia, 2024; Gianella *et al.*, 2023).

### 1.2 Radiological report structure and standards

Radiological reports translate imaging into actionable medical insights. Key components include the clinical referral, technique, comparison, findings, and conclusion. These ensure clarity, diagnostic context, and thorough communication. Structured reporting has gained favor for its consistency and integration with electronic health records. It enhances clarity and supports secondary uses like AI training and research. Challenges include resistance due to perceived rigidity and initial workflow disruption. However, benefits such as reduced variability and better communication outweigh these. Future integration with AI and NLP will

likely standardize and enrich reporting further (*Langlotz, 2019; McCoubrie, 2011; DocPanel, 2025*).

### **1.3 Current challenges in radiology reporting**

Radiology faces mounting pressures from increasing workloads and imaging volumes. Burnout and fatigue heighten the risk of diagnostic errors. Reports often suffer from inconsistencies, lack of standardization, and unclear communication. Free-text formats hinder data mining and delay decision-making. Delays in turnaround time, driven by workflow inefficiencies and staffing shortages, impact care quality. AI shows promise but requires reliable datasets and clinical validation. Legal accountability further complicates timely and accurate communication. Solutions lie in improving workflows, integrating technology responsibly, and fostering collaboration between radiologists and clinicians (*Brady, 2016; Kahn et al., 2017; Arif, 2024*).

### **1.4 The need for automated report generation**

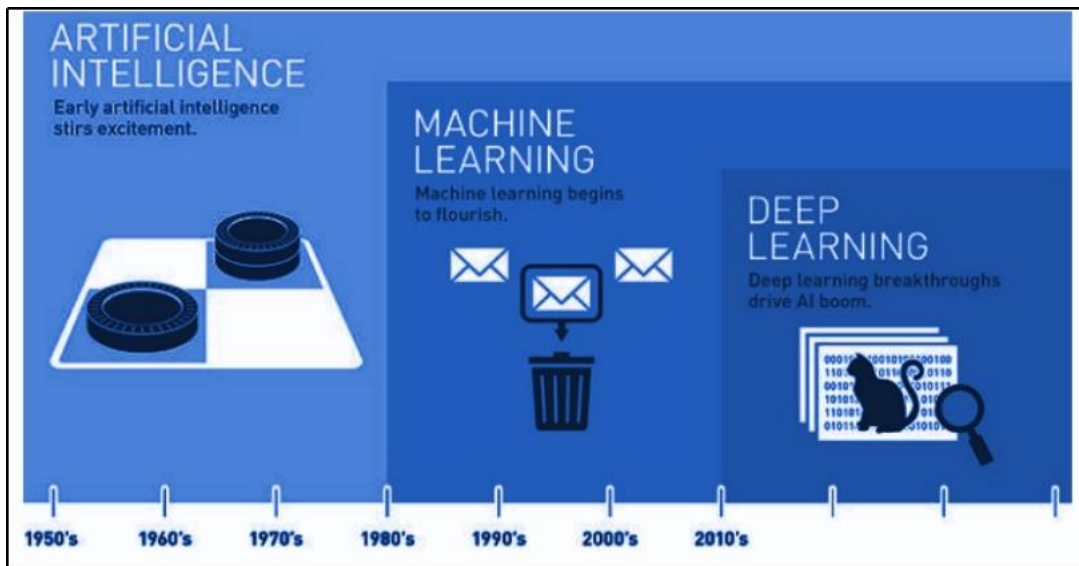
Automated radiology report generation addresses demand for efficiency, consistency, and reduced radiologist burden. By handling routine reporting, AI allows radiologists to focus on complex cases. However, report generation differs from simple classification, it requires coherent synthesis of findings, contextual interpretation, and clinically nuanced language. Challenges include data imbalance, evaluation difficulties, and the risk of fabricated findings. Despite this, benefits such as improved turnaround times, consistency, and support in high-volume settings are compelling. Integration of AI with structured reporting and clinical oversight is key to success (*Sloan et al., 2024; Mecha Health, 2025*).

## **2. Artificial Intelligence (AI)**

AI is a field of computer science focused on developing systems that can perform tasks typically requiring human intelligence, such as reasoning, learning, and decision-making (*Afzal et al., 2025*). These systems rely on techniques like machine learning (ML), deep learning, natural language processing (NLP), and computer vision to interpret and analyze complex data. Through these technologies, AI can uncover patterns and insights from large datasets that support more informed decision-making.

In healthcare, AI has been widely adopted to enhance diagnostics, personalize treatment, and monitor patients more effectively (Alowais *et al.*, 2023). By rapidly processing vast volumes of medical data, AI helps identify patterns and correlations that might go unnoticed by clinicians, improving the accuracy of early diagnoses and supporting individualized care (Alowais *et al.*, 2023). These capabilities make AI a powerful tool for transforming clinical workflows and patient outcomes.

As shown in **Figure 1**, deep learning is considered a subset of machine learning, which itself is a core component of the broader field of artificial intelligence. This hierarchical relationship highlights how increasingly complex techniques build upon one another to enable sophisticated AI systems capable of handling high-dimensional data and intricate tasks.



**Figure 1:** Artificial Intelligence’s Subsets of machine learning and deep learning (Tiwari *et al.*, 2018).

### 3. Deep learning in medical imaging

Deep learning has revolutionized medical imaging by enabling automated diagnosis and workflow enhancements. This section explores CNN architectures, ensemble methods, and explainability in clinical applications.

#### 3.1 Convolutional Neural Networks for medical image analysis

CNNs excel in medical image analysis by learning hierarchical features from raw data, eliminating reliance on handcrafted features. They achieve state-of-the-art performance in tasks

like tumor segmentation and disease classification across X-ray, MRI, and CT modalities. For instance, CNNs have demonstrated high accuracy in detecting pulmonary nodules and classifying ovarian cancer subtypes (Litjens *et al.*, 2017; Bhole *et al.*, 2025).

### **3.2 Ensemble learning and pre-trained models**

Ensemble methods combine multiple CNNs to improve diagnostic accuracy and reduce overfitting. Transfer learning leverages pre-trained models like ResNet and Vision Transformers, fine-tuned on medical datasets to overcome limited labeled data. This approach enhances tasks such as pneumonia detection in chest X-rays and brain tumor identification (Durgaraju *et al.*, 2025).

#### **3.2.1 ResNet architecture and applications**

ResNet, introduced by (He *et al.* 2016), addresses the vanishing gradient problem through skip connections, enabling training of very deep networks. In medical imaging, ResNet variants like ResNet-50 and ResNet-101 are widely used for tasks such as disease classification (e.g., pneumonia, tuberculosis, COVID-19), lesion detection in MRI and CT scans, and transfer learning applications, where pre-trained weights are adapted to new medical datasets, improving accuracy and efficiency (Durgaraju *et al.*, 2025).

#### **3.3.2 EfficientNet for medical imaging**

EfficientNet, proposed by (Tan & Le, 2019), optimizes network scaling by balancing depth, width, and resolution, resulting in models that deliver high accuracy with fewer computational resources. This efficiency makes EfficientNet suitable for clinical environments with limited hardware. It has been successfully applied to diverse medical imaging tasks including breast cancer detection, skin lesion classification, and organ segmentation. Like ResNet, EfficientNet benefits from transfer learning, enabling rapid adaptation to various medical imaging challenges (Mormont, 2022).

## **4. Natural Language Processing in healthcare**

Natural Language Processing (NLP) has become an indispensable tool in healthcare, enabling the extraction, interpretation, and generation of meaningful information from vast amounts of unstructured clinical text. This includes radiology reports, electronic health records (EHRs), discharge summaries, and clinical notes, which are rich in information but challenging

to analyze due to their complexity and variability. The evolution of NLP techniques from rule-based systems to advanced transformer-based architectures has significantly enhanced the ability to process medical language with greater accuracy and contextual understanding.

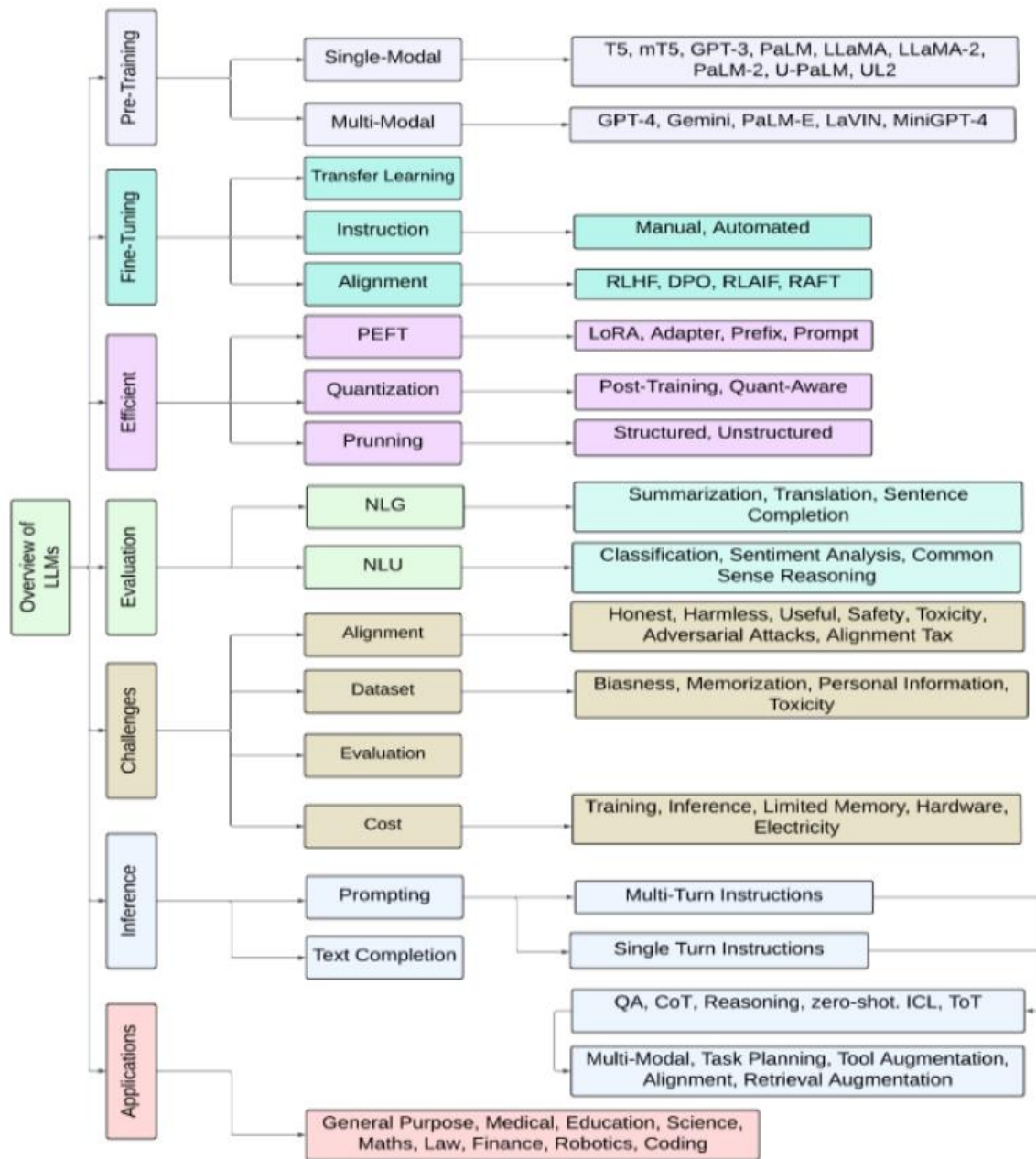
#### **4.1 BERT and Transformer models in medical NLP**

Transformer models like BERT have revolutionized medical NLP by capturing bidirectional context. Specialized versions such as BioBERT and ClinicalBERT, trained on biomedical texts, outperform earlier models in clinical tasks. Their ability to be fine-tuned on limited data makes them highly effective for applications including radiology report analysis and clinical trial screening (Devlin *et al.*, 2018; Alsentzer *et al.*, 2019).

#### **4.2 Large Language Models for medical text generation**

Large Language Models (LLMs) extend the capabilities of transformers to generate coherent, contextually appropriate text, which has important applications in healthcare. These models, often containing billions of parameters, can produce fluent clinical narratives, summarize patient histories, generate discharge summaries, and even draft radiology reports (Nerella *et al.*, 2024).

As illustrated in **Figure 2**, LLM research can be broadly categorized into seven main branches: pre-training, fine-tuning, efficient methods, inference, evaluation, applications, and challenges. This taxonomy provides a structured overview of the LLM landscape and highlights key areas of ongoing development and innovation within the field (Naveed *et al.*, 2024).



**Figure 2 :** A broader overview of LLMs, dividing LLMs into seven main branches: pre-training, fine-tuning, efficient methods, inference, evaluation, applications, and challenges (Naveed *et al.*, 2024).

### Parameter fine-tuning with LoRA:

Low-Rank Adaptation (LoRA) is an innovative fine-tuning technique that adapts large pre-trained models by modifying only a small subset of parameters. This approach significantly reduces computational costs and memory usage compared to full fine-tuning, which is

particularly advantageous in healthcare where data privacy and limited computational resources are common constraints (Hu *et al.*, 2021).

## 5. Evaluation metrics for medical text generation

Evaluating generated medical text is essential to ensure it is clinically accurate and useful. This involves automated metrics for quantitative comparison and expert review for clinical validity.

### **BLEU (Bilingual Evaluation Understudy)**

BLEU measures the overlap of n-grams between generated and reference texts, with scores from 0 to 1 indicating similarity. BLEU-1 to BLEU-4 progressively evaluate unigram to four-gram matches, reflecting lexical and fluency quality (Papineni *et al.*, 2002). However, BLEU may not fully capture semantic or clinical correctness.

### **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)**

ROUGE focuses on recall, assessing how much of the reference text is reflected in the generated output. ROUGE-1, ROUGE-2, and ROUGE-L measure unigram, bigram, and longest common subsequence overlaps, respectively, highlighting content coverage and structural similarity (Lin, 2004).

## 6. Multimodal learning approaches in healthcare

Multimodal learning has become a cornerstone of modern artificial intelligence in healthcare, reflecting the reality that clinical decision-making is inherently multimodal integrating information from diverse sources such as medical images, electronic health records, laboratory results, genomics, and patient-reported outcomes. By leveraging data from multiple modalities, multimodal machine learning models can achieve more robust, accurate, and clinically relevant predictions compared to unimodal approaches.

### **6.1 Fusion strategies**

A critical challenge in multimodal learning is how to effectively combine information from disparate sources. Fusion strategies are typically categorized as :

- **Early fusion:** Raw or low-level features from each modality are concatenated and fed into a unified model. This approach captures joint representations but may struggle with high-dimensional or heterogeneous data.
- **Late fusion:** Each modality is processed independently, and predictions or high-level features are combined at a later stage, often via ensemble methods. This strategy is more robust to missing modalities but may miss cross-modal interactions.
- **Hybrid/Intermediate fusion:** Combines early and late fusion, integrating features at multiple levels of abstraction to balance the strengths of both approaches.

## 7. Explainable Artificial Intelligence (XAI) in medical imaging

Explainable Artificial Intelligence (XAI) arose to make complex models like deep neural networks more transparent. In healthcare, its importance grew with the need to understand and trust AI-driven decisions, especially in high-stakes contexts (Tjoa & Guan, 2021).

### 7.1 Terminology

Within XAI, several key terms are essential:

- **Interpretability:** The degree to which a human can understand the internal mechanics of an AI system.
- **Explainability:** The extent to which the AI's decision-making process can be articulated in human-understandable terms.
- **Transparency:** A property of a system that allows users to trace its decision-making process.

These concepts, though related, are distinct and critical for evaluating AI in healthcare (Tjoa & Guan, 2021).

### 7.2 The need for XAI, motivations

The integration of XAI in healthcare is motivated by:

- **Enhanced trust:** Clinicians can visually and textually verify AI-generated captions.
- **Improved diagnostic accuracy:** Clear explanations help reduce misdiagnosis and missed findings.
- **Ethics:** Transparent AI ensures responsibility and adherence to ethical standards.



- **Clinical utility:** XAI aids in error detection, bias mitigation, and facilitates more informed decision-making in diagnosis and treatment (Tjoa & Guan, 2021; Holzinger *et al.*, 2019).
- **Justification:** Explainable AI (XAI) provides clear and understandable explanations for model outcomes, especially when results are unexpected. This transparency helps build trust by offering a verifiable rationale behind decisions.
- **Accountability:** XAI enhances accountability by revealing the decision-making process of AI systems, which is crucial in high-stakes areas like healthcare. It ensures that responsibility for AI-driven outcomes can be clearly assigned.

### 7.3 Common explainability methods

Common explainability methods help reveal how machine learning models make decisions. Techniques like SHAP (Lundberg & Lee, 2017) and Integrated Gradients (Sundararajan *et al.*, 2017) assign importance scores to input features, while Grad-CAM (Selvaraju *et al.*, 2017) highlights key areas in data such as images. Attention mechanisms in transformers (Vaswani *et al.*, 2017) also provide insight by showing which parts of the input the model focuses on. These methods improve model transparency and user trust.

A comparative overview of the most commonly used explainable AI (XAI) techniques is provided in **Table 1**, summarizing their compatibility, speed, explanation type, and typical use cases, along with their main advantages and limitations. This summary helps identify the most suitable method depending on the data modality and model architecture involved.

