Democratic People's Republic of Algeria.
Ministry of Higher Education and Scientific Research
University of May 8, 1945 - Guelma -
Faculty of Mathematics, Computer Science, and Material Sciences
Department of Computer Science

Master's thesis

Field: Computer Science

Option: Information and Communication Science and Technology

Title:

Development of an Intelligent System for Automatic Medical Report Generation

Presented By

HAMMOUDA Hiba Errahmen

Members of the Jury:

- **President:** Dr. FERKOUS Chokri
- **Supervisor:** Dr. SERIDI Ali
- **Co-supervisor:** Dr. BOURDJIBA Yamina
- **Examiner:** Pr. ZEDADRA Nawel

June 2025

Acknowledgements

This work marks the culmination of a long academic and personal journey, and I would like to express my heartfelt gratitude to all those who contributed to it, directly or indirectly.

All praise is due to Allah, who granted us the strength and patience to complete this project.

I would like to extend my deepest thanks to ***Dr. Seridi Ali*** *and* ***Dr. Bordjiba Yamina*** *for their invaluable guidance, support, and encouragement throughout this work. Their expertise and advice have been essential at every stage of this journey.*

I also extend my sincere gratitude to the members of the jury ***Dr. Ferkous Chokri & Pr. Zedadra Nawel*** *for taking the time to evaluate my work and for their constructive feedback and valuable remarks.*

I would also like to express my heartfelt thanks to ***Dr. Zin Eddin Kouahla****, Head of the Department, and to* ***Miss. Madiha Kharroubi****, engineer at the LabSTIC laboratory, for their constant support and availability. Their doors were always open, and their assistance was greatly appreciated.*

My sincere appreciation also goes to the ***teachers and students of the Computer Science Department****, whose support and collaboration helped create a rich and motivating academic environment.*

May Allah reward you all.

الإهداء

بسم الله الرحمن الرحيم

الحمد لله، والصلاة والسلام على رسول الله سيدنا محمد، صلى الله عليه وسلم.

أما بعد...

ها أنا أُطوي صفحة من صفحات العمر، وأرسي سفينتي على مرافئ الطمأنينة، بعد رحلة شقّها الأمل وسقاها الصبر.

فالحمد لله الذي بنعمته تتمّ الصالحات، والحمد لله الذي أنار طريقي، ووفقني لإنجاز هذا المشروع المبارك.

أهدي هذا العمل، الذي حفرته الأيام، وسفته الليالي بالدعاء والتعب،

إلى من كان لهم الفضل بعد الله فيما أنا عليه اليوم...

إلى والديّ العزيزين **حمودة إبراهيم** و **كريشيام عبلة**، يا من كنتما النور في عتمة دربي، والسند الذي لا يلين،

يا من كان دعاؤكما سرّ توفيقي، فلكما في القلب دعاء لا ينقطع.

إلى أخواتي الزهرات :**رحمة، إكرام، مودة، ونعمة**،

كنتنّ الفرح الذي يسبق كل إنجاز، والدفء في كل منعطف، قناديل دربي، وبهجة روحي في كل لحظة.

إلى ابن أختي الحبيب **إدريس**، وبنت أختي الغالية **رملة**، ضحكتكما تسرق منّي التعب، وتزرعان في قلبي أملاً لا يوصف.

إلى وزوجي **محمدي**، شكرا لك على دعمك.

إلى رفيقة دربي **منار**، وصديقات الجامعة **آية وهالة وأماني**،

كنتنّ نبض القلب في رحلته الجامعية، ونسمة راحة في زحمة الأيام.

إلى رفيقات المسجد **المعلمة ليندة** والأخريات، من جمعني بهنّ حب الله، يا من كان في صحبتكنّ زادٌ للروح، وطمأنينة تسكن القلب،

زاد كنّ الله سكينة كما كنتم سكينة لروحي.

وفي الختام...

أسأل الله أن يرحم من غادرونا وتركوا في القلب فراغًا لا يُملأ،

إلى جديّ وجدّتيّ، وعمّتي، وزوج خالتي، وابن عمي،

رحمكم الله رحمةً واسعة، وجعل قبوركم روضة من رياض الجنة، وذكراكم نورًا لا يخبو من القلب.

Abstract

Radiological reports play an essential role in the diagnostic process, particularly for thoracic pathologies visible on chest X-rays. The manual preparation of these reports by radiologists is a demanding, time-consuming task that is subject to subjective variations. The emergence of deep learning offers promising prospects for automating this task and improving clinical productivity.

In this work, we propose an automatic radiology report generation system based on a hybrid deep learning architecture. The system integrates a pre-trained convolutional neural network (EfficientNetB0) for visual feature extraction, coupled with a Transformer-based decoder for diagnostic text generation. The model is trained on the Indiana University Chest X-ray database, after structured pre-processing of the images and text reports. In order to improve the linguistic consistency and terminological accuracy of the reports generated, a post-processing phase is introduced, based on the BioGPT model, which specialises in the biomedical field. This step improves the fluidity, readability and clinical accuracy of the reports produced.

The experimental results obtained demonstrate the effectiveness of the system. The BLEU-4 score increased from 0.4191 (LSTM model) to 0.8286 (Transformer model with BioGPT), while the BERTScore reached 0.9628, reflecting strong semantic similarity with the reference reports. These performances confirm the potential of the proposed approach to assist radiologists and improve the quality of AI-assisted diagnoses.

Keywords: Radiology report generation, encoder-decoder architecture, deep learning, CNN-Transformer, EfficientNetB0, Transformer decoder, BioGPT, chest X-rays, natural language processing

Résumé

Les rapports radiologiques jouent un rôle essentiel dans le processus de diagnostic, en particulier pour les pathologies thoraciques visibles sur les radiographies du thorax. La préparation manuelle de ces rapports par les radiologues est une tâche exigeante, longue et sujette à des variations subjectives. L'émergence de l'apprentissage profond offre des perspectives prometteuses pour automatiser cette tâche et améliorer la productivité clinique.

Dans ce travail, nous proposons un système de génération automatique de rapports de radiologie basé sur une architecture hybride d'apprentissage profond. Le système intègre un réseau neuronal convolutionnel pré-entraîné (EfficientNetB0) pour l'extraction des caractéristiques visuelles, couplé à un décodeur basé sur Transformer pour la génération de texte diagnostique. Le modèle est entraîné sur la base de données de radiographie thoracique de l'Université de l'Indiana, après un prétraitement structuré des images et des rapports textuels. Afin d'améliorer la cohérence linguistique et la précision terminologique des rapports générés, une phase de post-traitement est introduite, basée sur le modèle BioGPT, spécialisé dans le domaine biomédical. Cette étape améliore la fluidité, la lisibilité et la précision clinique des rapports produits.

Les résultats expérimentaux obtenus démontrent l'efficacité du système. Le score BLEU-4 est passé de 0,4191 (modèle LSTM) à 0,8286 (modèle Transformer avec BioGPT), tandis que le score BERTS a atteint 0,9628, reflétant une forte similarité sémantique avec les rapports de référence. Ces performances confirment le potentiel de l'approche proposée pour aider les radiologues et améliorer la qualité des diagnostics assistés par l'IA.

Mots-clés Génération de rapports radiologiques, architecture codeur-décodeur, apprentissage profond, CNN-Transformateur, EfficientNetB0, décodeur transformateur, BioGPT, radiographies du thorax, traitement du langage naturel.

الملخص

تلعب تقارير الأشعة دورًا أساسيًا في عملية التشخيص، خاصةً بالنسبة للأمراض الصدرية التي تظهر في الأشعة السينية للصدر. ويُعد الإعداد اليدوي لهذه التقارير من قبل أخصائيي الأشعة مهمة شاقة وتستغرق وقتاً طويلاً وتخضع لتغيرات ذاتية. يوفر ظهور التعلم العميق آفاقاً واعدة لإعداد هذه المهمة آلياً وتحسين الإنتاجية العيادية.

في هذا العمل، نقترح نظاماً آلياً لإعداد تقارير الأشعة يعتمد على بنية التعلم العميق الهجين. يدمج النظام شبكة عصبية تلافيفية مدربة مسبقًا (EfficientNetB0) لاستخراج الميزات المرئية، إلى جانب وحدة فك ترميز قائمة على المحول لتوليد النص التشخيصي. تم تدريب النموذج على قاعدة بيانات جامعة Indiana University للأشعة السينية للصدر، بعد المعالجة المسبقة المنظمة للصور والتقارير النصية. ومن أجل تحسين التناسق اللغوي ودقة المصطلحات للتقارير التي تم إنشاؤها، يتم إدخال مرحلة ما بعد المعالجة، استنادًا إلى نموذجBioGPT ، المتخصص في المجال الطبي الحيوي. تعمل هذه الخطوة على تحسين السلاسة وسهولة القراءة والدقة الإكلينيكية للتقارير التي يتم إنتاجها.

. تُظهر النتائج التجريبية التي تم الحصول عليها فعالية النظام. وارتفعت درجة BLEU-4 من 0.4191 (نموذج LSTM) إلى 0.8286 (نموذج المحول مع BioGPT)، في حين بلغت درجة BERTScore 0.9628، مما يعكس تشابهًا دلاليًا قويًا مع التقارير المرجعية. يؤكد هذا الأداء إمكانات النهج المقترح لمساعدة أخصائيي الأشعة وتحسين جودة التشخيص بمساعدة الذكاء الاصطناعي

الكلمات المفتاحية :توليد التقارير الشعاعية، ، بنية التشفير والفك، التعلم العميق ، CNN-Transformer، EfficientNetB0، Transformer Decoder، BioGPT، صور الأشعة السينية للصدر، معالجة اللغة الطبيعية

Contents

List of Figures

List of Tables

General introduction

Radiology plays a fundamental role in modern clinical diagnosis, providing vital visual information to support medical decision-making. Among various imaging modalities, chest X-rays are the most frequently used due to their low cost, non-invasive nature, and diagnostic importance in detecting thoracic pathologies such as pneumonia, cardiomegaly, tuberculosis, and pulmonary edema. However, the interpretation and reporting of these images require significant expertise and time from radiologists, who are often overwhelmed by the growing volume of imaging data in clinical workflows.

Despite advances in digital health, radiology reports are typically generated manually by experts after a detailed visual analysis of X-ray images. This process is not only labor-intensive but also susceptible to inter-observer variability and reporting inconsistencies. In regions with limited access to trained radiologists, the quality and timeliness of radiological assessments are further compromised. These limitations have motivated the development of intelligent systems capable of automating the generation of diagnostic reports directly from medical images.

Artificial intelligence (AI), particularly deep learning, has shown remarkable progress in various medical image analysis tasks. Encoder-decoder architectures—comprising convolutional neural networks (CNNs) for image feature extraction and recurrent or transformer-based models for text generation—have been successfully applied to medical image captioning. However, generating accurate and clinically coherent radiology reports remains a challenging task due to several factors. These include the complexity and variability of medical language, the need for domain-specific knowledge, and the scarcity of large, high-quality annotated datasets.

In In this context, our work addresses this issue by proposing a system for automatically generating radiological reports from chest X-rays. The approach we have developed is based on a hybrid architecture combining a pre-trained visual encoder (EfficientNetB0) with a Transformer-type decoder to produce diagnostic text. To improve the linguistic quality and clinical accuracy of the reports, a post-processing module is integrated, based on the BioGPT biomedical language model, trained to correct and refine the texts generated.

This work is structured into four chapters, preceded by a general introduction and followed by a general conclusion.

The first chapter is devoted to the field of thoracic radiology. It presents the basics of medical imaging, the types of pathologies that can be detected by radiography, the formats of clinical reports, and the databases available for training AI systems in this field.

The second chapter introduces the foundations of artificial intelligence and deep learning. It details convolutional neural networks, sequential decoders (LSTM, Transformer), and biomedical language models such as BioGPT. A review of recent approaches to generating medical reports is also presented.

The third chapter describes our methodology. It describes the proposed pipeline, from data pre-processing to the design of the CNN-Transformer architecture, including tokenisation, image normalisation and model input formats.

Finally, the fourth chapter deals with the practical implementation of the system, model training, performance evaluation using metrics such as BLEU, ROUGE-L, METEOR and BERTScore, and analysis of the results. This chapter also discusses the limitations identified and suggests possible improvements.

.

Chapter I: Image Captioning Task

1. Introduction

The automatic generation of descriptive text from visual content, known as image captioning, represents a compelling interdisciplinary challenge at the intersection of computer vision and natural language processing. Over the past decade, natural image captioning (NIC) has achieved remarkable progress, fueled by the development of powerful deep learning architectures and large-scale annotated datasets. However, the extension of these techniques to the medical domain introduces unique complexities that go far beyond those encountered in general image captioning. Medical image captioning, particularly for diagnostic purposes, demands the generation of clinically accurate, coherent, and contextually grounded radiology reports—tasks which require not only visual understanding but also domain-specific biomedical reasoning. This chapter provides a comprehensive overview of the progression from traditional NIC techniques to contemporary radiology report generation methods. It presents a detailed survey of publicly available medical imaging datasets, explores the evolution of model architectures from retrieval-based systems to Transformer-based decoders, and critically examines both generic and radiology-aware evaluation metrics designed to assess the clinical validity of generated reports.

2. Natural Image Captioning

The automatic generation of image captions is a complex challenge at the intersection of computer vision (CV) and natural language processing (NLP). It requires a detailed understanding of the visual content of an image, as well as the ability to describe its objects, attributes and relationships in a fluid human language. While natural image captioning (NIC) uses a common vocabulary, medical image captioning (MIC) requires specialist knowledge and biomedical terminology that is often unfamiliar to the general public.

Natural image captioning, the task of automatically generating descriptive sentences for images, has seen significant advancements (Pan et al., 2024). Over the past few years, a wide range of approaches have been proposed to address this challenge.

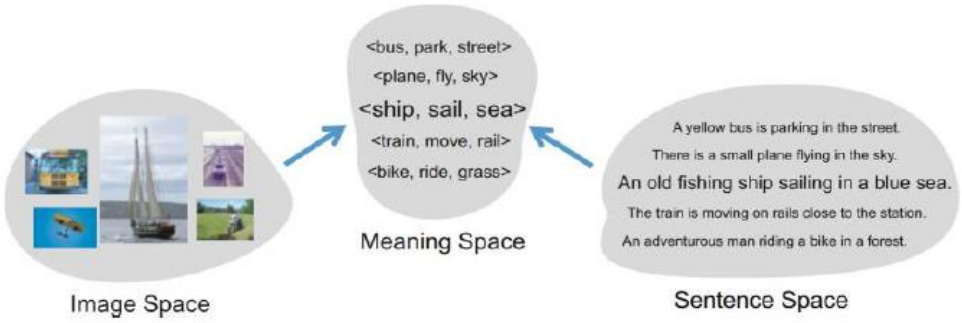


Figure 1: *Exemple of caption generation from image* (Hutchison et al., 2010)

A dominant architectural approach in image captioning is the encoder-decoder framework. These methods commonly employ a CNN as the encoder to extract visual features from the input image and an RNN as the decoder to generate the corresponding textual description(Pan et al., 2024). The Show-Tell model is a foundational example of this end-to-end neural network approach, where CNN-extracted image features are fed into an LSTM to produce captions(Vinyals et al., 2015).

Inspired by the human visual system, numerous methods have integrated attention mechanisms into the encoder-decoder framework. (Vinyals et al., 2015),(You et al., 2016) and (Lu et al., 2017). These attention mechanisms enable the model to automatically focus on the most relevant parts of the image while generating the caption(Pan et al., 2024). For instance, Lu et al. introduced an adaptive attention model that dynamically adjusts its focus between visual cues and the language model. (Lu et al., 2017). Similarly, (Anderson et al., 2018)proposed a combined bottom-up and top-down attention mechanism that computes attention at the level of objects and salient regions. Other research efforts have focused on improving the individual components of the captioning model(Pan et al., 2024). To enhance the image encoder, some methods explicitly model the relationships between different visual regions using Graph Convolutional Networks (GCNs) or scene graphs. (X. Yang et al., 2019; Yao et al., 2018). For improving the text decoder, hierarchical RNNs have been developed for paragraph generation, and novel attention mechanisms like the X-Linear attention block have been introduced to better utilize visual information. (Pan et al., 2020)The Transformer model, known for its powerful representation capabilities, has also been adopted as a replacement for RNNs in the text decoder. (Vaswani et al., 2023)Furthermore, Reinforcement Learning techniques have been applied to directly optimize non-differentiable captioning evaluation metrics(S. Liu et al., 2017; Pasunuru & Bansal, 2017, 2017).

While these advances in image annotation have been made, applying these methods directly to medical report generation often results in reduced performance. Pan, Y. et al. Generation of thoracic radiology reports based on multiscale feature fusion. This is primarily due to the unique characteristics and challenges associated with generating radiology reports.

3. Medical Image Diagnostic Captioning

The automated generation of a diagnosis from the study of one or more medical images of a patient is known as a diagnostic legend (DC)(Pavlopoulos et al., 2021). A medical report is nothing more than a factual, in-depth account of the important findings from medical imaging, drawn up by a professional (Monshi et al., 2020). The generation of these diagnostic reports is often considered to be a monotonous operation that can be automated(Yin et al., 2019).

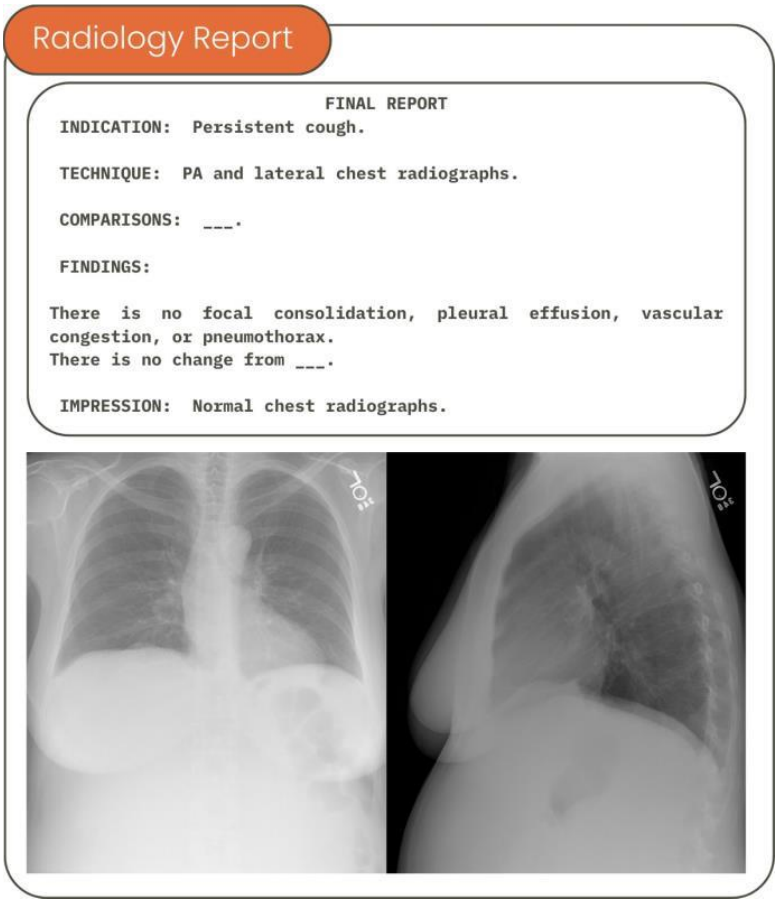


Figure 2: *Example of a normal finding in radiology report from the MIMIC-CXR Dataset* **(A. E. W. Johnson et al., 2019)**

4. Datasets for Radiology Report Generation

Systems for generating radiology reports based on deep learning depend significantly on extensive, labeled collections of medical images and their associated textual descriptions. Presented here is a summary of the key datasets, organized into families and extensions, along with other significant corpora.

4.1. Indiana University X-ray Dataset (IU-Xray):

The IU-Xray dataset (also called Open-i dataset), published in 2016, is among the first and most frequently utilized public datasets for research on generating radiology reports. It includes chest X-rays in both frontal and lateral perspectives from 3,955 patients, yielding 7,470 images. Every image is associated with a structured report composed in English, featuring sections like “Findings” and “Impression.” The dataset is especially useful for training and assessing models focused on generating reports at the sentence level. Its small dimensions and superior annotations render it a remarkable benchmark for proof-of-concept research(Demner-Fushman et al., 2016).

4.2. MIMIC-CXR Dataset Collection

The MIMIC-CXR dataset is a large-scale, resource widely used for training and evaluating deep learning models in radiology report generation. It contains hundreds of thousands of chest X-ray images paired with free-text reports. Its scale, diversity, and clinical depth make it a cornerstone for model development in this field.

4.2.1. Medical Information Mart for Intensive Care CXR (MIMIC-CXR):

MIMIC-CXR, released in 2019, is an extensive, anonymized dataset with 227,827 radiological examinations and 377,110 chest X-ray images from 65,379 individuals. It comprises both front and side perspectives, as well as English-written free-text reports. The dataset can be accessed under a limited access license for authorized researchers. It is commonly utilized in both classification and report creation activities(A. E. W. Johnson et al., 2019).

4.2.2. Medical Information Mart for Intensive Care - Annotated Biomedical Mention(MIMIC-ABM):

An extension of MIMIC-CXR, the MIMIC-ABM (Annotated Biomedical Mention) dataset was made available in 2020. It includes 38,551 entries labeled with biomedical entities to assist in entity recognition and relation extraction. While the number of patients isn't given, it aids in tasks such as medical named entity recognition (NER) and enhances the quality of visual-semantic embeddings(Ni et al., 2020).

4.2.3. Chest ImaGenome:

Launched in 2021, Chest ImaGenome expands on MIMIC-CXR by offering detailed image-level annotations for 242,072 frontal CXR images. These annotations

encompass spatial entities and connections, rendering the dataset valuable for training attention-driven and grounding models. (Wu et al., n.d.).

4.2.4. Chest X-Ray Pro (CXR-PRO):

Released in 2022, CXR-PRO is yet another limited-access extension of MIMIC-CXR. It includes 374,139 examinations from 65,379 individuals, preserving both frontal and lateral perspectives. It aims to minimize hallucinations by eliminating nonexistent earlier references in reports(Ramesh et al., 2022).

4.3. Multi-Source CXR Dataset Series

The Multi-Source CXR datasets were introduced to increase the robustness and generalizability of radiology models. By integrating images from diverse clinical settings and temporal contexts, they enable models to better handle variability across patient populations and institutions. These datasets are particularly valuable for longitudinal and comparative studies.

4.3.1. Multi-Source Chest X-ray (MS-CXR)

The MS-CXR dataset, launched in 2022, is a limited collection of chest X-rays sourced from various origin points. It consists of 1,047 frontal photographs from 851 individuals. Reports are in English and encompass diverse clinical settings(Boecking et al., 2022).

4.3.2. Multi-Source Chest X-ray – Temporal (MS-CXR-T)

MS-CXR-T, released in 2023, builds upon MS-CXR by incorporating a temporal aspect, comprising 1,326 frontal images from 800 different patients. It is especially beneficial for time-related reasoning and detecting changes in radiology(Bannur et al., 2023)..

4.4. Other Datasets

Several additional datasets complement the main collections by offering unique features such as different imaging modalities, languages, or regional healthcare contexts. These resources, though often smaller or restricted in access, support cross-lingual research and broaden the applicability of automated report generation across global healthcare systems.

- **PadChest:** is a restricted dataset published in 2019. It covers frontal and lateral CXRs from 67,625 patients, with 109,931 exams and 160,868 images. It is in Spanish (Bustos et al., 2020).
- **Chinese Hospital Chest X-ray (CH-Xray):** is a private dataset published in 2022, comprising frontal CXRs from 11,049 patients. It contains 11,049 images in Chinese (H. Zhao et al., 2021).

