AI-Powered Radiology Report Generation Using Multi-Model

By
Ridam Roy
213-15-4260

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements forthe **Degree of Bachelor of Science in Computer Science and Engineering**

Supervised by

Md. Sazzadur Ahamed

Assistant Professor
Department of Computer Science and EngineeringDaffodil International University

Co-Supervised by

Dr. Md. Zahid Hasan

Associate Professor
Department of Computer Science and EngineeringDaffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
Dhaka, Bangladesh

September 17, 2025

APPROVAL

This Project titled **AI-Powered Radiology Report Generation using Multi-Model**, submitted by **Ridam Roy**, ID No: **213-15-4260** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **17 September, 2025**.

BOARD OF EXAMINERS

Dr. Fernaz Narin Nur (FNN)
Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman

Dr. Abdus Sattar (AS)
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

17.09.25

Mr. Mohammad Jahangir Alam (MJA)
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

17.09.25

Dr. Sajeeb Saha (DSS)
Associate Professor
Department of Computer Science and Engineering
Jagannath University

External Examiner

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Md. Sazzadur Ahamed**, Assistant Professor, Department of Computer Science and Engineering, Daffodil International University. We also declare that neither this projectnor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Md. Sazzadur Ahamed
Assistant Professor
Department of Computer Science and EngineeringDaffodil International University

Co-Supervised by:

Dr. Md. Zahid Hasan
Associate Professor
Department of Computer Science and EngineeringDaffodil International University

Submitted by:

Ridam Roy

Ridam Roy
Student ID: 213-15-4260
Department of Computer Science and EngineeringDaffodil International University

ACKNOWLEDGEMENTS

This work would not have been possible without the support and contributions of many individuals over the past two semesters. We are deeply grateful to everyone who has assisted us in one way or another.

First, I express my heartfelt thanks and gratefulness to the almighty for His divine blessing making it possible for me to complete the **Final Year Design Project(FYDP)** successfully.

I am grateful and wish my profound indebtedness to **Md. Sazzadur Ahamed**, **Assistant Professor**, Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Deep knowledge and keen interest of our supervisor in the field of **Artificial Intelligence** to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

I would like to express my heartfelt gratitude to the Head of the Department of Computer Science and Engineering, for his kind help in finishing our project and also to other faculty members and the staff of the Department of Computer Science and Engineering, Daffodil International University.

I would like to thank our entire course-mates at Daffodil International University, who took part in this discussion while completing the coursework.

Finally, I must acknowledge with due respect the constant support and patience of my parents.

ABSTRACT

The rapid development of artificial intelligence (AI) in the field of medical imaging has presented unprecedented opportunities to diagnose with automated accuracy and clinical decision support, but serious challenges still lie in the development of systems that are both accurate and clinically reliable. In this paper we provide an AI-based classification of Chest X-rays and automated radiology report generation, which integrates the latest deep learning models with large language models (LLMs) to generate reliable, interpretable, and clinically meaningful results. The methodology is planned in four major steps. To optimize the predictive accuracy and generalizability with a wide range of imaging conditions, first, fine-tuned architectures such as InceptionV3, ResNet, and VGG are used to classify chest X-rays as COVID-19, pneumonia, or no diseases using transfer learning, adaptive training strategies, and large amounts of data augmentation. Second, Grad-CAM and Grad-CAM++ visualizations are used to draw attention to the most interesting parts of the photos, improving interpretability, visually explaining them, and helping clinicians trust the model more with the knowledge that it focuses attention on pertinent anatomical objects. Third, the classification outputs and the XAI heatmaps are fed into a domain-adapted LLM (Gemini-Flash 2.5), which can produce structured, radiology-style reports that are clinically coherent, readable, and consistent with standard reporting forms. Lastly, evaluation is a combination of quantitative measures, such as accuracy, precision, recall, and F1-score, and expert-in-the-loop validation, where radiologists evaluate the generated reports based on whether they contain factual correctness, structural quality, clinical completeness, interpretability, and utility. The results of the experiment show that InceptionV3 obtained the best results (AUROC = 0.9988, Accuracy = 97.7%), which were better than VGG16, VGG19, ResNet50, and ResNet101. Comparative analysis shows that the proposed system yields not just highly accurate predictions, but also clinically consistent and interpretable reports, which are essential to solve issues with hallucinations, cross-modal misalignment, and limited grounding found in previous research. Expert validation helps to underline the practical reliability and safety of the system and justify its possibility to be applied in practice in radiology processes, decrease the amount of diagnostic work, and accelerate, standardize, and streamline to a greater degree the process of patient care in clinical practice.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

This chapter presents a general overview of the project and forms the background of the research. It introduces the rationale of automating radiology report generation, outlines the purpose of the study and briefly explains the approach taken to do so. Moreover, it also describes the results that should be achieved and the structure of the report so that the reader could conveniently navigate through it.

1.1 Introduction

Radiology generation (also known as radiology AI) is rapidly emerging as a field that seeks to automate the production of diagnostic reports on medical images such as X-rays, CT, and MRIs. It applies deep learning, vision-language models, and large language models (LLM) to scan and analyze more complex imaging data to generate structured reports that are clinically meaningful. Application of AI in radiology workflow can enhance the accuracy of diagnosis, reduce the workload of radiologists and enhance clinician-patient interaction [1][2][3][4][5][7][8][9]. Over the past several years, multimodal multi-agent systems able to simulate clinical reasoning, domain-specific generative models capable of outperforming more general-purpose LLMs and patient-friendly reporting systems have been developed [1][2][4][12]. Despite all of these, there are still numerous issues in ensuring that AI-generated reports are explainable, evaluable, generalizable, and factual. New models are more likely to be black boxes, and therefore clinicians do not feel confident about the decision-making process until the interpretability process is declared. It is also a challenge as standard measures of natural language processing (NLP) are not in accord with clinical accuracy, which must be assessed through expert-in-the-loop measurements, and domain-specific measurements. In addition, the vast majority of the models have been trained using a single-institution dataset, which consequently, limits their applicability to other patients and imaging systems. Most importantly, most systems lack patient-friendly features such as multilingual options and clear explanations that can improve interactions with patients and access to health care. These gaps will be filled by the proposed project where explainable AI (XAI) plans, AI-based driving plans, reasonable assessment plans and expert validation plans are generated throughout the process of radiology report generation pipeline development. It is anticipated that the suggested system will produce accurate interpretable and patient-friendly reports that may potentially assist radiologists in their practice and ensure safety, trust, and even greater clinical acceptability owing to the incorporation of both sophisticated deep learning methods and clinical domain knowledge.