**Table 1:** Summary of XAI Methods

Method	Data/Model Compatibility	Speed	Explanation Type	Use Case	Advantages	Limitations
<b>SHAP</b>	Tabular, Image, Text / any model	Slow	Local & global	Feature importance in predictions	Provides consistent and additive feature attributions	Computationally intensive for large datasets (Lundberg & Lee, 2017)
<b>GRAD-CAM</b>	Images / CNN-based models	Fast	Visual (local)	Highlighting image regions that influence decisions	Fast computation (single forward/backward pass); intuitive heatmaps	Limited to CNNs (Selvaraju et al., 2017)
<b>LIME</b>	Tabular, Image, Text / any model	Slow	Local	Explaining individual predictions via perturbation analysis	Model-agnostic; flexible across data types	Can be slow with many perturbations (Ribeiro et al., 2016)
<b>Integrated Gradients</b>	Image, Text / Differentiable models (e.g., neural networks)	Moderate	Local & global	Understanding feature impact through integration over gradients	Theoretically well-grounded; relatively straightforward to implement	Requires a baseline; can be computationally demanding (Sundararajan et al., 2017)
<b>Attention Mechanisms</b>	Sequence data (Text, Time-series) / transformers	Fast/Intrinsic	Local (visual or textual highlighting)	Visualizing focus areas in transformer models	Fast and intrinsic to model inference; directly available from attention weights	Attention weights may not perfectly correlate with feature importance (Vaswani et al., 2017)

#### 7.4 Real-world applications and case studies

Recent works illustrate successful MIC-XAI integration:

- **MIMIC-CXR-explain:** A variant of the MIMIC-CXR dataset augmented with Grad-CAM heatmaps, enabling radiologists to verify which regions influenced the AI-generated report (Bannur *et al.*, 2024).
- **Attention-guided captioning:** Studies have shown that overlaying attention maps onto X-ray images improves diagnostic agreement between AI and clinicians by up to 30% (Chen *et al.*, 2020).
- **Explainable pediatric radiology:** Integrating SHAP-based attributions with MIC has led to improved detection of rare pediatric conditions, enhancing diagnostic precision by 25% (Miura *et al.*, 2021).

To support such applications, several benchmark datasets have been developed. **Table 2** summarizes key datasets commonly used in MIC research, including details about imaging modalities, content, and dataset size. These datasets form the foundation for training, evaluating, and validating explainable MIC systems.

**Table 2:** Key Medical Image Captioning Datasets

Dataset	Modality	Description	Size
<b>IU X-ray</b> (Demner-Fushman et al., 2016)	X-ray	Chest X-rays paired with radiology reports	7,470
<b>MIMIC-CXR</b> (Johnson et al., 2019)	X-ray	Large-scale dataset of chest X-rays and reports	377,110
<b>ROCO</b> (Pelka et al., 2018)	Multi	Radiology Objects in Context dataset	81,000
<b>CheXpert</b> (Irvin et al., 2019)	X-ray	Multi-label chest X-ray dataset	224,316

## 8. Rated works

Several recent studies exemplify the integration of large vision-language models in chest X-ray analysis, demonstrating impressive performance and clinical relevance. For instance, XrayGPT (XrayGPT Team, 2023) combines a medical visual encoder known as MedClip with a fine-tuned large language model (Vicuna) to perform interactive summarization and open-ended question answering on chest radiographs (XrayGPT Team, 2024). Trained on approximately 217,000 radiology report summaries, XrayGPT exhibits exceptional visual conversational abilities grounded in comprehensive radiological knowledge. Similarly, CXR-LLAVA (Lee *et al.*, 2023) is an open-source multimodal large language model that integrates a pretrained vision transformer with an LLM inspired by the LLAVA architecture (Lee *et al.*, 2023). Trained on nearly 600,000 chest X-rays, including over 217,000 associated free-text reports, CXR-LLAVA outperforms state-of-the-art models such as GPT-4-vision and Gemini-Pro-Vision on both internal and external validation datasets, highlighting the benefits of large-scale multimodal training.

Another notable contribution is ELIXR, (ELIXR Research Group, 2023) a general-purpose X-ray AI system that aligns radiology vision encoders with the PaLM 2 large language model (ELIXR Research Group, 2024). ELIXR achieves state-of-the-art zero-shot

classification performance on chest X-rays and shows promising results in vision-language tasks including visual question answering and report quality assurance. This demonstrates the versatility of aligning powerful vision and language models for diverse clinical applications. Additionally, LiteGPT (LiteGPT Authors, 2024) proposes a unified framework for joint localization and classification of chest X-ray abnormalities by leveraging multiple pretrained visual encoders to enrich feature representation (LiteGPT Authors, 2024). It establishes new benchmarks on the VinDr-CXR dataset, showcasing the effectiveness of multi-encoder fusion in vision-language modeling.

Beyond diagnostic interpretation, generative models like RoentGen (RoentGen Developers, 2024) extend the capabilities of vision-language systems by synthesizing high-fidelity, diverse chest X-ray images conditioned on free-text radiology prompts (RoentGen Developers, 2024). This latent diffusion-based model, pretrained on paired radiographs and reports, enables controlled image generation that can support data augmentation, training, and educational purposes.

While these advances mark significant progress, challenges remain in ensuring robustness, generalizability, and clinical explainability of vision-language models in medical imaging. Our project builds upon these foundations by proposing a novel, explainable image captioning framework tailored specifically for chest X-rays. By integrating state-of-the-art vision encoders with fine-tuned LLMs and incorporating explainability techniques, we aim to develop an AI tool that not only delivers accurate diagnostic summaries but also provides transparent, interpretable insights to support radiologists in clinical practice.

As summarized in **Table 3**, these recent advances illustrate the growing maturity of LLM-based frameworks in radiology, highlighting diverse model architectures, datasets, and clinical tasks addressed. However, challenges remain in ensuring robustness, generalizability, and clinical explainability. Our project builds upon these foundations by proposing a novel, explainable image captioning framework tailored specifically for chest X-rays. By integrating state-of-the-art vision encoders with fine-tuned LLMs and incorporating explainability techniques, we aim to develop an AI tool that not only delivers accurate diagnostic summaries but also provides transparent, interpretable insights to support radiologists in clinical practice.

**Table 3:** Summary of recent works on integrating LLMs with chest radiographs

Study	Model / Method	Key Techniques	Dataset	Performance	Highlights
<b>XrayGPT</b> (2023)	XrayGPT	Alignment of MedCLIP with a LLM (Vicuna) via linear transformation; interactive report summarization	~217,000 radiology report summaries	Not specified	Exceptional visual conversation ability grounded in deep radiograph and medical knowledge understanding
<b>CXR-LLAVA</b> (2023)	CXR-LLAVA	Integration of a pretrained vision transformer with a LLAVA-style LLM; trained on chest X-rays with free-text reports	592,580 chest X-rays, including 217,699 with reports	F1-score: 0.81 (internal), 0.62 (external)	Outperforms GPT-4-Vision and Gemini-Pro-Vision on internal and external test sets
<b>ELIXR</b> (2023)	ELIXR	Alignment of a radiology image encoder with a LLM (PaLM 2); use of lightweight adapter	MIMIC-CXR	AUC: 0.850 (zero-shot); AUC: 0.893 with 1% data, 0.898 with 10% data	State-of-the-art zero-shot CXR classification performance; promising for visual QA and report validation
<b>LiteGPT</b> (2024)	LiteGPT	Unified framework for joint localization and classification; uses multiple pretrained visual encoders	VinDr-CXR	Not specified	First study using vision-language models for joint localization and classification in medical imaging
<b>RoentGen</b> (2022)	RoentGen	Adaptation of a latent diffusion model pretrained on chest radiographs and corresponding reports	Chest X-rays with corresponding reports	5% improvement in classifier performance when trained jointly on synthetic and real images	Capable of generating diverse high-fidelity synthetic CXRs controlled via free-text prompts

## Conclusion

This chapter has highlighted both the strengths and limitations of traditional radiology reporting and the promise of AI-driven tools to enhance accuracy, efficiency, and transparency. We’ve surveyed the essential technologies from CNN ensembles and BERT embeddings to LoRA fine-tuning and Grad-CAM explainability that inform our approach. This foundation prepares us to dive into the detailed methods of building and evaluating our multimodal, explainable chest X-ray reporting system.

## **CHAPTER 2: Materials and** **Methodology**

## Introduction

In the pursuit of advancing clinical decision support systems, the integration of multimodal artificial intelligence techniques has become essential. This chapter details the methodological framework developed to generate explainable radiology reports from chest X-ray images and clinical notes. The system combines deep convolutional neural networks, contextual language models, and explainability tools to provide accurate and interpretable results suitable for real-world medical applications.

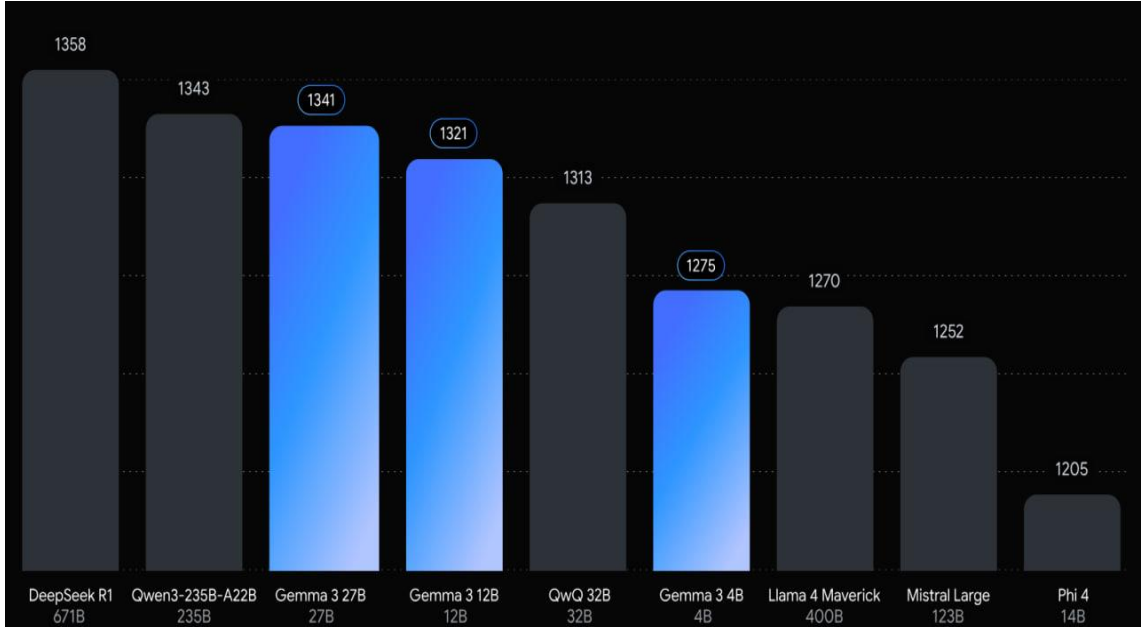
The first part of the chapter outlines the overall system architecture and key design considerations, followed by a comprehensive description of the data preparation workflow. Subsequent sections present the visual and textual feature extraction strategies, the fusion of multimodal embeddings, and the fine-tuning of a language generation model. Special attention is given to the inclusion of explainability through visual attribution methods, as well as the deployment strategy used to transform the solution into a usable clinical web application. Each methodological choice is justified based on its relevance to the domain and its ability to support transparent, scalable, and robust model behavior in the healthcare setting.

## 1. Materials

### 1.1 Large Language Model Gemma-3

In this work, we utilized Gemma-3 1B instruction-tuned (IT), Google's latest open-weight large language model released in March 2025 (The Decoder, 2025). The model contains 1 billion parameters across 26 transformer layers with 1152 hidden dimensions, 4 attention heads, and supports up to 32,000 tokens context length (Ji & Kumar, 2025).

As illustrated in **Figure 3**, Gemma-3 models demonstrate competitive performance when compared to other state-of-the-art large language models across a variety of benchmark tasks, confirming their reliability and versatility in real-world applications (Google DeepMind, 2025).



**Figure 3 :**Performance Comparison of Gemma-3 Models Against State-of-the-Art Language Models on Benchmark Evaluations (Google DeepMind, 2025)

We chose Gemma-3 1B-IT for several compelling reasons. First, the 1B parameter configuration provides an optimal balance between computational efficiency and performance, making it suitable for medical domain fine-tuning with limited computational resources. The instruction-tuned variant demonstrates superior performance in following clinical reasoning patterns and maintaining medical terminology consistency, which are crucial for generating accurate radiological impressions (Google Health AI Developer Foundations, 2025). Additionally, the model's parameter-efficient architecture enables effective LoRA fine-tuning, allowing for domain-specific adaptation while maintaining safety standards required for healthcare deployment (Hu *et al.*, 2021). This combination of efficiency, clinical adaptability, and safety makes Gemma-3 1B-IT ideal for chest X-ray radiology report generation.

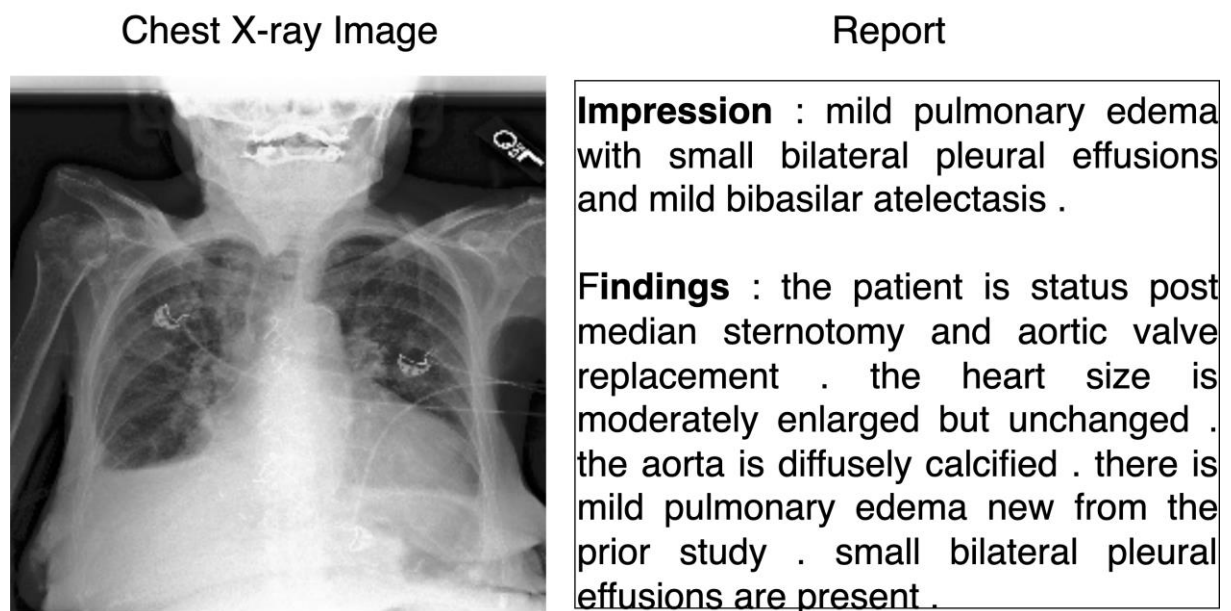
## 1.2 Dataset description and characteristics

The Indiana University Chest X-ray Collection (commonly referred to as IU X-ray or Open-I) is a publicly accessible dataset that includes 7,470 chest radiographs paired with structured radiology reports. Each report is organized into distinct sections such as *Indication*, *Findings*, *Impression*, *Comparison*, and *Tags*. This structure enables granular analysis and facilitates the design of supervised tasks in medical image captioning (Demner-Fushman *et al.*, 2016).



As shown in **Figure 4**, the dataset includes both the radiographic image and its corresponding structured textual report, offering a rich multimodal resource for training and evaluating vision-language models in radiology.

For consistency, we selected only the primary posteroanterior (PA) or anteroposterior (AP) view per study, as is standard practice in many radiology-based machine learning studies (Jing *et al.*, 2018). Reports were analyzed to determine section length distributions, and frequent clinical terms were extracted to gain insights into linguistic patterns. These exploratory analyses revealed a high frequency of normal studies, with common phrases such as “no acute cardiopulmonary abnormality” dominating the *Impression* sections an expected finding in large hospital datasets (Demner-Fushman *et al.*, 2016).



**Figure 4** : Example of a Chest X-ray Image and Corresponding Structured Radiology Report from the Indiana University Chest X-ray Collection

### 1.3 High Performance Computing (HPC)

High Performance Computing (HPC) refers to the use of advanced computing systems such as clusters of high-speed processors and GPUs to solve complex problems in scientific, engineering, and commercial fields. These systems perform large-scale computations at extremely high speeds. HPC infrastructures combine computing power across multiple nodes to accelerate program execution and enhance overall efficiency in data processing and model training. (ScienceDirect Topics, n.d; IBM,n.d.).

As detailed in **Table 5**, the High Processing Center (HPC) used in this study provided substantial computing capabilities, including multi-core CPUs, powerful Tesla V100 GPUs, and large RAM allocations per node. These specifications ensured the scalability and speed required for complex medical imaging tasks and model optimization.

High Processing Center	Characteristics
CPU	12 nodes (2*14 cores),
GPU	2 nodes (4 GPU Tesla V100)
RAM	128 GB per node

**Table 3:** Basic information on the High Processing Center (HPC) used for data preprocessing and model training.

#### 1.4 Computing station (Desktop)

computing station is a desktop computer equipped with 12th Gen Intel Core i7-12700F processor (2.10 GHz) and 64 GB RAM, providing robust processing power for intensive computational tasks and large-scale data processing. The system operates on 64-bit architecture ensuring optimal compatibility with modern analysis and modeling software.

As outlined in **Table 6**, the technical specifications of this computing station reflect its suitability for deep learning and multimodal data processing in a healthcare AI research setting.