- **Chinese Cross-institutional Chest X-ray (CX-CXR):** is a restricted dataset published in 2018. It contains CXRs in frontal and lateral views of 33,236 patients for a total of 45,598 images. Reports are in Chinese (F. Wang et al., 2021).
- **COVID-19 CT Report Dataset (COV-CTR):** is a public dataset published in 2022. It contains axial CTs of 728 patients with as many images, in English (M. Li et al., 2023).
- **Japanese Liver CT (JLiverCT):** is a private dataset published in 2023. It contains axial CTs of 1,083 patients with the same number of images. Data in Japanese (Nishino et al., 2022).
- **CT Radiology Annotated for Text and Entity Extraction(CT-RATE):** is a public dataset published in 2024. It contains axial CTs from 21,304 patients, with 25,692 examinations and 50,188 images. The reports are in English (Hamamci et al., 2025).

Dataset	**Year**	**Patients**	**Images / Exams**	**Views**	**Language**	**Access**
IU-Xray	2016	3,955	7,470	Frontal/Lateral	English	Public
MIMIC-CXR	2019	65,379	377,110 / 227,827	Frontal/Lateral	English	Restricted
MIMIC-ABM	2020	—	38,551 reports	—	English	Restricted
Chest ImaGenome	2021	—	242,072	Frontal	English	Restricted
CXR-PRO	2022	65,379	374,139 exams	Frontal/Lateral	English	Restricted
MS-CXR	2022	851	1,047	Frontal	English	Restricted
MS-CXR-T	2023	800	1,326	Frontal	English	Restricted
PadChest	2019	67,625	160,868	Frontal/Lateral	Spanish	Restricted
CH-Xray	2022	11,049	11,049	Frontal	Chinese	Private
CX-CXR	2018	33,236	45,598	Frontal/Lateral	Chinese	Restricted
COV-CTR	2022	728	728	Axial CT	English	Public
JLiverCT	2023	1,083	1,083	Axial CT	Japanese	Private

CT-RATE	2024	21,304	50,188 / 25,692	Axial CT	English	Public

Table 1: *Summary Of The Available Datasets For Radiology Report Generation.*

5. Methods

A wide range of methods have been developed for radiology report generation, evolving from rule-based systems to modern deep learning architectures. It can be categorized these approaches based on their design philosophy and underlying mechanisms. The progression reflects the increasing complexity and sophistication required for accurate clinical text generation.

5.1. Early Approaches

Early systems relied on hand-crafted rules, retrieval techniques, and template-based generation to produce diagnostic descriptions. While limited in flexibility and scalability, these approaches laid the groundwork for automated report generation. They remain useful for well-structured tasks with constrained vocabularies.

5.1.1. Retrieval-based Methods

Retrieval-based methods generate an image caption based on the analysis of similar images extracted from a database. According to established rules, the final caption corresponds to the closest image or to a combination of the best *k*-captions identified(Ayesha et al., 2021).

5.1.2. Template-based Methods

Template-based approaches use predefined structures with empty slots to be filled in, enabling captions to be generated in a syntactically and semantically controlled way. The method begins by detecting a set of visual descriptors. These concepts are then combined into complete sentences using specific sentence templates or grammatical rules(Y. Yang et al., 2011).

5.2. Generative Approaches

Generative methods leverage neural networks to learn the mapping between visual inputs and textual outputs. Architectures such as encoder-decoder models and attention mechanisms have enabled more expressive and context-aware report generation. These approaches support end-to-end learning from large-scale datasets.

5.2.1. Encoder-Decoder Architectures

Encoder-decoder (ED) architectures learn to extract features end-to-end. The encoder, often implemented as a convolutional neural network (CNN), extracts visual features from the image, which are then used by a language model (LM) to generate syntactically and semantically correct sentences(Sutskever et al., 2014).

5.2.2. Compositional Architectures

Compositional approaches are based on the assembly of several functional modules, trained separately. An image is first processed by a convolutional neural network (CNN) to extract visual features. These representations are then used by a language model (LM) to generate a set of candidate descriptions, which are re-evaluated and re-ranked using a deep multimodal similarity model. The best evaluated description is selected as the image caption(Reale-Nosei et al., 2024).

5.2.3. Attention-based Architectures

Encoder–Decoder and compositional methods generally overlook the spatial structure of the input image, generating captions based on the image as a whole. Attention-based approaches, however, dynamically focus on specific regions during caption generation, allowing for a more detailed and accurate description(Reale-Nosei et al., 2024).

5.2.4. Dense image captioning

While encoder-decoder and compositional methods generate captions by considering the image as a whole, and attention-based methods focus selectively on regions before merging them into a single output, both approaches ultimately produce a single overall description. This mono-captioning strategy can be subjective and insufficient to fully capture complex scenes. Dense captioning offers an alternative by producing multiple region-specific captions(Reale-Nosei et al., 2024).

One of the first models that implemented this idea was DenseCap, which uses a CNN for feature extraction, a dense region suggestion layer to determine regions of interest, and a language model (often an LSTM) to generate individual captions for each region(J. Johnson et al., 2016).

5.3. Hybrid Methods

Hybrid approaches combine the strengths of retrieval, template, and generative techniques to improve accuracy and adaptability. They offer a flexible framework that balances structure with creativity, making them particularly effective for medical domains where factual correctness is critical. These models aim to minimize errors while preserving clinical utility.

5.3.1. Template-based and Generative Models

In the field of caption generation for medical images, a promising approach is to combine template-based and generative methods. Given that medical reports often follow a fixed structure, template-based approaches initially seem well-adapted for Diagnostic Captioning (DC). However, their lack of flexibility can limit their applicability across diverse diagnostic scenarios. To overcome this, several studies have proposed hybrid approaches that combine the reliability of templates with the adaptability of generative models. For example, Gill et al. showed that the generation of context-specific frontal

images x-ray of the pelvis to detect hip fractures - a well-defined task - can be effectively managed using only two templates. In their approach, images are coded using a dense network and categorized as either positive or negative for fractures. For positive cases, an LSTM with an attention mechanism fills in the appropriate fields in the predefined template(Gale et al., 2019).

5.3.2. Retrieval-based and Generative Models

The advantages and disadvantages of retrieval-based models have been explored extensively. Similar to hybrid template-based strategies, retrieval-based approaches can be combined with generative models to better adapt previously generated reports to new imaging data. This combination alleviates some of the strict limitations seen in pure template-based approaches by offering the ability to generalize to unseen cases(Beddiar et al., 2023).

Furthermore, some studies have proposed using Reinforcement Learning (RL) to dynamically decide whether to reuse an existing report or create a new one from scratch(C. Y. Li et al., 2018; Xiong et al., 2019).

6. Language Evaluation Metrics

Human evaluations of machine translation are extensive but costly. They can take months to complete and involve human labor that cannot be reused. For this reason, the researchers have created the language evaluation metrics to assess the generated text which are simple to use and faster.

6.1. Natural Language Generation Evaluation Methods

To assess the quality of generated reports, researchers have adopted a variety of automatic evaluation metrics from natural language generation. These include n-gram overlap measures like BLEU and METEOR, as well as embedding-based scores such as BERTScore. Each metric offers different insights into linguistic quality and semantic fidelity.

6.1.1. Bilingual Evaluation Understudy (BLEU)

BLEU is a metric that evaluates the quality of the generated text by measuring the n-gram overlap between candidate and reference sentences, without the need for precise positional alignment. To discourage very short outputs, this metric includes a brevity penalty (BP) that penalizes captions that are shorter than the reference. A BLEU scores closer to 1 indicates better performance. BLEU-n is commonly used to specify the number of words considered in an n-gram comparison(Papineni et al., 2001).

The standard formula for the BLEU score for a corpus is :

$$BLEU = BP \times exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

With

- p_n is the modified accuracy of the n-grams (with saturation according to maximum occurrences)
- w_n is the weight of each n-gram (classically $w_n = \frac{1}{N}, with\ N = 4$)
- BP is the brevity penalty, calculated as follows:

$$B\rho = \begin{cases} 1 & if\ c > r \\ \exp\left(1 - \frac{r}{c}\right) & else \end{cases}$$

  - c the total length (in words) of the generated output
  - r the length of the nearest reference.

6.1.2. Metric for Evaluation of Translation with Explicit Ordering (METEOR)

METEOR is a metric developed to evaluate the correlation between automatically generated captions and those produced by humans, at the sentence level. It extends BLEU-1 by introducing the harmonic mean between precision and recall, called the Fβ score, with a recall-oriented weighting. The Fβ score is a generalization of the F1 score, in which the β parameter allows recall to be prioritized. When no n-grams correspond between the model output and the human reference, the METEOR score can be reduced by up to 50%. In machine translation, a score higher than 0.6 is often interpreted as surpassing human performance, as two humans generally do not produce a perfect match. On the other hand, a BLUE score close to 1 is often considered unrealistic and may indicate that the model has been overlearned(Banerjee & Lavie, 2005).

The global formula is :

$$\text{METEOR} = F_{mean} \cdot (1 - \text{Penalty})$$

is the weighted harmonic mean of precision (P) and recall (R):

$$F_{mean} = \alpha P + \frac{P \cdot R}{(1-\alpha)R}$$

In the original version α=0.9

$$\text{Penalty} = \gamma\left(\frac{ch}{m}\right)^{\beta}$$

with typical values : γ=0.5, β=3.

6.1.3. Recall Oriented Understudy for Gisting Evaluation (ROUGE)

ROUGE is a metric initially developed to assess the quality of automatic summaries, by measuring the n-gram overlap between the generated summary and a human reference. It is also used in tasks such as image caption generation to compare the descriptions produced with references(Lin, 2004).

$$\text{ROUGE-N} = \frac{\text{number of matched n-grams between candidate and reference}}{\text{total number of n-grams in the reference}}$$

6.1.4. Consensus-based Image Description Evaluation (CIDEr)

CIDEr (Consensus-based Image Description Evaluation) has been specifically designed to evaluate image descriptions. It measures the cosine similarity between the weighted TF-IDF representations of the n-grams of the generated and reference captions(Vedantam et al., 2015).

6.1.5. BERTScore

BERTScore is an automatic evaluation metric for text generation, based on contextual representations of pre-trained language models. It measures token-level similarity between a generated output and a reference, using cosine similarity in the embedding space. According to its authors, BERTScore correlates better with human judgments and improves performance in model selection compared with conventional n-gram-based metrics.(Zhang et al., 2020)

6.2. Clinical Efficacy (CE)

Standard Natural Language Generation (NLG) evaluation metrics are designed to assess fluency and coherence in human-like texts. However, in radiology, reports often contain specialised medical terminology, making these general-purpose metrics not enough. As a result, researchers have developed domain-specific evaluation methods that better capture clinical accuracy and relevance.

6.2.1. Radiology-Aware Model-Based Evaluation Metric for Report Generation

This metric is an adaptation of the COMET framework, originally designed for evaluating machine translation, applied here to the field of radiology. It uses pre-trained language models to independently encode the reference report (source) and the generated report (hypothesis), and then computes combined features from their semantic representations. These features are then processed by a regressor trained to predict a quality score by minimising the mean square error (MSE). This approach enables the quality of the reports generated to be assessed without the need for an explicit reference(Calamida et al., 2023).

6.2.2. MRScore

MRScore is a radiology-aware rating metric that combines GPT-4 with a fine-tuned Large Language Model (LLM) to assess the quality of generated radiology reports. GPT-

4 is used to rate reports based on clinically-informed parameters and functions as a human judgement proxy. These scores are used to train a reward model - built on the Mistral-7B-Instruct LLM and fine-tuned using Reinforcement Learning with Human Feedback (RLHF) - which learns to replicate GPT-4's preferences, providing efficient and human-aligned scores at inference time(Y. Liu et al., 2024).

6.2.3. RaTEScore

RaTEScore is a radiology-specific similarity metric that evaluates report quality by comparing extracted medical entities between a reference and a candidate report. It comprises three modules: medical named entity recognition (NER), synonym-aware embedding, and an affinity-based scoring function. While the scoring does not directly use large language models, GPT-4 was used during the creation of the RaTE-NER dataset — a large-scale, manually annotated corpus — to enrich the training data with nuanced and rare radiological conditions. This indirect use of GPT-4 helped improve the quality and coverage of the NER model used in RaTEScore's evaluation pipeline (W. Zhao et al., 2024)

6.2.4. GREEN: Generative Radiology Report Evaluation and Error Notation

GREEN is a comprehensive evaluation framework for radiology report generation that leverages large language models to detect, classify, and explain clinically significant and insignificant errors across six categories.

- **False Finding (Hallucination):** Reporting findings not present in the reference report.
- **Missing Finding (Omission):** Omitting clinically relevant findings present in the reference.
- **Incorrect Location:** Describing findings in the wrong anatomical location.
- **Incorrect Severity:** Misstating the clinical severity of a condition.
- **Incorrect Size:** Reporting an inaccurate size for a finding.
- **Incorrect Comparison:** Misrepresenting temporal changes, such as stability or progression.

It outputs both a numerical score — reflecting the accuracy and clinical relevance of a generated report — and a textual summary that highlights specific error types using clustering-based techniques(Ostmeier et al., 2024).

7. Conclusion

The task of radiology report generation stands at the frontier of medical artificial intelligence, demanding a synthesis of visual comprehension, linguistic fluency, and clinical precision. This chapter has outlined the conceptual and technical evolution of image captioning, tracing its adaptation from natural to medical contexts. It has shown that while traditional encoder-decoder frameworks and attention mechanisms provide a

foundational basis, the specialized nature of medical reporting necessitates innovations in both dataset curation and model design. Moreover, the emergence of domain-specific evaluation metrics underscores the inadequacy of generic language metrics in assessing diagnostic accuracy. As models become more complex and datasets increasingly diverse, the focus is shifting toward ensuring factual correctness, minimizing hallucinations, and aligning machine-generated reports with clinical expectations. Looking forward, the integration of multimodal reasoning, reinforcement learning, and large-scale foundation models holds promise for achieving clinically trustworthy radiology report generation.

Chapter II: Deep Learning Overview and Literature Review

1. Introduction

The rapid evolution of artificial intelligence has positioned deep learning as a foundational paradigm for automating complex cognitive tasks, including those in medical image interpretation and report generation. This chapter provides a comprehensive overview of deep learning models and their relevance to the field of radiology. It begins with fundamental concepts of neural networks, detailing convolutional and recurrent architectures, before progressing to recent innovations such as Transformers, Vision-Language Models (VLMs), and emerging alternatives like Mamba and State-Space Models (SSMs). The chapter also introduces essential techniques like transfer learning and large language models, which have significantly enhanced performance across vision and language domains. In its final sections, a thorough literature review is presented, highlighting the diverse modeling strategies employed in recent research to automate diagnostic captioning and clinical reporting. This foundational overview sets the stage for the development and justification of the proposed system in subsequent chapters.

2. Deep learning models

Machine learning is a branch of artificial intelligence that allows computers to learn from data and make forecasts without being specifically coded. Through the examination of data patterns, machine learning algorithms create models that evolve and enhance through experience, rendering them well-suited for tasks that require handling dynamic or intricate datasets. Essential methods encompass supervised, unsupervised, semi-supervised, and reinforcement learning, along with deep learning, transfer learning, and ensemble approaches. These methods are commonly employed in various sectors—from healthcare and finance to transportation and customer support—to improve decision-making, automate processes, identify anomalies, and tailor user experiences. The process of machine learning includes gathering data, training models, and assessing performance, backed by platforms that provide scalable computing capabilities and strong development tools(Azure Microsoft, 2025).

Deep learning is a branch of machine learning that employs multi-layer neural networks to gain insights from vast amounts of unstructured data like images, text, and sound. It allows machines to autonomously identify features and make choices without direct coding. Drawing inspiration from the human brain, deep learning models analyze data via layers of connected nodes to recognize patterns and produce predictions. Deep learning, underpinned by frameworks like TensorFlow and PyTorch, finds extensive

application in areas such as self-driving cars, healthcare diagnostics, and processing natural language. Its effectiveness is fueled by enhanced computational capabilities, extensive datasets, and adaptable model structures(Azure, 2025b).

2.1. Neural Networks

A neural network is a kind of machine learning model that mimics the operations of the human brain. It consists of layers of artificial neurons that are linked together, featuring an input layer, several hidden layers, and an output layer. Every neuron evaluates inputs according to designated weights and a threshold; when the output surpasses this threshold, the signal moves to the subsequent layer. Neural networks adapt based on training data, constantly refining their internal parameters to enhance performance. After training, they can swiftly execute intricate tasks like image classification and speech recognition with great precision. Referred to as Artificial Neural Networks (ANNs), these frameworks are essential to deep learning systems and are pivotal in contemporary AI applications, such as technologies like Google's search engine (IBM, 2025c).

2.1.1. Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are a type of deep neural network highly proficient in analyzing visual data, including tasks like image classification, object detection, and recognition. In contrast to conventional neural networks, CNNs are structured to analyze and learn from spatial hierarchies in data through the use of three fundamental layers: convolutional layers, pooling layers, and fully connected layers.

- **Convolutional layer** serves as the basic component, using trainable filters (kernels) that move across the input data to execute dot products, resulting in feature maps that identify patterns such as edges or textures. These filters utilize parameter sharing and connection sparsity, significantly lowering computational complexity.
- **Pooling layers**, usually max or average pooling, come after convolutional layers to downsample feature maps, minimizing dimensionality while maintaining important features, aiding in enhancing generalization and reducing overfitting.
- **Fully connected layers** integrate the extracted features and execute the final classification through activation functions such as softmax.

CNNs have a hierarchical structure: initial layers detect basic forms (such as lines, edges), whereas later layers identify intricate patterns (like faces, organs). This architecture enables CNNs to substitute manual feature extraction with learning from start to finish. CNNs have become the standard in computer vision and medical imaging tasks due to their scalability and performance(IBM, 2025a).

As an advancement of the convolutional neural network architecture, numerous architectures have arisen to boost and refine the performance of convolutional neural networks. These architectures are trained on millions of images like ImageNet, allowing

them to conserve resources and time while being applicable in various areas. Such as EfficientNet.

- **EfficientNet**

EfficientNet is a collection of convolutional neural networks aimed at maximizing accuracy and efficiency by consistently scaling model depth, width, and input resolution through a compound scaling approach. Developed by Google researchers, EfficientNet offers improved performance while utilizing fewer parameters and requiring less computation than conventional CNNs. The foundational model is identified via neural architecture search, and larger versions (such as EfficientNet-B0) expand upon it while ensuring balanced scaling. EfficientNet is commonly utilized in image classification and transfer learning because of its excellent accuracy, minimal memory usage, and rapid inference speed(Tan & Le, 2020).

2.1.2. Recurrent Neural Network (RNN)

A Recurrent Neural Network (RNN) is a deep learning architecture specifically aimed at processing sequential data or time series inputs, where the sequence of data points holds significant contextual value. In contrast to typical feedforward networks, RNNs possess internal memory through loops that enable information to carry over across time steps. This makes them ideal for tasks like language modeling, speech recognition, machine translation, sentiment analysis, and forecasting time series. An RNN can forecast upcoming flood levels by examining historical flood data and meteorological information, or create text descriptions for images by understanding patterns in word sequences. Although traditional RNNs are beneficial, they can have difficulties with long-term dependencies because of problems such as vanishing gradients, which more sophisticated versions like LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Unit) aim to address(IBM, 2025d).

- **Long Short-Term Memory (LSTM)**

A Long Short-Term Memory (LSTM) network is a specific kind of Recurrent Neural Network (RNN) created to more effectively capture long-term dependencies in sequential information. It resolves the gradient vanishing and exploding issues typical in traditional RNNs with a distinctive memory cell design that can maintain information over extended periods. This architecture features gates—input, forget, and output gates—that manage the information flow, enabling the model to determine what to retain, modify, or eliminate at each moment. Due to these abilities, LSTMs excel in tasks such as language modeling, speech recognition, machine translation, handwriting generation, and sequence prediction. LSTM's power comes from its capability to retain context over numerous steps, which makes it especially effective for intricate patterns in time-related data like text, audio, or video(NVIDIA, 2025a).

2.2. Transformer

The Transformer is a deep learning framework launched in 2017 that transformed natural language processing by removing recurrence in favor of self-attention techniques. Its encoder-decoder architecture facilitates the simultaneous processing of sequences, permitting the model to focus on every aspect of the input at once. This design enhances computational efficiency and more effectively captures long-range dependencies compared to conventional RNNs. The Transformer acts as the basis for numerous sophisticated models, such as BERT and GPT, by facilitating strong sequence comprehension and generation abilities(Vaswani et al., 2023).

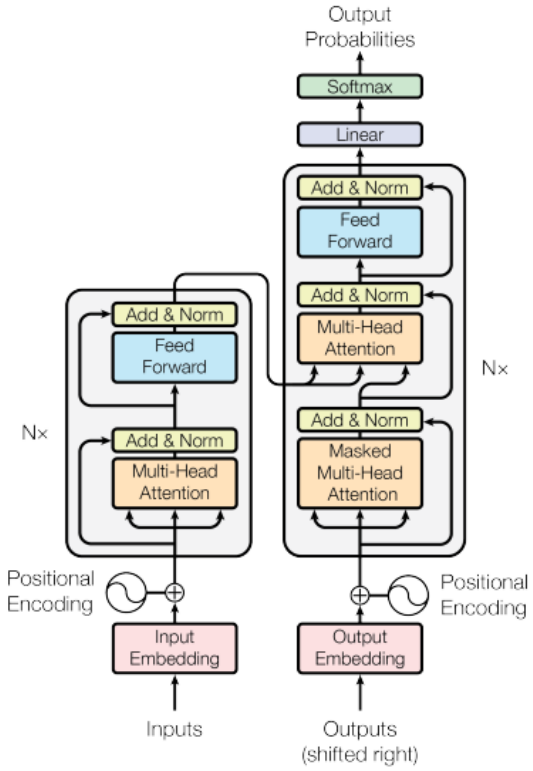


Figure 3: *The transformer model architecture*(Vaswani et al., 2023)..

- **GPT**

Generative Pretrained Transformers (GPTs) represent a series of extensive language models created by OpenAI, based on the transformer framework and fine-tuned for producing natural language. Since the launch of GPT-1 in 2018, the models have developed into robust, multimodal systems such as GPT-4o, which can process and create text, images, and audio. Trained on extensive datasets and optimized for particular tasks, GPT drives numerous AI applications such as chatbots, code creation, and data examination. These models can be accessed via APIs, allowing integration into various tools and services across different sectors (IBM, 2025g).

- **BERT**

BERT (Bidirectional Encoder Representations from Transformers) is a model for language representation that employs deep bidirectional Transformers to pre-train on extensive collections of unlabeled text. In contrast to previous models, BERT utilizes a masked language modeling objective and a next sentence prediction task to capture context from both sides, allowing for a deeper comprehension of language. It can be adapted with slight modifications to handle various NLP tasks like question answering, sentiment analysis, and natural language inference. BERT achieved new state-of-the-art performance on various benchmarks, showcasing the effectiveness of bidirectional pre-training for natural language understanding(Devlin et al., 2019).

- **Vision Transformer (ViT)**

Vision Transformers (ViTs) represent an innovative method for image recognition that utilizes the Transformer architecture on sequences of image patches, removing the requirement for convolutional layers. In ViT, an image is split into patches of fixed size, with each patch being linearly embedded and integrated with positional information, and the resulting sequence is input into a conventional Transformer encoder. A unique classification token is employed to generate predictions. In contrast to CNNs, ViTs do not rely on image-specific inductive biases such as translation invariance and locality; however, when trained on extensive datasets, they attain cutting-edge results in image classification. This renders ViTs a scalable and effective substitute for conventional convolutional networks, particularly in data-abundant situations(Dosovitskiy et al., 2021).

2.3. Large Language Models

Large Language Models (LLMs) are robust AI systems developed on extensive datasets to comprehend, produce, and engage in natural language. Based on transformer architectures, LLMs are capable of carrying out various tasks like responding to inquiries, summarizing texts, translating languages, coding, and producing coherent content. In contrast to traditional models designed for particular tasks, LLMs are versatile and applicable across various areas, positioning them at the heart of the current generative AI movement. Notable instances comprise OpenAI's GPT series, Google's BERT and PaLM, Meta's LLaMA, and IBM's Granite models. By understanding intricate language patterns with billions of parameters, LLMs are transforming areas ranging from customer service to research and content generation (IBM, 2025b).