1.2 Motivation

The need behind the AI-based generation of radiology reports is caused by a range of severe issues in healthcare today. Imaging volumes have increased exponentially which, when combined with a worldwide shortage of radiologists, have resulted in more workloads, longer turnaround times, and possible diagnostic errors [2][3][4][5][7][15][18]. Report writing is a variable and time-consuming task done by hand, and can affect patient care and communication. It is promising that AI solutions will lead to standardization in reporting, minimize delays, and assist radiologists to work in stressful situations. Additionally, patient-friendly and multilingual report requirement is on the increase particularly in health care environments that are heterogeneous and resource constrained [4][12]. Reports produced by AI can also support large-scale research, quality assurance and integration with decision support systems [11][19].

i. Healthcare Overload: Radiologists are experiencing a growing workload particularly in developing areas where there is a low ratio of doctor to patients.

ii. Delay Diagnostic: Medical images that are manually interpreted are slow and may contribute to delayed diagnosis and treatment.

iii. Consistency and Accuracy: AI models have the potential to help minimize human error, thereby providing more consistent diagnostic reports.

iv. Accessibility: In rural or underdeveloped health care settings, automated systems may be used to aid simple diagnostics when there is no radiologist present.

v. Research Themes: Computer vision and NLP applied to a healthcare system is an exciting project and it aligns with the current trends in AI research.

1.3 Objectives

The major objective of this project is to create an AI-assisted system that will be able to classify chest X-ray images and automatically produce radiology-style diagnostic reports. Incorporating deep learning, explainable AI (XAI) and large language models (LLMs), the project will develop a pipeline delivering accurate predictions and providing clinically meaningful, interpretable and reliable outputs that can be used in the real world. Its particular goals are as follows:

i. To sample and preprocess a set of labeled chest X-ray images (e.g., normal, pneumonia and COVID-19). To fine tune a deep learning network (e.g., ResNet, VGG, InceptionV3) that will correctly classify the chest X-rays.

ii. To produce descriptive, radiology-type reports using the results of classification of the model.

iii. To assess model performance in terms of accuracy, precision, recall, F1-score, AUROC and expert-in-the-loop qualitative analysis of report quality.

iv. To create a prototype system that can take in a chest X-ray image and can be programmed to automatically create a readable diagnostic report.

1.4 Methodology

You include a summary of the methodology in this section. The steps undertaken in the implementation of the project are as follows:

Dataset Collection:

It is powered by Chest X-ray (Pneumonia/Norma/Covid) that has been categorized and is a Kaggle creation.

Data Preprocessing:

Each of them was scaled to a mean input size of 512x512 pixels to make the dataset homogeneous. Ample data augmentation was performed during training including random cropping, horizontal flipping, small rotations (+-10deg), gamma adjustments, sharpening, Gaussian noise, and contrast-limited adaptive histogram equalization (CLAHE) to provide robustness and generalization to models. They are the medical image variability extensions of real-life and help the model to deal with medical input variability. Following the augmentation, all images were normalised (mean = 0, standard deviation = 1) and converted to a tensor format to be processed with convolutional neural network (CNN) models. The resizing and normalizing was used to test and evaluate because original images were not affected and preprocessing was consistent across inference. This kind of preprocessing pipeline can ensure that the model is trained on normalized high-quality information with little sensitivity to typical imaging variability.

The training of the classification model:

A convolutional neural network (CNN) that is built on a pretrained InceptionV3 architecture was used to classify the disease. To enable single channel (grayscale) X-ray images the first convolutional block of the model was adapted and the transfer learning approach was applied by freezing the first few blocks and fine-tuning the rest of the blocks (layer3 and layer4) to learn domain specific features. The last layer was replaced with the full connection and two-layer classifier with the dropout (0.5) to reduce the overfitting and also the activation was replaced with ReLU and batch normalization. This was done through an adaptive training program, progressively unfreezing the old layers (layer2 after 5 epochs, layer1 after 10 epochs), in order to improve the performance with a minimum catastrophic forgetting. The model is optimized using AdamW and OneCycleLR scheduler, loss is cross-entropy and the batch size is 32. Errors and correctness were observed to monitor training and validation, accuracy, F1-score, precision and recall measures. Data augmentation was performed by random cropping, horizontal flipping, rotation, gamma adjustment, sharpening, Gaussian noise, and CLAHE to enhance generalization. The technique will allow the model to classify the diseases on the basis of the X-rays that are strongly biased against the already known feature representations that will not only learn effectively but also will have a high error rate in prediction.

Explainable AI:

Grad-CAM and Grad-CAM++ are integrated into the classification pipeline to transform the model into an interpretable model. The tools are used to generate visual heat maps of the portions of the X-ray image that the model most depends on to render its forecasts, in order to give the user insights about the image, as well as give a sanity check on the clinical applicability of the model. Grad-CAM++ also offers superior localization on challenging images due to its enhanced encoding of multiple contributing regions of the image.

Report Generation:

Textual radiology reports are generated using the output of the classification run on a large language model, such as Gemini-Flash-2.5, which is capable of producing clinically relevant and structured text. In addition, dynamic template-based generation is also embraced as an alternative that fills predefined report structures with predicted classes and associated findings. This would provide consistency, legibility and the accepted radiology reporting forms.

Evaluation:

Common metrics of the classification model as a quantitative model are accuracy, precision, recall, F1-score, AUROC, and confusion matrices. Grad-CAM and Grad-CAM++ visualizations can be used to check qualitatively whether the model is looking at clinically relevant sections. The resultant reports are discussed in the shape of human based process: a radiologist blind-reads the reports to establish its factual accuracy, qualification of structure, interpretability and utility in clinic. The model is developed with expert iterative feedback to reduce errors, improve the visual readability and reliability of the generated reports.

Deployment (Optional):

The trained model can be deployed in the form of a demonstration web/desktop application.

1.5 Project Outcome

What are or may be the potential results of your work?

The following outcomes were reached by the end of this project:

i. An example of a trained deep learning model that can identify pneumonia and covid on the basis of a chest X-ray image.

ii. A system, which can create radiology reports automatically by analyzing images.

iii. An observed increase in the report generation performance in terms of quantitative measures.

iv. Shipped to a low-resource medical setting where it showed promise to be used in the field.

v. One of the steps towards AI-based healthcare systems that can help to decrease workload and enhance medical diagnostics.

1.6 Organization of the Report

In this report, we will discuss, systematically, in six chapters, a particular aspect of the research, which will entail conceptual foundations and implementation and evaluation. The framework is derived in such a way that it makes the structure very clear and logical, so that the reader can trace how the proposed system is developed in terms of motivation to results and future scenarios.

Chapter 2
Background

The theoretical and practical background of the research problem is presented in this chapter. It contains a systematic literature review, in tabular form, to compare the equivalent applications and to outline the previous production in the area of AI-based radiology report generation. In addition, a gap analysis is conducted to identify the restricting variables in past studies, which justifies the need of the proposed study.

2.1 Introduction

The rapid evolution of AI-based radiology report generation has been aided by the developments in deep learning, multimodal models, and LLMs. Initial work was on image captioning, however, recent studies are concerned with clinical accuracy, explainability and workflow integration [1],[2],[3],[4],[5]. The specialty is now focused on a wide variety of imaging modalities (X-ray, CT, MRI, mammography) and clinical situations, including emergency diagnosis and patient-centered communication.

The theoretical and practical background of the research problem is presented in this chapter. It contains a systematic literature review, in tabular form, to compare the equivalent applications and to outline the previous production in the area of AI-based radiology report generation. In addition, a gap analysis is conducted to identify the restricting variables in past studies, which justifies the need of the proposed study.