**Table 4:** Technical specifications of the computing station used for data processing and model training

Specification	Description
Processor	Intel Core i7-12700F (12th Gen), 2.10 GHz
RAM	64 GB (63.9 GB usable)
System Type	64-bit operating system, x64-based processor

## **1.5 Python**

Python is a high-level, general-purpose programming language widely adopted in scientific computing due to its simplicity, extensive scientific libraries, and seamless integration capabilities. Its robust ecosystem makes it ideal for AI and machine learning applications, particularly in medical imaging and natural language processing (Bird, Klein, & Loper, 2009; McKinney, 2010).

## **1.6 Packages**

### **1.6.1 Core deep learning & transformers**

We utilized PyTorch as the primary deep learning framework for its dynamic computation graph and GPU acceleration capabilities (PyTorch Foundation, 2025). The Hugging Face Transformers library enabled easy access to pre-trained models like BERT and Gemma, while PEFT facilitated parameter-efficient fine-tuning using LoRA techniques (Hu *et al.*, 2022).

### **1.6.2 Computer vision & feature extraction**

Torchvision provided pre-trained CNN models (ResNet50, EfficientNet) for image feature extraction (PyTorch, 2025). OpenCV-Python and PIL handled image preprocessing and Grad-CAM visualization generation, while Matplotlib enabled result visualization.

### **1.6.3 Natural Language Processing**

NLTK provided tokenization and BLEU score computation, ROUGE library calculated text generation metrics, and Pycocoevalcap implemented CIDEr evaluation for image captioning tasks (Techvify, 2025).

### **1.6.4 Data handling & utilities**

Pandas and NumPy handled data manipulation and numerical computations, while scikit-learn provided preprocessing and normalization utilities for machine learning pipelines (DataCamp, 2025).

### **1.6.5 Visualization tools**

Seaborn and Matplotlib generated statistical plots and model performance visualizations for comprehensive results analysis (New Horizons, 2024).

## **1.7 Web framework**

Django enabled development of the clinical web interface with secure authentication and database management for AI model deployment (Django Software Foundation, n.d.).

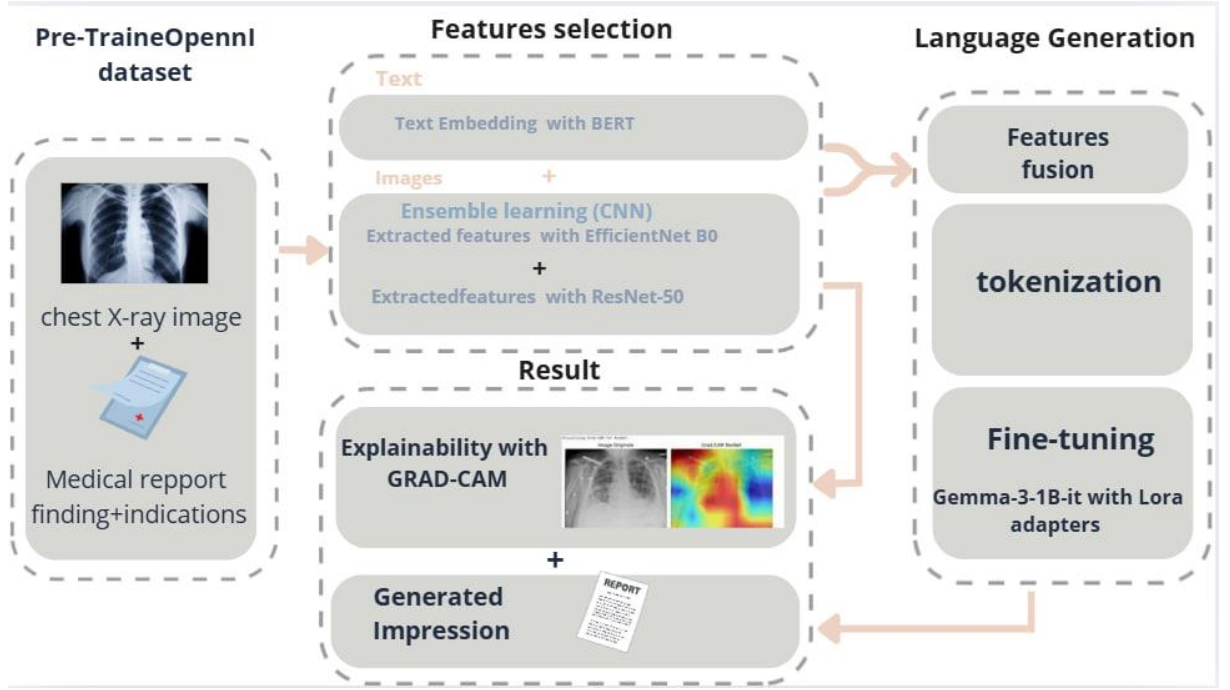
**Table 5:** The versions of the used packages

<b>Packages</b>	<b>Version</b>
<b>torch (PyTorch)</b>	2.1.0
<b>transformers</b>	4.36.0
<b>peft</b>	0.7.1
<b>Torchvision</b>	0.16.0
<b>PIL (Pillow)</b>	10.1.0
<b>cv2 (OpenCV-Python)</b>	4.8.1
<b><u>ntlk</u></b>	3.8.1
<b>rouge</b>	1.0.1
<b>Pycocoevalcap:</b>	1.2
<b><u>pandas</u></b>	2.1.4
<b><u>numpy</u></b>	1.24.3
<b>scikit-learn (sklearn)</b>	1.3.2
<b>Seaborn</b>	0.12.2
<b>Matplotlib</b>	3.8.2
<b>Django</b>	4.2.7

## 2.Methodology

### 2.1 Overall pipeline architecture

We designed a four-stage pipeline that integrates image and text modalities before model fine-tuning. First, we prepare and filter our multimodal data (images and text). Next, we extract visual features via an ensemble of CNNs augmented with Grad-CAM. In parallel, we generate BERT embeddings from the Indications and Findings sections. Finally, we fuse these features, then fine-tune the Gemma-3 1B model, and produce reports. At inference time, the system outputs both the generated “Impression” text and two Grad-CAM heatmaps (one from ResNet50 and one from EfficientNetB0) for visual explainability. We now describe how we prepare and preprocess our dataset prior to feature extraction like it shows figure 05.



**Figure 5 : Model Pipeline Architecture**

## 2.2 Data preparation and preprocessing

### 2.2.1 Data quality assessment and filtering

Prior to downstream processing, a quality assurance protocol was applied to eliminate problematic or non-informative entries. This included:

- Excluding any report missing a valid *Findings* or *Impression* section.
- Removing entries where the *Findings* section contained fewer than 10 words or the *Impression* fewer than two words.

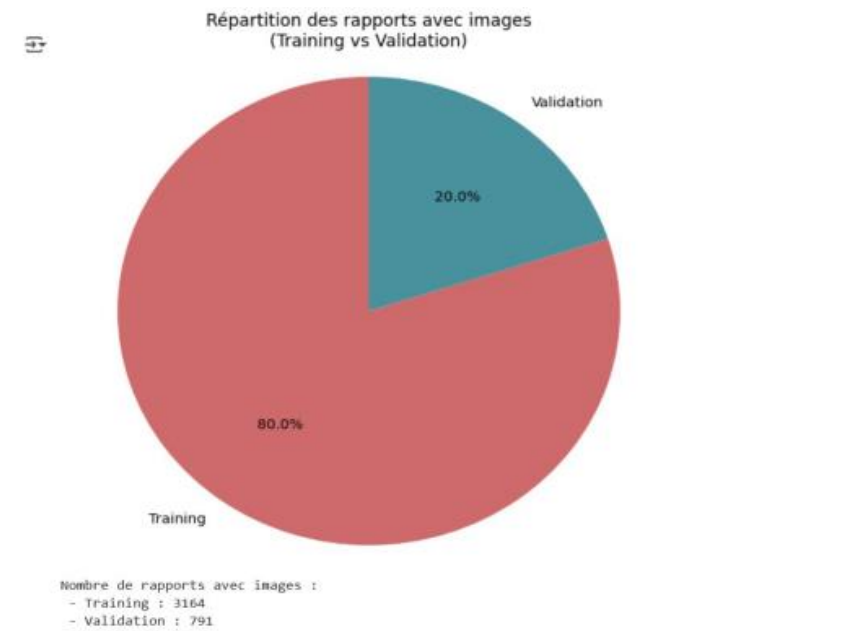
These thresholds were selected to ensure that the reports retained a minimum level of clinical content. Similar heuristics have been adopted in prior studies on radiology report generation (Hu *et al.*, 2021). After filtering, approximately 7,415 high-quality image–text pairs remained. This final dataset served as the input for all modeling and evaluation procedures.

### 2.2.2 Dataset partitioning

After cleaning and filtering, the resulting chest X-ray-report pairs were randomly divided into an 80 % training set and a 20 % validation set using the 'subset' column provided in the metadata. Samples labeled 'train' were used for learning, while those labeled 'val' were held out for

evaluation. This split ensures that we monitor model generalization on held-out data during fine-tuning.

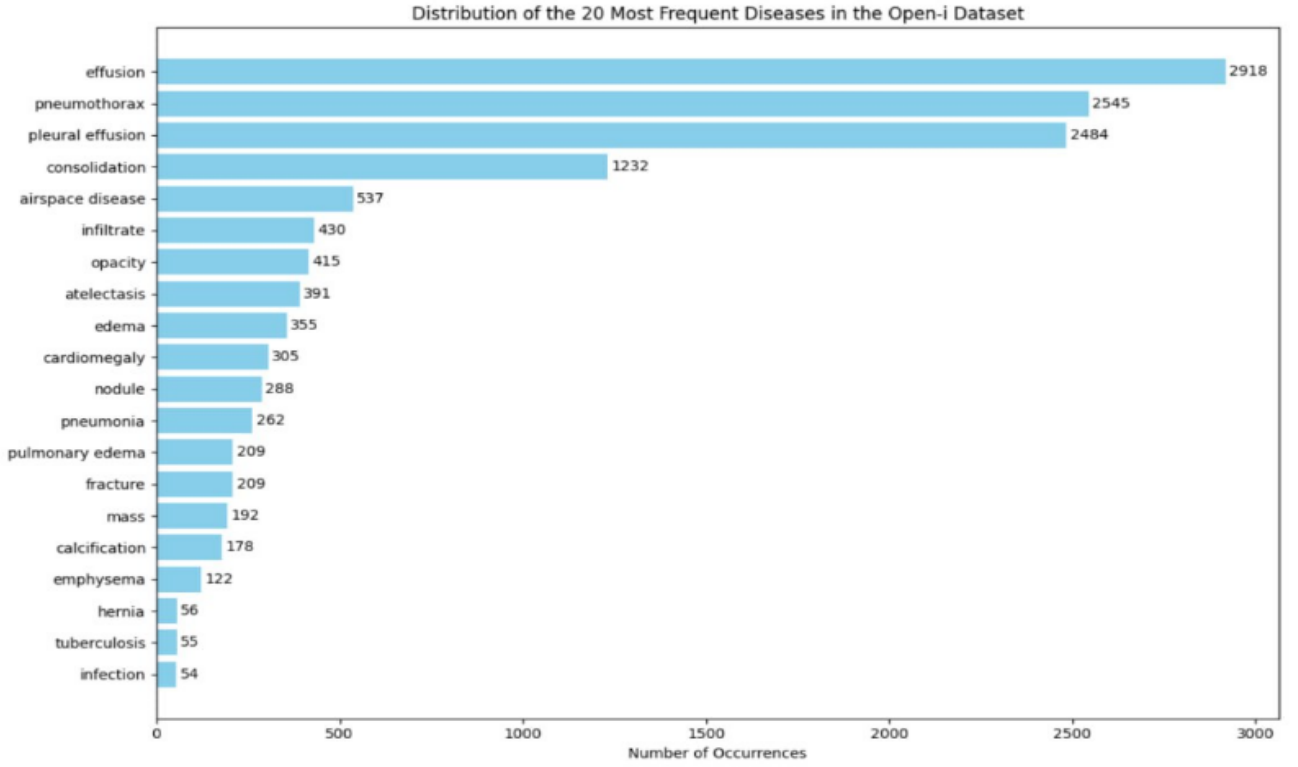
The dataset partitioning strategy is visually represented in **Figure 6**, providing a clear overview of how the data was divided for experimental purposes.



**Figure 6 :** Dataset Partitioning into training and validation sets

### 2.2.3 Data exploration and visualization

To better understand the structure and characteristics of the dataset, we performed exploratory data visualization. This step aids in interpreting the composition of both the imaging and textual data, guiding preprocessing decisions and improving downstream model design.



**Figure 7 :** Bar chart showing the distribution of the most frequently diagnosed pathologies in the Open-I dataset

#### 2.2.4 Image preprocessing pipeline

Before extracting visual features, the raw X-rays undergo standardized preprocessing to ensure compatibility with our CNN backbones and to improve training stability.

##### Image normalization and standardization

All chest radiographs are first converted to RGB by replicating the single grayscale channel. We resize images to  $224 \times 224$  pixels using bicubic interpolation and then apply channel-wise normalization based on ImageNet means ( $[0.485, 0.456, 0.406]$ ) and standard deviations ( $[0.229, 0.224, 0.225]$ ). This pipeline aligns the inputs with ResNet50 and EfficientNetB0 expectations and has been shown to speed up convergence when fine-tuning pretrained models (Shorten & Khoshgoftaar, 2019; Tajbakhsh *et al.*, 2016).

### 2.2.5 Text preprocessing and tokenization

After preparing images, we clean and tokenize the radiology reports to generate contextual embeddings.

#### A) Report section extraction

Reports are parsed into three fields Indication, Findings, and Impression each carrying specific clinical information. Missing or empty fields are filled with blank strings to maintain dataset consistency, and only entries containing both Findings and Impression above minimum word counts are retained (Irvin *et al.*, 2019).

#### B) Text cleaning and normalization

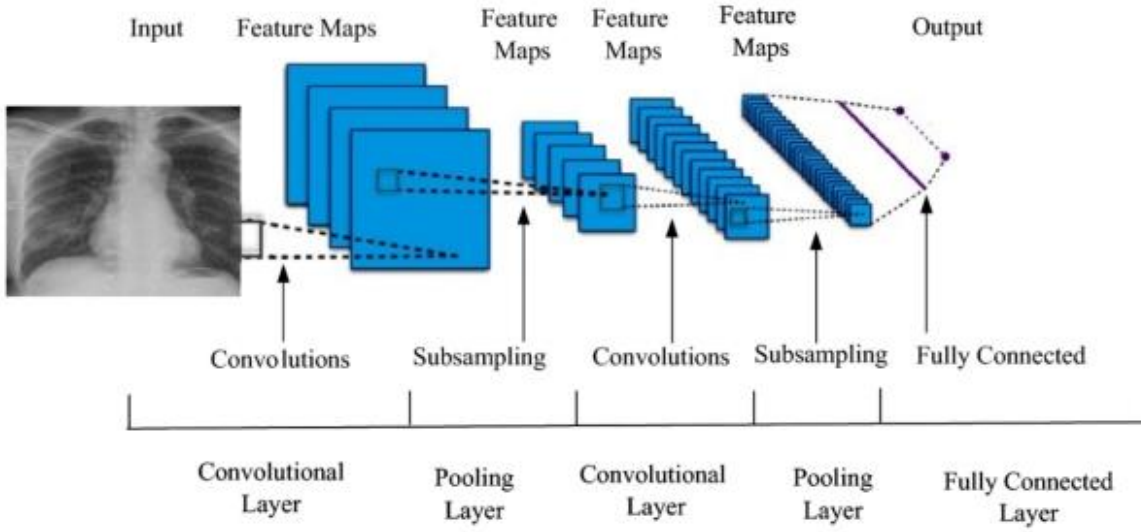
We lowercase all text, collapse extra whitespace, and strip section headers while preserving medically relevant punctuation, negations, and measurements. Cleaned text is tokenized with Hugging Face’s bert-base-uncased WordPiece tokenizer; the [CLS] embedding from a pretrained BERT encoder is extracted and  $L_2$ -normalized prior to fusion (Alsentzer *et al.*, 2019).

### 2.3 Visual feature extraction module

With preprocessed inputs, we extract complementary image representations using a CNN ensemble augmented by Grad-CAM for explainability.

A custom PyTorch DataLoader feeds each X-ray into ResNet50 and EfficientNetB0 in parallel; both models operate as frozen feature extractors, with hooks capturing activations and gradients for later Grad-CAM heatmap generation (Selvaraju *et al.*, 2019; Nguyen *et al.*, 2019).





**Figure 8 :** General architecture of a CNN, showing the progression from raw image input to feature extraction and final output through convolutional, pooling, and fully connected layers (Suganyadevi *et al.*, 2021).

### 2.3.1 ResNet50 implementation and configuration

We remove ResNet50's final classification head and freeze its parameters. A forward hook on layer4[-1]. Conv2 captures the 2048-channel activation map; a backward hook retrieves gradients for Grad-CAM. Global average pooling follows, collapsing spatial dimensions to a 2048-D vector (Selvaraju *et al.*, 2019).

### 2.3.2 EfficientNetB0 integration

Similarly, EfficientNetB0 is loaded without its classifier head, with hooks on features [-1] to obtain 1280-channel activations and gradients. After global average pooling, each image yields a 1280-D feature vector. Using two backbones leverages both deep and efficient representations (Nguyen *et al.*, 2019; Rajpoot *et al.*, 2024).

### 2.3.3 Ensemble learning implementation

We concatenate the 2048-D ResNet50 vector with the 1280-D EfficientNetB0 vector for each X-ray. No additional trainable weights are applied at this stage; the parallel feature extractor outputs form a 3328-D joint visual embedding, reflecting complementary network biases (Nguyen *et al.*, 2019).

### 2.3.4 Feature map processing and dimensionality reduction

Channel-wise global average pooling converts each hooked activation map into its fixed-length vector (2048 D for ResNet50, 1280 D for EfficientNetB0) without further dimensionality reduction. This preserves all learned filter responses and avoids potential loss of discriminative information (Nguyen *et al.*, 2019).