2.4. Vision-Language Models (VLMs)

Vision-language models (VLMs) are sophisticated multimodal AI systems that integrate computer vision and natural language processing to comprehend and produce text based on visual information. They comprise two primary elements: a vision encoder—typically utilizing Vision Transformers (ViTs)—that converts visual information into embeddings, and a language encoder, generally employing transformer models such as

BERT or GPT, to manage textual data. VLMs facilitate tasks like image captioning, visual question answering, and object recognition by understanding the intricate connections between images and language through learning (IBM, 2025f).

2.5. Beyond Transformers (Mamba, SSM)

The prevalence of transformers in deep learning has initiated a quest for architectures that provide enhanced efficiency and scalability. Options investigate linear attention, recursion, and convolution, frequently blending features to maintain expressiveness while lessening computational demand. These architectures after transformers seek to overcome constraints such as memory bottlenecks and inefficiencies in handling long sequences(Schneider, 2024).

- **State-Space Models (SSMs)**

State-space models represent a category of sequence modeling frameworks that describe hidden states changing over time based on learned dynamic systems. In contrast to transformers that depend on global self-attention, SSMs capture temporal dependencies via continuous or discretized updates, facilitating efficient linear-time processing for extremely long sequences. Their framework, grounded in control theory, provides robust inductive biases, rendering them highly effective for tasks that involve intricate temporal patterns like speech, genomics, and long-context language modeling(Schneider, 2024).

- **Mamba**

Mamba is a modern neural architecture that enhances the state-space model framework by integrating a selective scanning mechanism that dynamically determines which input data to prioritize in sequence processing. This design enables Mamba to attain competitive accuracy on benchmark tasks while preserving the linear-time complexity typical of SSMs. By closing the performance-efficiency divide with transformers, Mamba shows that structured sequence models can provide both computational benefits and substantial representational capability(Gu & Dao, 2024).

2.6. Transfer Learning

Transfer learning is a machine learning technique where knowledge gained from one task or dataset is reused to improve performance on a related task. It is especially valuable in deep learning, where training models from scratch can be costly and require vast amounts of labeled data. By starting with a pre-trained model, transfer learning reduces training time, enhances generalization, and performs well even with limited data. However, it works best when the source and target tasks are similar; otherwise, it risks negative transfer, which can degrade model performance. Proper task alignment is therefore key to its success (IBM, 2025e).

3. Literature Review

The development of automated radiology report generation has been supported by a growing body of research exploring deep learning techniques. This section reviews prior work, emphasizing the models and strategies proposed to bridge visual and textual modalities. Highlighting key advancements, it helps identify existing challenges and inform future improvements.

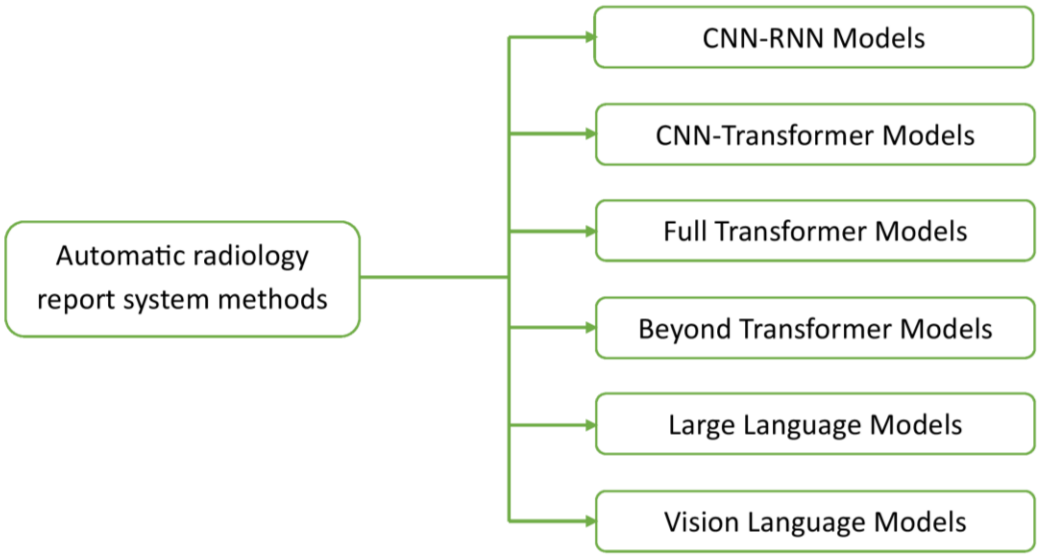


Figure 4: *Summary of the categories of radiology report generation methods*

3.1. CNN-RNN Models

Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have established themselves as essential components in generating automated radiology reports, allowing intricate medical images to be converted into organized natural language

(Sirshar et al., 2022) utilizes the CNN-RNN approach, particularly integrating convolutional neural networks (CNNs) for image encoding with recurrent neural networks (RNNs), implemented here with LSTM units, for generating text. The authors suggest a comprehensive model for generating automated radiology reports, incorporating an attention mechanism within a CNN-LSTM framework. The encoder section employs VGG-16 to derive features from chest X-ray (CXR) images, transforming them into compact vector representations. These visual embeddings are subsequently fed into an LSTM-based decoder that produces textual medical reports one word at a time. A significant improvement in this model is the attention mechanism, which actively directs the decoder's focus to particular areas of the image while generating the report. This reflects how a radiologist would highlight pathological areas when reporting observations. The model underwent training utilizing two datasets: the IU X-Ray from Indiana University

and MIMIC-CXR. The training was carried out on Google Colab utilizing an NVIDIA Tesla K80 GPU. Assessment was conducted using standard metrics for image captioning. The system reached BLEU-1 to BLEU-4 scores of 0.580, 0.342, 0.263, and 0.155 respectively on the IU X-Ray dataset, indicating the attention layer's role in generating more semantically aligned and coherent results. Still, constraints remain. The LSTM element has difficulty processing lengthy and intricate sequences, frequently resulting in a loss of contextual coherence among report sentences. Although VGG-16 has demonstrated effectiveness, it may not possess the same level of expressive capability as newer CNN architectures such as ResNet or EfficientNet. Additionally, the model does not possess clear mechanisms for modeling coherence at the paragraph level, which could lead to reports that are syntactically accurate yet clinically fragmented. Ultimately, the scale of the dataset was restricted, and larger, more varied training datasets could improve performance even more. Notwithstanding these constraints, the method represents a significant contribution to vision-language modeling within radiology.

(X. Wang et al., 2018) creates a new model TieNet (Text-Image Embedding Network), another method that built on a CNN-RNN paradigm, enhances this architecture by integrating visual and textual data streams within a single framework. The approach integrates convolutional neural networks for extracting spatial features at the image level with recurrent neural networks to capture semantic information from unstructured radiology reports. Its originality stems from the incorporation of multi-level attention mechanisms that improve the interpretability and efficiency of disease classification and report generation tasks. The architecture is trained on matched image-text datasets, utilizing radiology reports not only as output targets but also as a type of guidance. The model carries out two complementary functions: creating detailed reports and executing multi-label classification for thoracic conditions. While training, gradients from the two tasks affect shared parameters, allowing the model to better align visual attributes with text meanings. Experiments were performed on the extensive ChestX-ray14 dataset and enhanced by OpenI’s radiology dataset. TieNet showcased impressive performance, reaching an average AUC above 0.9 in disease classification and surpassing baseline metrics in report generation with a BLEU-1 of 0.2860, BLEU-4 of 0.0736, METEOR of 0.1076, and ROUGE-L of 0.2263. These metrics validate the model’s ability to produce medically pertinent and linguistically smooth results. Regardless of its advantages, TieNet shows specific limitations. The model has difficulty with intricate linguistic elements like negation, hedging, and uncertainty—characteristics often present in clinical narratives. Furthermore, although the multi-level attention enhances alignment, it does not provide detailed reasoning regarding pathological concepts, which restricts its capacity to differentiate nuanced disease variants. These concerns indicate future paths, like integrating graph-based or transformer-based elements for enhanced structured reasoning. Nonetheless, TieNet distinguishes itself as a scalable, multi-task system for extracting

knowledge from real-world PACS data and progressing toward semi-automated radiological analysis.

The study of (Jing et al., 2018) also follows the CNN-RNN framework but introduces a hierarchical LSTM model refined through co-attention. In contrast to conventional captioning systems that produce brief phrases, this research addresses the more challenging aim of creating comprehensive medical reports featuring a coherent layout and detailed content, closely resembling narratives penned by radiologists. The suggested architecture integrates visual and semantic attention methods and presents a two-phase report generation approach to harmonize content organization and sentence-level coherence. At the heart of the model is a multi-task framework: it concurrently forecasts medical keywords (tags) and produces descriptive paragraphs. A co-attention mechanism that integrates visual and semantic information allows the model to concentrate on pertinent areas of the image and related medical terminology. The hierarchical LSTM initially chooses a topic at the sentence level ("what to convey") and then constructs the sentence word by word ("how to articulate it"), thus enhancing logical coherence throughout the document. The model underwent training and evaluation using two datasets: IU X-Ray (radiology reports) and PEIR Gross (descriptions of pathological images). In both datasets, the system consistently surpassed traditional CNN-RNN and visual attention baselines on BLEU, METEOR, ROUGE, and CIDEr metrics. Qualitative assessment additionally indicated that the produced reports demonstrated a strong level of clinical significance and stylistic resemblance to texts written by humans. However, the method does have its limitations. It relies significantly on the correctness of anticipated tags; mistakes at this point can propagate and diminish the quality of the overall report. Moreover, the system's resilience weakens when faced with noisy or low-quality images, potentially interfering with both attention alignment and content organization. Finally, the architecture might gain from better modeling of inter-sentence relationships to boost narrative coherence. This research represents a major progress in structured report creation by incorporating document-level modeling into medical image description.

(Moradi et al., 2018)'s work is related to the CNN-RNN family but emphasizes multimodal localization instead of immediate report creation. The authors explore techniques for automatically generating visual annotations (region of interest – ROI) on chest X-ray images by utilizing data from existing free-text reports. This study tackles a major limitation in medical AI: the lack of extensive, manually labeled datasets for supervised training. Two designs are suggested. The initial model is a complete CNN-LSTM framework, which extracts image features and combines them with LSTM-generated textual embeddings to forecast polygonal ROIs. The second is a modular pipeline: DenseNet, trained on ChestX-ray14, extracts visual features; Doc2Vec obtains textual semantics; and both vectors are input into a multi-layer perceptron for coordinate regression. The objective is to forecast polygonal bounding boxes that emphasize

irregularities described in the related reports. The two architectures were assessed using a dataset of 494 chest X-rays that radiologists had manually annotated. The modular approach surpassed the end-to-end model, attaining a Dice coefficient of 61% (compared to 46%) and decreasing centroid error to 5.1% of image width (versus 7.2%). These findings emphasize the significance of independent processing and reveal the essential function of semantic information derived from text inputs. Nonetheless, the approach encounters distinct constraints. The ROIs are represented as basic quadrilaterals, limiting the system's capacity to capture atypical lesion forms. The pipeline assumes that every image displays one abnormality, which restricts its use in complicated, multi-disease situations. Future efforts might investigate more detailed segmentation results (e.g., masks) and implement transformer-based language encoders to enhance semantic grounding. Still, the method provides a viable route for utilizing current clinical reports to create useful annotation datasets with little human involvement.

ARTICLE	AUTHOR	MODEL	METHOD	RESULTS	DATASET	LIMITATIONS
Attention-based Radiology Report Generation	(Sirshar et al., 2022)	CNN-LSTM + Attention	VGG-16 encoder, LSTM decoder, attention module	BLEU-1: 0.580, BLEU-4: 0.155	IU X-Ray, MIMIC-CXR	Long sequence issues, outdated CNN, limited coherence
TieNet	(X. Wang et al., 2018)	CNN-RNN + Multi-level Attention	Text-image embedding, multi-task learning	AUC > 0.9, BLEU-4: 0.0736	ChestX-ray14, OpenI	Struggles with negation, lacks fine-grained control
On the Automatic Generation of Medical Imaging Reports	(Jing et al., 2018)	Hierarchical LSTM + Co-attention	Keyword prediction + sentence-level generation	BLEU-4=0.247 on IU X-Ray	PEIR Gross, IU X-Ray	Sensitive to tag quality, poor noise handling
Bimodal Network Architectures for Automatic Generation of Image Annotatio	(Moradi et al., 2018)	CNN + Doc2Vec	DenseNet + MLP for ROI prediction	average centroid distance of 7.8 ± 7.7 % compared to	494 X-rays with manual ROIs	Simple shapes, assumes single anomaly

n from Text				5.1 ± 4.0 %		

Table 2: : *Summary of Studies on Radiology Report Generation Based On CNN-RNN Methods.*

3.2. CNN-Transformer methods

Situated within the CNN-Transformer framework, (Aksoy et al., 2023)'s study improves report generation by integrating contextual, non-visual data into the modeling approach. Aksoy et al. suggest a multimodal Transformer architecture integrating visual characteristics obtained from chest X-rays with organized patient demographic information like age, gender, and ethnicity. The visual element is obtained via an EfficientNet encoder, while demographic characteristics are integrated into semantic vectors. These two modalities are subsequently processed together by a Transformer encoder-decoder to produce radiology reports informed by context. The main innovation consists of integrating demographic characteristics with visual data, recognizing that radiologists typically take this context into account during actual diagnostic processes. The model underwent training and evaluation using the MIMIC-CXR and MIMIC-IV datasets, assessing various configurations: image-only input, image along with one demographic variable, and image together with multiple demographics. The findings indicated that adding ethnicity by itself resulted in the most significant enhancement in BLEU and BERTScore. Nonetheless, merging various demographics (such as gender and ethnicity) did not uniformly improve performance, likely because of data imbalance or redundancy in features. Although incorporating contextual metadata signifies progress in personalizing automated diagnosis, the model encounters multiple obstacles. At times, it generates redundant text or fabricates results, especially regarding uncommon conditions. Moreover, the visible effect of ethnicity on performance brings crucial issues regarding fairness and bias within medical AI systems. These problems highlight the necessity of thoroughly assessing demographic factors regarding their predictive advantages and ethical considerations. The study shows that generating reports based on individual patient context can enhance output quality and sets the stage for more refined, patient-centered report creation in future research.

Another research of (Z. Wang et al., 2023a) marks a progression in CNN-Transformer techniques, tackling the limitations of "single-expert" attention frameworks employed in automated radiology report creation. The METransformer framework presents an innovative idea of "multi-expert joint diagnosis," mimicking cooperative decision-making by combining various trainable expert tokens in the Transformer encoder and decoder. Every token is crafted to focus on distinct spatial areas of the image, directed by an orthogonal loss that promotes variety and complementarity among expert representations. While decoding, cross-attention mechanisms enable the expertise of each individual to affect the text generation process, using a metric-based voting system to

determine the final output. The model was assessed utilizing the IU-Xray and MIMIC-CXR datasets. In comparison to traditional Transformer and CNN-RNN baselines, METransformer showed better performance on standard metrics for natural language generation, such as BLEU, ROUGE, and CIDEr. These findings confirm that well-coordinated ensemble-style attention can result in the generation of reports that are more accurate and clinically significant. Although it has its advantages, the model shows some weaknesses. The design lacks domain-specific medical expertise, like structured ontologies or diagnostic guidelines, that could improve the clinical clarity and factual precision of the produced reports. Moreover, the complexity brought on by several expert pathways escalates the model's computational expenses and training demands. Although METransformer highlights the advantages of collaborative reasoning in a Transformer setting, its dependence only on visual cues presents opportunities for enhancement by incorporating external medical knowledge. However, it establishes a hopeful standard for utilizing diversity in focus to more accurately replicate expert-level diagnostic methods.

Within the domain of CNN-Transformer models,(Quigley et al., 2025) create a new model named RadTex which brings a transition from contrastive learning to a generative approach that is more effective in grasping the intricate semantics of radiology. Conventional medical vision-language pretraining (MVLP) techniques such as ConVIRT and GLoRIA employ contrastive objectives to synchronously align image and text features at global or local levels. Nonetheless, these techniques frequently face challenges in achieving the detailed, sentence-level alignment needed for producing coherent radiology reports. To address this, RadTex utilizes a bidirectional captioning-focused pretraining method that prioritizes language modeling rather than image-text contrast. RadTex includes a convolutional encoder to extract image features and a Transformer decoder to produce reports. The model is trained using matched chest X-ray images alongside their related radiology reports. It employs next-token prediction bidirectionally to create deeper semantic connections between image areas and text descriptions. This design enables the model to generate interpretable and clinically significant reports, even when trained on minimal data. Even with a smaller CNN encoder and a limited training dataset, RadTex attains impressive results: a CheXpert macro-AUC of 89.4% and a macro-F1 score of 0.349 in generating reports. Its design is streamlined enough for single GPU deployment, and its adaptability to prompting techniques enables fine-tuning for various clinical situations. Nonetheless, constraints remain. The model's effectiveness is limited by the breadth and variety of its pretraining data. It also does not include clear integration of structured clinical knowledge, which could enhance both interpretability and factual accuracy. In addition, although encouraging, depending on generative techniques necessitates cautious management to prevent hallucinations or excessive confidence. Nonetheless, RadTex signifies a notable advancement in MVLP, demonstrating how generative captioning can outshine contrastive pretraining in radiology use cases.

ARTICLE	AUTHOR	MODEL	METHOD	RESULTS	DATASET	LIMITATIONS
METransformer	(Z. Wang et al., 2023a)	Transformer + Expert Tokens	Multi-expert attention, orthogonal loss, voting mechanism	BLEU-4= 0.172	IU-Xray, MIMIC-CXR	No medical knowledge, computational overhead
Transformer with Demographics	(Aksoy et al., 2023)	Efficient Net + Transformer	Visual + demographic embeddings (age, gender, ethnicity)	BLEU-4= 0.091 ± 0.002	MIMIC-CXR, MIMIC-IV	Bias risk, hallucinations, rare finding errors
RadTex	(Quigley et al., 2025)	CNN + Transformer Decoder	Bidirectional captioning-based pretraining	AUC 89.4%, F1 0.349	Chest X-ray pairs (CheXpert-like)	Small encoder, limited data, no clinical priors

Table 3: *Summary of Studies on Radiology Report Generation Based On CNN-Transformer Methods.*

3.3. Full Transformer-based methods

(Agarwal & Verma, 2025)'s research illustrates a comprehensive Transformer-based method by combining a Vision Transformer (ViT) encoder with a GPT-4 language decoder to produce intricate and context-sensitive radiology reports. The suggested framework, CrossViT-GPT4, substitutes conventional convolutional and recurrent components with a transformer-exclusive architecture that can identify spatial patterns and generate linguistically detailed descriptions. The ViT encoder transforms chest X-rays into embeddings at the patch level, maintaining spatial and positional context. A cross-modal attention mechanism subsequently associates these image representations with pertinent medical terminology, enabling the GPT-4 decoder to produce fluent and clinically cohesive narratives. The Indiana University (IU) and NIH Chest X-ray datasets were used for training and evaluating the model. CrossViT-GPT4 delivered exceptional

outcomes: it recorded a BLEU-1 of 0.854, CIDEr of 0.883, METEOR of 0.759, and ROUGE-L of 0.712 on the IU dataset. In the NIH dataset, it achieved BLEU scores reaching 0.825 and a CIDEr score of 0.857. These findings highlight its enhanced capability compared to conventional CNN-RNN and hierarchical LSTM models, especially in preserving long-range cohesion and merging visual semantics with text output. Although it has strengths, the model encounters various challenges. Dependence on extensive computation results in high costs for training and inference, complicating deployment in resource-limited settings. Moreover, the lack of annotated data for radiology reports restricts its reliability across various clinical situations. The model demonstrates lower performance when encountering noisy or low-quality inputs. However, by leveraging the language comprehension of GPT-4 and the spatial accuracy of ViTs, CrossViT-GPT4 represents a notable progression in transformer-oriented medical imaging technologies and sets the stage for more in-depth, automated diagnostic instruments.

Set within the complete Transformer-based category, (Shisu et al., 2024) presents a new vision-language framework that combines the advantages of Vision Transformers (ViTs) with an evolutionary algorithm (EA)-influenced design to enhance medical image classification named EATFormer. The motivation arises from the constraints of human-driven diagnostics and conventional CNN systems, which frequently lack consistency and do not effectively grasp global image context. The architecture of EATFormer improves feature extraction by integrating several specialized modules within a hierarchical transformer structure. The architecture features a tailored transformer block termed the Enhanced EA-based Transformer (EAT), which incorporates several essential elements: the Multi-Scale Region Aggregation (MSRA) module for combining features across varying scales; the Global and Local Interaction (GLI) module to improve attention with spatial accuracy; and the Modulated Deformable Multi-Scale Attention (MD-MSA) mechanism to flexibly adjust to irregularities in medical images. Moreover, the Task-Related Head (TRH) customizes results to align with the particular classification goals. These improvements enable EATFormer to identify both detailed and broader patterns without relying significantly on positional encoding. When assessed using the Chest X-ray and Kvasir datasets, EATFormer showed enhanced classification accuracy and faster prediction speed relative to conventional CNN and standard ViT models. Its tokenization system based on patches and four-tiered hierarchical framework also enhance scalability and processing efficiency. Although the document does not specifically list any limitations, some conclusions can be drawn. The model's intricate nature and the incorporation of various specialized modules probably require significant computational power and may impede real-time implementation. Moreover, the ability to generalize to other imaging fields or clinical environments might rely on the diversity of the dataset. Nonetheless, EATFormer exemplifies the increasing sophistication of ViT-based models in clinical AI and highlights the effectiveness of hybrid approaches that combine neural architecture advancements with bio-inspired optimization.

ARTICLE	AUTHOR	MODEL	METHOD	RESULTS	DATASET	LIMITATIONS
CrossViT-GPT4	(Agarwal & Verma, 2025)	ViT + GPT-4	Cross-modal attention, ViT encoder, GPT-4 decoder	BLEU-1: 0.854, CIDEr: 0.883 (IU); BLEU: 0.825 (NIH)	IU X-Ray, NIH CXR	High compute cost, poor robustness to noise, data scarcity
EATFormer	(Shisu et al., 2024)	ViT + EA modules	MSRA, GLI, MD-MSA, TRH, hierarchical ViT	Higher accuracy = 0.9533	Chest X-ray, Kvasir	Likely compute-intensive, generalizability not tested

Table 4: *Summary of Studies on Radiology Report Generation Based On Full Transformer Methods.*

3.4. Vision Language Multimodal (VLMs)

In the evolving Vision-Language Multimodal (VLM) framework, CoDiXR presents a versatile generative model aimed at producing chest X-ray images along with their associated radiological narratives created by (Molino et al., 2025). Rooted in the Composable Diffusion (CoDi) framework, CoDiXR facilitates "any-to-any" cross-modal creation, for instance, generating side views from front images or creating clinical narratives from X-ray images. This flexibility meets increasing needs for artificial medical data to aid data augmentation, safeguard privacy, and enhance diagnostic tool creation in AI-powered healthcare. The system utilizes a combination of Latent Diffusion Models, contrastive learning, and self-supervised methods. Utilizing the MIMIC-CXR dataset, it processes images with uniform resizing and normalization, considering frontal and lateral X-rays as separate modalities to improve cross-view consistency. The model generates high-quality results assessed through numerical metrics: it demonstrates excellent performance on Fréchet Inception Distance (FID) for image quality and BLEU scores for text creation. In downstream disease classification tasks, the synthetic data produced by CoDiXR matches or surpasses the performance of real-world data, indicating its potential value in clinical AI workflows. Regardless of these advantages, CoDiXR has significant shortcomings. Its effectiveness decreases when depending exclusively on visual inputs, suggesting a possible imbalance in its cross-modal training approach. Furthermore, although assessment through proxy tasks is promising, there is no formal clinical validation

to guarantee the safety or diagnostic accuracy of the results. These issues highlight the significance of fine-tuning for specific domains when modifying general-purpose generative models for healthcare. Nevertheless, CoDiXR establishes an important benchmark for multimodal synthesis in radiology, providing innovative avenues for scalable, privacy-aware data creation. Future directions involve enhancing cross-view generalization and performing expert assessments to evaluate practical applicability.