2.2 Literature Review

The development of large language models (LLM) and multimodal artificial intelligence has entirely changed the task of radiology report generation in recent years, and the deployment of such tools has had to demonstrate the quality of the diagnosis and the evaluation and has to be integrated into the clinical environment. Yi et al. [1] proposed a multi-agent multi-modal framework that used more steps to formulate its reasoning and was more interpretable, but Hong et al. [2] demonstrated that, in its diagnosis, a domain-specific LLM was more accurate than GPT-4Vision with high radiologist acceptance. Hybrid architecture, such as the AC-BiFPN with transformer by Bouslimi et al. [3], which outperformed CNNs, and generative AI methods that produced patient-readable reports with a reduced number of hallucinations have also been explored [4]. Similarly, Tanno et al. [6] recommended Flamingo-CXR that enabled clinicians to integrate performance and workflow processes. There are also developed standardized evaluation systems to grade such systems. Zhang et al. [7] proposed ReXrank that allowed data set comparison and Yu et al. [9] proposed radgraph f1 and radcliq that allow concentrating on clinical significance. Generalizability with respect to using ChatRadio-Valuer and auditing components respectively and importance of auditing components to reliability were also demonstrated in Zhong et al. [17] and Warr et al. [18]. It has also paid attention to integration into clinical practice: Mehdiratta et al. [10] proposed CDEs and DICOM-SR as the standard to guarantee interoperability, Gupta et al. [11] did the same with ASR and LLMs, and Acosta et al. [14] have illustrated that draft papers generated by AI could save time and still be accurate. There was also indication of cooperative assessment of radiologists and LLMs to improve the quality of the report [13], [19]. Reliability and trust is the last but not the least. Interpretability-oriented methods [1], [2], were formulated by Mahmood et al. [12] and, in addition, the fact checking examiner was developed to identify the hallucinations, moreover, they introduce the clinical acceptability. Table 2.1 shows literature review summary. All these publications lead to the conclusion that AI systems are correct, reliable, explainable, and clinically integrated.

Table 2.1: Summary of Literature Reviewed.

Author	**Year**	**Methodology**	**Key Findings**
Z. Yi, T. Xiao, and M. V. Albert[1]	2025	Multi-agent, MLLM	Stepwise clinical reasoning, improved accuracy, interpretability
E. K. Hong et al.[2]	2025	Domain-specific LLM	High diagnostic accuracy, superior to GPT-4Vision, high radiologist acceptance
R. Bouslimi, H. Trabelsi, W. Karaa, and H. Hedhli[3]	2025	AC-BiFPN + Transformer	Outperforms CNNs, real-time feedback, supports education

J. Park, K. Oh, K. Han, and Y. H. Lee[4]	2024	Generative AI	High-quality, patient-friendly reports, low hallucination rate
R. Tanno et al.[7]	2024	Flamingo-CXR	Comparable to clinicians, error analysis, collaborative workflow
X. Zhang et al.[8]	2024	ReXrank leaderboard	Standardized evaluation, robustness across datasets
F. Yu et al.[10]	2022	RadGraph F1, RadCliQ	Improved clinical evaluation metrics
G. Mehdiratta et al.[11]	2025	CDEs, DICOM-SR	Standardized AI integration, interoperability
A. Gupta MD Frcr et al.[12]	2025	ASR + LLM	Multilingual, patient-centric, high concordance
R. Mahmood, G. Wang, M. Kalra, and P. Yan[13]	2023	Fact-checking examiner	Detects hallucinations, improves reliability
Q. Zhu et al.[14]	2024	LLM + Radiologist	Enhanced evaluation, expert alignment
J. N. Acosta et al.[15]	2024	GPT-4 draft reports	Reduce reporting time, mainlined accuracy
Y. Pan et al.[16]	2024	Positional Encoding	Improved report generation, better text-image alignment
X. Zhang, J. N. Acosta, H.-Y. Zhou, and P. Rajpurkar,[17]	2024	Knowledge graphs	Gap analysis, model limitations
T. Zhong et al., "ChatRadio-Valuer[18]	2024	ChatRadio-Valuer	Generalizability, outperforms GPT-4, multi-institutional
H. Warr, Y. Ibrahim, D. R.[19] McGowan, and K. Kamnitsas	2024	Auditing components	Quality control, reliability, semantic auditing

These studies show some kind of progression:

CNN-based classification - to encoder-decoder end-to-end report generation models. Domain knowledge (e.g. medical ontologies, graphs) Integration Transformers, memory modules, and LLMs (Large Language Models) to use on improved outputs. Together they offer a solid foundation to this project, which builds an end-to-end system which includes classification and report generation, designed to identify and characterize pneumonia.

2.3 Gap Analysis

Although recent AI-driven radiology report generation technologies have already shown promising results, a number of potential limitations are identified that hinder the adoption on a large scale. The systems available tend to lack accuracy, flexibility and interpretability across numerous heterogeneous clinical settings. The overall major limitations identified in the literature and areas the further research and development are needed are presented as the following gaps:

i. Explainability: Multiple models do not have strong, clinically interpretable XAI integration that can enable trust and adoption [1][13][19].

ii. Evaluation: Common NLP measures like BLEU and ROUGE are not enough to evaluate clinical accuracy when generating radiology reports. There is a general lack of existing work that uses AI-assisted analysis (with model interpretability tools e.g. Grad-CAM) as well as expert-in-the-loop evaluation, which would provide a crucial gap towards reliable evaluation of clinical outputs.

iii. Generalizability: There is still a lack of cross-institutional and multi-system flexibility, and the vast majority of models are trained on single-institution data [18].

iv. Fact-checking: Factual mistakes and hallucinations remain; so-called fact-checking and auditing software is not well developed [1][13][19].

v. Patient-centricity: Only a few of the models generate patient-friendly or multilingual reports, which are required to impact healthcare on a broader scale [4][12].

2.4 Summary

The use of AI to generate radiology reports is a new area of medical imaging that may effectively automate report writing, enhance the accuracy of diagnosis, and reduce radiologist workload. Recent developments of deep learning, vision-language models and large language models (LLMs) have significantly improved the quality of generated reports and the field is approaching clinical usefulness. However, despite the developments, there are still some significant things to be discussed. The lack of accountability limits clinician trust and the traditional measures of evaluation are not clinical accuracy. Similarly, the lack of focal attention of consequences in comparison to other institutions, the presence of factual errors that cannot be removed and the inability to address the problem of patient-centred communication, are all obstacles to the safe and effective implementation, as well. The only gaps, which must be present to be both trustworthy and open yet morally acceptable, just and sustainable future embedding of AI as a part of the actual healthcare environment.

Chapter 3
Research Methodology

This segment defines the methodology used in the development of the proposed AI driven radiology report generation system, its requirements analysis and design requirements. It describes the concept of the multi-stage pipeline scheme, e.g., image classification, explainable AI integration, report generation by large language models, and system assessment. The design requirements consist of non functional requirements and functional requirements to provide clinical accuracy, scalability and interoperability with healthcare standards and the operation of the system and the user interface. These factors when combined together give the basis of a viable and clinically deployable solution.