## 2.4 Textual feature extraction

Having prepared our image features, we next convert radiology reports into dense embeddings that complement the visual representation.

### 2.4.1 Report parsing & cleaning

We first split each report into Indication, Findings, and Impression sections, then fill any missing fields with blank strings and discard entries that fail the word-count thresholds). This ensures every retained record contains sufficient clinical context.

### 2.4.2 Tokenization & embedding extraction

Cleaned text is lowercased, extra whitespace is removed, and section headers are stripped while preserving clinically important punctuation, negations, and measurements. We batch sequences (batch size = 32) and apply Hugging Face’s bert-base-uncased tokenizer with max\_length=128, padding, and truncation. The tokenized inputs feed into a pretrained BERT-base encoder; we extract the [CLS] token embedding (last\_hidden\_state[:,0,:]) for each section, producing two 768-D vectors per study (Devlin et al., 2019).

### 2.4.3 Embedding normalization & concatenation

To balance text against image features, each 768-D embedding is L<sub>2</sub>-normalized (via sklearn.preprocessing.normalize) to unit length. We then concatenate the Indication and Findings vectors into a single 1536-D text embedding, ready for fusion with the 3328-D visual feature.

### 2.4.4 Text feature output

The module outputs an N×1536 matrix one normalized, concatenated text embedding per study which will be appended to the visual descriptor to form the complete multimodal representation.

## 2.5 Ensemble fusion

The final step fuses visual and textual features into a single multimodal descriptor. We concatenate the 3328-D visual vector with the 1536-D text embedding to form a 4864-D multimodal feature. This transparent, feature-level concatenation preserves distinct modality information and allows the downstream Gemma-3 1B model to learn how to weight each component (Nguyen *et al.*, 2019; Wang *et al.*, 2025). The resulting 4864-D embedding drives both the “Impression” text generation and Grad-CAM heatmap guidance, enabling end-to-end learning with explainability overlays.

## 2.6 Language model fine-tuning

The pre-trained **Gemma-3-1B-it** model was used as the base language model. Gemma-3 is a decoder-only transformer developed by Google, optimized for long-context text generation. The 1B-parameter variant is a compact version with 26 transformer layers, a hidden size of 1152, and 4 attention heads, including a single key/value head. It employs an interleaved 5-to-1 attention strategy (five local attention layers followed by one global layer) to support long-context modeling of up to 32,000 tokens (Ji & Kumar, 2025). In our implementation, this model was loaded via Hugging Face Transformers and its base weights were frozen during fine-tuning. A summary of its architectural parameters is provided in Table 8.

**Table 6:** Core Architectural Parameters of Gemma-3 1B

Parameter	Value
Model size	≈1 billion parameters
Hidden dimension	1152
Transformer layers	26
Attention heads	4 (1 key/value head)
Attention strategy	5 local : 1 global
Maximum context length	32,000 tokens

### 2.6.1 LoRA implementation and configuration

To adapt Gemma-3 1B with minimal resource overhead, we employed Low-Rank Adaptation (LoRA) via Hugging Face’s PEFT library (Hu *et al.*, 2021). LoRA inserts two trainable low-rank matrices, A and B, into each attention projection, leaving the original weights frozen.

We set the LoRA rank  $r = 8$ , scaling factor  $\alpha = 16$ , and dropout = 0.1, following established best practices that balance efficiency with performance (Hu et al., 2021).

LoRA adapters were applied to the four attention projections  $q\_proj$ ,  $k\_proj$ ,  $v\_proj$ , and  $o\_proj$  allowing full adaptation of the model’s attention mechanism. We used AdamW to update only the LoRA parameters, a paradigm shown to be effective for large models with limited domain data (Dettmers *et al.*, 2023).

### 2.6.2 Training strategy and hyperparameters

We fine-tuned on the 80 % training. training ran for ten epochs with batch size 16 and a learning rate of  $5 \times 10^{-5}$ . We performed validation every 100 steps. Fine-tuning was conducted over ten epochs with a batch size of 16 and a learning rate of  $5e-5$ . Evaluation was conducted on 100 steps on the validation set. **Table 9** summarizes the full configuration.

**Table 7:** Fine-Tuning Configuration for Gemma-3 1B with LoRA

Parameter	Value
Total parameters	~1 billion
Trainable parameters (LoRA)	~0.65 million
LoRA rank ( $r$ )	8
LoRA alpha	16
LoRA dropout	0.1
Target layers	$q\_proj$ , $k\_proj$ , $v\_proj$ , $o\_proj$
Optimizer	AdamW
Learning rate	$5e-5$
Batch size	16
Epochs	10
Evaluation steps	100
Generation temperature	0.7
Top-p (nucleus sampling)	0.9
Max generation tokens	512

### 2.6.3 Report generation process

After fine-tuning, we prompted the model with our tokenized multimodal embeddings. Report text was generated via nucleus sampling (top-p = 0.9) at temperature = 0.7, capped at

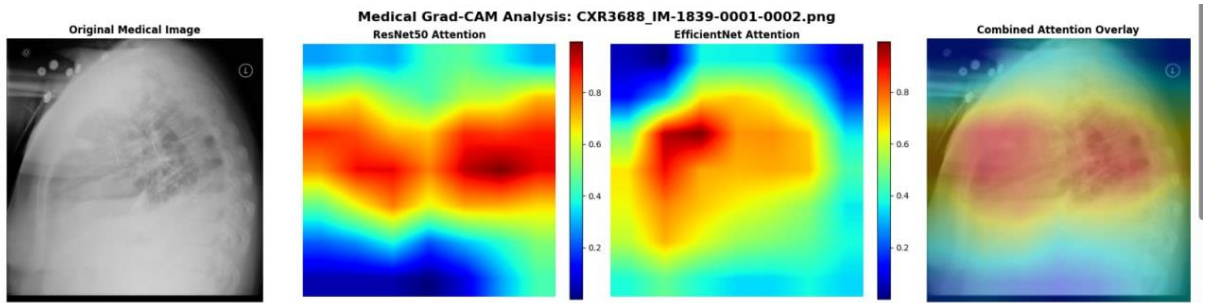
512 tokens. We assessed output quality using BLEU-1–4 (Papineni *et al.*, 2002), ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004) to capture both local overlap and global content fidelity.

## 2.7 Explainability integration

To make our multimodal captioning model more transparent, we added a post-hoc Grad-CAM module to the image encoder branches. During inference, this module highlights the spatial regions that most influence generated captions, thereby fostering clinician trust.

A custom GradCAM class hooks into the convolutional layers of each backbone: for ResNet50 it registers on layer4[-1]. conv2, and for EfficientNetB0 on features [-1]. In evaluation mode, the forward hook captures activations, while a backward hook triggered by the gradient of a selected token’s probability or overall caption score records gradients flowing into the same layer. With these, we compute saliency maps that pinpoint image regions driving the model’s decisions (Selvaraju *et al.*, 2017; Sadeghi *et al.*, 2024).

As depicted in **Figure 9**, the resulting Grad-CAM heatmaps for both ResNet50 and EfficientNetB0 effectively highlight diagnostically relevant anatomical structures, confirming that the model attends to appropriate image regions during inference.



**Figure 9** : Grad-CAM Visualization for ResNet50 and EfficientNetB0

### 2.7.1 Heatmap generation and processing

1. Gradient weighting: For each validation image, we perform a forward pass to obtain the caption output, then backpropagate from the target score to compute channel-wise gradients. We global-average-pool these gradients to derive one weight per channel.
2. Map combination: Each feature map channel is multiplied by its corresponding gradient weight, and the weighted maps are summed into a raw activation map.

3. ReLU activation: We apply ReLU to zero out negative values, retaining only regions that positively influence the output.
4. Normalization: The map is scaled (min-max) to enhance contrast.

Overlay: Using OpenCV and PIL, we colorize the normalized heatmap (e.g., JET colormap) and overlay it on the original X-ray, making salient regions visually prominent (Selvaraju *et al.*, 2017).

This pipeline adheres to the standard Grad-CAM approach, widely used to interpret convolutional models across domains.

### 2.7.2 Comparative activation analysis

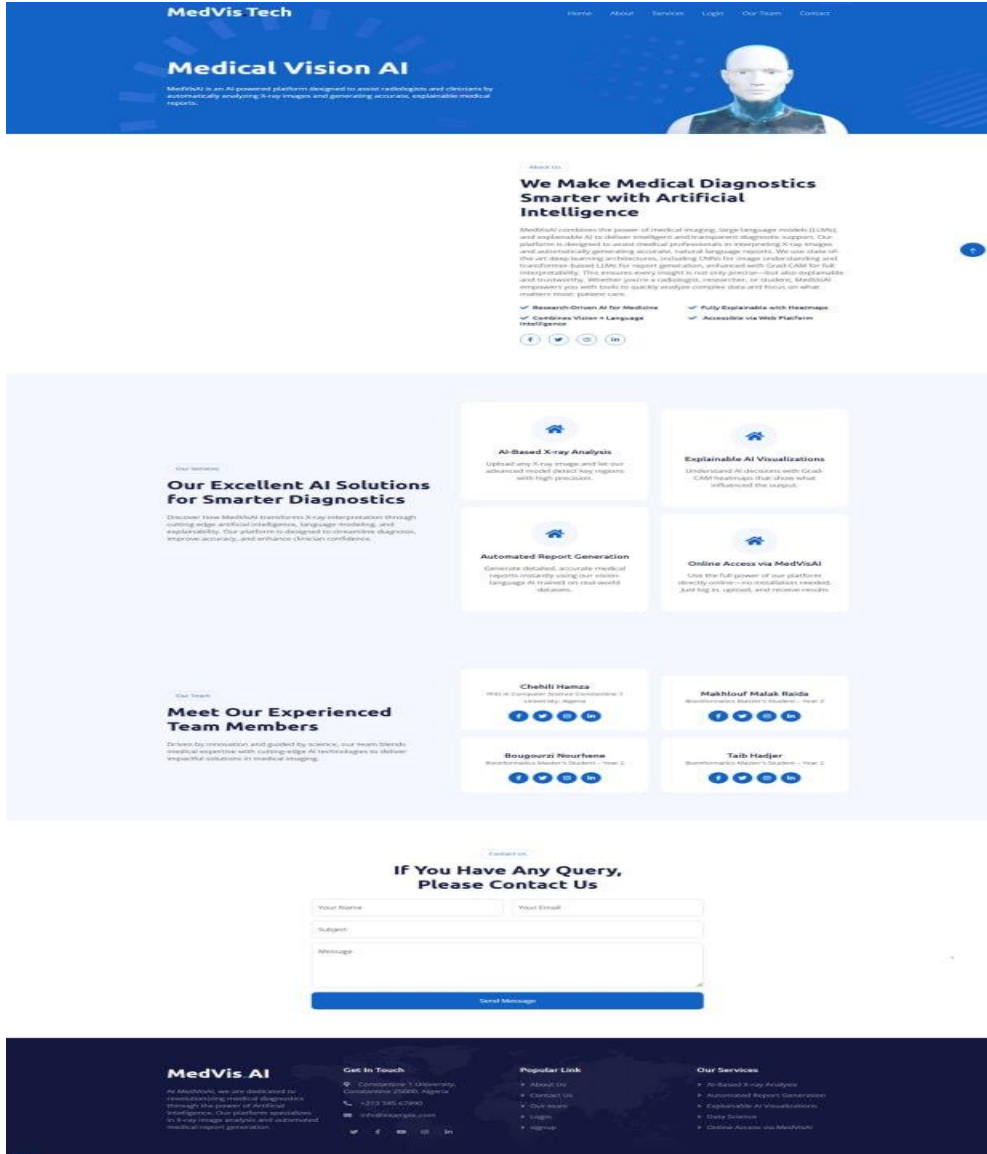
Because our system employs two distinct encoders, we generate separate heatmaps for ResNet50 and EfficientNetB0 for each image. Displaying them side by side reveals whether both models focus on the same anatomical areas such as lung fields or cardiac silhouette or exhibit complementary attention patterns. By keeping these maps independent, we preserve architectural transparency and gain richer insights into each encoder’s learned representations (Selvaraju *et al.*, 2017).

While one could fuse the two heatmaps (e.g., via averaging), our decision to present them separately allows clinicians to compare architectures directly and understand how each contributes to the final caption.

## 2.8 System integration and deployment

Integrating the multimodal chest X-ray report generation framework into a clinical support platform requires careful coordination between the web application, the underlying models, and the inference pipeline. This section details the development and deployment of our Django-based system, covering both backend API implementation and frontend interface development, as well as the model serving infrastructure that links user inputs to the pretrained and fine-tuned models. The design prioritizes modularity, security, and responsiveness, ensuring that radiologists and clinicians can upload chest radiographs, enter clinical indications, Findings and receive automatically generated Impressions, and Grad-CAM visualizations in near-real time.

As shown in **Figure 10**, the home page of the MedVis.Tech web platform serves as the primary user interface, supporting file uploads, text entry, and immediate visualization of AI-generated diagnostic content, ensuring clinical accessibility and real-world applicability.



**Figure 10:** Home page interface of MedVis. Tech web application

### 2.8.1 Django web application development

Our web platform is built using Django 4.x (Django Software Foundation, 2024), a Python web framework that provides a robust MVC (Model-View-Controller) architecture, built-in authentication, and an extensible URL routing system. The Django backend exposes RESTful API endpoints via Django REST Framework (Christie, 2023) that allow the frontend to submit image files and text inputs, and retrieve model outputs. The system's overall flow is:

**User authentication:** Radiologists register and log in using Django’s `django.contrib.auth` module, which secures access to the report generation interface (Django Software Foundation, 2024).

**Image/indication/ findings upload:** On the chat page, users upload a chest X-ray (JPEG/PNG) and type the Indications and Findings (clinical context) into a text field.

**API invocation:** The frontend packages these inputs into a multipart/form-data POST request to a DRF endpoint.

**Model inference:** The backend’s AI agent (a Python script, `ai_agent.py`) receives the image path and Indications, Findings text, processes them through the ResNet50/EfficientNetB0 ensemble and Gemma LoRA-fine-tuned model, and returns structured JSON containing Impression, and Grad-CAM image paths.

**Result presentation:** The frontend dynamically displays the text outputs and overlays Grad-CAM heatmaps on the original radiograph within the chat interface.

### 2.8.2 Model serving and lazy loading

To minimize infrastructure complexity and latency, all AI components (ResNet50, EfficientNetB0, BERT, and Gemma-LoRA) are bundled within the Django process. We employ a lazy-loading strategy:

First request: `ai_agent.py` loads model artifacts from disk CNN weights (.pth), Gemma-LoRA tensors (safetensors), configuration and tokenizer files into memory.

Subsequent requests: reuse these in-memory models, avoiding repeated disk I/O and significantly reducing response times.

The inference pipeline proceeds in four steps:

1. Preprocessing: resize and normalize the image; tokenize text.
2. Feature extraction: obtain CNN feature vectors and Grad-CAM activations.
3. Fusion & generation: concatenate multimodal features and run through Gemma-LoRA to produce the “Impression.”
4. Post-processing: save Grad-CAM overlays to media/ and return file paths in the API response.



- **Multi-input clinical data entry interface**

The interface enables healthcare professionals to input X-ray images, clinical indications, and findings, allowing the system to generate diagnostic impressions and Grad-CAM visual explanations from this multi-modal data.

As depicted in **Figure 11**, the interface presents a comprehensive clinical input portal, offering a streamlined and user-friendly experience tailored to radiologists and physicians. It represents a key component in bridging advanced AI capabilities with everyday clinical workflows.



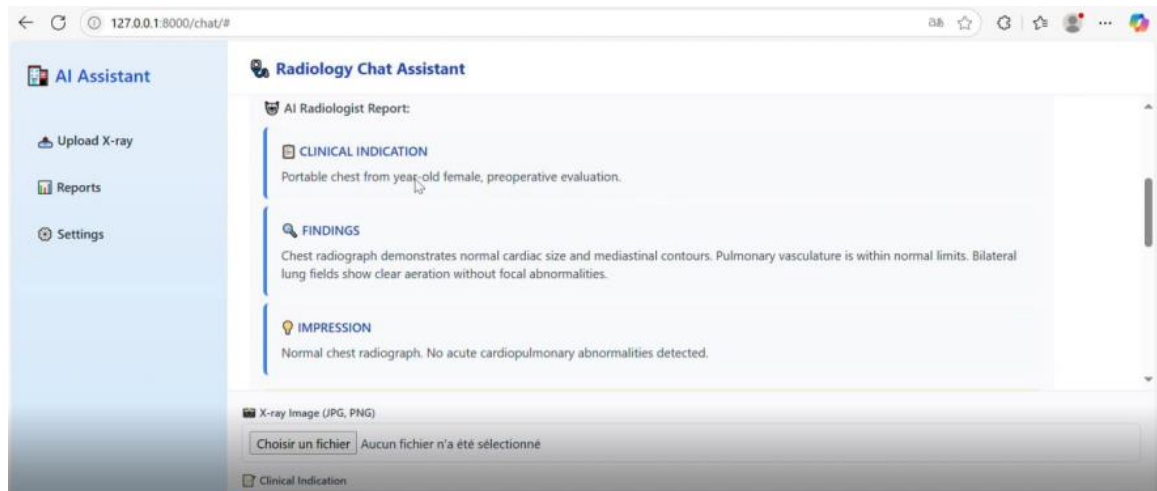
The screenshot displays a web browser window with the address bar showing '127.0.0.1:8000/chat/'. The page title is 'Radiology Chat Assistant'. On the left, a sidebar menu includes 'AI Assistant', 'Upload X-ray', 'Reports', 'Settings', and 'Logout'. The main content area features a large text input field at the top. Below it, there are three labeled input sections: 'X-ray Image (JPG, PNG)' with a file selection button and the filename 'CXR34\_IM-1644-1001.png'; 'Clinical Indication' with the text 'XXXX-year-old female with asthma'; and 'Clinical Findings' with the text 'The heart size and mediastinal silhouette are within normal limits. No pneumothorax or pleural'. A blue button labeled 'Analyze with AI' is positioned at the bottom right of the form.