(Pellegrini et al., 2025) creates another method based on VLMs, RaDialog represents a novel category of Vision-Language Multimodal Models (VLMs) aimed at both producing radiological reports and facilitating interactive discussions with healthcare professionals. The model is designed as an AI assistant with human input, improving radiology processes by allowing clinicians to create, edit, and discuss reports instantly. RaDialog integrates visual data from chest X-rays, structured pathology labels obtained through a CheXpert classifier, and the linguistic abilities of a large language model (LLM). A prompt-engineering component combines these resources into adaptable directives that steer multi-task results like report creation, editing, and responding to questions. Training and assessment are performed on the MIMIC-CXR dataset, supplemented by a semi-automatically annotated instructional dataset designed for radiology. The training data accommodates various conversational styles and tasks, allowing the model to sustain domain-specific dialogue while preserving general LLM abilities. To enhance computational efficiency, RaDialog employs parameter-efficient fine-tuning rather than complete retraining. This enables efficient domain adaptation without the substantial expenses usually linked to large-scale LLMs. In terms of performance, RaDialog shows a 7.3% enhancement in diagnostic effectiveness, surpassing both general models like MedPaLM and radiology-focused benchmarks like ELIXR. Its interactive features enhance collective decision-making, making it an effective tool for clinical teamwork. Nonetheless, there are still limitations: the system presently accommodates only single-view images, and any mistakes made by the CheXpert classifier may carry over into the ultimate output. Additionally, the model's effectiveness has yet to be confirmed through clinical trials. Potential advancements could include the incorporation of multi-view input capability, the addition of more detailed patient metadata, and enhanced clinical testing. Nonetheless, RaDialog signifies a significant advancement in the development of interactive, explainable AI systems for medical imaging.

ARTICLE	**AUTHOR**	**MODEL**	**METHOD**	**RESULTS**	**DATASET**	**LIMITATIONS**
CoDiXR	(Molino et al., 2025)	CoDi-based VLM	Latent diffusion, contrastive/self-	High BLEU-4= 0.22	MIMIC-CXR	Struggles with image-only input,

			supervised, any-to-any synthesis			lacks clinical validation
RaDialog	(Pellegrini et al., 2025)	LLM + visual labels	Prompt-based dialogue, CheXpert + LLM + efficient fine-tuning	+7.3% diagnostic accuracy, superior interactivity	MIMIC-CXR + instruction set	Single-view only, label noise propagation, no clinical trials

Table 5: *Summary of Studies on Radiology Report Generation Based On VLMs.*

3.5. Large Language Models (LLMs) + Prompting methods

KARGEN illustrates a large language model method that improves the creation of radiology reports by integrating organized domain expertise into a static large language model. Instead of refining the LLM (LLaMA) created by (Y. Li et al., 2024), KARGEN enhances its input by merging visual characteristics from chest X-ray images with disease-specific information derived from a medical knowledge graph. This approach enables the model to generate clinically relevant narratives without altering the foundational language model. The architecture comprises four elements: a Swin Transformer that gathers spatially aware visual embeddings, a Graph Convolutional Network (GCN) that represents interactions among disease concepts, a fusion module that merges visual and graph-based features through either element-wise gating or modality-wise expert weighting, and a report generator driven by a frozen LLaMA decoder. These elements collaborate to synchronize textual output with visual signals and advanced medical associations. KARGEN was assessed using the IU-Xray and MIMIC-CXR datasets. It consistently exceeded traditional baselines and other LLM-only architectures across standard metrics—BLEU, METEOR, ROUGE, and CIDEr. Incorporating specialized disease knowledge enhanced both the accuracy and contextual richness of the produced reports. Nonetheless, the constraints of KARGEN arise from the breadth and depth of the medical knowledge graph utilized. Broadening the graph's scope to capture more intricate clinical connections might boost model effectiveness and versatility. Furthermore, although the fusion strategies are effective, they add architectural complexity that could affect inference speed in clinical environments. Nevertheless, KARGEN establishes a significant benchmark for integrating structured medical information with LLMs, underscoring the promise of prompt-augmented LLM workflows in clinical NLP use cases.

(Z. Wang et al., 2023b)'s model R2GenGPT is part of the new category of prompt-driven frameworks that utilize static large language models (LLMs) to connect the modality divide between image inputs and text-based diagnostic reports. It offers a modular

approach where only a simple visual-to-text mapping layer undergoes training, enabling LLMs to understand image features without requiring significant retraining. The architecture consists of three main elements: a visual encoder (e.g., Swin Transformer), a visual mapper, and a static LLM like LLaMA or BioClinicalBERT. The visual mapper translates visual attributes into the LLM's word embedding space, allowing the direct input of image-derived "tokens" into the fixed decoder. R2GenGPT investigates three mapping approaches: Shallow Alignment (training is limited to the projection layer), Delta Alignment (slightly modifies the LLM, about 0.07% of parameters), and Deep Alignment (involves more extensive training). The model was evaluated on the IU-Xray and MIMIC-CXR datasets, achieving performance that equaled or surpassed leading models in BLEU, METEOR, ROUGE, and CIDEr. Its training was remarkably efficient in computation, needing little memory and reaching convergence rapidly. These benefits render R2GenGPT appropriate for clinical settings where resources could be scarce. Nonetheless, certain constraints persist. Due to the model depending on mapping into a static embedding space, it might not entirely leverage subtle visual semantics. Interpretability poses a challenge, since the intermediate alignment mechanism functions like a "black box." Furthermore, the decoder-only configuration might restrict its ability to represent inter-sentence coherence. Despite these concerns, R2GenGPT offers a scalable, sophisticated approach that utilizes the linguistic capabilities of LLMs while ensuring efficiency and modularity.

(Zeng et al., 2024) creates RadCouncil which presents an innovative multi-agent framework for generating radiology reports using LLMs, highlighting the importance of collaboration and specialized tasks. Instead of depending on one model to handle every step, RadCouncil divides the impression writing process into three agent roles that reflect actual radiology workflows. This method closely aligns with the movement toward prompt-based, retrieval-augmented generation (RAG) strategies in large language modeling. The system comprises three agents: a Retrieval Agent that employs vector similarity to identify pertinent previous reports; a Radiologist Agent that generates impressions from current discoveries and retrieved cases; and a Reviewer Agent that assesses and improves the output for diagnostic precision and stylistic uniformity. The architecture prevents complete model retraining through prompt-based task assignment, enabling swift deployment and tailored domain customization. Assessment of chest X-ray datasets (precise sources not detailed) integrated standard NLG metrics (BLEU, ROUGE, BERTScore) alongside qualitative evaluations from GPT-4. RadCouncil surpassed individual agent baselines in diagnostic accuracy, adherence to radiological standards, and linguistic fluency. The Reviewer Agent was instrumental in minimizing hallucinations and enhancing factual accuracy, confirming the effectiveness of the system's multifaceted error-checking framework. Nevertheless, RadCouncil encounters constraints concerning memory and reasoning abilities. The RAG pipeline is limited by the size of its context window, restricting the amount of information that can be retrieved in each instance.

Additionally, multi-agent coordination increases complexity and might necessitate stronger memory management for long-term scenarios. However, RadCouncil illustrates the promise of collaborative, agent-based LLM systems in clinical settings and paves the way for more sophisticated, understandable, and interactive medical AI solutions.

ARTICLE	AUTHOR	MODEL	METHOD	RESULTS	DATASET	LIMITATIONS
KARGEN	(Y. Li et al., 2024)	Frozen LLaMA + GCN + Swin	Visual + knowledge graph fusion with gated/expert pathways	BLEU-4= 0.180 on IU-Xray	IU-Xray, MIMIC-CXR	Limited graph coverage, fusion complexity
R2GenGPT	(Z. Wang et al., 2023b)	Frozen LLM + Visual Mapper	Shallow/Delta/ Deep alignment from vision to LLM embeddings	BLEU-4= 0.173 on IU-Xray	IU-Xray, MIMIC-CXR	Bottleneck in alignment, limited coherence, interpretability
RadCouncil	(Zeng et al., 2024)	Multi-agent LLMs + RAG	Retrieval + Draft + Review agents	BLEU = 24.22	Chest X-ray	Context window constraints, agent orchestration complexity

Table 6: *Summary of Studies on Radiology Report Generation Based On LLMs.*

3.6. Beyond Transformers methods

R2Gen-Mamba presents a hybrid framework that diverges from conventional Transformer-exclusive architectures by combining a Mamba encoder—rooted in state-space modeling—with a Transformer-based decoder created by(Sun et al., 2024). This design tackles the computational inefficiencies present in Transformer architectures, especially their quadratic complexity in managing lengthy sequences. Mamba, featuring linear time complexity, enables effective sequence modeling while maintaining contextual comprehension, making it exceptionally appropriate for real-time medical uses. The architecture of the model is designed so that the Mamba encoder converts chest X-ray images into spatially aware representations, subsequently fed into a Transformer decoder that creates the relevant diagnostic report. This setup achieves a balance between semantic

depth and resource efficiency, targeting the requirements for precise clinical analysis while providing scalability in resource-limited settings. Assessments carried out on the IU X-ray and MIMIC-CXR datasets indicate that R2Gen-Mamba surpasses conventional Transformer-based baselines in important metrics like BLEU, METEOR, ROUGE, and CIDEr. These findings emphasize its capability to sustain or improve language quality while lessening computational requirements. Significantly, its hybrid characteristics preserve the contextual power of Transformers in language generation while attaining a remarkable acceleration in feature extraction. Nonetheless, the model's reliance on a Transformer decoder indicates it has not entirely avoided the limitations linked to attention mechanisms. Moreover, the degree to which the hybrid framework can be applied to other medical imaging types is still unclear. Nonetheless, R2Gen-Mamba establishes essential foundations for investigating effective, high-performance options in radiology report creation, particularly for scenarios where infrastructure limitations hinder the implementation of conventional Transformer models.

SERPENT-VLM is (Sun et al., 2024) model. It marks a notable shift from traditional static Transformer models by incorporating a self-improving, feedback-oriented framework for generating radiology reports. Instead of depending on a rigid sequence-to-sequence process, the model progressively adjusts its output through an innovative loss function that correlates the visual elements of an image with the meaning of the produced report. This method minimizes hallucinations and enhances factual accuracy—critical issues in clinical AI systems. The process starts with a static visual encoder to obtain high-dimensional characteristics from chest X-rays. These attributes act as input for a large language model (LLM), which generates a preliminary report. A self-tuning loss function subsequently evaluates the combined visual representation against the embedding of the produced text, prompting the model to modify and enhance its output for improved alignment. This loss enhances conventional causal language modeling goals, enabling improved semantic management. Assessed on the IU X-ray and ROCO datasets, SERPENT-VLM delivers top-tier outcomes, exceeding the performance of sophisticated models like BiomedGPT and LLaVA-Med on various benchmarks. It demonstrates notable resilience, preserving functionality even with noisy or low-quality images frequently found in clinical datasets. In addition, its rapid inference speed makes it suitable for implementation. Nonetheless, the model has constraints. It has been assessed solely on a limited variety of datasets, raising doubts about its applicability to other radiological situations. Additionally, its feedback system might enhance training complexity, creating difficulties for reproducibility and scaling. Nonetheless, SERPENT-VLM establishes an impressive benchmark for adaptive, self-repairing systems in medical vision-language modeling and may motivate a new wave of progressive, dependable clinical report generators.

(X. Wang et al., 2024) built MambaXray-VL which is a next-generation radiology report generation model that bypasses traditional Transformer architectures by adopting the Mamba framework—a state-space sequence model known for its linear scalability and memory efficiency. This model is particularly designed to handle long-sequence data efficiently, making it ideal for clinical environments where computational resources are constrained.At the heart of MambaXray-VL is a non-Transformer vision encoder based on the Mamba architecture. It is paired with pretrained LLMs such as BioClinicalBERT or LLaMA2 for report decoding. The training pipeline consists of three phases: self-supervised autoregressive modeling from image segments, contrastive learning to align X-ray images with their textual reports, and supervised fine-tuning using standard evaluation metrics. This modular and scalable approach facilitates more robust visual-textual alignment without relying on quadratic attention.The model is assessed using the newly introduced CXPMRG-Bench benchmark, which includes 19 competing systems—14 LLM-based and 2 vision-language models—evaluated across datasets like CheXpert Plus, IU X-ray, and MIMIC-CXR. MambaXray-VL outperforms all baseline models in both language generation and interpretability metrics, demonstrating its suitability for high-stakes clinical applications.Nonetheless, certain limitations exist. While benchmark results are strong, the model's real-world utility is yet to be verified through clinical trials or radiologist validation. Additionally, the novel benchmark, while comprehensive, could limit comparability with prior work. Still, MambaXray-VL reaffirms the viability of state-space models in radiology and sets a new standard for future non-Transformer vision-language systems, offering a computationally lean alternative without sacrificing performance or clinical relevance.

ARTICLE	**AUTHOR**	**MODEL**	**METHOD**	**RESULTS**	**DATASET**	**LIMITATIONS**
R2Gen-Mamba	(Sun et al., 2024)	Mamba Encoder + Transformer Decoder	Hybrid sequence modeling for efficiency and fluency	BLEU-4 = 0.176 on IU-Xray	IU X-ray, MIMIC-CXR	Transformer decoder bottleneck, generalizability
SERPENT-VLM	(Kapadnis et al., 2024)	Self-Refining LLM + Vision Encoder	Self-adjusting loss aligns image/text	BLEU-4= 0.190 on IU-Xray	IU X-ray, ROCO	Limited dataset scope, training complexity

			iteratively			
MambaXray-VL	(X. Wang et al., 2024)	Mamba Vision Encoder + LLM Decoder	3-stage pretraining (AR, contrastive, supervised)	BLEU-4 = 0.185 on IU-Xray	CheXpert Plus, IU X-ray, MIMIC-CXR	Benchmark dependency, lacks clinical trials

Table 7: *Summary of Studies on Radiology Report Generation Based On Beyond Transformer*

4. Conclusion

This chapter has presented a structured examination of deep learning techniques that underpin automated radiology report generation. From foundational architectures like CNNs and RNNs to advanced models such as Transformers, BERT, GPT, and Vision-Language Models, each method was analyzed in terms of its capabilities, limitations, and suitability for complex multimodal tasks. Emerging frameworks like Mamba and State-Space Models were also discussed, reflecting the ongoing quest for more efficient and scalable alternatives to traditional attention-based models. The literature review synthesized past research efforts, categorizing them by architectural approach and highlighting their contributions to the field. Together, these insights provide the theoretical and empirical foundation necessary for designing an effective deep learning-based diagnostic captioning system, which is addressed in the following chapter.

Chapter III: Conception

1. Introduction

Artificial intelligence is transforming medicine, making it possible to create systems that help doctors, or even automate certain tasks. The automatic generation of radiology reports is a key example: it requires a good understanding of medical images and the ability to write accurate clinical reports. The design of such a system represents a major multidisciplinary challenge, lying at the intersection of computer vision, automatic natural language processing (ANLP) and medical expertise.

This chapter is devoted to the detailed design of our system for automatically generating radiology reports. We explore the evolution of our architectural approach, from an initial version based on a Long Short-Term Memory (LSTM) decoder that served as a foundation for sequence extraction, to the final, optimized implementation. The latter incorporates a Transformer-type architecture for more robust modelling of contextual dependencies in text, complemented by an innovative post-processing module based on a Large Language Model (LLM). This evolution was necessary to correct text quality and consistency problems encountered at the outset.

We will detail the steps involved in this design: how we prepared the data (images and text), how we extracted the visual information using a network called EfficientNet-B0, and how the report is first generated and then enhanced by the LLM. Each choice was made for technical and medical reasons, in order to create high-quality automatic reports that are useful in hospitals.

2. Objective

This work aims to propose a robust approach for the automatic generation of radiological reports by combining the power of visual deep learning with advanced natural language processing. The system was initially designed using a CNN-LSTM architecture, in which visual features extracted by a convolutional encoder (EfficientNet-B0) were fed to an LSTM decoder responsible for producing the report sequentially. Although this first version produced globally consistent reports, it revealed several limitations, particularly in terms of linguistic fluency, semantic coverage, and terminological consistency.

To overcome these weaknesses, we replaced the LSTM decoder with a Transformer architecture, better suited to modelling complex sequences. This new CNN-Transformer system allows more structured and contextually relevant generation of radiology reports from thoracic images. However, despite this improvement, grammatical errors, lexical approximations and a lack of clinical fluidity persist in certain text productions.

To address these shortcomings, we introduce a final post-processing phase, provided by a large language model (LLM), such as BioGPT. This step acts as a stylistic and linguistic refinement layer, correcting imperfections in the raw report generated by the Transformer, while respecting the diagnostic content. It aims to improve readability, terminological accuracy and compliance with medical writing standards.

3. System architecture

A deep learning-based system for automatically generating radiological reports is proposed in this master's thesis. Our system aims to generate a radiology report from chest X-ray image input.

The proposed system uses several interdependent components to generate clinically useful written reports from medical images. It incorporates a crucial dataset preparation phase, an encoder to ensure accurate classification, which will serve as input to a report generator. Our system architecture is illustrated in the following Figure 3.1.

3.1. Dataset Preparation Phase

This phase is dedicated to pre-processing the data, both the images and the associated radiology reports. In terms of text, the reports are tokenized and cleaned up to eliminate any unnecessary symbols that could interfere with the learning process. As for the images, they undergo augmentation operations (such as random rotation or flipping), scaling, and normalization. These transformations are designed to improve the model's generalization capacity and ensure optimum compatibility with our deep learning architecture.

3.2. Encoding Phase

The system is based on a hybrid encoder-decoder architecture. The encoding module is based on EfficientNet-B0, a pre-trained convolutional neural network responsible for extracting high-level visual representations from chest X-rays. These representations are then used by a decoder to generate the corresponding text report. The model training is based on a double loss function: a binary cross-entropy for multi-label classification (of pathologies) and a standard cross-entropy for text generation. This dual objective enables the coder to capture rich visual features that are both informative for diagnostic classification and consistent with the semantics of radiological reports.

3.3. Report Generation Phase

In an initial version of the system, radiology reports were generated by an LSTM-type decoder, which modelled the report as a sequence of words generated successively from the visual features extracted by the convolutional encoder. Although this approach produced globally consistent reports, it had several limitations: lack of lexical diversity, omission of some important clinical information, and a limited ability to model long-term dependencies.

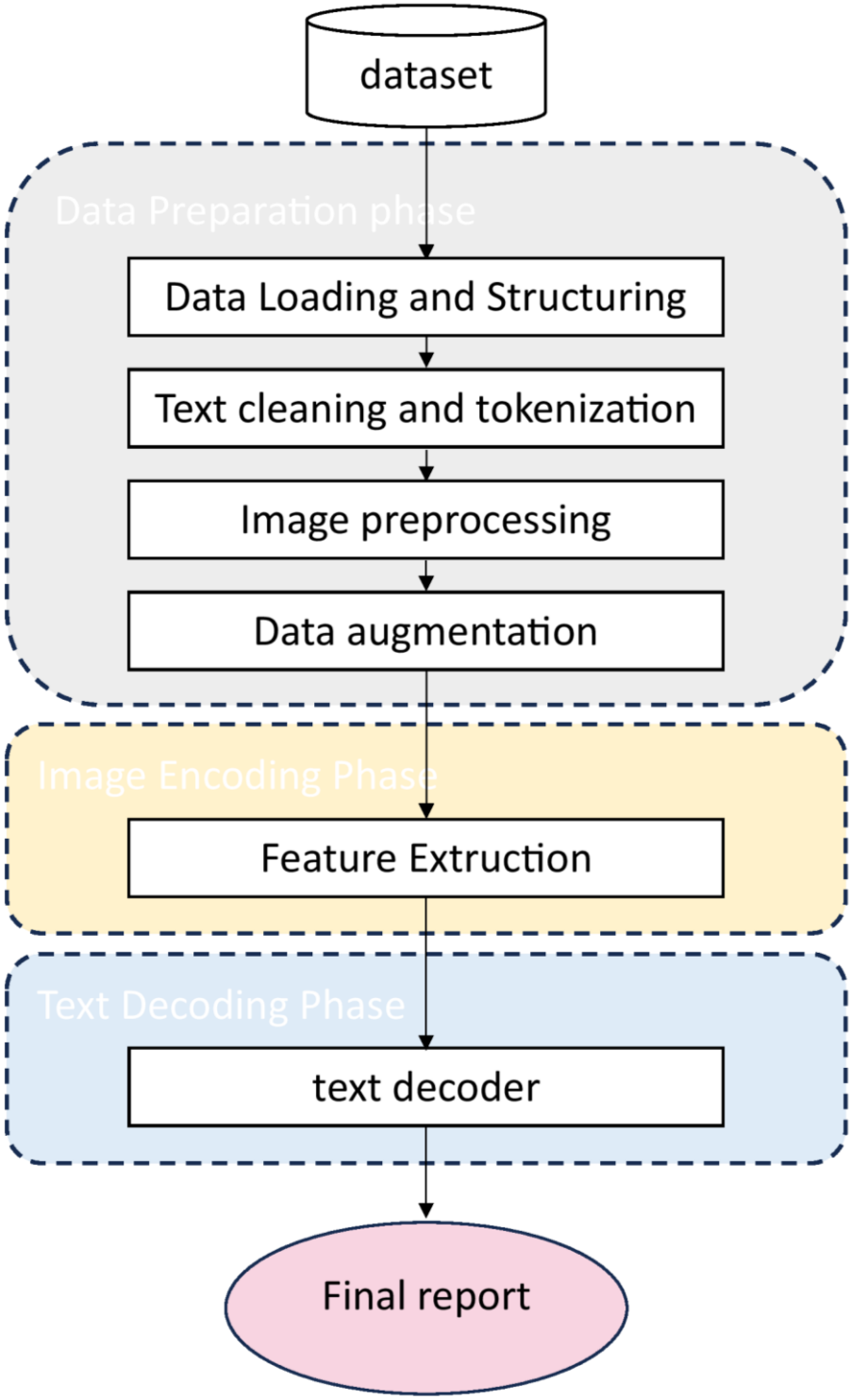


Figure 5: *Automatic Radiology Report Generation System Architecture*

To overcome these weaknesses, the LSTM was replaced by a Transformer decoder, which is better able to model the complex linguistic structure of medical reports. The new decoder generates the text word by word, starting with a token at the beginning of a sequence, and stopping when an end token appears or at a maximum length. The decoding used is greedy decoding, for reasons of computational simplicity, although more advanced alternatives (beam search, nucleus sampling) could be envisaged to improve fluidity and accuracy. Diagnostic labels extracted during the classification phase are injected into the decoder input in order to guide the generation towards medically relevant content.

However, a qualitative analysis of the reports generated revealed recurring errors, including omissions, lexical errors and imprecise wording. To remedy this, a post-processing phase was introduced. This is based on a large language model (LLM), such as BioGPT, used as a stylistic and linguistic refinement layer. The LLM only intervenes on the generated text, without reconsidering the image, and aims to improve legibility, syntax and compliance with medical writing standards. This final module produces reports that are more professional, more natural and better aligned with clinicians' expectations. Specific LLM training on a radiology corpus could constitute a further improvement prospect.

4. Data Preparation, Preprocessing, and Augmentation

To ensure high-quality, correctly structured input to our automatic radiology report generation system, we have implemented a rigorous data preparation pipeline. This process combines image augmentation, report tokenization, image normalization and structured text formatting, making it easy to train the encoder-decoder architecture.