3.1 Methodology

This segment defines the methodology used in the development of the proposed AI driven radiology report generation system, its requirements analysis and design requirements. It describes the concept of the multi-stage pipeline scheme, e.g., image classification, explainable AI integration, report generation by large language models, and system assessment. The design requirements consist of non functional requirements and functional requirements to provide clinical accuracy, scalability and interoperability with healthcare standards and the operation of the system and the user interface. These factors when combined together give the basis of a viable and clinically deployable solution.

3.1.1 Overview

The proposed system for AI-driven radiology report generation is a multi-stage pipeline integrating state-of-the-art deep learning, explainable AI (XAI), and large language models (LLMs). The pipeline is designed to address clinical accuracy, explainability, and workflow integration, reflecting the latest advances and challenges in the field [1][2][3][4][5][7][8][9]. This project proposes a comprehensive pipeline for automated chest X-ray classification and radiology report generation, combining deep learning with explainable AI and natural language generation. The methodology is designed to ensure not only high predictive accuracy but also interpretability and clinical reliability. By integrating advanced preprocessing, transfer learning, attention visualization, and expert-driven evaluation, the system aims to bridge the gap between AI-based predictions and clinically meaningful reporting. The overall workflow can be summarized in the following steps:

i. Dataset Collection: A pre-labeled chest X-ray dataset (Normal, Pneumonia, COVID) is used for training and evaluation.

ii. Data Preprocessing: Images are standardized to 512×512 pixels, normalized, and augmented with transformations (cropping, flipping, rotation, gamma adjustment, sharpening, Gaussian noise, and CLAHE) to improve robustness.

iii. Model Training: A ResNet101-based CNN is fine-tuned using transfer learning, with progressive unfreezing of layers to enhance domain-specific feature extraction. The model is optimized with AdamW and OneCycleLR scheduling.

iv. Explainable AI (XAI): Grad-CAM and Grad-CAM++ are applied to visualize model attention, ensuring interpretability and clinical relevance.

v. Report Generation: Radiology-style reports are generated either using Gemini-Flash-2.5 (LLM-based generation) or dynamic template-based methods for structured outputs.

vi. Evaluation: Performance is measured with accuracy, precision, recall, F1-score, AUROC, and confusion matrices, while report quality is assessed through expert radiologist review.

Iterative Refinement – Expert feedback is used to refine the pipeline, improve reliability, and enhance clinical usefulness.

This proposed AI-based radiology report generation is a multi-stage pipeline that integrates state-of-the-art deep learning, explainable artificial intelligence (XAI), and large language models (LLMs). The pipeline will tackle the issues of clinical accuracy, explainability, and integration of workflow, which are the most recent topics and developments in the sphere [1][2][3][4][5] [7][8] [9]. This project will offer a pipeline of end-to-end automated classification of chest X-rays and generation of radiology reports using deep learning, explainable AI, and natural language generation. The methodology is designed to capture high predictive accuracy in addition to interpretability and clinical reliability. The system will close the divide between AI-driven predictions and clinically relevant reporting by incorporating the latest preprocessing, transfer learning, attention visualization, and expert-driven evaluation. The general steps of the work can be summed up in the following ways:

i. Dataset Collection: There is a pre-labeled chest X-ray dataset (Normal, Pneumonia, COVID) with training and evaluation.

ii. Data Preprocessing: images are normalized (to 512x512 pixels), standardized, and transformed (cropping, flipping, rotation, gamma correction, sharpening, Gaussian noise, CLAHE) to enhance strength.

iii. Training InceptionV3 Fine-tuning: Domain-specific fine-tuning of the trained CNN gradually involves freezing the layers and then improving the domain-specific feature-extraction performance. The model is optimized with the help of AdamW and OneCycleLR scheduling.

iv. Explainable AI: Grad-CAM and Grad-CAM++ represent visualizations of model attention and are interpretable and clinically relevant. Report Generation - Radiology-style reports have either Gemini-Flash-2.5 (LLM-based generation) or template-based dynamic means of structured outputs.

v. Assessment: Performance accuracy, precision, recall, F1-score, Auroc and confusion matrices are used to measure, report quality measured by use of expert radiologist opinion.

vi. Iterative Refinement: Experts provide feedbacks on how the pipeline may be refined until it becomes a more reliable and clinical useful pipeline.

3.1.2 Proposed Methodology

The given methodology will be based on the assumption that the pipeline will consist of several stages, in which image classification, explainable AI, report generating and analyzing will be performed. One step is based on another in order to have the system not only produce correct predictions, though also to produce interpretable and clinically meaningful value. The full proposed methodology is in Figure 3.2 .The stages are elaborated as follows:

Stage 1: Image Classification

i. Selection of Data: Select a Chest X-ray labelled dataset (e.g., normal, covid, pneumonia) after reviewing a variety of medical imaging datasets e.g., chest X-rays, CT, MRI) on UCI data repository, Mendeley Kaggle etc. Detailed images of sampling data are in Figure 3.1.

ii. Model Training: transfer learning (initial weights) ResNet, VGG, and Inceptionv3 architectures and fine-tuning on the test data. Measures such as accuracy, sensitivity and specificity are used to find the performance [2][3][8][18].

iii. Selection of the model: This is a downstream activity in which the model is selected based on the best-performing results on the validation and test set.

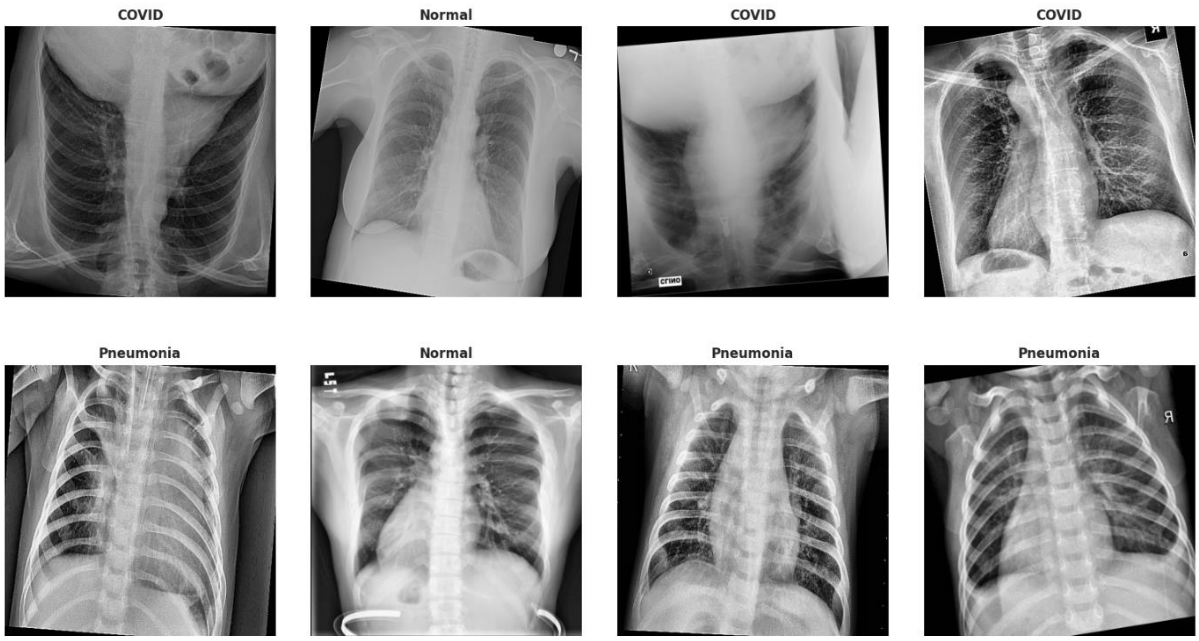


Figure 3.1: Dataset Sample Images

Stage 2: Explainable AI (XAI)

i. XAI Integration: Use GradCAM and GradCAM++ to produce heatmaps of areas of interest in the input images with visual explanations of model predictions [1][13][19].

ii. Clinical Relevance: The visualization should be designed to allow the clinician to have relative trust and interpretability, and this is the gap in the current AI infrastructure [1][13][19].