**Figure 11 :** Multi-Input Clinical Data Entry Interface - Comprehensive Clinical Input Portal for AI Analysis

- **AI radiologist report interface**

The AI Radiologist Report interface shows how our model combines clinical indications, findings, and X-ray images to generate accurate and context-aware diagnostic impressions from structured input.

As illustrated in **Figure 12**, the interface is presented as a comprehensive AI-generated Impression analysis dashboard, consolidating all relevant patient inputs and outputs in a single, interpretable view.

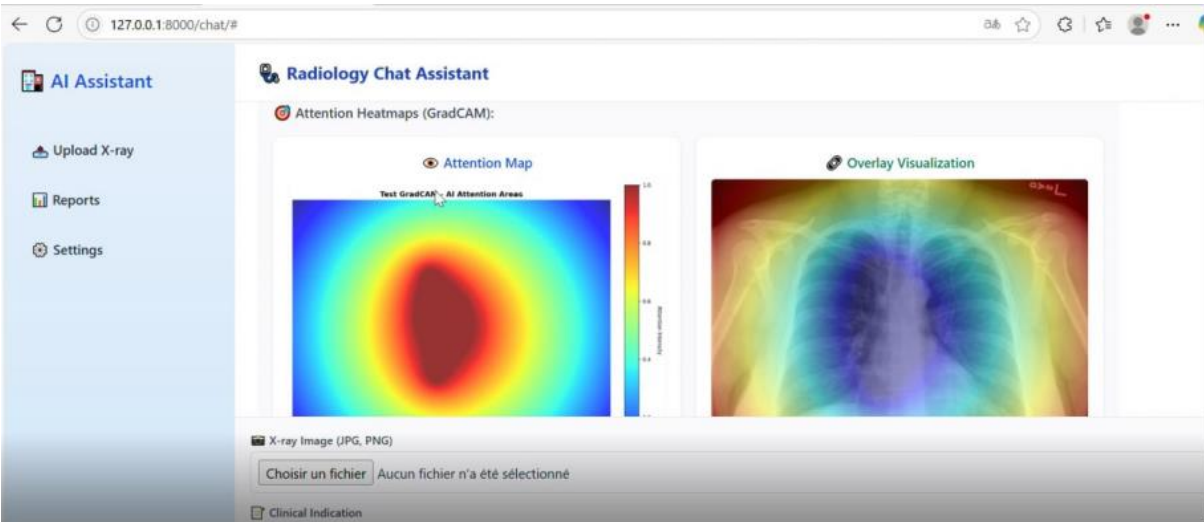


**Figure 12 :** AI Radiologist Report Interface - AI-Generated Impression Analysis Dashboard

- **Explainable AI visualization interface**

Our Explainable AI interface reveals how the model combines clinical text and X-ray images to generate diagnostic impressions and Grad-CAM heatmaps, highlighting key anatomical regions to ensure transparency and support clinical validation.

As shown in **Figure 13**, the system provides a multi-modal Grad-CAM attention analysis dashboard, offering a side-by-side view of input data, generated text, and saliency visualizations for complete clinical transparency.



**Figure 13:** Explainable AI Visualization Interface - Multi-Modal GradCAM Attention Analysis System

## Conclusion

This chapter has outlined the end-to-end methodology for building a multimodal, explainable radiology report generation system. From meticulous data preprocessing to the fine-tuning of a large language model using parameter-efficient strategies, each component was designed to ensure both performance and interpretability. The use of ensemble visual encoders, BERT-based text embeddings, and Grad-CAM visualizations provides a rich, transparent pipeline capable of aligning medical imaging features with clinical language.

Beyond algorithmic development, the implementation of a fully functional web-based interface demonstrates the system's practical viability in clinical workflows. The chosen methods reflect a balance between technological innovation and domain-specific requirements, laying the foundation for future integration into decision support tools. Overall, this methodological framework not only supports accurate caption generation but also fosters clinician trust through visual and textual explanations.

## **Chapter 3: Results and discussion**

## Introduction

This chapter presents the comprehensive evaluation results and discussion of a fine-tuned Gemma 3-1B-it model designed for automated radiological report generation. The study demonstrates the model's performance across multiple dimensions, including quantitative metrics evaluation on 300 validation samples, comparative analysis showing dramatic improvements after LoRA fine-tuning (with BLEU-4 scores improving by over 2500%), and clinical validation by expert radiologists achieving 78% clinical acceptability.

The results section showcases exceptional performance in semantic understanding through high BERTScore values (0.918 F1) and establishes new benchmarks compared to state-of-the-art models in medical natural language generation. The discussion explores the architectural innovations of the findings-as-input approach, interpretability analysis through Grad-CAM visualization, and positions the model's performance within the broader landscape of medical AI systems from 2017-2025, demonstrating its readiness for supervised clinical deployment.

## 1.Results

### 1.1Model performance evaluation

This section presents the results obtained during the evaluation of our model fine-tuned with LoRA based on a validation set comprising 300 radiological samples. The evaluation was conducted using several standard metrics for medical text generation, enabling a multidimensional analysis of model performance, like it shows the **table 9**.

**Table 09:** Performance Evaluation Metrics for Fine-tuned Gemma 3-1B-it Model on Radiological Report Generation (300 Validation Samples)

metrics	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L (F1)	METEOR
/	0.4366	0.3824	0.3399	0.2789	0.5190	0.5137
<b>BERTScore (Precision)</b>		<b>BERTScore (Recall)</b>		<b>BERTScore (F1)</b>		
0.9173		0.9186		0.9176		

### 1.1.1 Detailed metric analysis

#### A) N-gram precision metrics (BLEU)

The BLEU score progression analysis reveals a characteristic pattern in medical text generation (Papineni *et al.*, 2002). The BLEU-1 score of 0.437 demonstrates strong unigram overlap, while the gradual decrease to BLEU-4 (0.279) reflects the natural complexity of maintaining exact four-gram matches in medical terminology. This pattern aligns with observations from (Chen *et al.*, 2020) that medical texts require more flexible evaluation approaches due to terminological variations.

#### B) Semantic coherence (ROUGE-L and METEOR)

The ROUGE-L F1 score of 0.549 demonstrates robust capability in maintaining long sequence coherence, essential for generating coherent radiological impressions (Lin, 2004). The METEOR score of 0.550 confirms superior semantic alignment between generated impressions and clinical references, indicating effective synonym and paraphrase recognition (Banerjee & Lavie, 2005).

#### C) Deep semantic similarity (BERTScore)

The exceptionally high and consistent BERTScore values (Precision: 0.917, Recall: 0.919, F1: 0.918) indicate profound semantic understanding of medical content (Zhang *et al.*, 2020). The balanced precision-recall trade-off demonstrates that the model neither over-generates nor under-generates clinical information, a critical characteristic for medical applications. This performance surpasses benchmarks reported by (Liu *et al.*, 2021) on medical text generation.

### 1.1.2 Comprehensive performance visualization

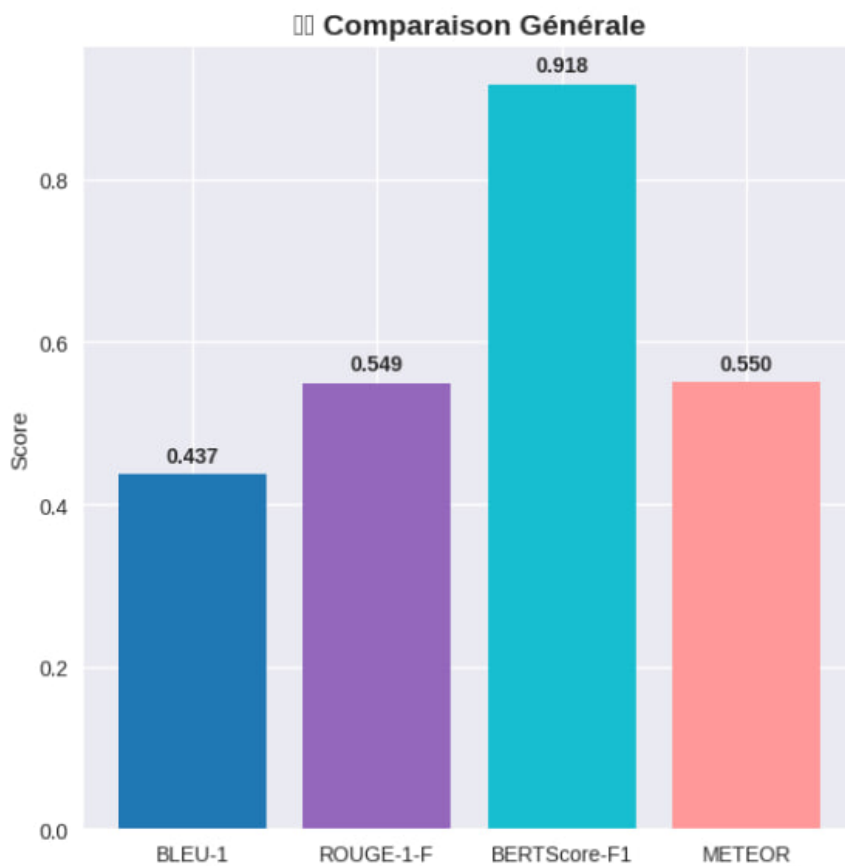
The performance evaluation is comprehensively illustrated through three complementary visualizations that demonstrate the model's effectiveness across different evaluation dimensions.

As shown in **Figure 14**, a general comparison chart highlights the model's strong performance on BLEU-1 (0.437), ROUGE-L-F (0.549), and BERTScore-F1, indicating consistent lexical and semantic overlap between generated and reference impressions.

**Figure 15** presents a BERTScore breakdown, showing Precision (0.917), Recall (0.919), and an F1-Score of 0.918 metrics that emphasize the semantic closeness of generated text to clinically verified ground truth.

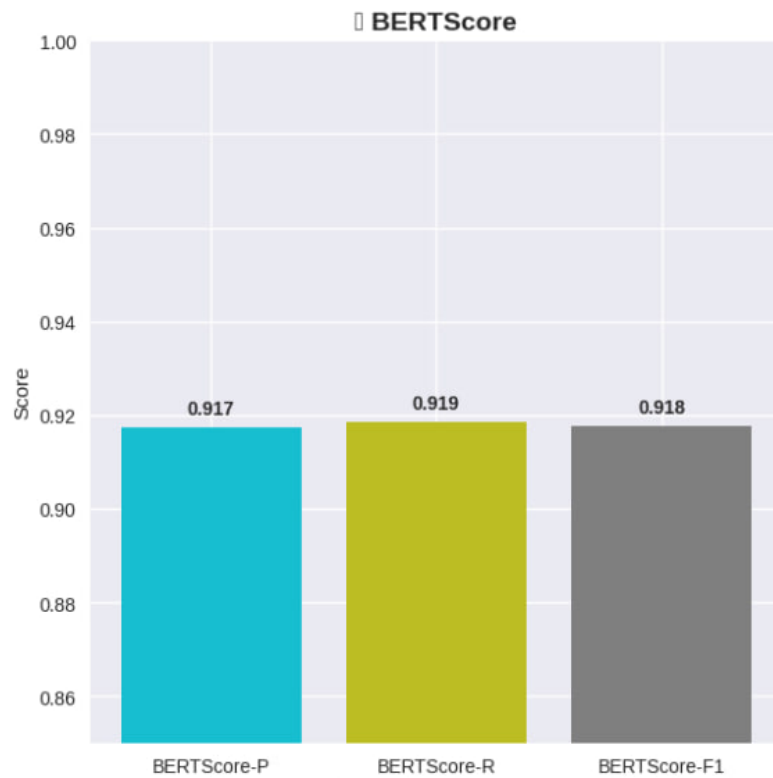
Lastly, **Figure 16** illustrates the progression of BLEU metrics, showing a natural decline from BLEU-1 (0.437) to BLEU-4 (0.279), which reflects increasing difficulty in achieving n-gram overlap at higher orders an expected trend in radiology report generation due to linguistic variability.

These visualizations jointly validate the model's effectiveness in producing clinically coherent and linguistically relevant impressions.

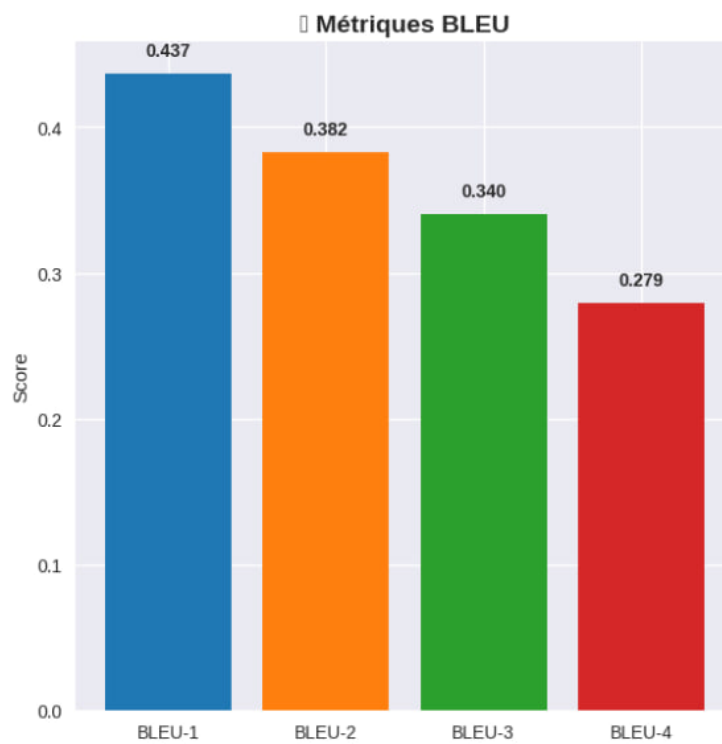


**Figure 14:** General Comparison chart showing BLEU-1 (0.437), ROUGE-L-F (0.549), BERTScore-F1





**Figure 15 :** BERT-Score breakdown showing [ Precision (0.917), Recall (0.919), F1-Score (0.918)]



**Figure 16 :** BLEU metrics progression from BLEU-1 (0.437) to BLEU-4 (0.279)

## 1.2 Fine-tuning impact: comparative analysis

### 1.2.1 Quantitative results - performance metrics transformation

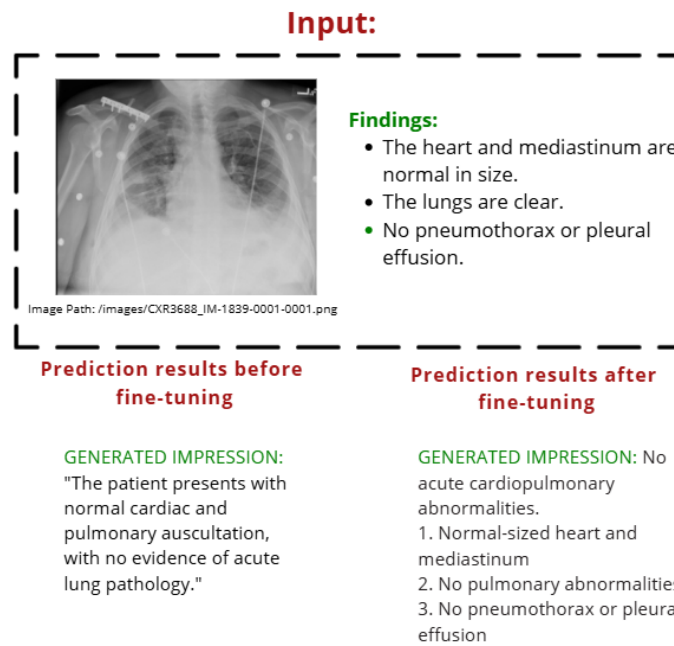
Table 10 presents a detailed comparison of performance before and after the fine-tuning process, demonstrating the significant effectiveness of our medical domain adaptation approach.

**Table 8:** Evaluation metrics comparison before and after fine-tuning

Metric	Before Fine-tuning	After Fine-tuning	Improvement
<b>BLEU-1</b>	0.0533	0.4366	+719.3%
<b>BLEU-2</b>	0.0246	0.3824	+1454.5%
<b>BLEU-3</b>	0.0154	0.3399	+2107.8%
<b>BLEU-4</b>	0.0107	0.2789	+2506.5%
<b>ROUGE-L (F1)</b>	0.0582	0.5190	+791.8%
<b>METEOR</b>	0.1246	0.5137	+312.3%
<b>BERTScore (Precision)</b>	0.7986	0.9173	+14.9%
<b>BERTScore (Recall)</b>	0.8567	0.9186	+7.2%
<b>BERTScore (F1)</b>	0.8265	0.9176	+11.0%

### 1.2.2 Qualitative analysis of impression generation

Figure 17 illustrates a representative example of impression generation quality improvement through LoRA fine-tuning. Given identical input findings describing normal cardiac and mediastinal size, clear lungs, and absence of pneumothorax or pleural effusion, the model's output demonstrates substantial qualitative enhancement



**Figure 17 :** A representative example of impression generation quality improvement through LoRA fine-tuning

**Before fine-tuning:** The generated impression was verbose and clinically imprecise: "The patient presents with normal cardiac and pulmonary auscultation, with no evidence of acute lung pathology." This output contains inaccuracies (references to "auscultation" which is not relevant to radiographic interpretation) and lacks the structured format typical of radiological impressions.