4.1. Data Loading and Structuring

The data used in this study comes from the Indiana University Chest X-ray Collection, which provides chest X-rays accompanied by their medical reports. The dataset also includes two metadata files:

- `indiana_projections.csv`, associates the unique identifiers (uid) with the standardized names of the image files.
- `indiana_reports.csv`, contains the text of the reports, divided into two sections: `impression` and `findings`.

To obtain a coherent and usable dataset, the following steps were taken:

- **Filename Matching:** each ‘uid’ from the reports was matched with its corresponding image using the projection file.

- **Report Cleaning:** the text in the findings and print columns was pre-processed to remove excessive punctuation, standardize white spaces and remove bogus strings (such as strings of letters “X”).
- **Report formatting**: the two text sections have been concatenated into a single string, framed by structuring tokens:

Findings:

{findings}

Impression:

{impression}

This structuring enables the decoder to distinguish between the different parts of the report while ensuring syntactic and semantic consistency. Only samples with both a valid image and non-empty reports were retained. The final dataset is represented as a table containing two columns, `image_path` and `report`.

4.2. Text Tokenization

The reports were tokenized using the “Keras Tokenizer” module, configured as follows:

- **No Filtering:** all characters, including punctuation, were retained to preserve useful semantic clues.
- **Case Sensitivity:** capitalization was maintained (lower=False) to distinguish acronyms and anatomical names.
- **Management of rare words:** words absent from the learned vocabulary have been replaced by the special <unk> token.

After fitting the tokenizer on the full corpus of reports:

- Each report was converted into a sequence of word indices.
- The longest sequence length was determined and used as the maximum length for padding and truncation.
- Two sequences were generated per report:
    - `decoder_inputs`: all tokens except the last one
    - `targets`: all tokens except the first

This strategy allows the use of teacher forcing during training, by providing the model with the previous tokens as a reference for predicting the next one.

4.3. Image Preprocessing

To ensure optimal compatibility and to fully exploit the capabilities of the EfficientNet convolutional encoder, which forms the backbone of our system, all chest X-ray images have undergone a standardized pre-processing process. These steps are crucial for normalizing the input data and aligning it with the format expected by a model pre-trained on large image datasets. The specific transformations applied are detailed below:

- **Image resizing**: Each image was loaded from its storage location and systematically resized to a resolution of 224×224 pixels. This ensures a uniform input size for the EfficientNet network, which is optimized for this dimensionality, and contributes to the consistency of batch processing.
- **Conversion from greyscale to RGB**: X-rays (single channel greyscale) were replicated to three channels (RGB) to fit the pre-trained architecture on ImageNet, without semantic alteration of the original pixels.
- **Pixel normalization**: pixel values were adjusted and standardized according to the parameters of the EfficientNet pre-training on ImageNet (normalization specific to this model).
- **Tensor conversion**: the images were converted to float32 format to enable them to be processed by the deep learning model.

4.4. Data Augmentation

Data augmentation is an essential tool for enhancing the generalization capability of a deep learning model. By exposing the model to a variety of transformations applied to training images, this technique enables it to learn to detect target features under diverse conditions, thereby improving accuracy on test data and limiting overfitting (Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6 (60).).

In this work, we implemented an augmentation pipeline based on the Keras Sequential API. Random transformations were applied to the images during training, without altering the total number of samples, to ensure a diverse input stream for the model. The transformations used include:

- **Random Horizontal Flipping:** Simulates variations in image orientation.

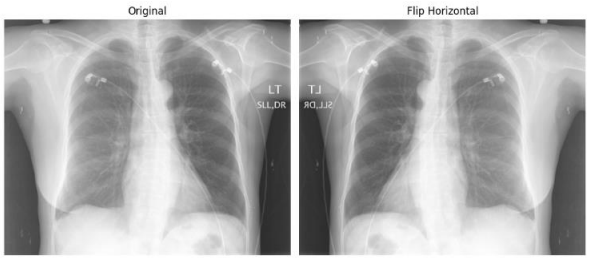


Figure 6: *Flipping of an image from dataset.*

- **Random Rotation:** Accounts for slight differences in patient positioning.

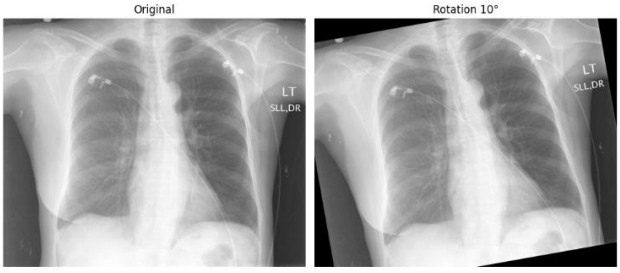


Figure 7: *Rotation of an image from dataset.*

- **Random Zooming:** Introduces spatial variability while preserving core structures.

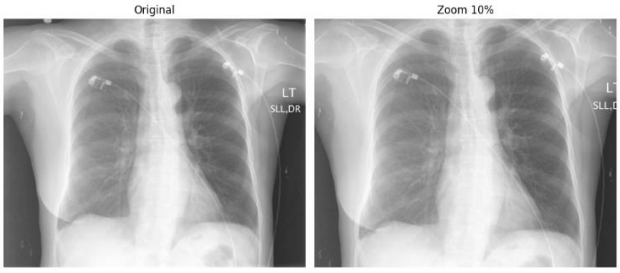


Figure 8: *Rotation of an image from dataset.*

These transformations are safe for chest X-rays and augment the training set with realistic variations. No augmentation was applied to validation or test samples.

5. Model Architecture

The design of the proposed system is based on a hybrid architecture combining a convolutional neural network (CNN) encoder and a decoder. This configuration enables the model to efficiently capture the spatial characteristics of medical images and generate structured text sequences capable of fully and coherently describing a diagnostic report. The CNN encoder is specifically dedicated to extracting discriminating visual information by highlighting the spatial and morphological aspects of the radiographic content. The decoder, meanwhile, is designed to model long-range dependencies within text sequences,

an essential capability for producing radiology reports made up of several interconnected, medically relevant sentences.

The model operates in two main phases:

- **The image encoding phase**, during which visual features are extracted from a chest X-ray and transformed into a compact, informative representation.
- **The text decoding phase**, in which the decoder uses this visual representation to generate the corresponding report, producing the tokens one by one, according to a sequential logic. The model functions in two primary phases:

This modular architecture not only offers great flexibility for adaptation to other types of data or medical tasks, but also greater transparency in the training, inference and development processes for future extensions. As a result, it provides a solid foundation for the production of automated reports aligned with the requirements of the clinical field.

5.1. Image Encoder (EfficientNetB0)

Our system's image encoder is based on the EfficientNet-B0 architecture, a model recognized for its optimal balance between performance, accuracy and computational efficiency. This backbone makes it possible to extract deep, discriminating visual features from chest X-rays.

- **EfficientNetB0 Backbone**: EfficientNet-B0 is the core of the encoder. Its design is based on coordinated scaling of depth, width and resolution, optimizing the use of resources for maximum accuracy. The model is pre-trained on ImageNet and the convolutional layers are frozen (not re-trained) to limit the risk of overfitting on our specific medical dataset. The complete architecture of EfficientNet-B0 is illustrated in Figure 3.5.
- **Input Layer:** The encoder expects input images of dimensions (224, 224, 3), conforming to standard EfficientNet specifications.
- **Convolutional Layers:** The EfficientNet-B0 network is composed of inverted residual blocks incorporating squeeze-and-excitation modules. These layers enable hierarchical and progressive feature extraction, from local details to abstract representations.
- **GlobalAveragePooling2D:** This layer is applied to the last convolutional feature map (typically $7 \times 7 \times 1280$). It averages the activations over the spatial dimensions, producing a global vector of 1280 dimensions summarizing the entire image
- **BatchNormalization:** This layer standardizes the resulting vector in order to speed up convergence during training and stabilize weight updates. It adjusts the mean and variance of activations for each batch of data.
- **Dropout:** A regularization mechanism is introduced via a dropout that randomly deactivates 30% of activations during training, reducing the risk of overfitting.

- **Dense Layer:** This transformation reduces the 1280-dimensional representation to a more compact 512-dimensional vector. ReLU activation introduces a non-linearity that is essential for capturing complex patterns.
- **Final Dropout:** A second dropout is applied after the dense layer to reinforce the regularization of the final image representation.

The encoder thus delivers a 512-dimensional dense vector representation, which is an abstract, compressed synthesis of the visual content of the chest X-ray. This vector acts as a conditional input for the Transformer decoder responsible for generating the text report.

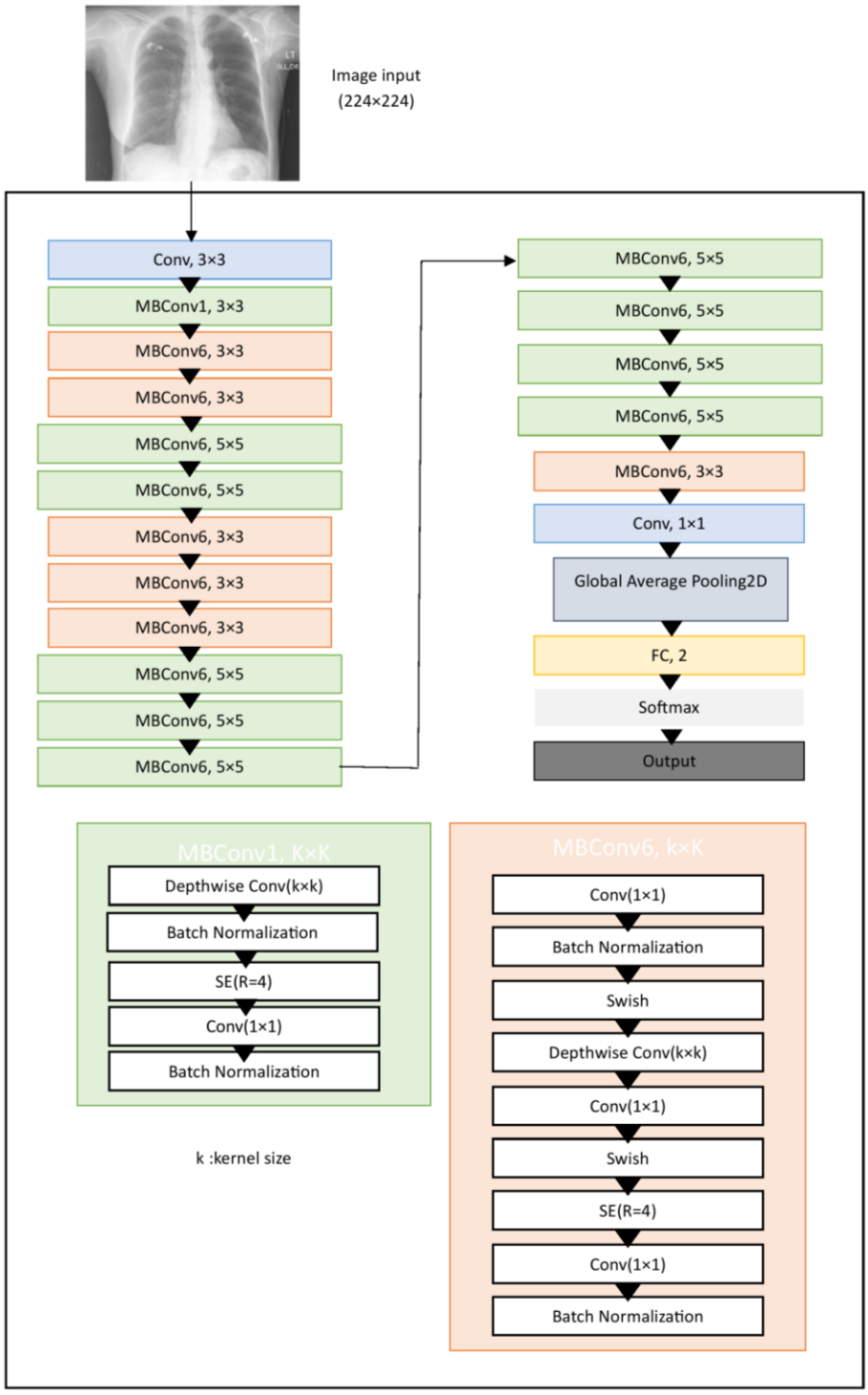


Figure 9: *The architecture of the used EfficientNetB0 pre-trained model in our system.*

5.2. Transformer Decoder

The generation of the text report is based on a decoder that transforms the visual features extracted from the image into a sequence of words forming a structured medical report. Two architectures were studied and implemented in this project: an initial version based on an LSTM-type decoder, and a second, more advanced version using a Transformer.

- **Initial version: LSTM decoder**

The first decoder implemented was based on a Long Short-Term Memory (LSTM) model, designed to generate the report text sequentially, one word at a time.

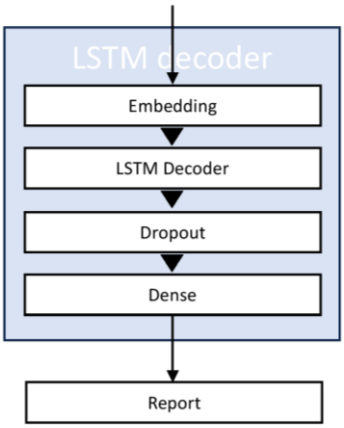


Figure 10: *The architecture of the used LSTM decoder.*

Although the LSTM decoder was able to generate understandable reports, it showed limitations in modelling long dependencies and producing fluent text over several complex sentences.

- **Improved version: Transformer decoder**

To overcome these limitations, a decoder based on the Transformer architecture was developed, capable of better handling the long-range relationships between words in the report. It is designed to generate sequences. It analyses visual data as well as previous words to predict the next word in the report.

- **Input Embedding:** Transforms each token index into a trainable 512-dimensional embedding vector. The embedding matrix is of size (vocab_size, 512), with each row representing a word vector.

- **Positional Encoding:** Positional encodings are included with each embedded token to integrate temporal order, enabling the model to distinguish between tokens

within a sequence. Sinusoidal encodings offer a distinct representation for every position.

- **Image Modification:**
    - The 512-dimensional image feature is linearly transformed to align with the input size of the decoder.
    - An additional axis is introduced to adjust the shape to (1, 512), aligning with the sequence dimension.
    - This visual token is added before the text token embeddings, allowing the model to consider image context from the beginning.
- **Transformer Blocks (2 layers):** Every block consists of:
    - **Multi-Head Self-Attention:** Each head focuses on distinct subspaces of the input. Attention weights are determined using scaled dot-product attention.
    - **Dropout (rate=0.2):** Used following attention to avoid overfitting.
    - **Incorporate & LayerNorm:** A residual link combines the attention output with the input, succeeded by layer normalization for consistency.
    - **Feed-Forward Network (FFN):** Two Dense layers, each with 512 units, interspersed with a ReLU activation function. FFNs incorporate non-linear transformations and enhance representational power.
    - **Second Dropout** and residual normalization complete the block.
    - **Final Linear Layer:** A TimeDistributed Dense layer maps the output from the last decoder layer to the size of the vocabulary. A softmax function subsequently transforms logits into probabilities for every word in the vocabulary.

This decoder architecture enables the model to produce reports one token at a time, taking into account both image features and the series of words previously generated. The attention mechanism enables selective emphasis on significant tokens and the visual.

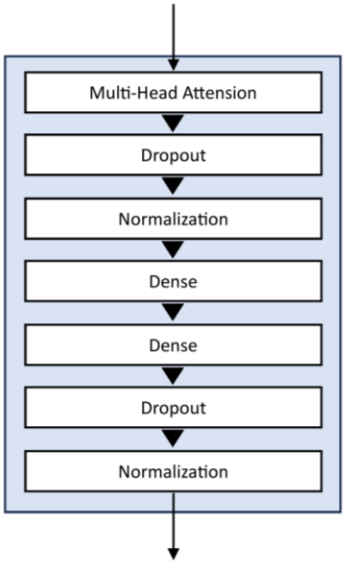


Figure 11: *The architecture of the used Transformer decoder in our system.*

6. Post-processing

After the initial generation of reports using the CNN-Transformer architecture, a complementary post-processing phase is integrated to improve the linguistic quality, semantic consistency, and diagnostic fidelity of the texts produced. This phase has two main components: inference on the test set and correction of the reports generated using a large specialized language model, BioGPT.

- **Inference with CNN-Transformer**

The aim of inference is to evaluate the system's performance on unprecedented data, reflecting use in real-life conditions. Each image-report pair in the test set is processed in batches, with the decoder generating a report in the form of a sequence of tokens. These tokens are then translated into text using the saved tokenizer. Special tokens (start, end and padding) are removed to produce a clean, readable final text.

The dataset produced at the end of this phase contains three aligned columns:

- image_path: the file path to the X-ray image;
- original_report: the reference ground truth (reference report provided by an expert);
- generated_report: the model-generated diagnostic text

This set is saved for subsequent evaluation tasks and quality analyses.

6.1. Correction Using BioGPT

In order to increase the clinical accuracy and readability of the generated reports, a correction phase based on a pre-trained language model, BioGPT, is applied. This step is based on a prompt-driven approach, where the generated reports are enriched by constructing input examples combining the original reports and the model predictions.

The main steps in this phase are:

- Tokenization of prompts and outputs using BioGPT's specific tokenizer;
- Optimization of the model in causal language modelling mode, using the AdamW optimize

Experiments are carried out with two alignment strategies - shallow alignment and deep alignment - aimed at harmonising visual and linguistic information.

Corrected reports are generated using beam search decoding to improve text diversity and consistency. Linguistic and semantic metrics are then calculated to quantify the final quality of the reports.

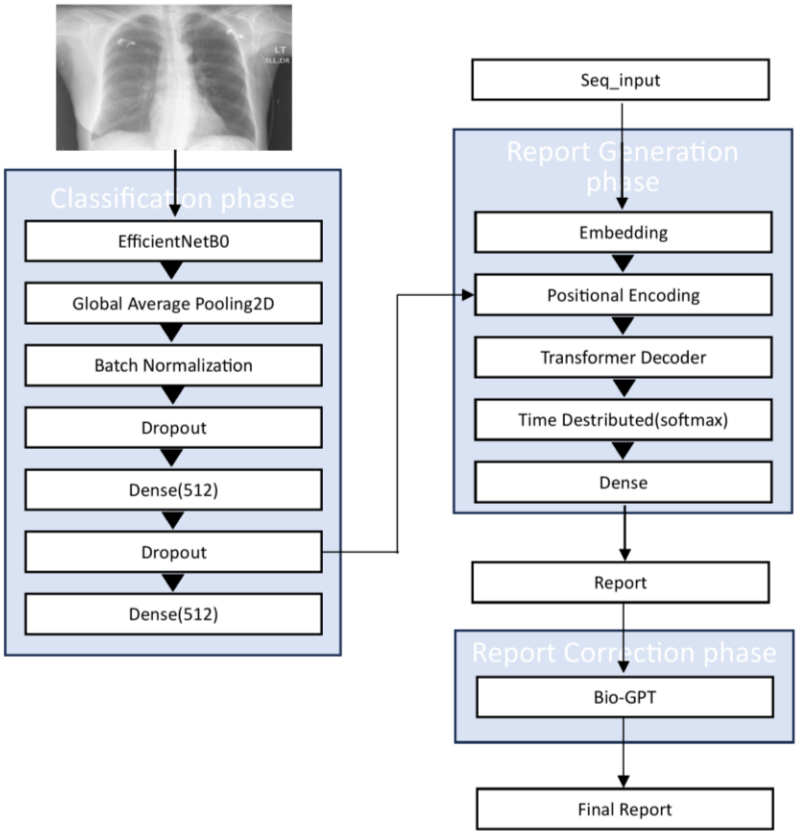


Figure 12: *The architecture of the full system.*

7. Conclusion

The automatic generation of radiological reports is a major challenge for AI in medicine, with the aim of optimizing workflows, standardizing practices and improving the quality of care in the face of increasing numbers of examinations. The challenge is to translate complex images into precise medical language and to guarantee the robustness of the model.

This chapter has detailed the architectural design of the system. It presented the evolution of an LSTM decoder towards a more advanced hybrid architecture: an EfficientNet-B0 encoder for visual extraction, and a Transformer decoder for text generation. To refine linguistic and terminological quality, a post-processing module based on a Large Language Model (LLM) has been integrated. The importance of data pre-processing (text and image) and dynamic data augmentation for model robustness was also highlighted.

Chapter IV: Implementation and Realization

1. Introduction

This chapter describes the practical implementation of our radiology report generation system, transforming the conceptual framework into an operational deep learning pipeline. We detail the development environment, tools (Python, Jupyter Notebook, Kaggle) and libraries used.

The chapter then explores the detailed implementation of the system components. This includes data pre-processing, report tokenization and image formatting for EfficientNetB0, which acts as a visual encoder. The architecture of the model is presented, with a custom Transformer decoder for language generation. Attention is given to training procedures, hyperparameters, and the integration of BioGPT for post-processing of generated reports. Overall, a complete and specialized pipeline for the automatic generation of medical imaging reports is presented.

2. Environment and Tools

2.1. Programming Language

- **Python:**

Python is a high-level, interpreted, and object-oriented programming language recognized for its straightforward syntax and dynamic semantics. Its inherent data structures, along with dynamic typing and binding, render it perfect for quick application development and scripting assignments. Python encourages code reuse and modularity by supporting modules and packages, with its comprehensive standard library accessible on various major platforms. A major advantage of Python is its quick edit-test-debug cycle, which boosts developer efficiency. Errors are managed via exceptions instead of segmentation faults, and debugging is supported by an interactive source-level debugger along with Python's introspective features. For numerous tasks, straightforward print-based debugging continues to be remarkably efficient because of Python's quick feedback(Foundation, 2025).

2.2. Development Environment

- **Jupyter Notebook**

Jupyter Notebook is a free web application that allows users to generate and distribute interactive documents, previously referred to as IPython Notebooks. It offers an online

platform for Python coding, enabling users to create and run code in organized sections that can be integrated with descriptive text and data. A "notebook" may denote the Jupyter web interface, the core Python server, or the final document produced. Used extensively in different fields, Jupyter facilitates activities like data cleaning, numerical simulation, statistical modeling, and machine learning (DataScientest, 2025)

- **Kaggle**

Kaggle is a prominent online platform for data science and machine learning, featuring a worldwide community of more than 500,000 members from 194 nations. It provides a robust, configuration-free setting for building models with Jupyter Notebooks, featuring access to free GPUs and a wealth of community resources. Users have access to more than 50,000 public datasets and 400,000 notebooks to aid their projects. Kaggle is relied upon by large corporations such as Walmart and Facebook, allowing users to join competitions, exchange code, and work together with others. Subjects cover a broad spectrum of areas—from healthcare forecasts to emotion interpretation—creating an active environment for education, skill development, and networking with professionals (DataScientest, 2024)

2.3. Model Construction Tools

- **TensorFlow**

TensorFlow is a popular open-source platform for machine learning that functions through data flow graphs. In this framework, nodes symbolize mathematical operations, while edges denote tensors—multidimensional data arrays—moving between them. This architecture allows for the development and training of machine learning models on CPUs, GPUs, and TPUs, from mobile devices to robust servers, without modifying the foundational code. Initially created by Google's Brain Team for deep learning studies, TensorFlow has evolved into a flexible resource embraced by data scientists, developers, and educators for various machine learning applications(NVIDIA, 2025c).

- **Keras**

Keras is an advanced deep learning API developed in Python that is compatible with various backends, such as TensorFlow, PyTorch, and JAX. Created for simplicity and adaptability, it enables users to construct intricate models with little coding while also permitting sophisticated customizations. Keras 3 enhances this flexibility by allowing developers to train and deploy the identical model across various frameworks without changes. It accommodates multiple data formats including NumPy, Pandas, TensorFlow datasets, and PyTorch DataLoaders. Keras is extensively utilized in research and industry, facilitating rapid development, wide compatibility, and effective deployment on various platforms(K. Team, 2024).