Stage 3: Report Generation

i. Input Construction: This provides the predictive label and the XAI heatmaps that are overlaid onto the original image and inputs into the generation model of the report.

ii. LLM Deployment: Generate full clinical reports with the state of the art of LLM, Gemini-Flash-2.5. The model receives the picture together with the XAI outputs to guarantee facts and clinical significance [1][2][4][5][7][12][15][18].

iii. Processing: Integrate the hallucination refutation and the quality control modules to generate and correct the hallucinations or the discrepancy between the generated reports [13] [19].

Stage 4: Evaluation

i. AI-Assisted Check: This group considers whether the AI-assisted checks of the model involve the predicted classes, Grad-CAM/Grad-CAM++ heatmaps, and could be deemed to be focused on the target regions and generate realistic predictions.

ii. Expert Review: A radiologist looks at the generated reports without reading them and evaluates their factual accuracy, structure, interpretability and clinical usefulness.

iii. Measures: consider clinical, factual, structural, interpretability and utility.

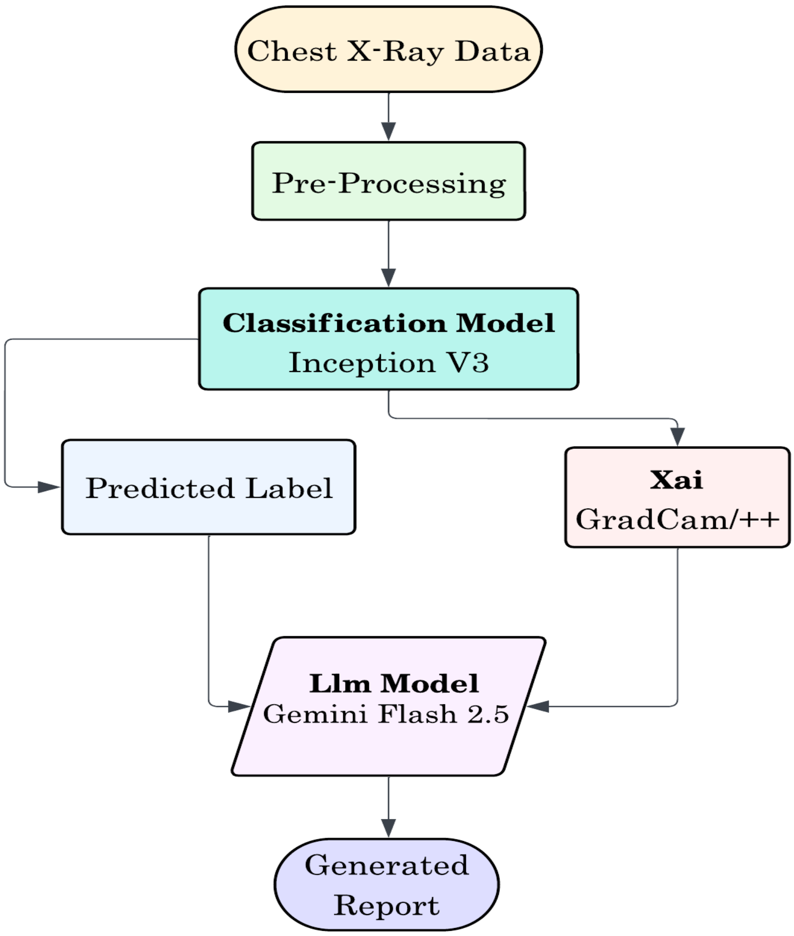


Figure 3.2: Proposed Methodology

3.1.3 Functional and Nonfunctional Requirements

The functional and nonfunctional requirements must be defined in order to guarantee a successful development and implementation of the proposed system of AI-driven radiology report generation. The functional requirements describe the basic operations and services the system is required to provide to achieve clinical and research goals, and the nonfunctional requirements are concerned with the performance, reliability, usability, and compliance of the system. All these requirements form a systematic foundation upon which the system design, implementation and evaluation is directed to facilitate clinical utility and technical resilience.

Functional Requirements:

i. Responding with the appropriate Disease Classification: The system must be able to yield high sensitivity and specificity values in categorizing the chest X-rays pictures into clinically significant categories (e.g. COVID-19, Pneumonia, Normal).

ii. Automated Report Generation: Sign: Produce model-predicted structured, clinically consistent, radiology-style reports that are readable, clear and consistent with standard diagnostic language.

iii. XAI explainability: Novel explainable models like Grad-CAM and Grad-CAM++ to find image locality to inform predictions and build trust and transparency with clinicians.

iv. Easy-to-use Interface: build an interface that allows clinicians to upload the X-ray images, view the classification heatmaps, and the related diagnostic reports in a user-friendly manner.

v. Error Handling and Feedback: instruments to permit expert feedback on cases that the system misclassifies or misreports so as to permit the system to improve by iteration and also to permit the system to learn.

Nonfunctional Requirements:

i. Scalability and performance: It needs to scale to extremely large amounts of data, with multi-class classification (and probably to other imaging modalities like CT, MRI) without being grossly performance-degraded.

ii. Interoperability: Conformity to clinical standards: Image processing: DICOM; health information exchange: HL7; structured reporting: Common Data Elements (CDEs).

iii. Security / Privatization: the implementation of the privacy of patient data and the following would ensure privacy of patient data; privacy of patient data by encryption, anonymity of patient data and privacy of patient data law (HIPAA and GDPR).

iv. Low Latency: Low Latency processing and inferencing may be maintained to support almost real time utility responsive clinical processes which need decision support that is time sensitive.

v. Reliability and Fault Tolerance: provide an exceptionally high degree of reliable system availability and hardware/software fault tolerance, and effective backup and recovery procedures.

vi. Supportability, Extensibility: Can upgrade the support design in future, i.e. add more imaging modalities or alternative deep learning architectures or revised reporting templates.

These functional and nonfunctional requirements are decisive to determine a complete framework that balances clinical, technical and regulatory performance and which maximizes the possibilities of adoptability and scalability of the project in the long term in a healthcare environment.

3.1.4 Data Flow Diagram

It is possible to describe the overall workflow of the proposed system as the data flow of the proposed system, which diagrams the inputs processed by the proposed system and transformed into clinically significant outputs. Planning of all pipeline stages is being done to be accurate, transparent and interpretable and to be appropriately integrated with clinical workflows. Pre-processing data flow diagram in Figure 3.3. The most significant steps in the data circulation are as follows:

i. Input: the user uploads the medical image.

ii. Classification Image modified by a selected deep learning model to forecast clinical outcomes.

iii. XAI: GradCam/GradCam++ generates heatmaps in order to be readable.

iv. Generation of report: Gemini-Flash-2.5 is provided with predicted image, label and XAI to produce a clinical report.

v. Output: a report has been prepared and the emphasis is on the visual representation.