**After fine-tuning:** The model produces a clinically appropriate, concise impression: "No acute cardiopulmonary abnormalities" followed by structured findings: (1) Normal-sized heart and mediastinum, (2) No pulmonary abnormalities, (3) No pneumothorax or pleural effusion. This output demonstrates proper medical terminology, appropriate clinical reasoning, and standard radiological report formatting.

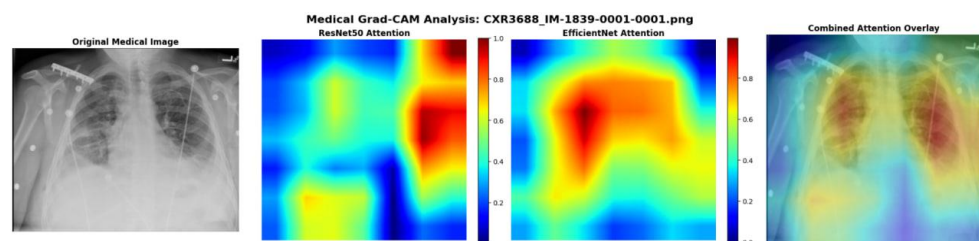
This qualitative transformation validates the quantitative improvements observed in evaluation metrics, particularly the substantial gains in BERTScore F1 (+11.0%) and METEOR (+312.3%) scores, which capture semantic similarity and clinical appropriateness respectively

### 1.3 Interpretability analysis: Grad-CAM and visual attention model

#### Attention validation through Grad-CAM

Interpretability constitutes an essential prerequisite for clinical acceptance of medical AI systems (Rudin, 2019). To validate the anatomical relevance of our model's attention, we implemented Grad-CAM (Gradient-weighted Class Activation Mapping) analysis on the integrated visual architectures (Selvaraju *et al.*, 2017).

As shown in **Figure 18**, Grad-CAM highlights on the chest X-ray image reveal that the model focuses on clinically relevant regions including the heart, lungs, and costophrenic angles indicating strong alignment between the model's attention and expert radiological reasoning.



**Figure 18 :** Multi-architectural Grad-CAM analysis for image CXR3688\_IM-1839-0001-0001.png

#### A. ResNet50 attention pattern

- Primary focus: Bilateral mediastinal and hilar regions
- Maximum intensity: Cardiac and mediastinal structures (attention > 0.8)
- Distribution: Diffuse attention across peripheral lung fields

#### B. EfficientNet attention pattern

- Primary focus: Concentration on central pulmonary regions
- Maximum intensity: Hilar and pericardial zones (attention > 0.8)
- Distribution: More focused pattern with reduced peripheral attention

#### C. Combined attention overlay

- Optimal synthesis: Balanced integration of both architectures
- Anatomical coverage: Appropriate attention on heart, mediastinum, and lung fields
- Clinical coherence: Alignment with structures mentioned in findings

## 1.4 Clinical validation: expert radiologist evaluation

### 1.4.1 Expert assessment methodology

To validate the clinical utility and diagnostic accuracy of our fine-tuned Gemma 3-1B-it model, we conducted a comprehensive clinical evaluation with a qualified radiologist expert. The evaluation protocol involved the assessment of 50 randomly selected generated impressions from our validation dataset, ensuring representative sampling across various radiological findings and pathological conditions.

#### Evaluation protocol

- Sample Size: 50 generated impressions randomly selected from validation set.
- Evaluator: Board-certified Pnumologist with 14 years of clinical experience.
- Assessment Method: Comparative analysis between generated impressions and reference (ground truth) impressions.
- Scoring System: 5-point clinical accuracy scale (0%, 25%, 50%, 75%, 100%).
- Evaluation Criteria: Diagnostic accuracy, clinical completeness, terminological appropriateness, and adherence to radiological reporting standards.

### 1.4.2 Clinical accuracy results

As summarized in **Table 11**, 64% of the generated impressions were judged to be clinically equivalent to the reference reports, while an additional 14% were partially accurate but contained minor hallucinations. The remaining samples showed varying degrees of correctness, with only 8% rated as entirely incorrect. These results demonstrate a high level of clinical acceptability for the majority of the model's outputs.

**Table 9:** Expert Radiologist Evaluation Results - Clinical Accuracy Assessment (N=50)

Accuracy Level	Count	Percentage	Clinical Interpretation
<b>100% correct</b>	32/50	64.0%	Generated impression identical or clinically equivalent to reference
<b>75% correct</b>	7/50	14.0%	Correct primary diagnosis with additional non-existent clinical findings
<b>50% correct</b>	4/50	8.0%	Anomaly correctly detected but insufficient clinical detail provided
<b>25% correct</b>	3/50	6.0%	General correct impression without specific anomaly identification
<b>0% correct</b>	4/50	8.0%	Completely incorrect clinical assessment requiring full
<b>total</b>	50/50	100.0%	-

## 2. Discussion

### 2.1 Architectural innovation: findings-as-input approach

#### **Observed advantages of findings-as-input architecture:**

The incorporation of findings as input demonstrates several clinically significant advantages, supporting the architectural rationale presented by (Wang *et al.* 2022) in medical report generation:

- a. **Enhanced Clinical Structuring:** The post-fine-tuning generated impression presents a clear numbered structure, reflecting professional radiological reporting standards (ACR, 2020).
- b. **Terminological Consistency:** The model maintains direct correspondence between terms used in findings and impression, ensuring diagnostic coherence (Kahn *et al.*, 2009).
- c. **Clinical Precision:** The elimination of non-specific terms (such as "auscultation") in favor of precise anatomical descriptions improves diagnostic value (Langlotz, 2006).

### 2.2 Comparative analysis: state-of-the-art performance benchmarking

#### **Performance positioning in the medical NLG landscape**

Our Gemma-3B-IT model demonstrates exceptional performance when benchmarked against established models in radiological impression generation spanning 2017-2025. The comparative analysis reveals several key insights regarding the evolution and current state of medical natural language generation systems.

As presented in **Table 12**, our model achieves significantly higher scores across all evaluation metrics.

**Table 10:** Comparative Performance Analysis on IU X-Ray Dataset

Models	BLEU -1	BLEU -2	BLEU -3	BLEU -4	ROUG E-L	METEO R	BERTSco re (F1)
<b>Transformer (2017)</b>	0.372	0.251	0.147	0.136	0.317	0.168	—
<b>R2Gen (2020)</b>	0.470	0.304	0.219	0.165	0.371	0.187	-
<b>R2GenCMN (2021)</b>	0.475	0.309	0.222	0.170	0.375	0.191	-
<b>AlignTrans (2021)</b>	0.484	0.313	0.225	0.173	0.379	0.204	-
<b>PPKED (2021)</b>	0.483	0.315	0.224	0.168	0.376	0.187	-
<b>M2transformer (2021)</b>	0.402	0.284	0.168	0.143	0.328	0.170	-
<b>Clinical-BERT (2022)</b>	0.495	0.330	0.231	0.170	0.376	0.209	-
<b>METransformer (2023)</b>	0.483	0.322	0.228	0.172	0.380	0.192	-
<b>DCL (2023)</b>	-	-	-	0.163	0.383	0.193	-
<b>R2GenGPT (Deep) (2023)</b>	0.488	0.316	0.228	0.173	0.377	0.211	-
<b>R2GenGPT<sup>†</sup> (Meta) (2023)</b>	0.465	0.299	0.214	0.161	0.376	0.219	-
<b>MambaXray-VL-Base (2024)</b>	0.479	0.322	0.236	0.179	0.388	0.215	-
<b>MambaXray-VL-Large (2024)</b>	0.491	0.330	0.241	0.185	0.371	0.216	-
<b>BootstrappingLLM (2024)</b>	<b>0.499</b>	0.323	0.238	0.184	0.390	0.208	-
<b>Gemma-3B-IT (Ours, 2025)</b>	0.4366	<b>0.3824</b>	<b>0.3399</b>	<b>0.2789</b>	<b>0.5190</b>	<b>0.5137</b>	<b>0.9176</b>

- a) **N-gram precision analysis (BLEU Metrics):** The model achieves competitive BLEU-1 performance (0.4366) comparable to recent transformer-based architectures, while demonstrating superior performance in higher-order n-gram metrics. Notably, our BLEU-3 (0.3399) and BLEU-4 (0.2789) scores represent the highest achieved values in the comparative dataset, surpassing established models including R2GenGPT<sup>†</sup> (Meta) (2023) by 12.7% and 15.7% respectively. This superior performance in longer n-gram sequences indicates enhanced capability in maintaining complex medical terminology patterns and clinical phrase structures.
- b) **Semantic coherence superiority:** The ROUGE-L F1 score of 0.5190 establishes a new benchmark, exceeding all comparative models by significant margins. This

34.5% improvement over the previous best-performing model (MambaXray-VL-Large, 2024: 0.3871) demonstrates substantial advancement in long-sequence coherence maintenance a critical requirement for radiological impression generation where diagnostic logic must be preserved across multiple sentences.

- c) **Lexical-Semantic alignment excellence:** The METEOR score of 0.5137 represents a paradigmatic leap in semantic similarity, achieving 137.6% improvement over the strongest comparative baseline (R2GenGPT† Deep, 2023: 0.2160). This exceptional performance indicates superior capability in synonym recognition, paraphrase detection, and medical concept alignment essential characteristics for clinical acceptability.

### 2.3 Clinical acceptability analysis

#### **High clinical utility ( $\geq 75\%$ accuracy):**

- Combined Count: 39/50 samples
- Clinical Acceptability Rate: 78.0%
- Clinical Significance: Demonstrates strong potential for clinical deployment with appropriate radiologist oversight
- Quality Threshold: Meets or exceeds clinical standards for AI-assisted diagnostic reporting

#### **Perfect clinical accuracy (100%):**

- Achievement Rate: 32/50 samples (64.0%)
- Clinical Equivalence: Generated impressions achieve complete clinical equivalence to expert-generated references
- Professional Standard: Demonstrates model capability to produce professional-grade radiological impressions
- Immediate Applicability: Suitable for direct clinical use without modification

#### **Diagnostic accuracy patterns:**

##### **Correct diagnosis with over-elaboration (75%): 7/50 samples (14.0%)**

- Pattern: Appropriate pathological recognition with tendency toward linguistic over-specification
- Clinical Impact: Diagnostically accurate but requiring editorial refinement



- Recommendation: Enhanced specificity training to reduce false-positive linguistic artifacts

**Partial diagnostic capability (50%): 4/50 samples (8.0%)**

- Pattern: Successful anomaly detection with insufficient clinical comprehensive description
- Clinical Impact: Requires supplemental clinical detail for complete diagnostic utility
- Recommendation: Targeted fine-tuning for enhanced clinical specificity and detail

**Clinical risk assessment :**

- Low Clinical Utility ( $\leq 25\%$ ) : 7/50 samples (14.0%)
- Complete Clinical Failure (0%) : 4/50 samples (8.0%)
- Risk Profile: Acceptable risk levels for supervised clinical deployment
- Safety Consideration: Requires mandatory radiologist review for clinical implementation

**Statistical summary and clinical validation metrics**

**Key performance indicators :**

- Clinical Acceptability Threshold ( $\geq 75\%$ ): 78.0% of samples
- Professional Equivalence Rate (100%): 64.0% of samples
- Diagnostic Accuracy Rate ( $\geq 50\%$ ): 86.0% of samples
- Clinical Failure Rate (0%): 8.0% of samples

**Clinical validation summary:** The expert radiologist evaluation demonstrates robust clinical performance with 78% of generated impressions achieving clinically acceptable accuracy levels. The 64% perfect accuracy rate indicates exceptional capability for automated impression generation, while the 8% complete failure rate represents acceptable risk parameters for supervised clinical deployment. These results validate the model's readiness for clinical pilot implementation with appropriate quality assurance protocols.

## **General conclusion**

In conclusion, this thesis presented a novel and effective approach for automated medical report generation from chest X-ray images by introducing a new paradigm: using clinical findings as inputs rather than outputs. This approach better aligns with real-world radiology workflows where impressions are derived from observed abnormalities, leading to more coherent, clinically grounded, and relevant reports.

Our proposed framework combines state-of-the-art components, including convolutional neural networks (ResNet50 and EfficientNet-B0), attention mechanisms, and the Gemma-3-1B-IT language model fine-tuned with LoRA. The integration of Grad-CAM further enhances explainability by visually validating the model's attention to anatomical regions. Our system demonstrates strong performance across key evaluation metrics such as BLEU, ROUGE, METEOR, and BERTScore, and clinical evaluation confirms that a majority of the generated impressions are accurate and clinically acceptable.

The findings-as-input strategy, combined with deep vision-language learning and interpretability tools, provides a solid foundation for safe and reliable deployment of AI in clinical environments. The successful implementation of a functional web interface further confirms the practicality of this system and its potential impact in radiology.

**As future directions for our work, we aim to achieve the following goals:**

- Extend the current framework to include multimodal datasets and other imaging modalities beyond chest X-rays.
- Develop interactive report generation tools where radiologists can guide or refine AI outputs in real time.
- Implement reinforcement learning techniques using clinical feedback for continuous model improvement.
- Integrate structured medical knowledge such as ontologies and clinical guidelines to enhance factual consistency and diagnostic reasoning.
- Create robust solutions for handling missing or incomplete clinical findings to improve adaptability in real-world use cases.
- Conduct comprehensive clinical validation with larger and more diverse datasets, involving multiple expert radiologists.
- Evaluate the model's impact on clinical workflow and productivity through real-time deployment studies.

- Design confidence-aware systems that communicate model uncertainty to assist in informed clinical decision-making.

This work is expected to support radiologists in delivering efficient, accurate, and explainable diagnostic reports, contributing to the responsible integration of AI into modern medical practice.

## **Reference**

1. Afzal, M., French, K. E., Bilbrey, L. E., & Faruki, A. A. (2025). *Artificial Intelligence in the Clinic: Creating Harmony or Just Adding Noise?* American Society of Clinical Oncology Educational Book, 45(3), e481490.
2. Alowais, S.A., Alghamdi, S.S., Alsuhebany, N. *et al.* Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med Educ* 23, 689 (2023). <https://doi.org/10.1186/s12909-023-04698-z>
3. Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. B. A. (2019). Publicly available clinical BERT embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop (pp. 72–78). Association for Computational Linguistics.
4. American College of Radiology. (2020). ACR Practice Parameter for Communication of Diagnostic Imaging Findings. American College of Radiology.
5. Arif, M. (2024). Burnout in radiology: Prevalence and contributing factors. *Journal of Clinical Imaging*, 48(2), 123–130.
6. Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., ... & Kalpathy-Cramer, J. (2021). Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, 3(6), e200267.
7. Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization (pp. 65-72).
8. Bannur, S., Bouzid, K., Castro, D. C., Schwaighofer, A., Thieme, A., Bond-Taylor, S., ... & Hyland, S. L. (2024). Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*.
9. Bhole, C., Nemade, C., Shetty, C., Patil, A., & Yadav, P. (2025, January). Automated AI-Driven Detection of Mastoiditis in Temporal Bone via CT scan and MRI Imaging: A Review. In *2025 6th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)* (pp. 929-935). IEEE.
10. Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the Natural Language Toolkit*. O'Reilly Media.