- **PyTorch**

PyTorch is a free deep learning framework created by Facebook AI Research, highly valued for its adaptability, ease of use, and integration with Python. Created for constructing neural networks, PyTorch facilitates dynamic computation graphs (define-by-run), rendering it perfect for quick prototyping and research. It includes reverse-mode automatic differentiation, robust GPU acceleration, and effortless compatibility with well-known Python libraries such as NumPy. PyTorch is widely preferred in both academic and industrial settings because of its reliable API, straightforward debugging, and strong support for distributed training, ONNX export, and visualization resources such as TensorBoard. Its dynamic community and growing ecosystem position it as a top option for deep learning advancement (NVIDIA, 2025b)

2.4. Preprocessing Tools

- **NumPy**

NumPy is an essential Python library for scientific calculations that provides robust tools for managing extensive, multi-dimensional arrays and matrices. Central to NumPy is the ndarray object, enabling effective storage and handling of homogeneous data. In contrast to regular Python lists, NumPy arrays maintain fixed sizes and require a uniform data type, allowing for efficient computations via compiled code. NumPy offers a broad array of functionalities, including linear algebra, statistics, sorting, and Fourier transforms. Due to its effectiveness and integration, it underpins numerous scientific Python packages, rendering it crucial for data analysis, numerical computing, and simulation (N. Developers, 2024).

- **Pandas**

Pandas is a robust and versatile Python library created for effective data manipulation and analysis, particularly for working with labeled or tabular datasets. It offers two primary data structures: Series for one-dimensional data and DataFrame for two-dimensional data, allowing for intuitive management of datasets akin to SQL tables or Excel spreadsheets. Layered on NumPy, pandas streamlines processes like managing missing values, aligning data sets, grouping and summarizing, reshaping, and combining. It additionally provides strong assistance for time series data and different file formats, such as CSV and Excel. Pandas is a fundamental component of Python's data science ecosystem, extensively utilized in areas such as finance, statistics, and engineering (P. Developers, 2024)

- **PIL**

Pillow is a popular Python library for image manipulation, acting as the approachable and actively supported fork of the original Python Imaging Library (PIL). It provides comprehensive assistance for different image file formats and includes effective tools for loading, editing, and saving images. Engineered for efficiency, Pillow offers rapid pixel-

level access and features a variety of robust capabilities including image filtering, transformations, color modifications, and format changes. It serves as an essential element in numerous Python-driven image analysis and computer vision projects, providing a strong base for developing image processing tools and workflows (Contributors, 2025).

2.5. Plotting Tools

- **Matplotlib**

Matplotlib is a library for visualizing data in Python, developed by Michael Droettboom and others, which first launched in 2003. It features an object-oriented API for generating high-quality, publishable charts and graphs. This system can handle various kinds of plots, such as line graphs, scatter plots, bar charts, histograms, and more. It also enables the customization of visual styles, layout, and saving in various file formats (M. Developers, 2025)

2.6. Evaluation and NLP Tools

- **NLTK:**

The Natural Language Toolkit (NLTK) is an extensive Python framework created for handling human language data and developing natural language processing (NLP) applications. It provides convenient access to more than 50 corpora and lexical resources like WordNet, coupled with robust libraries for activities such as text classification, tokenization, stemming, tagging, parsing, and semantic analysis. NLTK features interfaces for strong NLP libraries and benefits from ongoing community discussions. It is commonly utilized in education, research, and industry because of its straightforward documentation and practical tutorials, which make it a great resource for novices as well as seasoned developers in computational linguistics (N. L. T. Team, 2025).

3. Dataset Preparation

3.1. Dataset Description

- **Indiana University CXR Dataset**

The Indiana University Chest X-ray dataset (IU X-ray) is a publicly accessible benchmark collection aimed at facilitating research in automated radiology report creation. Created and launched by the U.S. National Library of Medicine, this dataset features a comprehensive combination of chest X-ray images along with corresponding narrative radiology reports. It has seen extensive application in natural language processing (NLP), computer vision, and medical AI tasks for developing and assessing image-to-text generation models(Demner-Fushman et al., 2016).

  - **Dataset Structure**

The dataset includes 7,470 chest X-ray images from 3,955 distinct patient studies, featuring a range of frontal and lateral perspectives. Every examination includes a diagnostic report written by hand and created by a radiologist. The images come from the medical imaging archives at the Indiana University School of Medicine and were anonymized to protect patient privacy before being released publicly(Demner-Fushman et al., 2016).

- **Report Structure**

Every radiology report is organized in a uniform format and generally includes these sections:

- **Findings:** A brief diagnostic assessment or clinical overview that usually emphasizes the key findings.
- **Impression:** A comprehensive account outlining the visual findings derived from the X-ray images, generally arranged by anatomical areas or radiological importance.
- **Comparison:** This section refers to previous imaging studies when relevant and emphasizes changes that have occurred over time.
- **Indication:** This part details the reasoning for the imaging examination, including patient symptoms, history, or clinical concerns.

This organized structure allows for detailed analysis and modeling of various elements of clinical reporting, ranging from descriptive specifics to overarching summaries.(Demner-Fushman et al., 2016)

- **Data Format and Accessibility**

The original dataset includes radiographs in DICOM (Digital Imaging and Communications in Medicine) format, but the version utilized in this study is hosted on Kaggle, offering the same dataset with images converted to PNG format and normalized to uniform resolution. This preprocessing enhances user-friendliness in deep learning processes, especially when training convolutional neural networks (CNNs) that need consistent input sizes. Transforming DICOM to PNG guarantees compatibility with common Python libraries (such as PIL, OpenCV, TensorFlow), and normalization improves pixel-level uniformity throughout the dataset.

- **Research Purpose and Utility**

The IU X-ray dataset acts as an important standard for assessing vision-language models, particularly those designed to create radiology reports from imaging data. Its fairly small size, organized reporting format, and publicly available nature render it suitable for initial-stage prototyping, guided learning, and performance evaluation. Specifically, the

dataset's diagnostic depth and narrative intricacy enable the investigation of both extractive and generative tasks, such as report summarization, impression creation, disease identification, and image captioning within medical fields(Demner-Fushman et al., 2016).

3.2. Dataset Preparation and Structuring

The dataset used in this system is the Indiana University Chest X-ray collection, which is publicly accessible. It contains de-identified radiographic images (in PNG format) along with structured text reports. Every image is linked to a distinct identifier (UID), which serves to connect it with its related radiology report.

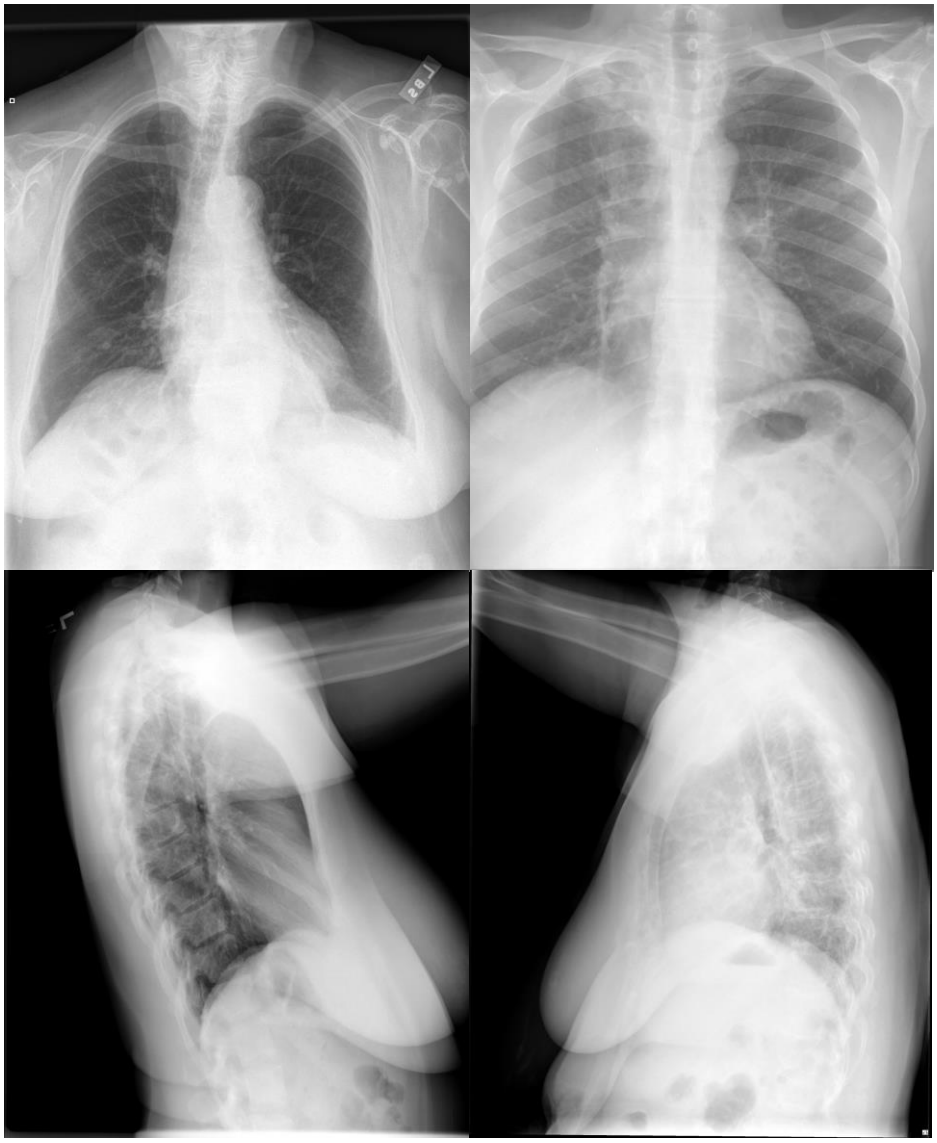


Figure 13: *Samples from the Indiana University Chest X-ray*

The unprocessed reports were analyzed to extract the two clinically relevant sections:

- Findings: describing the features observed on the radiographic image;
- Impression: summarizing the radiologist's diagnostic conclusions.

To ensure the quality of the training data all incomplete samples - i.e. those lacking at least one of the two sections - were discarded. The text of the reports was then cleaned up using regular expressions to remove redundant spaces, repetitive punctuation (e.g. serial dots) and formatting artefacts such a placeholder characters like 'XXXX'.

Finally, each report is framed by special tokens explicitly marking the start and end of the sequence:

<start> Findings:\n{findings}\n\nImpression:\n{impression} <end>

This structured format plays an essential role in guiding the Transformer decoder, providing it with clear cues about the logical and syntactic boundaries of the content to be generated.

3.3. Tokenization and Vocabulary

Text tokenization is performed using the Keras Tokenizer class. This tokenizer is configured to:

- Case sensitive: Preserve case sensitivity (lower=False), as medical terminology often relies on case distinctions.
- Management of rare words: Include an out-of-vocabulary token <unk> for rare or unseen words.
- Punctuation preservation: Avoid filtering out punctuation, which can be semantically significant in medical reports.

The tokenizer is fitted on all structured reports, yielding a vocabulary of unique tokens (including punctuation, words, and special tokens). All reports are then converted into integer sequences and padded to the length of the longest report.

As part of the Transformer decoder training, these sequences are split into two parts:

- **Decoder Inputs**: All tokens in the sequence, except the last.
- **Targets**: all the tokens in the sequence, except for the first.

This technique enables the model to learn the probability of each subsequent word, conditioned on the previous ones.

- **Tokenizer Configuration**:

Component	Value
Tokenizer Tool	Keras Tokenizer
Filters	None
Lowercase	False
OOV Token	<unk>
Max Sequence Length	Dynamically computed from corpus
Vocabulary Size	Varies depending on corpus (e.g. ~4,000)

Table 8: *Summary of Text preparation step.*

3.4. Image Preprocessing and Data Augmentation

Each chest X-ray image is resized to 224x224 pixels to align with the input size expected by EfficientNetB0, the chosen image encoder. Preprocessing includes:

- Conversion to NumPy arrays.
- Pixel normalization using efficientnet.preprocess_input, which performs mean subtraction and BGR channel reordering.

Data augmentation is applied only to the training set to improve model generalization. Augmentations include:

- Random horizontal flipping
- Small random rotations (±10%)
- Random zooming (±10%) These are implemented using TensorFlow's Sequential augmentation pipeline.

Step	Method
Resize	224×224 pixels
Color Preprocessing	EfficientNetB0 preprocessing (BGR shift)
Augmentation	Flip (horizontal), Rotation (±10°), Zoom (±10%)

Table 9: *Summary of Image Preparation Step.*

4. System Implementation

4.1. CNN Encoder

This system employs a hybrid architecture combining visual and textual processing modules:

Image Encoder: The visual backbone is EfficientNetB0, pretrained on ImageNet. The model is frozen to retain general visual features and avoid overfitting. The output of the CNN is:

- Pooled using GlobalAveragePooling2D
- Normalized with BatchNormalization
- Passed through a dropout layer (0.3)
- Projected to a 512-dimensional vector

Layer	Description
Base Network	EfficientNetB0 (frozen)
Output Shape	(None, 7, 7, 1280)
Global Pooling	GlobalAveragePooling2D
Normalization & Dropout	BatchNorm + Dropout(0.3)
Dense Layer	512 units + Dropout(0.3)

Table 10: *Summary of Image Encoder Architecture*

4.2. LSTM Decoder

A custom LSTM decoder is built using:

- A token embedding layer that transforms input sequences from vocab_size to 256-dimensional embeddings.
- A single-layer LSTM with 512 hidden units, initialized using image features as both the hidden and cell states.
- A dropout layer to regularize the LSTM output.
- A final Dense layer with softmax activation applied at each time step to produce the probability distribution over the vocabulary.

Layer	Description
Input	Token Embedding (vocab_size → 256)
Image Features	Projected (Dense 512) + used as initial LSTM states
LSTM Layer	Single-layer LSTM (512 units)
Dropout	Dropout(0.3) after LSTM output
Output Layer	Dense(vocab_size, softmax) applied per time step

Table 11: *Summary of Text Decoder Architecture*

The CNN encoder (EfficientNetB0) extracts visual features, which are globally pooled and projected to 512 units to initialize the decoder's hidden state. This enables the model to condition text generation on the input image.

Hyperparameter	Value
Optimizer	Adam
Learning Rate	1e-4
Loss Function	Sparse Categorical Crossentropy
Batch Size	16
Epochs	75
Metrics	Accuracy

Table 12: *Summary of the CNN-LSTM model Training hyperparameters*

4.3. Transformer decoder

A custom decoder is built using:

- Token embedding layer (vocab_size → 512)
- Positional encoding (sinusoidal, as in Vaswani et al., 2017)
- Concatenation of image embedding to the start of the token sequence
- 2 Transformer decoder layers, each with:
    - Multi-head attention (4 heads)
    - Feed-forward network (512 units)
    - Layer normalization and residual connections

The decoder output is passed through a TimeDistributed Dense layer with softmax activation to produce a probability distribution over the vocabulary.

Layer	Description
Input	Token Embedding + Projected Image Feature
Positional Encoding	Applied to combined sequence
Decoder Layers	2 Transformer layers
Attention Heads	4 heads per layer
FFN Dimension	512 units
Output Layer	TimeDistributed(Dense(vocab_size, softmax))

Table 13: *Summary of Transformer Decoder Architecture.*

- **Hyperparameters**

The model is compiled with the Adam optimizer and trained using the sparse categorical crossentropy loss. Training and validation metrics are plotted to monitor convergence and detect potential overfitting.

Hyperparameter	Value
Optimizer	Adam
Learning Rate	1e-4
Loss Function	Sparse Categorical Crossentropy
Batch Size	32
Epochs	50
Metrics	Accuracy

Table 14: *Summary Of CNN-Transformer Model Training Hyperpameters*

5. Model Training and Evaluation

The model was trained on the augmented dataset, with a batch size of 32 and over 50 epochs. The evolution of the metrics was monitored on the validation set at each epoch in order to monitor convergence and detect any over- or under-learning. The results were visualized using the matplotlib library, enabling the loss and accuracy curves to be analyzed over the iterations.

5.1. Training Evaluation Metrics:

- **Accuracy** measures the proportion of correctly predicted tokens.
- **Loss** (Sparse Categorical Crossentropy) captures how well the predicted probability distribution aligns with the target token distribution.

Equations:

- **Accuracy** $= \frac{\text{Number of correct token predictions}}{\text{Total number of tokens}}$
- **Loss** $= -\sum y_i \log(\hat{y}_i)$, where y_i is the true class and $\hat{y}_i$ is the predicted probability for token

  - **For LSTM Decoder Model**

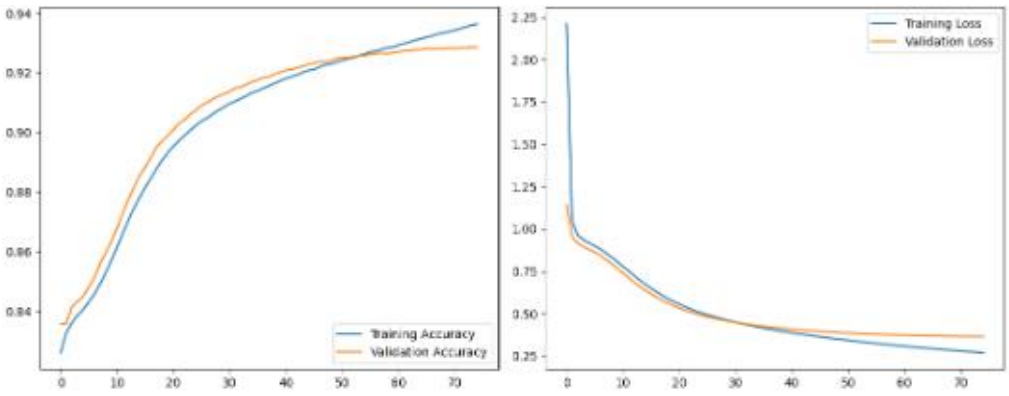


Figure 14:*Accuracy And Loss Graphs for LSTM Decoder Model*

  - **For Transformer Decoder Model**

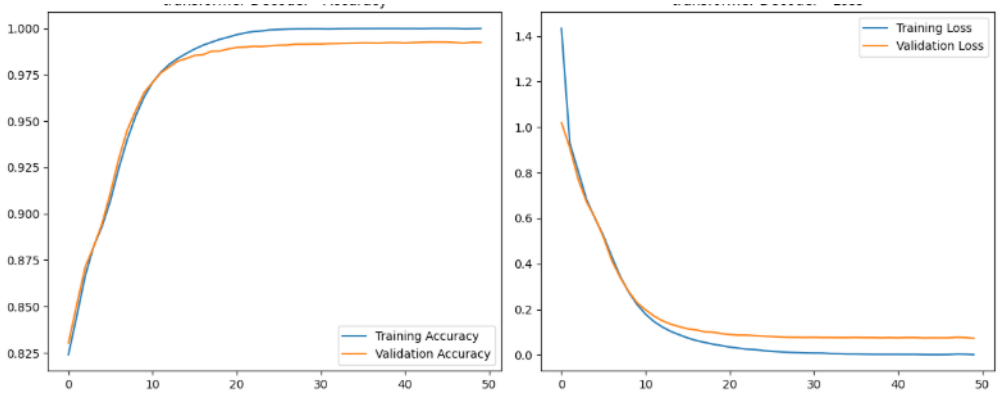


Figure 15: *Accuracy And Loss Graphs For Transformer Deoder Model*

Reported Scores:

Metric	Transformer Decoder	LSTM Decoder
Training Accuracy	99.98%	93.26%
Validation Accuracy	99.15%	92.16%
Training Loss	00.97%	28.86%
Validation Loss	07.71%	42.73%

Table 15: *The Model Metrics*

6. Post-Processing Phase: Correction model training

In order to improve the linguistic quality and clinical fidelity of the reports generated, a post-processing phase based on BioGPT is integrated into the pipeline. The correction model is fine-tuned on the basis of examples made up of pairs associating the generated reports (input) and the original reference reports (target). The structure of the prompts used follows the following format:

Correct the following radiology report: <generated_report>Corrected report: <original_report>

This prompt format enables BioGPT to learn the distribution and structure of realistic radiology reports written by experts. The correction model training is based on the following elements:

- Model used: BioGptForCausalLM from the HuggingFace Transformers library.
- Token masking: application of masking with a value of -100 on the tokens in the input prompt.
- Metrics tracking: accuracy and loss tracked at each epoch, with output in a format similar to Keras to make it easier to interpret the results.

Fine-tuning Component	Value
Model	BioGptForCausalLM
Pretrained Source	microsoft/biogpt
Prompt Format	“Correct the following radiology report: …”
Learning Rate	5e-5
Optimizer	AdamW
Batch Size	4
Epochs	5

Evaluation Metrics	Accuracy, Loss (Token-wise)

Table 16: *Hyperparameters of the correction model*

7. Evaluation and Metrics

7.1. Evaluation Methodology

The evaluation of generated and corrected reports is based on a combination of quantitative and qualitative approaches.

- Quantitative evaluation: classic natural language generation metrics are calculated in order to estimate the linguistic and semantic fidelity of the reports generated in relation to the reference reports.
- Qualitative evaluation: examples of generated reports are compared with the original reports to analyze clinical relevance, linguistic fluency and terminological consistency.

7.2. Results Overview

System performance was measured on the test set in different configurations: an uncorrected LSTM decoder, an uncorrected Transformer decoder and a Transformer decoder with BioGPT post-processing.

The table below summarizes the average scores obtained:

Metric	LSTM decoder without correction	Transformer Decoder without Correction	Transformer decoder with Correction
Bleu-avg	0.4191	0.6071	0.8286
Bleu-1	0.4536	0.7349	0.8789
Bleu-2	0.4315	0.6802	0.8601
Bleu-3	0.4235	0.6395	0.8445
Bleu-4	0.4191	0.6071	0.8286
Rouge-L	0.4861	0.7455	0.9248
METEOR	0.0553	0.5950	0.8933
BertScore	0.8258	0.9121	0.9628

Table 17: *Natural Language Generation Evaluation Metrics Values.*

These results illustrate a significant improvement in performance when switching from an LSTM decoder to a Transformer, as well as a significant gain in linguistic and semantic quality thanks to the integration of the post-processing phase with BioGPT.

An in-depth analysis of the results shows a significant improvement in the performance of our automatic radiology report generation system, thanks in particular to the evolution of its architecture. Replacing the LSTM decoder with a Transformer proved

to be a key factor in this progress, increasing BLEU-4 scores by 45%. This substantial improvement is attributable to the Transformer's superior ability to model long-term dependencies in text, enabling more consistent and structurally accurate report generation.

The integration of a post-processing phase via BioGPT has also had a significant impact on the linguistic and semantic quality of the reports produced. This step significantly improved the fluidity of the text, as evidenced by a 50% increase in the METEOR score, and considerably improved medical accuracy, with a BERTScore of over 0.96. In particular, this correction made it possible to reduce terminological inconsistencies, bringing the reports generated closer to clinical standards.

The concrete examples presented in Table 18 confirm that the clinical impressions generated by the model are structurally very close to the reference reports, validating the system's ability to capture the diagnostic essence. However, residual errors remain, including misspellings of specific technical terms (e.g. 'granulomatous disease'), highlighting areas for improvement.

Despite these advances, certain limitations have been identified. The model shows a data bias, struggling to generate accurate descriptions for rare conditions (e.g. partial pneumothorax), suggesting the need for a more diverse dataset or targeted augmentation techniques. In addition, the addition of the BioGPT module, although beneficial for quality, leads to an increase in inference time of around 20%, a factor to be considered for real-time integration in a clinical environment.

- **Exemples**

Here are some examples comparing the reports generated with the reference reports:

CNN-LSTM Model		**Ground Truth Report:**	Findings: The aortic is mildly tortuous. The cardiomediastinal silhouette and pulmonary vasculature are within normal limits. There is no pneumothorax or pleural effusion. There are no focal areas of consolidation. There are T-spine osteophytes. Large body habitus. Impression: No acute cardiopulmonary abnormality.