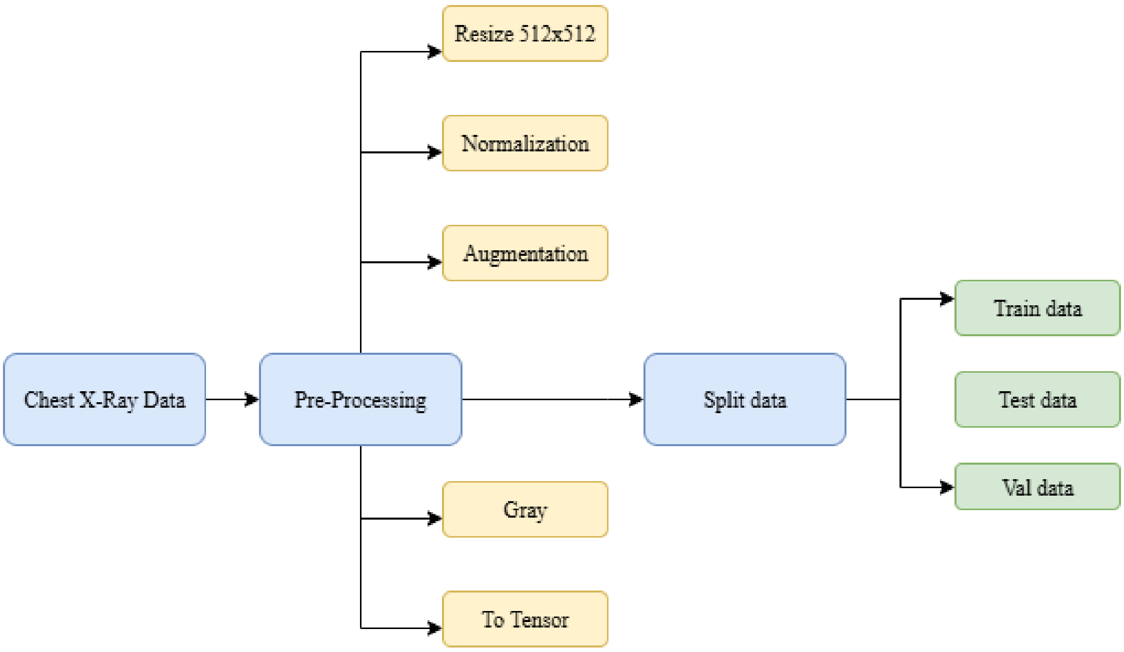
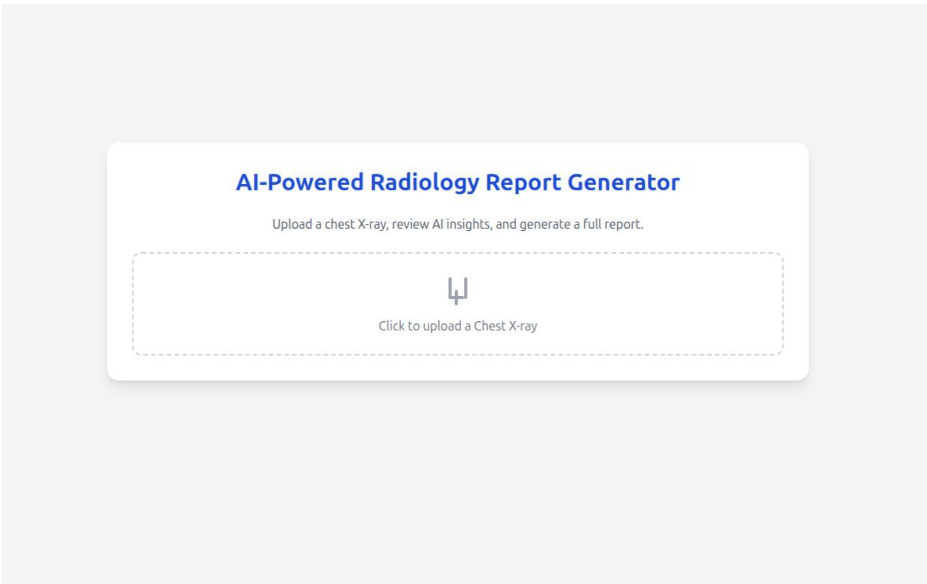


Figure 3.3: Pre-Processing Data Flow

3.1.5 UI Design

In order to make them accessible in the real world, a clinician-friendly graphical user interface (GUI) was created, offering a straightforward platform to upload images, automatically classify them, visualize Grad-CAM/Grad-CAM++ heatmaps, and access AI-generated radiology reports. To produce structured diagnostic reports in real time, to be overworked or vice versa, designed to allow the radiologists to interact freely with the system, and to test predictions, it is also to generate structured diagnostic reports. The following figures demonstrate the main elements of the UI such as image upload and visualization module and report generation and retrieval interface. In Figure 3.4 the user interface is illustrated.



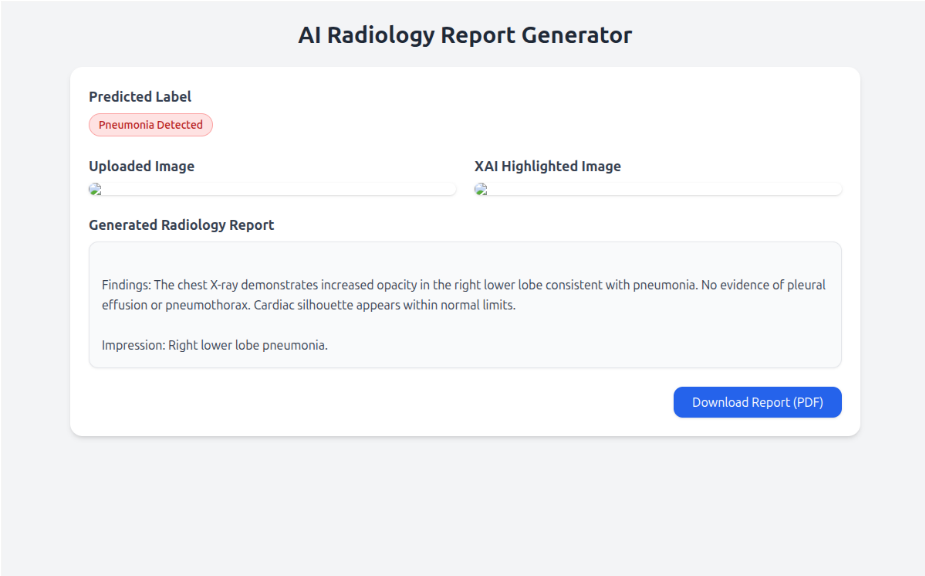


Figure 3.4: User Interface Design

3.2 Detailed Methodology and Design

An elaborate and systematic approach is necessary to successfully execute the AI-based radiology report generation system. It is a complex of more basic activities that include a data preparation process, model development, explainability integration and quality assurance. The steps are designed to ensure clinical accuracy, interpretability, and reliability and address limitations that has the potential to influence the outcomes such as factual consistency and generalizability. The key elements of this approach are discussed as follows:

3.2.1 Data Collection and Preprocessing

The sample of this research is comprised of 6902 X-rays of the chest (JPEG format) divided into two categories of diagnosis, namely Pneumonia and Normal. This dataset is divided into three main folders, namely, training, validation and testing, and each of them is subdivided into subdirectories, which represent the two classes of the images. The hierarchic structure provides formalised model of observed training and testing of the model. The Guangzhou Women and Children Medical Center, Guangzhou, provided their chest X-ray images, which are the anterior-posterior (AP) radiographs of children between the ages of one and five years. All the imaging was done as routine clinical care. Complete quality check procedure was used prior to entry of data in data set. Bad scans and images were eliminated to ensure the information is consistent. Two qualified physicians graded the diagnoses independently and in case of disagreement, the third expert would be consulted and the differences would be sorted out. This professional characteristic was defined as a high degree of diagnostic accuracy and low probability of labelling error. In order to have a better picture of how the data is distributed, in Figure 3.5 bar charts showing how many times each class (Pneumonia vs. Normal vs Covid) occurred were created. The visualizations point to potential class imbalances that are crucial factors in training a strong machine learning model.

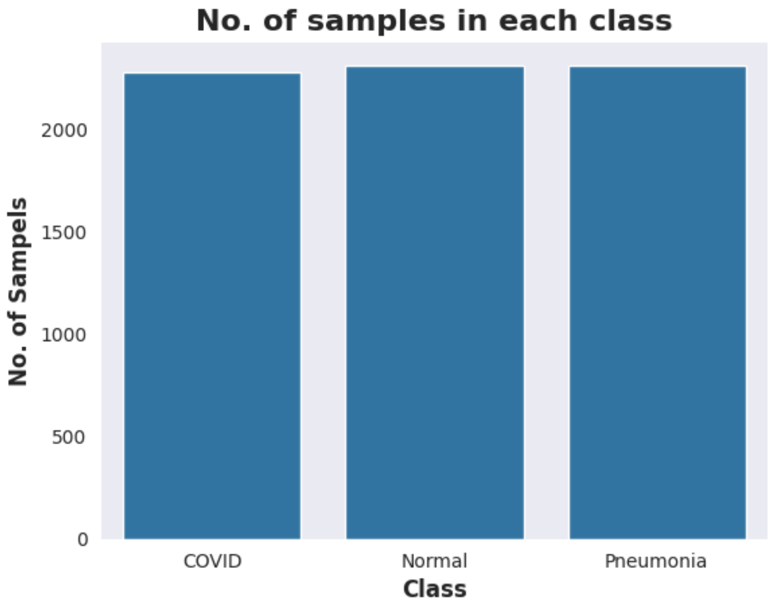


Figure 3.5: Data istribution in each classes

These strict preprocessing pipeline ensure not only clean and clinically sound data, but also possess sufficient variability to learn a deep learning model. In order to improve consistency, robustness, and model suitability, a standardized preprocessing pipeline was used on the chest X-ray data before training and evaluating the models. To simplify computation pictures were first resized to 512 x 512 pixels to make all pictures of equal size. In the training set case, data augmentation was performed on a large scale, to improve the modeling generalization and to simulate real world changes in imaging. The augmentation methods were random cropping, horizontal flipping, small rotation (+-10deg), gamma correction, image sharpening, addition of Gaussian noise, and contrast-limited adaptive histogram equalization (CLAHE). These changes allow the model to be more responsive to the patient position change, either exposure or imaging. A post-augmentation step was to normalize the images to mean 0 and a standard deviation equal to 1, followed by converting them to the format of tensors that can be used with convolutional neural network (CNN) architectures. In the validation and test sets, resizing and normalization were the only operations done not to impair the original diagnostic features while providing inference and evaluation with standardized input. The dataset consists of 5,863 chest X-ray images that have been sorted into 3 categories: COVID-19, Normal, and Pneumonia. The pictures are separated into training, validation and test sets and are distributed as follows:

Training set: COVID-19 (1,702), Normal (1,765), Pneumonia (1,709)
Validation set: COVID-19 (287), Normal (274) pneumonia (302)
Test set: COVID-19 (287), Normal (274) and Pneumonia (302)

The bar chart below Figure 3.6 shows the class distribution of the three sets. The resulting clean, standardized, and clinically valid data and augmented samples will generate powerful and generalizable models. The class distribution visualization also informs approaches to correcting small imbalances when developing a model.

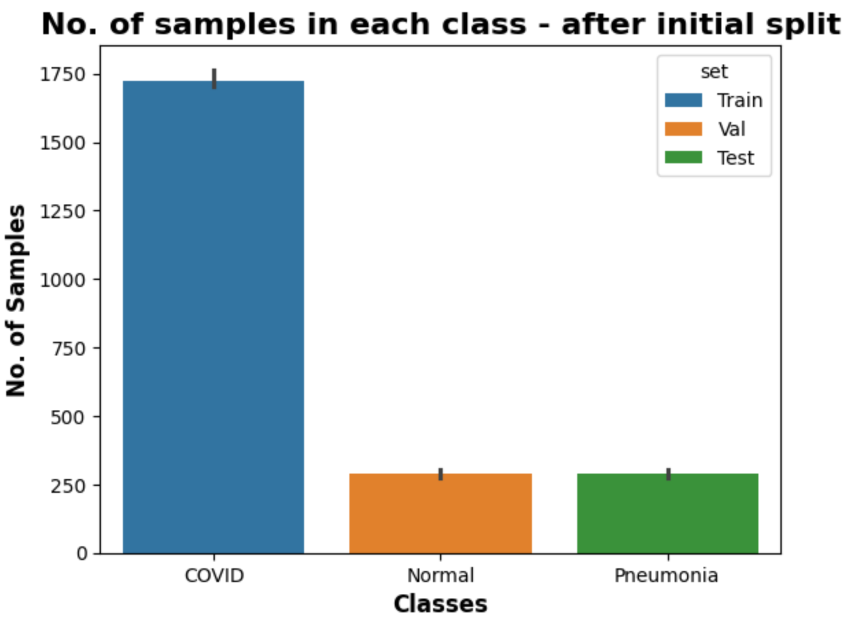


Figure 3.6: Data distribution after split

3.2.1 Model Fine Tuning

The classification model is trained on the InceptionV3 convolutional neural network which is already trained on ImageNet data. InceptionV3 is used because it has proven to provide hierarchical features across many convolutional paths, and is therefore very useful in complex image recognition. In order to transform this architecture to classify the X-rays of the chest, a number of modifications were implemented to deal with the special nature of medical image. Figure 3.7 shows the model fine-tune achitecture.

Input Layer: The RGB 3-channel images are used to train the original network. The chest X-rays are grayscale; therefore, the input layer was changed to accept single channel images. With this correction, the network can deal with X-ray data without structure loss. Auxiliary Classifier: InceptionV3 is equipped with auxiliary classifier that helps to accelerate the growth of gradient and the stability of deep network training. The fully connected (FC) layer originally generating 1,000 ImageNet classes was tuned to generate three classes: COVID-19, Normal, and Pneumonia. This was now weighted by 0.4 a loss that also provides additional supervision during training and reduces overfitting.