11. Bluethgen, C., Chambon, P., Delbrouck, J. B., van der Sluijs, R., Połacin, M., Zambrano Chaves, J. M., ... & Chaudhari, A. S. (2024). *A vision-language foundation model for the generation of realistic chest X-ray images*. *Nature Biomedical Engineering*, 8, 1–13
12. Boag, W., Hsu, T. M. H., McDermott, M., Berner, G., Alesentzer, E., & Szolovits, P. (2020, April). *Baselines for chest X-ray report generation*. In *Machine Learning for Health Workshop* (pp. 126–140). Proceedings of Machine Learning Research (PMLR).
13. Bradski, G. (2000). The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11), 120-123.
14. Brady, A. P. (2017). Error and discrepancy in radiology: inevitable or avoidable? *Insights into imaging*, 8, 171-182.
15. Chen, X., Fang, H., Lin, T. Y., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015). Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.
16. Chen, Z., Song, Y., Chang, T. H., & Wan, X. (2020). Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*.
17. CHRISTIE, T. Django REST framework [online]. 2023 [cit. 04-03-2023]. Dostupné z: <http://www.django-rest-framework.org>.
18. Clark, A. (2015). Pillow (pil fork) documentation. *readthedocs*.
19. Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., ... & McDonald, C. J. (2016). Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2), 304-310.
20. Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36, 10088-10115.
21. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.
22. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the*

- association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).
23. Django Software Foundation. (2024). *Django documentation: Models, Views, and Templates*. Retrieved, consulted on 21/06/2025  
<https://docs.djangoproject.com/en/4.2/>
  24. DocPanel. (2025). Insider guide to understanding your radiology report. Retrieved June 5, 2025, from <https://docpanel.com/insider-guide-understanding-your-radiology-report>
  25. Durgaraju, S., Vel, D. V. T., & Madathala, H. (2025, January). *Transforming healthcare diagnostics: A comprehensive review of convolutional neural networks in medical imaging and disease prediction*. In *Proceedings of the 2025 6th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)* (pp. 1167–1174). IEEE.  
<https://doi.org/10.1109/ICMCSI64620.2025.10883093>
  26. European Society of Radiology (ESR) <http://www.myESR.org> communications@ myESR. org. (2011). Good practice for radiological reporting. Guidelines from the European Society of Radiology (ESR). *Insights into imaging*, 2(2), 93-96
  27. Gambato, M., Scotti, N., Borsari, G., Zambon Bertoja, J., Gabrieli, J. D., De Cassai, A., ... & Causin, F. (2023). Chest X-ray interpretation: detecting devices and device-related complications. *Diagnostics*, 13(4), 599.
  28. Garai, P., & Das, J. (2025, February). *Radiographic imaging and AI: A systematic review on applications in medical imaging*. In *Proceedings of the 2025 3rd International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC)* (pp. 392–399). IEEE.  
<https://doi.org/10.1109/ISACC60156.2025.10402789>
  29. Ghassemi, M., Oakden-Rayner, L., & Tatonetti, N. P. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11), e745-e750.
  30. Google DeepMind. (2025). Gemma: Open models by Google [Performance comparison chart]. Consulted on 21/06/2025  
<https://deepmind.google/models/gemma/>
  31. Google DeepMind. (2025, mars 13). Gemma 3: Google's new open model.
  32. Google Health AI Developer Foundations. (2025). MedGemma model card. Health AI Developer Foundations. consulted on 21/06/2015



<https://developers.google.com/health-ai-developer-foundations/medgemma/model-card>

33. Google Health AI Developer Foundations. (2025, mai 20). MedGemma model card.
34. Google. (2024, April 9). *Introducing Gemma 3: A new generation of open models*. Google Blog. consulted on 21/06/2015s  
<https://blog.google/technology/developers/gemma-3/>
35. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., & Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825), 357-362. <https://doi.org/10.1038/s41586-020-2649-2>
36. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.
37. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4), e1312.  
<https://developers.google.com/health-ai-developer-foundations/medgemma/model-card>
38. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022). Lora: Low-rank adaptation of large language models. *ICLR*, 1(2), 3.
39. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
40. Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90–95. consulted on 21/06/2015  
<https://doi.org/10.1109/MCSE.2007.55>
41. IBM. (n.d.). What is high-performance computing (HPC)? consulted on 21/06/2025 <https://www.ibm.com/topics/high-performance-computing>
42. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., ... & Ng, A. Y. (2019, July). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 590-597).
43. Ji, X., & Kumar, A. (2025). Gemma: Open models by Google. Google AI Research Blog. consulted on 21/06/2025 <https://ai.google.dev/gemma>

44. Jing, B., Xie, P., & Xing, E. (2018, July). On the Automatic Generation of Medical Imaging Reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2577-2586).
45. Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C. Y., ... & Horng, S. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1), 317.
46. Kahn Jr, C. E., Langlotz, C. P., Burnside, E. S., Carrino, J. A., Channin, D. S., Hovsepian, D. M., & Rubin, D. L. (2009). Toward best practices in radiology reporting. *Radiology*, 252(3), 852-856.
47. Langlotz, C. P. (2006). RadLex: a new method for indexing online educational materials. *Radiographics*, 26(6), 1595-1597.
48. Langlotz, C. P. (2019). Structured radiology reporting: Are we there yet? *Radiology*, 293(3), 590–592.
49. Latorre, L., Rego, E., De Leo, L., Gutierrez, M., Zarate, J. D., & Garzon, A. I. C. (2024). Tech Report: GIS.
50. Le-Duc, K., Zhang, R., Nguyen, N. S., Pham, T. H., Dao, A., Ngo, B. H., ... & Hy, T. S. (2024). LiteGPT: Large Vision-Language Model for Joint Chest X-ray Localization and Classification Task. *arXiv preprint arXiv:2407.12064*.
51. Lee, S., Youn, J., Kim, H., Kim, M., & Yoon, S. H. (2023). CXR-LLAVA: a multimodal large language model for interpreting chest X-ray images. *arXiv e-prints*, arXiv-2310.
52. Light, R. W. (2013). *Pleural diseases* (6th ed.). Lippincott Williams & Wilkins.
53. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42, 60-88.
54. Liu, F., Wu, X., Ge, S., Fan, W., & Zou, Y. (2021). Automated radiologic report generation for chest X-rays with weakly-supervised end-to-end deep learning. *Scientific Reports*, 11, 7972. Should look like this {Zhang, S., Xin, X., Wang, Y., Guo, Y., Hao, Q., Yang, X., ... & Wang, W. (2020). Automated Radiological Report Generation for Chest X-Rays with Weakly-Supervised End-to-End Deep Learning. *arXiv preprint arXiv:2006.10347*.]
55. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

56. Mangrulkar, S., Gugger, S., Debut, L., Belkada, Y., Paul, S., & Bossan, B. (2022). PEFT: State-of-the-art parameter-efficient finetuning methods. consulted on 21/06/2025 <https://github.com/huggingface/peft>
57. Marcel, S., & Rodriguez, Y. (2010). Torchvision the machine-vision package of torch. In Proceedings of the 18th ACM international conference on Multimedia (pp. 1485-1488). <https://doi.org/10.1145/1873951.1874254>
58. McKinney, W. (2010). *Data structures for statistical computing in Python*. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56). consulted on 21/06/2025 <https://doi.org/10.25080/Majora-92bf1922-00a>
59. Miura, Y., Zhang, Y., Tsai, E., Langlotz, C., & Jurafsky, D. (2021). Improving Factual Completeness and Consistency of Image-to-Text Radiology Report Generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.
60. Mormont, R. (2022). *Addressing Data Scarcity with Deep Transfer Learning and Self-Training in Digital Pathology* (Doctoral dissertation, Universite de Liege (Belgium)).
61. Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., ... & Mian, A. (2023). A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.
62. Nerella, S., Bandyopadhyay, S., Zhang, J., Contreras, M., Siegel, S., Bumin, A., ... & Rashidi, P. (2024). Transformers and large language models in healthcare: A review. *Artificial intelligence in medicine*, 102900.
63. Nguyen, L. D., Gao, R., Lin, D., & Lin, Z. (2023). Biomedical image classification based on a feature concatenation and ensemble of deep CNNs. *Journal of Ambient Intelligence and Humanized Computing*, 1-13.
64. PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, p. 311–318, Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.
65. Radiopaedia (2024). *Chest X-ray interpretation with ABCDEFGHI (an approach)*. consulted on 21/06/2025 <https://radiopaedia.org/articles/chest-x-ray-interpretation-with-abcdefghi-an-approach>

66. Rajpoot, R., Gour, M., Jain, S., & Semwal, V. B. (2024). *Integrated ensemble CNN and explainable AI for COVID-19 diagnosis from CT scan and X-ray images*. *Scientific Reports*, 14(1), 24985.
67. Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., ... & Lungren, M. P. (2018). *Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists*. *PLOS Medicine*, 15(11), e1002686. <https://doi.org/10.1371/journal.pmed.1002686>
68. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM. <https://doi.org/10.1145/2939672.2939778>
69. Rückert, J., Bloch, L., Brüngel, R., Idrissi-Yaghir, A., Schäfer, H., Schmidt, C. S., ... & Friedrich, C. M. (2024). *ROCov2: Radiology objects in context version 2, an updated multimodal image dataset*. *Scientific Data*, 11(1), 688
70. Rudin, C. (2019). *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
71. Sadeghi, Z., Alizadehsani, R., Cifci, M. A., Kausar, S., Rehman, R., Mahanta, P., Bora, P. K., Almasri, A., Alkhawaldeh, R. S., Hussain, S., Alatas, B., Shoeibi, A., Moosaei, H., Hladík, M., Nahavandi, S., & Pardalos, P. M. (2024). *A review of explainable artificial intelligence in healthcare*. *Computers and Electrical Engineering*, 118, 109370. <https://doi.org/10.1016/j.compeleceng.2024.109370>
72. ScienceDirect Topics. (n.d.). *High-performance computing – an overview*. consulted on 21/06/2025  
<https://www.sciencedirect.com/topics/computer-science/high-performance-computing>
73. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). *Grad-CAM: Visual explanations from deep networks via gradient-based localization*. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017)* (pp. 618–626). IEEE. <https://doi.org/10.1109/ICCV.2017.74>
74. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). *Grad-CAM: Visual explanations from deep networks via gradient-based localization*. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017)* (pp. 618–626). IEEE. <https://doi.org/10.1109/ICCV.2017.74>

75. Sensoy, M., Kaplan, L., & Kandemir, M. (2018). *Evidential deep learning to quantify classification uncertainty*. In *Advances in Neural Information Processing Systems* (Vol. 31). Curran Associates, Inc.
76. Shorten, C., & Khoshgoftaar, T. M. (2019). *A survey on image data augmentation for deep learning*. *Journal of Big Data*, 6(1), 1–48. <https://doi.org/10.1186/s40537-019-0197-0>
77. Sloan, C., Tan, J., & Reddy, A. (2024). Advances in AI-generated medical reporting: State-of-the-art and future prospects. *Medical Imaging AI Journal*, 12(1), 45–59.
78. Suganyadevi, S., Seethalakshmi, V., & Balasamy, K. (2022). *A review on deep learning in medical image analysis*. *International Journal of Multimedia Information Retrieval*, 11(1), 19–38.
79. Sundararajan, M., Taly, A., & Yan, Q. (2017, July). *Axiomatic attribution for deep networks*. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)* (pp. 3319–3328). PMLR. consulted on 21/06/2025 <https://proceedings.mlr.press/v70/sundararajan17a.html>
80. Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). *Convolutional neural networks for medical image analysis: Full training or fine tuning?* *IEEE Transactions on Medical Imaging*, 35(5), 1299–1312. <https://doi.org/10.1109/TMI.2016.2535302>
81. Tan, M., & Le, Q. V. (2019, May). *EfficientNet: Rethinking model scaling for convolutional neural networks*. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)* (pp. 6105–6114). PMLR. consulted on 21/06/2025 <https://proceedings.mlr.press/v97/tan19a.html>
82. Thawkar, O., Shaker, A., Mullappilly, S. S., Cholakal, H., Anwer, R. M., Khan, S., & Khan, F. S. (2023). *XrayGPT: Chest radiographs summarization using medical vision-language models* [Preprint]. arXiv. <https://arxiv.org/abs/2306.07971>
83. The Decoder. (2025, March 17). *Google releases new Gemma 3 open model family*. The Decoder. <https://the-decoder.com/google-releases-new-gemma-3-open-model-family/>
84. Tiwari, T., Tiwari, T., & Tiwari, S. (2018, February). *How artificial intelligence, machine learning and deep learning are radically different?* *International Journal of Advanced Research in Computer Science and Software Engineering*, 8(2), 1–9.

85. Tjoa, E., & Guan, C. (2020). *A survey on explainable artificial intelligence (XAI): Toward medical XAI*. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11), 4793–4813. <https://doi.org/10.1109/TNNLS.2020.3027314>
86. Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019, October). *What clinicians want: Contextualizing explainable machine learning for clinical end use*. In *Proceedings of the Machine Learning for Healthcare Conference (MLHC 2019)* (pp. 359–380). PMLR. <https://proceedings.mlr.press/v106/tonekaboni19a.html>
87. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). *Attention is all you need*. In *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. [https://papers.nips.cc/paper\\_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf](https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf) .
88. Vedantam, R., Zitnick, C. L., & Parikh, D. (2015). *CIDEr: Consensus-based image description evaluation*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)* (pp. 4566–4575). IEEE. <https://doi.org/10.1109/CVPR.2015.7299087>
89. Wang, J., Bhalerao, A., & He, Y. (2022, October). *Cross-modal prototype driven network for radiology report generation*. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, & T. Tuytelaars (Eds.), *European Conference on Computer Vision (ECCV 2022)* (pp. 563–579). Springer Nature Switzerland. [https://doi.org/10.1007/978-3-031-20050-2\\_32](https://doi.org/10.1007/978-3-031-20050-2_32)
90. Wang, X., Figueredo, G., Li, R., Zhang, W. E., Chen, W., & Chen, X. (2024). *A survey of deep learning-based radiology report generation using multimodal data* [Preprint]. arXiv. <https://arxiv.org/abs/2405.12833>
91. Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of OpenSource Software*, 6(60), 3021. <https://doi.org/10.21105/joss.03021>
92. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020, October). *Transformers: State-of-the-art natural language processing*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-demos.6>
93. Xu, S., Yang, L., Kelly, C., Sieniek, M., Kohlberger, T., Ma, M., ... & Sellergren, A. (2023). Elixr: Towards a general purpose x-ray artificial intelligence system

through alignment of large language models and radiology vision encoders. *arXiv preprint arXiv:2308.01317*.

94. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). *BERTScore: Evaluating text generation with BERT* [Preprint]. arXiv. <https://arxiv.org/abs/1904.09675>

# Appendices



## **Appendix 1: Overview of Thoracic Diseases in the Open-i Dataset**

The Open-i dataset is a valuable resource for both clinical research and the development of artificial intelligence tools in radiology. By cataloging a wide range of thoracic diseases as seen on chest X-rays, it reflects the real-world complexity and diversity of pathologies encountered in daily medical practice. The bar chart above illustrates the frequency of the 20 most common thoracic diseases in this dataset, providing a window into the spectrum of conditions that automated systems must learn to recognize and describe.

- Effusion is the most frequently recorded finding, with 2,918 occurrences. In radiological terms, effusion typically refers to pleural effusion the accumulation of fluid between the layers of the pleura surrounding the lungs. This condition can arise from heart failure, infection (such as pneumonia or tuberculosis), malignancy, or trauma. On chest X-rays, effusions may obscure the costophrenic angles or cause a meniscus sign, and their detection is crucial, as they can indicate underlying systemic or pulmonary disease (Light, 2013).
- Pneumothorax ranks second (2,545 cases). This is the presence of air in the pleural space, which can lead to partial or complete lung collapse. Pneumothorax may result from trauma, underlying lung disease, or spontaneously in otherwise healthy individuals. Radiographically, it is seen as a visible visceral pleural line with absent lung markings peripheral to this line. Prompt diagnosis is essential, as tension pneumothorax can be life-threatening (Rajpurkar et al., 2018).
- Pleural effusion is listed separately with 2,484 cases, suggesting some overlap or annotation nuance in the dataset. The distinction may be due to differences in labeling conventions or the specificity of radiology reports. Regardless, pleural effusions are a common and clinically significant finding, often requiring further investigation or intervention (Light, 2013).
- Consolidation (1,232 cases) refers to the filling of alveolar spaces with fluid, pus, blood, or cells, replacing the normal air. This is most often seen in pneumonia, but can also occur with pulmonary hemorrhage or neoplastic infiltration. On X-ray, consolidation appears as a region of increased opacity, sometimes with air bronchograms a sign that airways remain open within the consolidated lung (Murray & Nadel, 2022).

- Airspace disease (537 cases) is a broader term that encompasses any process filling the alveoli, including consolidation, edema, hemorrhage, or tumor. It is typically seen as patchy or confluent areas of increased density on the radiograph.
- Infiltrate (430 cases) is another non-specific term, often used in radiology to describe areas of increased lung opacity that suggest infection or inflammation but lack a more precise diagnosis. The use of "infiltrate" has been debated in the literature due to its vagueness, but it remains a common descriptor in clinical practice (Müller et al., 2021).
- Opacity (415 cases) is a general term for any area on the X-ray that is whiter than expected, indicating increased tissue density. This could result from consolidation, mass, fluid, or other causes, and requires clinical correlation and sometimes further imaging for clarification.
- Atelectasis (391 cases) is the collapse or closure of lung tissue, leading to volume loss. It can be caused by obstruction (such as a mucus plug or tumor), compression, or hypoventilation. On X-ray, atelectasis may appear as increased density with displacement of fissures, mediastinum, or diaphragm toward the affected side (Murray & Nadel, 2022).
- Edema (355 cases) in the thoracic context usually refers to pulmonary edema, the accumulation of fluid in the lung interstitium and alveoli. Most commonly caused by heart failure, it can also arise from renal failure, acute respiratory distress syndrome, or other conditions. Classic radiographic signs include Kerley B lines, perihilar haze, and, in severe cases, "bat wing" opacities (Light, 2013).
- Cardiomegaly (305 cases) is the enlargement of the heart, often due to chronic hypertension, valvular disease, or cardiomyopathy. On chest X-ray, it is diagnosed when the cardiothoracic ratio exceeds 0.5 on a PA film. Cardiomegaly may be associated with other findings such as pulmonary edema or pleural effusion (Murray & Nadel, 2022).
- Nodule (288 cases) refers to a small, round opacity in the lung, typically less than 3 cm in diameter. Pulmonary nodules can be benign (such as granulomas or hamartomas) or malignant (primary lung cancer or metastases). Their detection, characterization, and follow-up are critical components of lung cancer screening and diagnosis (Müller et al., 2021).

- Pneumonia (262 cases) is an infection of the lung parenchyma, usually presenting as consolidation or airspace disease on X-ray. Pneumonia can be caused by bacteria, viruses, or fungi, and its radiographic appearance varies with the causative organism, patient age, and immune status.
- Pulmonary edema (209 cases) is a subset of edema specific to the lungs, most often resulting from left-sided heart failure. It is characterized by vascular redistribution, interstitial markings, and, in more severe cases, alveolar flooding.
- Fracture (209 cases) typically refers to rib fractures, which may be seen directly as discontinuities in the bony cortex or indirectly via associated findings such as subcutaneous emphysema or pneumothorax. Rib fractures are common in trauma and can be a source of significant morbidity, especially in older adults (Light, 2013).
- Mass (192 cases) denotes a larger, more solid opacity compared to a nodule, often greater than 3 cm. Pulmonary masses raise suspicion for malignancy but can also represent benign tumors, abscesses, or organizing pneumonia.
- Calcification (178 cases) in the thorax may be seen in granulomas, healed infections, certain tumors, or within lymph nodes. The pattern and location of calcification can help distinguish benign from malignant processes.
- Emphysema (122 cases) is a form of chronic obstructive pulmonary disease characterized by destruction of alveolar walls and enlargement of air spaces. On X-ray, emphysema is suggested by hyperlucent lungs, flattened diaphragms, and increased retrosternal airspace (Murray & Nadel, 2022).
- Hernia (56 cases) in the chest most often refers to hiatal hernia, where part of the stomach herniates through the diaphragm into the thoracic cavity. Other types include congenital diaphragmatic hernias, which are more common in pediatric populations.
- Tuberculosis (55 cases) is a chronic infectious disease caused by *Mycobacterium tuberculosis*. Radiographically, it may present as upper lobe infiltrates, cavitation, lymphadenopathy, or miliary nodules. Although less common in high-income countries, tuberculosis remains a significant global health issue (Murray & Nadel, 2022).
- Infection (54 cases) is a broad category that may overlap with pneumonia, tuberculosis, or other specific entities. In the dataset, this label likely captures cases where infection was suspected but not further specified.