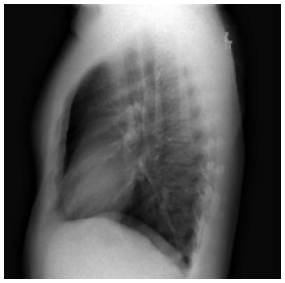
		Generated Report:	Generated Report: Findings: The heart is normal enlarged. The cardiomediastinal silhouette is pulmonary vasculature are within normal limits There is no pneumothorax or pleural effusion. There are no focal areas of consolidation. Impression: There are no osteophytes. There degenerative in degenerative acute bony abnormality. .
Cnn-Transformer + Correction		**Ground Truth Report:**	Findings: Heart size borderline enlarged. No focal alveolar consolidation, no definite pleural effusion seen. No typical findings of pulmonary edema. Dense nodule in the right lower lobe suggests a previous granulomatous process. Impression: Borderline heart size, no acute pulmonary findings
		Generated Report:	Findings: Heart size borderline enlarged. No focal alveolar consolidation, no definite pleural effusion seen. No typical findings of pulmonary edema. Calcific nodule in the right lower lobe suggests a previous granulomatous disease. Impression: Negative heart size, no acute pulmonary finding
		After Correction Report:	Findings: Heart size borderline enlarged. No focal alveolar consolidation, no definite pleural effusion seen. No typical findings of pulmonary edema. Calcific nodule in the right lower lobe suggests a previous granulomatous disease. Impression: Negative heart size, no acute pulmonary findings

Table 18 : *Exmples of generated reports*

8. Final Discussion

The results obtained in this study highlight the considerable potential of hybrid deep learning architectures for the automatic generation of radiology reports. In both quantitative and qualitative terms, the performances demonstrate a significant advance in the ability to transform complex visual information into precise, structured medical text descriptions that comply with clinical requirements.

The architectural evolution of the system has been a key factor in improving performance. Moving from the LSTM decoder to a Transformer architecture proved to be a major strategic choice. The 45% increase in the BLEU-4 score illustrates the superiority of Transformers in modelling long-range dependencies and taking into account the contextual subtleties of the language. This advance has made it possible to generate texts with a more rigorous syntactic structure and greater semantic coherence, thus better meeting the expectations of specialists in the field of radiology.

The integration of a post-processing module based on BioGPT also played a key role as a linguistic refinement layer. This component raised the final quality of the reports by correcting lexical imperfections, harmonizing medical terminology and improving the fluidity of the texts generated. The 50% increase in the METEOR score and a BERTScore in excess of 0.96 testify to the system's ability to produce reports that are stylistically and semantically close to those written by experts. This module has proved essential in guaranteeing the readability, reliability and compliance of the reports generated with current medical standards.

The examples presented (see Table 18) provide a concrete illustration of the system's ability to generate clinical impressions whose structure and content are remarkably aligned with those of the reference reports. These results validate the system's effectiveness in extracting key diagnostic information and rendering it in a form that can be used in clinical practice.

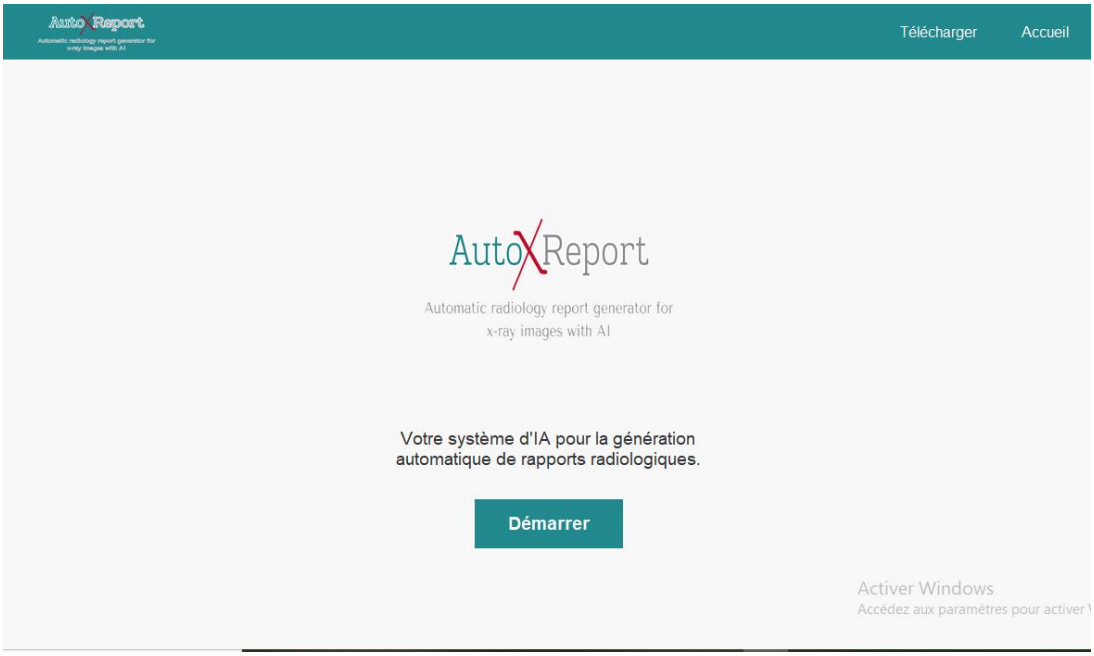
However, the study also highlighted certain limitations and identified areas for improvement. Despite the encouraging results, errors remain, notably lexical approximations on specific technical terms or difficulties in generating rare words (e.g. ‘granulomatous disease’). These findings highlight the need to refine the generation mechanisms to better manage specialized vocabulary. In addition, a bias linked to the data was observed: the model struggles to accurately describe certain rare pathologies, such as partial pneumothorax. The integration of more diversified data sets or the use of augmentation techniques targeted at these poorly represented cases could constitute promising avenues for remedying this limitation.

Finally, although the addition of BioGPT significantly improved the quality of the reports, it was accompanied by an increase in inference time of around 20%. Although this

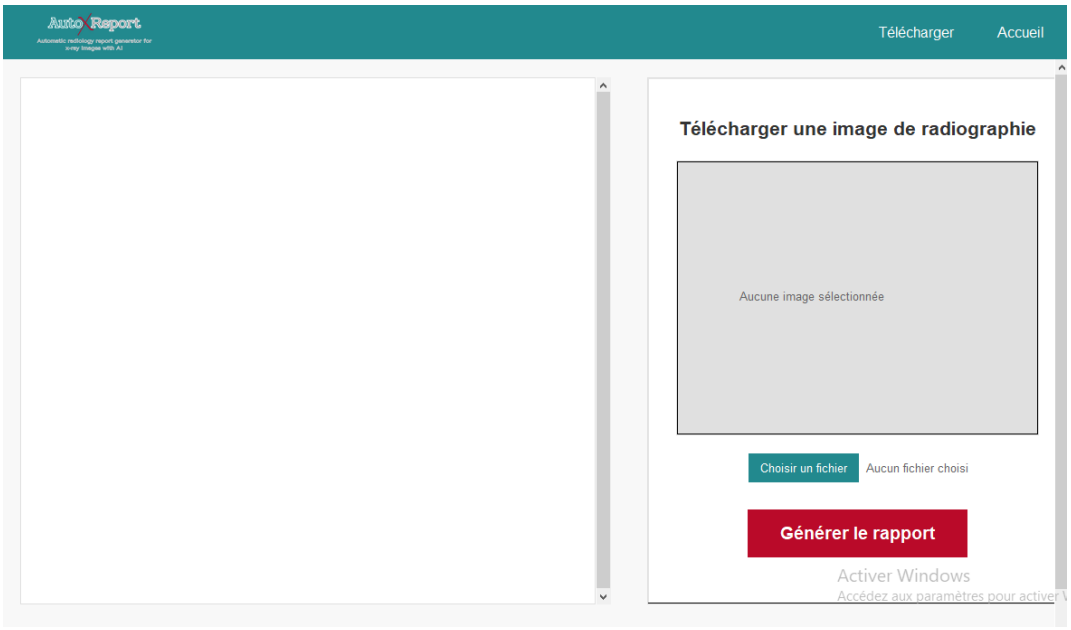
additional cost is tolerable in an experimental setting, it is a point of caution when it comes to integrating the system into a clinical environment, where responsiveness is essential. Future optimizations should therefore aim to reduce this latency without compromising the linguistic and diagnostic quality of the reports.

These improvements would evolve the existing research prototype into a more functional, flexible, and internationally implementable AI system for radiology clinical assistance.

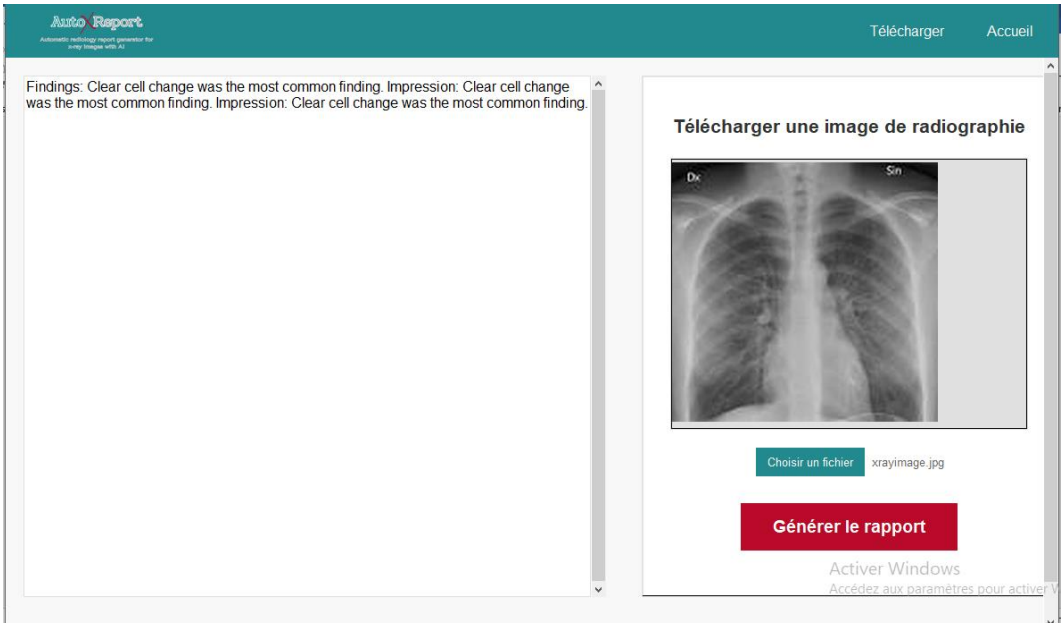
9. System Interface



10. Figure 16: *Home Page of The system*



11. **Figure 17:** *Upload Image Page*



12. **Figure 18:** *Example*

13.Conclusion

In this chapter, we presented the implementation of our system for the automatic generation of radiological reports, which integrates visual and linguistic processing within a deep learning framework. The system is based on a pre-trained CNN (EfficientNet-B0)

for extracting X-ray features, combined with a Transformer decoder for generating diagnostic texts.

We have described the main stages of the implementation, including the tools used, data pre-processing techniques, model training and the addition of a post-processing module with BioGPT to improve the linguistic and clinical quality of the reports. Although the results are promising, certain limitations remain, opening up prospects for future work, particularly in terms of data diversity, multilingual support and adaptation to other imaging modalities.

General conclusion and prospects

Radiological report writing is an essential part of the medical diagnostic process. However, this task, performed manually by radiologists, is facing increasing challenges: a high workload, inherent inter-observer variability, and increased time pressure. In this context, automating the generation of these reports, particularly for chest X-rays widely used in the detection of pulmonary and cardiovascular pathologies, is a promising way of significantly improving clinical efficiency, standardising medical reports and, ultimately, optimising the quality of care.

In this work, we proposed a comprehensive end-to-end system for the automatic generation of radiology reports using a deep learning pipeline. The system combines convolutional neural networks (EfficientNetB0) for image feature extraction and a custom Transformer decoder for report generation, effectively applying the encoder-decoder paradigm to the domain of medical image captioning. A major contribution of this work lies in the integration of a post-processing module based on the BioGPT biomedical model. This is used at the end of the pipeline to fine-tune the linguistic consistency, language fluidity and compliance with medical standards of the reports generated, a crucial stage for their clinical acceptability. The system has been rigorously trained and evaluated on the Indiana University Chest X-ray dataset.

A critical aspect of our methodology involved careful data preparation. This included matching chest X-ray images to their corresponding reports, cleaning textual data using regular expressions, and formatting reports with special tokens to define clear start and end points. The resulting dataset was then tokenized, and the image inputs were preprocessed with normalization and data augmentation techniques to improve generalization.

The experimental results obtained are particularly encouraging and testify to the robustness and effectiveness of the proposed approach. The final model performed very satisfactorily, with a BLUE-4 score of 0.8286, a RED-L of 0.9248 and a BERTScore of 0.9628. These high metrics confirm a strong semantic and lexical similarity between the automatically generated reports and those written by professionals. The analysis of the architectural contributions showed that the switch from an LSTM decoder to a Transformer architecture, as well as the strategic addition of post-processing by BioGPT, were decisive in achieving significant gains in terms of quality, semantic accuracy and readability, enabling the system to produce reports that are structured, relevant and close to professional standards.

The prospects opened up by this study are vast and pave the way for future developments. To further enhance the system's performance and broaden its field of application, several areas of improvement are envisaged:

- Refine linguistic post-processing by integrating more specialised or multi-lingual biomedical models (e.g. BioMedGPT, ClinicalT5).
- Extend the system to other imaging modalities, such as computed tomography (CT) or magnetic resonance imaging (MRI), by adapting the architecture and pre-processing.
- Enhance learning on rare cases, using targeted data augmentation techniques or generative models (e.g. GANs).
- Carry out clinical validation under real conditions, in collaboration with radiologists, to assess the relevance and acceptability of the reports generated.
- Optimise inference time to enable seamless integration into real-time hospital environments.

Bibliography

Agarwal, L., & Verma, B. (2025). *Advanced Chest X-Ray Analysis via Transformer-Based Image Descriptors and Cross-Model Attention Mechanism* (No. arXiv:2504.16774). arXiv. https://doi.org/10.48550/arXiv.2504.16774

Aksoy, N., Ravikumar, N., & Frangi, A. F. (2023). Radiology report generation using transformers conditioned with non-imaging data. In B. J. Park & H. Yoshida (Eds.), *Medical Imaging 2023: Imaging Informatics for Healthcare, Research, and Applications* (p. 23). SPIE. https://doi.org/10.1117/12.2653672

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6077–6086. https://doi.org/10.1109/CVPR.2018.00636

Ayesha, H., Iqbal, S., Tariq, M., Abrar, M., Sanaullah, M., Abbas, I., Rehman, A., Niazi, M. F. K., & Hussain, S. (2021). Automatic medical image interpretation: State of the art and future directions. *Pattern Recognition*, *114*, 107856. https://doi.org/10.1016/j.patcog.2021.107856

Azure, M. (2025a). *What is a Machine Learning Platform?* https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-machine-learning-platform/

Azure, M. (2025b). *What is Deep Learning?* https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-deep-learning

Banerjee, S., & Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. https://aclanthology.org/W05-0909

Bannur, S., Hyland, S., Liu, Q., Pérez-García, F., Ilse, M., Castro, D. C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., Schwaighofer, A., Wetscherek, M., Lungren, M. P., Nori, A., Alvarez-Valle, J., & Oktay, O. (2023). Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15016–15027. https://doi.org/10.1109/CVPR52729.2023.01442

Beddiar, D.-R., Oussalah, M., & Seppänen, T. (2023). Automatic captioning for medical imaging (MIC): A rapid review of literature. *Artificial Intelligence Review*, *56*(5), 4019–4076. https://doi.org/10.1007/s10462-022-10270-w

Boecking, B., Usuyama, N., Bannur, S., Castro, D. C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., Poon, H., & Oktay, O. (2022). Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, & T. Hassner (Eds.), *Computer Vision – ECCV 2022* (Vol. 13696, pp. 1–21). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-20059-5_1

Bustos, A., Pertusa, A., Salinas, J.-M., & De La Iglesia-Vayá, M. (2020). PadChest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, *66*, 101797. https://doi.org/10.1016/j.media.2020.101797

Calamida, A., Nooralahzadeh, F., Rohanian, M., Fujimoto, K., Nishio, M., & Krauthammer, M. (2023). *Radiology-Aware Model-Based Evaluation Metric for Report Generation* (No. arXiv:2311.16764). arXiv. https://doi.org/10.48550/arXiv.2311.16764

Contributors, P. (2025). *Pillow (PIL Fork) Documentation*. https://pillow.readthedocs.io/en/stable/

DataScientest. (2024). *Kaggle: All About This Platform*. https://datascientest.com/en/kaggle-all-about-this-platform

DataScientest. (2025). *Jupyter Notebook: Tout savoir sur cet outil incontournable*. https://datascientest.com/jupyter-notebook-tout-savoir

Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., Thoma, G. R., & 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. https://doi.org/10.1093/jamia/ocv080

Developers, M. (2025). *Matplotlib: Visualization with Python*. https://matplotlib.org/

Developers, N. (2024). *What is NumPy? —NumPy v2.0.dev0 Manual*. https://numpy.org/devdocs//user/whatisnumpy.html

Developers, P. (2024). *Getting Started—Pandas Documentation*. https://pandas.pydata.org/docs/getting_started/overview.html

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (No. arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* (No. arXiv:2010.11929). arXiv. https://doi.org/10.48550/arXiv.2010.11929

Foundation, P. S. (2025). *The History of Python (The Python Blurb)*. https://www.python.org/doc/essays/blurb/

Gale, W., Oakden-Rayner, L., Carneiro, G., Palmer, L. J., & Bradley, A. P. (2019). Producing Radiologist-Quality Reports for Interpretable Deep Learning. *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 1275–1279. https://doi.org/10.1109/ISBI.2019.8759236

Gu, A., & Dao, T. (2024). *Mamba: Linear-Time Sequence Modeling with Selective State Spaces* (No. arXiv:2312.00752). arXiv. https://doi.org/10.48550/arXiv.2312.00752

Hamamci, I. E., Er, S., Wang, C., Almas, F., Simsek, A. G., Esirgun, S. N., Doga, I., Durugol, O. F., Dai, W., Xu, M., Dasdelen, M. F., Wittmann, B., Amiranashvili, T., Simsar, E., Simsar, M., Erdemir, E. B., Alanbay, A., Sekuboyina, A., Lafci, B., … Menze, B. (2025). *Developing Generalist Foundation Models from a Multimodal Dataset for 3D Computed Tomography* (No. arXiv:2403.17834). arXiv. https://doi.org/10.48550/arXiv.2403.17834

IBM. (2025a). *Convolutional Neural Networks*. https://www.ibm.com/think/topics/convolutional-neural-networks

IBM. (2025b). *Large Language Models*. https://www.ibm.com/think/topics/large-language-models

IBM. (2025c). *Neural Networks Explained*. https://www.ibm.com/think/topics/neural-networks

IBM. (2025d). *Recurrent Neural Networks*. https://www.ibm.com/think/topics/recurrent-neural-networks

IBM. (2025e). *Transfer Learning in AI*. https://www.ibm.com/think/topics/transfer-learning

IBM. (2025f). *Vision-Language Models*. https://www.ibm.com/think/topics/vision-language-models

IBM. (2025g). *What is GPT?* https://www.ibm.com/think/topics/gpt

Jing, B., Xie, P., & Xing, E. (2018). On the Automatic Generation of Medical Imaging Reports. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2577–2586. https://doi.org/10.18653/v1/P18-1240

Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C., Mark, R. G., & Horng, S. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, *6*(1), 317. https://doi.org/10.1038/s41597-019-0322-0

Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). DenseCap: Fully Convolutional Localization Networks for Dense Captioning. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4565–4574. https://doi.org/10.1109/CVPR.2016.494

Kapadnis, M. N., Patnaik, S., Nandy, A., Ray, S., Goyal, P., & Sheet, D. (2024). *SERPENT-VLM: Self-Refining Radiology Report Generation Using Vision Language Models* (No. arXiv:2404.17912). arXiv. https://doi.org/10.48550/arXiv.2404.17912

Li, C. Y., Liang, X., Hu, Z., & Xing, E. P. (2018). *Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation* (No. arXiv:1805.08298). arXiv. https://doi.org/10.48550/arXiv.1805.08298

Li, M., Liu, R., Wang, F., Chang, X., & Liang, X. (2023). Auxiliary signal-guided knowledge encoder-decoder for medical report generation. *World Wide Web*, *26*(1), 253–270. https://doi.org/10.1007/s11280-022-01013-6

Li, Y., Wang, Z., Liu, Y., Wang, L., Liu, L., & Zhou, L. (2024). *KARGEN: Knowledge-enhanced Automated Radiology Report Generation Using Large Language Models* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2409.05370

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In A. Nenkova & O. Rambow (Eds.), *Text Summarization Branches Out* (pp. 74–81). Association for Computational Linguistics. https://aclanthology.org/W04-1013

Liu, S., Zhu, Z., Ye, N., Guadarrama, S., & Murphy, K. (2017). Improved Image Captioning via Policy Gradient optimization of SPIDEr. *2017 IEEE International Conference on Computer Vision (ICCV)*, 873–881. https://doi.org/10.1109/ICCV.2017.100

Liu, Y., Wang, Z., Li, Y., Liang, X., Liu, L., Wang, L., & Zhou, L. (2024). *MRScore: Evaluating Radiology Report Generation with LLM-based Reward System* (No. arXiv:2404.17778). arXiv. https://doi.org/10.48550/arXiv.2404.17778

Lu, J., Xiong, C., Parikh, D., & Socher, R. (2017). Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3242–3250. https://doi.org/10.1109/CVPR.2017.345

Molino, D., Feola, F. di, Shen, L., Soda, P., & Guarrasi, V. (2025). *Any-to-Any Vision-Language Model for Multimodal X-ray Imaging and Radiological Report Generation* (No. arXiv:2505.01091). arXiv. https://doi.org/10.48550/arXiv.2505.01091

Monshi, M. M. A., Poon, J., & Chung, V. (2020). Deep learning in generating radiology reports: A survey. *Artificial Intelligence in Medicine*, *106*, 101878. https://doi.org/10.1016/j.artmed.2020.101878

Moradi, M., Madani, A., Gur, Y., Guo, Y., & Syeda-Mahmood, T. (2018). Bimodal Network Architectures for Automatic Generation of Image Annotation from Text. In A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, & G. Fichtinger (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018* (Vol. 11070, pp. 449–456). Springer International Publishing. https://doi.org/10.1007/978-3-030-00928-1_51

Ni, J., Hsu, C.-N., Gentili, A., & McAuley, J. (2020). Learning Visual-Semantic Embeddings for Reporting Abnormal Findings on Chest X-rays. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1954–1960. https://doi.org/10.18653/v1/2020.findings-emnlp.176

Nishino, T., Miura, Y., Taniguchi, T., Ohkuma, T., Suzuki, Y., Kido, S., & Tomiyama, N. (2022). Factual Accuracy is not Enough: Planning Consistent Description Order for Radiology Report Generation. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 7123–7138. https://doi.org/10.18653/v1/2022.emnlp-main.480

NVIDIA. (2025a). *Discover LSTM (Long Short-Term Memory)*. https://developer.nvidia.com/discover/lstm

NVIDIA. (2025b). *What is PyTorch? | NVIDIA Glossary*. https://www.nvidia.com/en-us/glossary/pytorch/

NVIDIA. (2025c). *What Is TensorFlow? | NVIDIA Glossary*. https://www.nvidia.com/en-eu/glossary/tensorflow/

Ostmeier, S., Xu, J., Chen, Z., Varma, M., Blankemeier, L., Bluethgen, C., Md, A. E. M., Moseley, M., Langlotz, C., Chaudhari, A. S., & Delbrouck, J.-B. (2024). GREEN:

Generative Radiology Report Evaluation and Error Notation. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 374–390. https://doi.org/10.18653/v1/2024.findings-emnlp.21

Pan, Y., Liu, L.-J., Yang, X.-B., Peng, W., & Huang, Q.-S. (2024). Chest radiology report generation based on cross-modal multi-scale feature fusion. *Journal of Radiation Research and Applied Sciences*, *17*(1), 100823. https://doi.org/10.1016/j.jrras.2024.100823

Pan, Y., Yao, T., Li, Y., & Mei, T. (2020). X-Linear Attention Networks for Image Captioning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10968–10977. https://doi.org/10.1109/CVPR42600.2020.01098

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 311. https://doi.org/10.3115/1073083.1073135

Pasunuru, R., & Bansal, M. (2017). *Reinforced Video Captioning with Entailment Rewards* (No. arXiv:1708.02300). arXiv. https://doi.org/10.48550/arXiv.1708.02300

Pavlopoulos, J., Kougia, V., Androutsopoulos, I., & Papamichail, D. (2021). *Diagnostic Captioning: A Survey* (No. arXiv:2101.07299). arXiv. https://doi.org/10.48550/arXiv.2101.07299

Pellegrini, C., Özsoy, E., Busam, B., Navab, N., & Keicher, M. (2025). *RaDialog: A Large Vision-Language Model for Radiology Report Generation and Conversational Assistance* (No. arXiv:2311.18681). arXiv. https://doi.org/10.48550/arXiv.2311.18681

Quigley, K., Cha, M., Barua, J., Chauhan, G., Berkowitz, S., Horng, S., & Golland, P. (2025). *Improving Medical Visual Representations via Radiology Report Generation* (No. arXiv:2310.19635). arXiv. https://doi.org/10.48550/arXiv.2310.19635

Ramesh, V., Chi, N. A., & Rajpurkar, P. (2022). *Improving Radiology Report Generation Systems by Removing Hallucinated References to Non-existent Priors* (No. arXiv:2210.06340). arXiv. https://doi.org/10.48550/arXiv.2210.06340

Reale-Nosei, G., Amador-Domínguez, E., & Serrano, E. (2024). From vision to text: A comprehensive review of natural image captioning in medical diagnosis and

radiology report generation. *Medical Image Analysis*, *97*, 103264. https://doi.org/10.1016/j.media.2024.103264

Schneider, J. (2024). *What comes after transformers? -- A selective survey connecting ideas in deep learning* (No. arXiv:2408.00386). arXiv. https://doi.org/10.48550/arXiv.2408.00386

Shisu, Y., Mingwin, S., Wanwag, Y., Chenso, Z., & Huing, S. (2024). *Improved EATFormer: A Vision Transformer for Medical Image Classification* (No. arXiv:2403.13167). arXiv. https://doi.org/10.48550/arXiv.2403.13167

Sirshar, M., Paracha, M. F. K., Akram, M. U., Alghamdi, N. S., Zaidi, S. Z. Y., & Fatima, T. (2022). Attention based automated radiology report generation using CNN and LSTM. *PLOS ONE*, *17*(1), e0262209. https://doi.org/10.1371/journal.pone.0262209

Sun, Y., Lee, Y. Z., Woodard, G. A., Zhu, H., Lian, C., & Liu, M. (2024). *R2Gen-Mamba: A Selective State Space Model for Radiology Report Generation* (No. arXiv:2410.18135). arXiv. https://doi.org/10.48550/arXiv.2410.18135

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). *Sequence to Sequence Learning with Neural Networks* (No. arXiv:1409.3215). arXiv. https://doi.org/10.48550/arXiv.1409.3215

Tan, M., & Le, Q. V. (2020). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks* (No. arXiv:1905.11946). arXiv. https://doi.org/10.48550/arXiv.1905.11946

Team, K. (2024). *About Keras*. https://keras.io/getting_started/about/

Team, N. L. T. (2025). *NLTK — Natural Language Toolkit*. https://www.nltk.org/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need* (No. arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Vedantam, R., Zitnick, C. L., & Parikh, D. (2015). CIDEr: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4566–4575. https://doi.org/10.1109/CVPR.2015.7299087

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3156–3164. https://doi.org/10.1109/CVPR.2015.7298935

Wang, F., Liang, X., Xu, L., & Lin, L. (2021). *Unifying Relational Sentence Generation and Retrieval for Medical Image Report Composition* (No. arXiv:2101.03287). arXiv. https://doi.org/10.48550/arXiv.2101.03287

Wang, X., Peng, Y., Lu, L., Lu, Z., & Summers, R. M. (2018). TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-Rays. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9049–9058. https://doi.org/10.1109/CVPR.2018.00943

Wang, X., Wang, F., Li, Y., Ma, Q., Wang, S., Jiang, B., Li, C., & Tang, J. (2024). *CXPMRG-Bench: Pre-training and Benchmarking for X-ray Medical Report Generation on CheXpert Plus Dataset* (No. arXiv:2410.00379). arXiv. https://doi.org/10.48550/arXiv.2410.00379

Wang, Z., Liu, L., Wang, L., & Zhou, L. (2023a). METransformer: Radiology Report Generation by Transformer with Multiple Learnable Expert Tokens. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11558–11567. https://doi.org/10.1109/CVPR52729.2023.01112

Wang, Z., Liu, L., Wang, L., & Zhou, L. (2023b). *R2GenGPT: Radiology Report Generation with Frozen LLMs* (No. arXiv:2309.09812). arXiv. https://doi.org/10.48550/arXiv.2309.09812

Wu, J., Agu, N., Lourentzou, I., Sharma, A., Paguio, J., Yao, J. S., Dee, E. C., Mitchell, W., Kashyap, S., Giovannini, A., Celi, L. A., Syeda-Mahmood, T., & Moradi, M. (n.d.). *Chest ImaGenome Dataset* (Version 1.0.0) [Dataset]. PhysioNet. https://doi.org/10.13026/WV01-Y230

Xiong, Y., Du, B., & Yan, P. (2019). Reinforced Transformer for Medical Image Captioning. In H.-I. Suk, M. Liu, P. Yan, & C. Lian (Eds.), *Machine Learning in Medical Imaging* (Vol. 11861, pp. 673–680). Springer International Publishing. https://doi.org/10.1007/978-3-030-32692-0_77

Yang, X., Tang, K., Zhang, H., & Cai, J. (2019). Auto-Encoding Scene Graphs for Image Captioning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10677–10686. https://doi.org/10.1109/CVPR.2019.01094

Yang, Y., Teo, C. T., Daumé III, H., & Aloimonos, Y. (2011). Corpus-Guided Sentence Generation of Natural Images. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 444–454. https://aclanthology.org/D11-1041

Yao, T., Pan, Y., Li, Y., & Mei, T. (2018). Exploring Visual Relationship for Image Captioning. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer Vision – ECCV 2018* (Vol. 11218, pp. 711–727). Springer International Publishing. https://doi.org/10.1007/978-3-030-01264-9_42

Yin, C., Qian, B., Wei, J., Li, X., Zhang, X., Li, Y., & Zheng, Q. (2019). Automatic Generation of Medical Imaging Diagnostic Report with Hierarchical Recurrent Neural Network. *2019 IEEE International Conference on Data Mining (ICDM)*, 728–737. https://doi.org/10.1109/ICDM.2019.00083

You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016). Image Captioning with Semantic Attention. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4651–4659. https://doi.org/10.1109/CVPR.2016.503

Zeng, F., Lyu, Z., Li, Q., & Li, X. (2024). *Enhancing LLMs for Impression Generation in Radiology Reports through a Multi-Agent System* (No. arXiv:2412.06828). arXiv. https://doi.org/10.48550/arXiv.2412.06828

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). *BERTScore: Evaluating Text Generation with BERT* (No. arXiv:1904.09675). arXiv. https://doi.org/10.48550/arXiv.1904.09675

Zhao, H., Chen, J., Huang, L., Yang, T., Ding, W., & Li, C. (2021). Automatic Generation of Medical Report with Knowledge Graph. *2021 10th International Conference on Computing and Pattern Recognition*, 1–1. https://doi.org/10.1145/3497623.3497658

Zhao, W., Wu, C., Zhang, X., Zhang, Y., Wang, Y., & Xie, W. (2024). RaTEScore: A Metric for Radiology Report Generation. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 15004–15019. https://doi.org/10.18653/v1/2024.emnlp-main.836

Annex: Startup Project

Project Idea

The project falls within the medical and healthcare sector, specifically targeting the modernization of radiology workflows using artificial intelligence. This innovative system seeks to automate the generation of radiology reports from chest X-ray images by combining advanced deep learning techniques with natural language processing.

The idea originated from observing the repetitive, time-consuming nature of report writing in radiology departments and the shortage of expert radiologists in many regions. The aim is to support clinical staff by automating descriptive reporting, enhancing consistency, and saving time for more critical diagnostic decisions.

To achieve this, we developed an end-to-end pipeline that uses a pre-trained convolutional neural network (EfficientNetB0) to extract visual features from X-ray images, followed by a Transformer-based decoder that generates textual reports in English. The generated reports are further refined using a domain-specific large language model, BioGPT, to ensure medical accuracy and fluency

Proposed Values

- **Modernity**

The system introduces a novel approach to radiology report generation by leveraging the latest advances in computer vision and natural language generation. By replacing manual report writing with AI-assisted tools, we offer a disruptive innovation for medical imaging.

- **Performance**

Our CNN-Transformer architecture, combined with BioGPT, provides high accuracy in generating semantically and clinically relevant English reports for chest X-ray images. The inclusion of fine-tuning and correction mechanisms ensures robust and consistent outputs.

- **Task Accomplishment**

The system automates key radiology tasks including image interpretation, findings summarization, and impression generation. This helps clinicians by reducing workload and enabling faster decision-making in high-throughput environments.

- **Design**

The platform is designed with usability in mind. From model inference to report correction, the system supports seamless integration into hospital information systems. User interfaces can be adapted to the needs of radiologists with minimal technical interaction.

- **Cost Reduction**

The system is designed to minimize development and operational costs, aligning with the economic constraints of the Algerian healthcare market. Automation of the reporting process leads to significant reductions in personnel workload and optimizes radiologist time usage.

- **Risk Reduction**

By reducing manual input, the system minimizes human error and ensures consistent report formatting. The correction stage using BioGPT further helps in aligning output with clinical standards, thus reducing the risk of misdiagnosis due to report inconsistencies.

- **Accessibility**

We aim to make AI-assisted radiology available to hospitals and clinics with limited access to expert radiologists. Through scalable and cost-efficient deployment models (e.g., local or cloud APIs), even smaller or rural facilities can benefit from AI diagnostics.

- **Ease of Use**

With straightforward deployment and intuitive input-output workflows, the system ensures that healthcare professionals can use the tool without needing advanced technical training. Reports can be generated with minimal interaction, increasing clinical efficiency.

Project Objectives

Our primary objective is to become a leader in the field of automated radiology report generation using deep learning and natural language processing. Within the next five years, we aim to establish our solution as a reference system in both clinical and academic radiology environments.

Implementation Timeline

Project Stage	**1m**	**2m**	**3m**	**4m**	**5m**	**6m**	**7m**	**8m**	**9m**
Preliminary Studies	✓	✓							
Algorithm Development		✓	✓						

Software Development			✓	✓	✓				
Integration & Testing					✓				
Pilot Phase						✓	✓		
Deployment							✓	✓	
Marketing et promotion								✓	✓

Innovative Aspects

The integration of deep learning-based computer vision models and transformer-based language models for the automatic generation of medical reports represents a significant advancement over traditional radiology workflows. Unlike conventional manual dictation or template-based systems, this approach enables dynamic, patient-specific report generation based on image content.

This project opens a new market segment for AI in radiology, specifically targeting diagnostic support in environments with limited access to expert radiologists. By adopting a continuous improvement strategy based on clinician feedback and advances in machine learning, the system will remain relevant and effective over time.

Regular updates to the model architecture and language output capabilities will ensure clinical alignment and technical competitiveness. This iterative enhancement process is crucial for maintaining the system's value in a fast-evolving AI and healthcare landscape.

Strategic Market Analysis

Market Sector Overview

In Algeria, the medical sector is undergoing digital transformation, particularly in the field of radiology, which is a critical component of diagnostic medicine. The demand for radiological services is increasing due to population growth, the rise in chronic diseases, and the scarcity of trained radiologists in remote regions.

Artificial intelligence in medical imaging is experiencing accelerated growth, driven by the need for efficient diagnostic tools and the potential of machine learning to automate and enhance clinical workflows. Automated radiology report generation systems offer cost-effective, scalable solutions that address both efficiency and quality in clinical documentation.

Key Market Characteristics:

- **High Demand for AI Tools:** Clinics and hospitals seek intelligent solutions to streamline diagnosis and reduce manual workloads.
- **Healthcare Digitization Initiatives:** Government and private initiatives are increasingly supporting digital health technologies.
- **Emerging Multilingual Needs:** In multilingual countries like Algeria, solutions that support multiple languages (Arabic, French, English) are especially relevant.

Key Market Segments:

- **Public Hospitals:** Seeking scalable tools to improve diagnostic accuracy under constrained resources.
- **Private Clinics:** Interested in cutting-edge technology for competitive advantage.
- **Medical Training Institutions:** Looking for tools to assist in radiology education and training.

Market Competition Intensity

The AI radiology space is moderately competitive, with several international players offering AI-powered diagnostics. However, few provide multilingual, domain-specific report generation with end-to-end pipelines integrated with correction mechanisms.

Main Competitors:

- **Template-based Reporting Tools:** Rigid and lacking adaptation to image content.
- **Foreign AI Solutions:** Powerful but often generalized, expensive, or lacking linguistic and clinical adaptation for local settings.
- **Manual and Semi-automated Approaches:** Labor-intensive, less scalable, and more error-prone.

Competitive Forces Analysis:

- **Entry Barriers:** High development cost and requirement of medical data present strong entry barriers.
- **Customer Bargaining Power:** High demand for accuracy and regulatory compliance puts pressure on solution providers to deliver robust performance.

- **Rivalry Among Providers:** Ongoing innovation in AI healthcare drives constant pressure to improve model accuracy, adaptability, and explainability.

Marketing Strategies

To maximize our market penetration and attract our target audience, we will implement a flexible subscription-based pricing strategy tailored to different institutional needs and budgets.

Subscription Plans:

Plan	Pricing Tier	Usage Limit	Key Features
Free Trial	Free	Up to 3 report generations	Allows new users to test the system on a limited number of images.
Weekly Plan	Low	Up to 20 reports per week	Suitable for short-term evaluations or low-volume clinics.
Monthly Plan	Medium	Scalable to hundreds of reports	Designed for mid-sized clinics or research teams with continuous usage.
Annual Plan	High	Unlimited usage	Ideal for hospitals and enterprise users; includes multilingual & priority support.

Communication Strategies:

1. **Digital Marketing:**
    - Develop an informative website with case studies, demo videos, and subscription details.
    - Use SEO and paid campaigns to reach targeted healthcare providers.
    - Engage medical professionals through LinkedIn and specialized forums.
2. **Conferences and Health Tech Expos:**
    - Present our system at medical technology events and radiology conferences.
    - Conduct live demos to demonstrate report generation and correction accuracy.
3. **Strategic Partnerships:**
    - Collaborate with radiology networks, health institutions, and medical device suppliers.

- Offer affiliate incentives to encourage resellers and referrers.

Sales Strategies:

- Launch promotional trials or free access periods for early adopters.
- Provide discounts for long-term or institutional licenses.
- Build a reseller and integrator network to reach hospitals and clinics.
- Offer post-deployment support, training resources, and feedback loops to improve user satisfaction.

Client Analysis

Our potential clients include:

- **Independent Radiologists and Small Clinics:** Looking for efficient tools to streamline reporting.
- **Hospital Radiology Departments:** Needing scalable AI assistance for large patient volumes.
- **Telemedicine and Teleradiology Companies:** Benefiting from rapid and accurate automated reports.
- **Medical NGOs and Government Health Initiatives:** Seeking scalable diagnostic tools in underserved areas.

System Development:

- **Data Collection and Cleaning:** Gathering radiology images and reports, preprocessing with tokenization and cleaning routines.
- **Model Development:** Training the CNN-Transformer model for report generation and fine-tuning BioGPT for correction.
- **Software Integration:** Developing a user interface and API for clinical use, enabling hospitals to access the system easily.

Testing and Validation:

- **Unit Testing:** Verifying individual components like tokenizers, decoder, and attention modules.
- **End-to-End Testing:** Running complete inference pipelines to check output coherence.

- **Clinical Validation:** Collaborating with radiologists for expert evaluation of generated reports.

Deployment and Maintenance:

- **Model Packaging:** Exporting the trained model in .keras format and tokenizer files for easy integration.
- **Infrastructure:** Hosting the solution on cloud platforms with secure access for hospitals.
- **Monitoring and Updates:** Regular performance reviews and model updates based on user feedback.

Material Resources

Digital Components:

- Pre-trained CNNs (EfficientNetB0)
- Transformer-based decoder
- BioGPT model weights
- TensorFlow and PyTorch libraries

Computational Infrastructure:

- GPU-based servers for training
- Cloud services for hosting APIs
- Data storage systems for medical datasets

Human Resources

Development Team:

- Deep Learning Engineers specialized in computer vision
- Natural Language Processing Engineers for report generation
- Backend Software Developers for API and system integration

Clinical Collaboration Team:

- Certified Radiologists for medical validation

- Clinical Reviewers for expert feedback

Operational Team:

- Site Reliability Engineers responsible for system deployment and scaling
- API and Integration Managers
- Technical Customer Support Personnel

Project Management Team:

- Artificial Intelligence Project Coordinator
- Quality Assurance Analysts
- Medical Artificial Intelligence Program Manager

Financial study

Estimated Startup Capital

Expense Category	Estimated Cost (USD)	Estimated Cost (DZD) (1 USD ≈ 140 DZD)
Software development (initial)	20,000	2,800,000
Software licenses and API tools	5,000	700,000
Cloud infrastructure and servers	10,000	1,400,000
Dataset acquisition	3,000	420,000
Secure data storage	2,000	280,000
UI design and integration	5,000	700,000
Initial user training	3,000	420,000
Launch marketing and promotion	5,000	700,000
Technical support setup	2,000	280,000
General operating costs	5,000	700,000
Total Estimated Capital	**60,000**	**8,400,000**

Monthly Operating Costs

Expense Category	Monthly Cost (USD)	Monthly Cost (DZD)
AI developer salaries	8,000	1,120,000

Clinical expert consultants	2,000	280,000
Technical support staff	2,000	280,000
Cloud hosting and maintenance	1,500	210,000
NLP/API service usage	1,000	140,000
Communication & internet	500	70,000
Digital marketing & campaigns	1,000	140,000
Office rent and utilities	1,000	140,000
Total Monthly Operating Cost	**17,000**	**2,380,000**

Three-Year Financial Projections

Year	Projected Revenue (USD)	Projected Revenue (DZD)	Operating Costs (USD)	Operating Costs (DZD)	Net Profit (USD)	Net Profit (DZD)
Year 1	200,000	28,000,000	204,000	28,560,000	**–4,000**	**–560,000**
Year 2	300,000	42,000,000	204,000	28,560,000	**96,000**	**13,440,000**
Year 3	400,000	56,000,000	204,000	28,560,000	**196,000**	**27,440,000**
Total (3 yrs)	**900,000**	**126,000,000**	**612,000**	**85,680,000**	**288,000**	**40,320,000**

Financial Analysis

Aspect	Details
Initial Investment	$60,000 / 8,400,000 DZD
Profitability	Break-even point expected by Year 2. Revenue increases due to subscription plans and low infrastructure scaling costs.
Cost Optimization	Cloud infrastructure and automation reduce operational costs. Efficient use of personnel and external tools keeps expenses under control.
Market Potential	High demand in public hospitals, academic institutions, and private clinics for scalable radiology automation, particularly in multilingual contexts.
Subscription Model	Revenue generated through tiered subscription plans (Free plan for 3 reports, Weekly, Monthly, Yearly) tailored to different medical centers' needs.

Fixed Costs (per year)

Item	Cost (USD/year)	Cost (DZD/year)
Office Rent	$12,000	1,680,000 DZD
Cloud Infrastructure (Base Plan)	$12,000	1,680,000 DZD

Admin & General Expenses	$5,000	700,000 DZD
Software Licenses	$6,000	840,000 DZD
Total Fixed Costs	$35,000	4,900,000 DZD

Variable Costs (per report)

Item	Cost (USD/report)	Cost (DZD/report)
GPU/API Compute Usage	$2.00	280 DZD
Technical Support	$1.00	140 DZD
Token/API Calls (e.g., GPT)	$0.50	70 DZD
Total Variable Cost	$3.50	490 DZD

Depreciation Calculation

Asset	Cost (USD)	Useful Life (Years)	Annual Depreciation	Depreciation (DZD)
GPU Accelerator	$20,000	5	$4,000	560,000 DZD

Break-even Units (Reports)

Break-even Point	Value in USD	Value in DZD
Revenue	$53,846	7,538,440 DZD
Reports	5,384	5,384

Revenue Scenarios (Yearly)

Scenario	Reports Sold	Revenue (USD)	Revenue (DZD)	Total Costs (USD)	Total Costs (DZD)	Net Profit (USD)	Net Profit (DZD)

Pessimistic	4,000	$40,000	5,600,000 DZD	$49,000	6,860,000 DZD	-$9,000	-1,260,000 DZD
Realistic	6,000	$60,000	8,400,000 DZD	$56,000	7,840,000 DZD	$4,000	560,000 DZD
Optimistic	10,000	$100,000	14,000,000 DZD	$70,000	9,800,000 DZD	$30,000	4,200,000 DZD

Annual Provisions for Risk (5% of Revenue)

Scenario	Revenue (USD)	Revenue (DZD)	Provision (5%)	Provision (DZD)
Pessimistic	$40,000	5,600,000 DZD	$2,000	280,000 DZD
Realistic	$60,000	8,400,000 DZD	$3,000	420,000 DZD
Optimistic	$100,000	14,000,000 DZD	$5,000	700,000 DZD

Summary Equations

- **Total Cost**:

$$Total\ Cost = Fixed\ Cost + (Variable\ Cost \times Units\ Sold)$$

- **Depreciation**:

$$Depreciation/year = \frac{Asset\ Price}{Useful\ Life(year)}$$

- **Break-Even Point (Units)**:

$$\frac{\text{Fixed Costs}}{\text{Unit Price - Variable Cost Per Unit}}$$

Business Model Canvas – AI Radiology Report Generation System

Key Partners

- Hospitals and radiology centers
- AI cloud providers (e.g., AWS, GCP)
- Open-source communities
- Medical data repositories (e.g., MIMIC-CXR)
- Medical associations for validation

Key Resources

- Pre-trained CNN and Transformer models
- Annotated medical datasets
- Compute infrastructure (e.g., GPUs, cloud instances)
- NLP correction modules (BioGPT)
- Skilled AI and medical staff

Key Activities

- Training and fine-tuning AI models
- Integrating CNN-Transformer pipeline
- Ensuring clinical validation
- BioGPT report correction

Value Propositions

- Automated radiology report generation
- Time and cost savings for radiologists
- Consistency in clinical documentation
- Improved diagnostic workflow in low-resource settings

Customer Relationships

- B2B technical integration and support
- Medical training workshops
- Feedback loop for fine-tuning
- Clinical deployment assistance

Channels

- API access via secure web interface
- On-premise deployment for hospitals
- Direct partnerships and field demonstrations
- Clinical software integration (e.g., PACS/RIS systems)

Customer Segments

- Public and private hospitals
- Radiology departments
- Clinics in underserved regions
- Medical schools and research centers

Cost Structure

- Model development and training
- Infrastructure and cloud compute
- Data acquisition and annotation
- Clinical evaluation and audits
- Support and updates
- Marketing and documentation

Revenue Streams

- SaaS licensing (per exam or monthly)
- Institution-level deployment licenses
- Custom integration fees
- Education licenses for universities