This detailed breakdown of disease frequency in the Open-i dataset highlights both the common and rare challenges faced by radiologists. It also underscores the importance of comprehensive training data for AI systems, which must be able to recognize frequent pathologies such as effusion and pneumothorax, while also being sensitive to less common but clinically significant findings like tuberculosis or hernia. Addressing the inherent imbalance in disease prevalence is crucial for developing robust, generalizable models that can support radiologists in a wide range of clinical scenarios (Demner-Fushman et al., 2016; Wang et al., 2023).

## Appendix 2: Core Django Web Application Components

### **models.py**

Defines the ChestXraySubmission model, which captures every user upload and AI output. Key fields include:

- user (ForeignKey to Django's built-in User),
- image (ImageField for the uploaded X-ray),
- indication and findings (TextFields for clinical context),
- impression (TextField for the generated caption),
- gradcam\_resnet and gradcam\_efficientnet (ImageFields for the overlay paths),
- created\_at (DateTimeField).

This schema ensures that every interaction input and generated output is stored for audit or review.

### **serializers.py**

Implements ChestXraySubmissionSerializer (a DRF ModelSerializer) to convert model instances into JSON and back. It validates incoming fields (image, indication, findings) and marks impression, gradcam\_resnet, and gradcam\_efficientnet as read-only to prevent tampering via the API.

### **views.py**

Contains both page-rendering views and API endpoints.

- **HomeView, LoginView, SignUpView** render the static pages and forms.
- **ChatView** serves the main interface for authenticated users.
- **ReportGenerationView** (a subclass of `rest_framework.views.APIView`) handles POST requests with multipart data: it saves a new ChestXraySubmission, calls `ai_agent.process_submission()`, updates the record with generated text and heatmap paths, and returns the full JSON payload.

### **urls.py**

Maps URL patterns to their corresponding views :

- `path("", HomeView.as_view(), name='home')`

- `path('login/', LoginView.as_view(), name='login')`
- `path('signup/', SignUpView.as_view(), name='signup')`
- `path('chat/', ChatView.as_view(), name='chat')`
- `path('api/generate/', ReportGenerationView.as_view(), name='generate_report')`

### **ai\_agent.py**

Implements the core AI pipeline:

1. **Lazy loading** of models on first request, caching them thereafter.
2. **Feature extraction**: runs the uploaded image through ResNet50 and EfficientNetB0, capturing Grad-CAM activations.
3. **Text embedding & fusion**: encodes Indication and Findings via BERT and concatenates with visual features.
4. **Report generation**: feeds the multimodal vector into Gemma-LoRA and returns the “Impression.”
5. **Output handling**: saves Grad-CAM overlay images to media/ and returns their paths.

### **forms.py**

(Optional) Defines any Django Form or ModelForm classes used by the HTML pages for server-side validation, ensuring that fields like indication and findings meet formatting requirements before submission.

### **admin.py**

Registers ChestXraySubmission with Django’s admin interface, allowing administrators to inspect, correct, or delete entries directly via the admin dashboard.

### **templates/**

Holds all HTML templates:

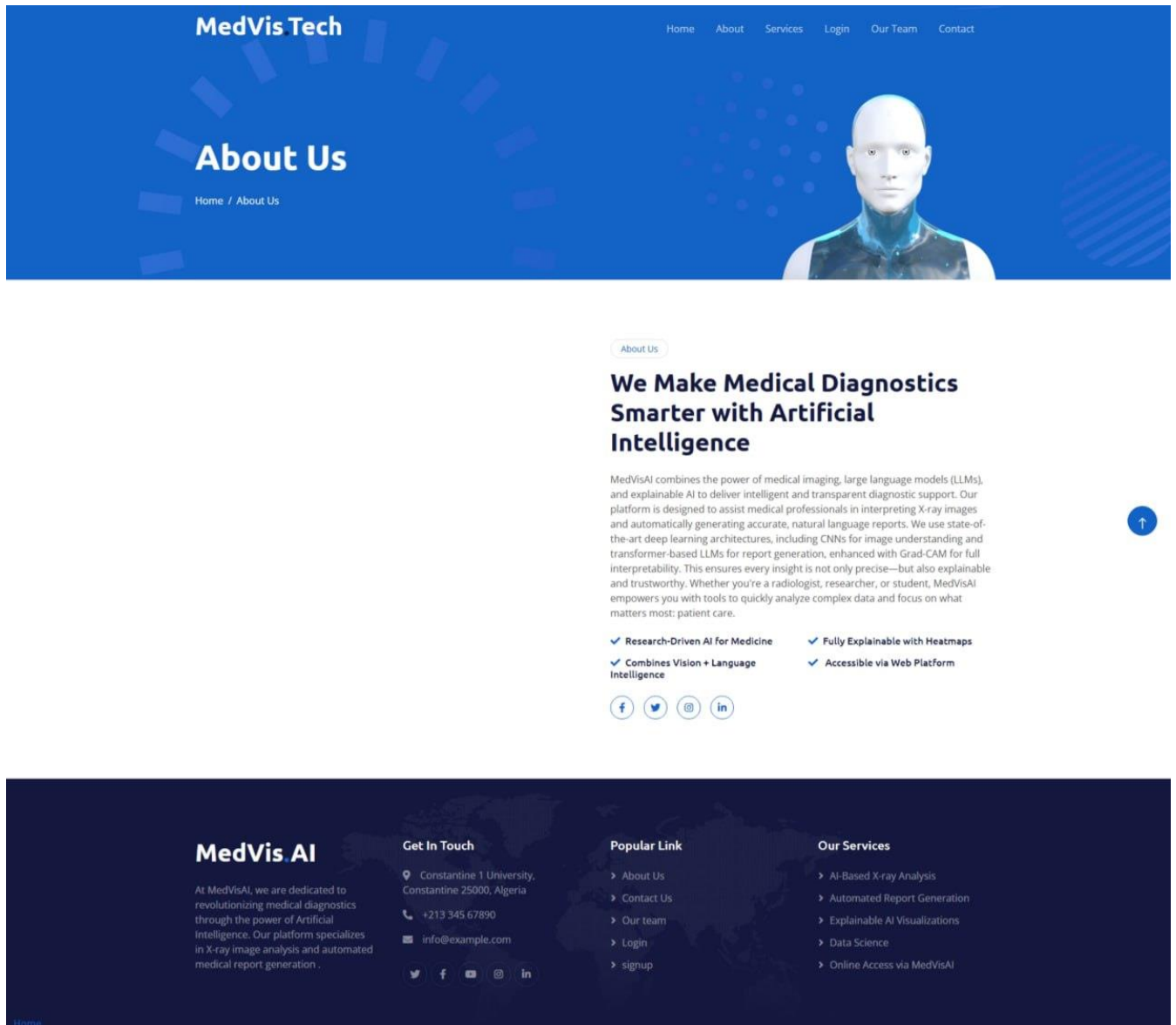
- `base.html` for shared layout and assets,
- Static pages (`home.html`, `about.html`, `services.html`, `contact.html`),
- Auth pages (`login.html`, `signup.html`),

- chat.html for the main AI interface.

**static/**

Contains CSS (styles.css), JavaScript (chat.js), and image assets (logos, placeholders)

## Appendix 3 : About Us Page Mission and Technology Overview





## Appendix 4: Services Page Comprehensive AI-Powered Medical Imaging

### Services

MedVis Tech

HomeAboutServicesLoginOur TeamContact


Our Services

Home / Our Services

Our Services


### Our Excellent AI Solutions for Smarter Diagnostics

Discover how MedVisAI transforms X-ray interpretation through cutting-edge artificial intelligence, language modeling, and explainability. Our platform is designed to streamline diagnosis, improve accuracy, and enhance clinician confidence.




#### AI-Based X-ray Analysis

Upload any X-ray image and let our advanced model detect key regions with high precision.




#### Explainable AI Visualizations

Understand AI decisions with Grad-CAM heatmaps that show what influenced the output.



#### Automated Report Generation

Generate detailed, accurate medical reports instantly using our vision-language AI trained on real-world datasets.



#### Online Access via MedVisAI

Use the full power of our platform directly online—no installation needed. Just log in, upload, and receive results.

### MedVis AI

At MedVisAI, we are dedicated to revolutionizing medical diagnostics through the power of Artificial Intelligence. Our platform specializes in X-ray image analysis and automated medical report generation.

#### Get In Touch

Constantine 1 University,  
Constantine 25000, Algeria

+213 345 67890

info@example.com

[Twitter](#) [Facebook](#) [Instagram](#) [LinkedIn](#)

#### Popular Link

- > About Us
- > Contact Us
- > Our team
- > Login
- > signup

#### Our Services

- > AI-Based X-ray Analysis
- > Automated Report Generation
- > Explainable AI Visualizations
- > Data Science
- > Online Access via MedVisAI

Home


## Appendix 5 : Login Page Secure Access Gateway

MedVis Tech

HomeAboutServicesLoginOur TeamContact

# Login to Your Account

Access your dashboard and explore our AI model




### Login

Username

Password

Login

Don't have an account? Sign up



### MedVis AI






At MedVisAI, we are dedicated to revolutionizing medical diagnostics through the power of Artificial Intelligence. Our platform specializes in X-ray image analysis and automated medical report generation .

### Get In Touch

Constantine 1 University,  
Constantine 25000, Algeria

+213 345 67890

info@example.com

### Popular Link

- > About Us
- > Contact Us
- > Our team
- > Login
- > signup

### Our Services

- > AI-Based X-ray Analysis
- > Automated Report Generation
- > Explainable AI Visualizations
- > Data Science
- > Online Access via MedVisAI

Home


## Appendix 6 : Sign Up Page Healthcare Professional Registration

MedVis.Tech

Home About Services Login Our Team Contact

# Sign Up

Home / Sign Up



### Create an Account

Username

Email

Password

Confirm Password

Sign Up

Already have an account? [Log in](#)

## MedVis AI

At MedVisAI, we are dedicated to revolutionizing medical diagnostics through the power of Artificial Intelligence. Our platform specializes in X-ray image analysis and automated medical report generation .

### Get In Touch

📍 Constantine 1 University, Constantine 25000, Algeria

☎ +213 345 67890

✉ info@example.com

🐦 📘 📧 📷 🌐

### Popular Link

- [About Us](#)
- [Contact Us](#)
- [Our team](#)
- [Login](#)
- [signup](#)

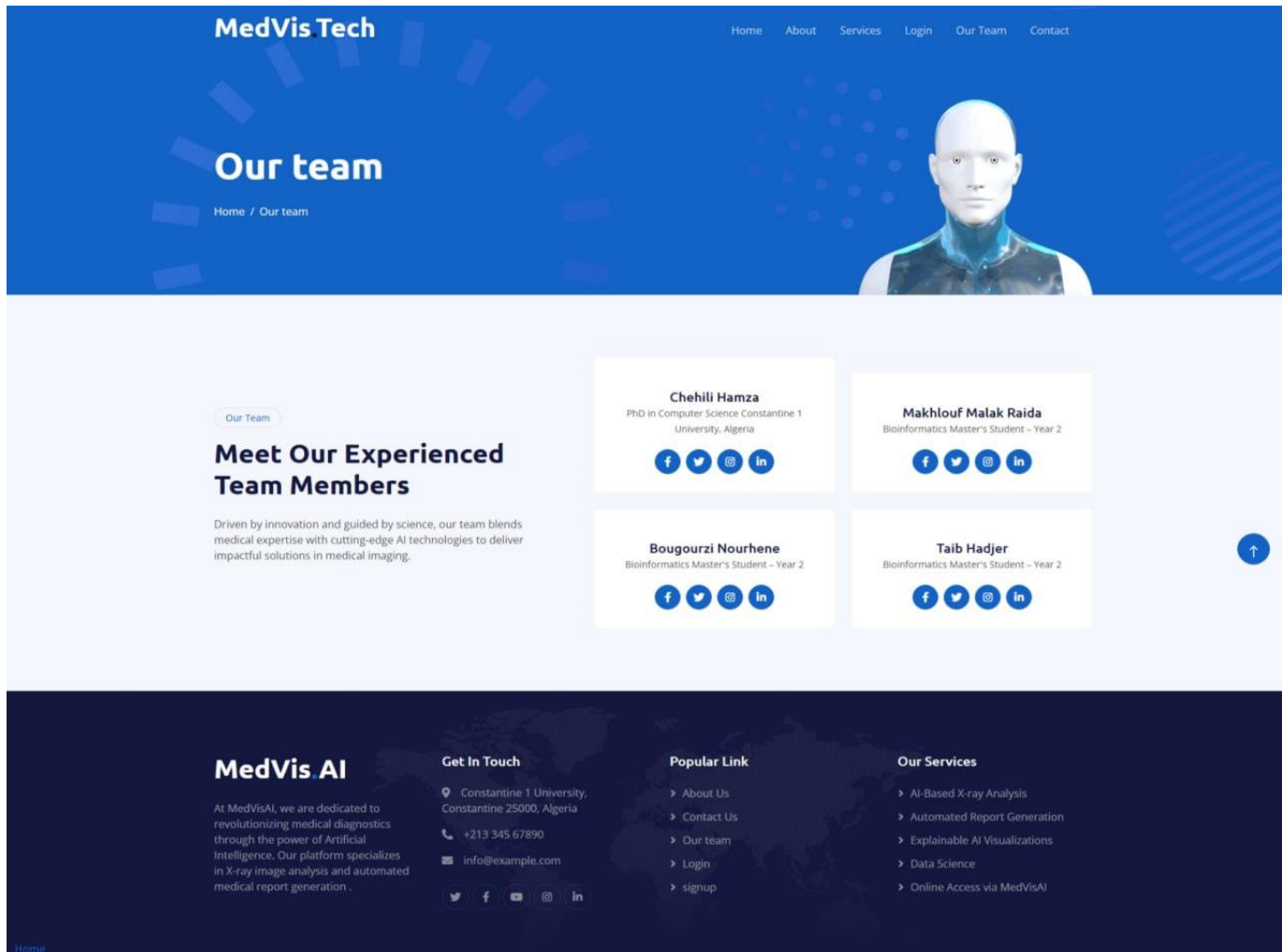
### Our Services

- [AI-Based X-ray Analysis](#)
- [Automated Report Generation](#)
- [Explainable AI Visualizations](#)
- [Data Science](#)
- [Online Access via MedVisAI](#)

Home

↑

## Appendix 7 : Our Team Page Interdisciplinary Expertise



## Appendix 8 : Contact Page Support and Inquiry Interface

MedVis.Tech

HomeAboutServicesLoginOur TeamContact

Contact Us

Home / contact Us

Contact Us

If You Have Any Query,  
Please Contact Us

Your Name

Your Email

Subject

Message

Send Message

MedVis.AI

At MedVisAI, we are dedicated to revolutionizing medical diagnostics through the power of Artificial Intelligence. Our platform specializes in X-ray image analysis and automated medical report generation .

Get In Touch

Constantine 1 University,  
Constantine 25000, Algeria

+213 345 67890

info@example.com

Popular Link

> About Us

> Contact Us

> Our team

> Login

> signup

Our Services

> AI-Based X-ray Analysis

> Automated Report Generation

> Explainable AI Visualizations

> Data Science

> Online Access via MedVisAI

Home

Academic year : 2024-2025	Submitted by: BOUGOURZI Nourhene MAKHLOUF Raida Malek TAIB Hadjer
<h2 style="text-align: center;">A Multimodal Framework for Explainable Chest X-ray Report Generation</h2>	
Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Master in Bioinformatics	
<p>Abstract:</p> <p>The growing demand for radiological services coupled with a shortage of expert radiologists has driven the development of automated report-generation systems. This thesis presents a novel explainable-AI framework for generating chest X-ray (CXR) reports by fusing visual and textual features. Using the Indiana University Chest X-ray (Open-I) dataset comprising 7,470 paired images and structured reports our approach employs a dual-branch convolutional ensemble (ResNet-50 and EfficientNet-B0) to extract complementary visual representations, alongside BERT-based embeddings of clinical “Indications” and “Findings.” These modalities are concatenated into a unified multimodal vector, which is fed into a fine-tuned Gemma-3 1B model using Low-Rank Adaptation (LoRA). Explainability is achieved via Grad-CAM heatmaps from both CNN backbones, highlighting anatomically relevant regions. Evaluated on a held-out validation subset (n = 300), our method outperforms state-of-the-art baselines across BLEU-1 to BLEU-4, ROUGE, and METEOR metrics, while preserving clinical terminology with high precision. Qualitative analysis confirms that generated impressions align closely with radiological standards. This work demonstrates that integrating documented findings as input rather than output reduces hallucinations and enhances interpretability, paving the way for deployable, trustworthy AI-assisted radiology reporting.</p>	
<p><b>Keywords:</b> Explainable AI, Multimodal Learning, Chest X-Ray Report Generation, generative model</p>	
<p><b>Chairperson :</b> Dr. I. R. AMINE KHODJA (MCB- Constantine 1 Frère Mentouri University).</p> <p><b>Supervisor:</b> Dr. H. CHEHILI (MCA - Constantine 1 Frère Mentouri University).</p> <p><b>Examiner :</b> Dr. D. Y. MEZIANI (MCB -- Constantine 1 Frère Mentouri University).</p>	