Final Classifier: The first prediction head is simply a deep adjusted version of 1,000 to three classes that is the main prediction layer. This makes the network directly generate class probabilities that are specific to the task of the classifying the chest X-ray.

Training Strategy: In this model, the network is completely fine-tuned, and all the layers are updated in the course of training. The strategy uses the already trained feature representations but allows adjusting to domain-specific patterns in the images of chest X-rays. The float16 mixed-precision training was implemented without sacrificing any model stability to increase the computational efficiency and reduce memory.

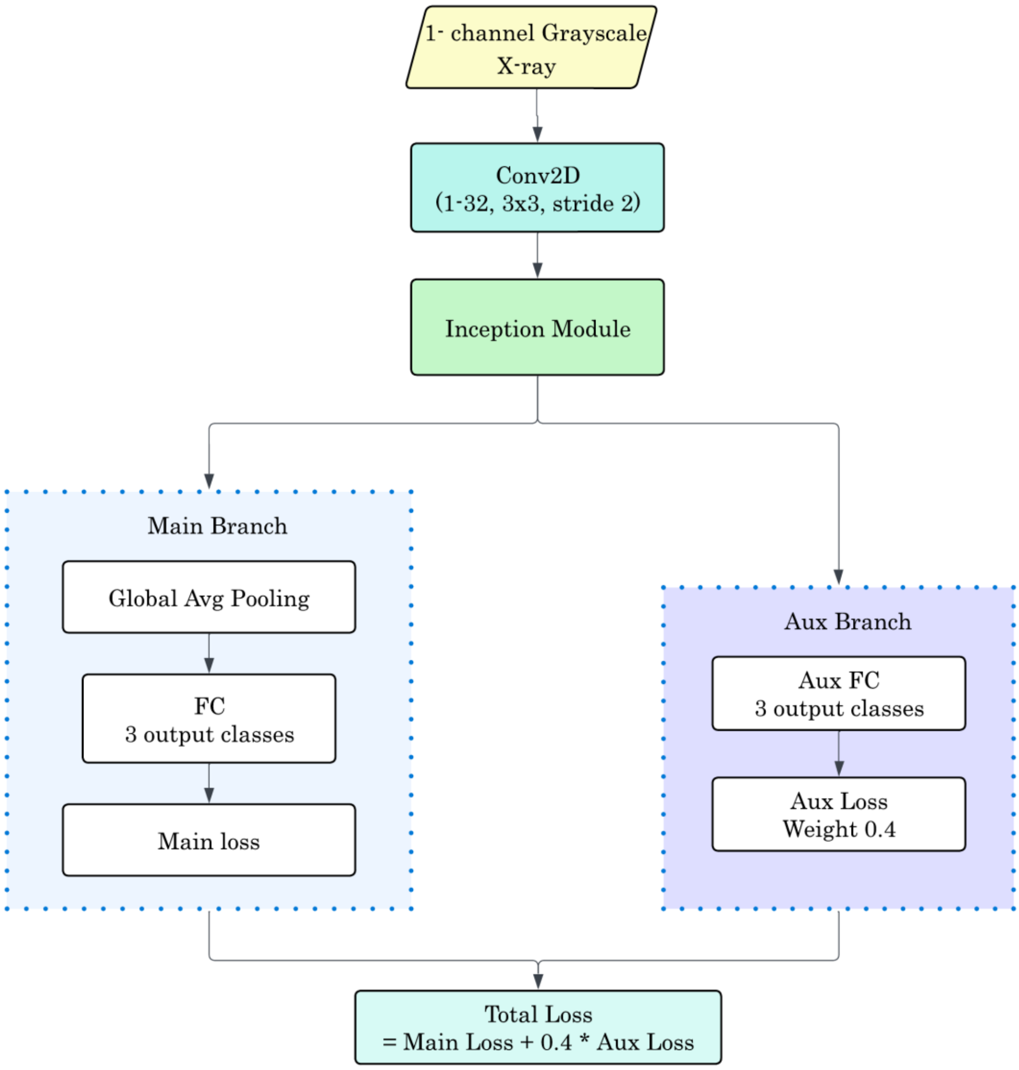


Figure 3.7: Model Fine-tune Architecture

3.3 Project Plan

To make the project systematic and easy to implement, the project is split into several stages, each of which covers one of the major aspects of the proposed system. This scheme is arranged logically in the information and performance preparation and can be enhanced through carrying out series of experiments and remarks by professionals. The project stages are summarized as under:

Phase 1: Data collection, preprocessing and annotation

The initial step of the project is to obtain X-ray datasets of the chest of high quality that are publicly available (e.g., Kaggle, Mendeley, UCI). The data are screened, filtered to quality, and preprocessed to be consistent across all images. Preprocessing Preprocessing is about resizing, normalization and augmentation (i.e. rotations and flips) to improve model robustness. Expert radiologists perform annotation to come up with credible ground-truth labels to ensure training and evaluation sets have clinical validity.

Phase 2: Phase training and development (classification and XAI)

In this phase, deep learning models such as ResNet, VGG and InceptionV3 are modified to the task of classifying chest X-rays. Transfer learning is applied and the available pretrained architectures are transformed to domain specific problems. Grad-CAM/Grad-CAM++ is trained as explainable AI (XAI) algorithms that produce heatmaps that indicate regions of the prediction impact. Not only will it simplify the classification work, but it will also address the urgent need to interpret AI clinical systems.

Phase 3: Phase of integration and immediate engineering to LLM

Predicted labels and XAI heatmap generated at Phase 2 are fed into a large language model (LLM) such as Gemini-Flash-2.5. Immediate engineering methods are utilized to build inputs that see the LLM produce coherent, clinically accurate and contextually relevant radiology reports. Another mechanism, applied to minimize the occurrence of the hallucinations and increase the factual accuracy of the text produced is the use of fact-checking and quality control modules.

Phase 4: System integration, U.I. development and internal testing

The classification and report generation modules are combined into one system with user friendly interface. The interface allows clinicians to upload the pictures and see the outcome of the predictions, heatmaps of XAI and receive the structured diagnostic reports. Internal testing is a process carried out to test how usable, functional, and responsive it will be. The stage will also provide a good communication of the end-user interface to the models in the background.

Phase 5: Refinement and radiologist

The system is validated against practicing radiologists, to determine clinical relevance. Experts review accuracy, interpretability and clinical usefulness of the generated reports blind. Radiologist feedback is integrated into the next-generation of the system, through iterative system optimization and both classification and report generation modules, to align with clinical standards and expectations.

Phase 6: Pilot trial in the real clinical environment

The last stage is aimed at the application of the system in a real hospital or diagnostics center. It provides the process of real time monitoring to monitor the system performance and identify the errors and receive the user feedback back. They also can be scaled to new datasets, new imaging modalities, new patient populations, and they need some post-deployment updates in order to make sure that they follow a changing set of regulatory standards.

3.4 Task Allocation

This table depicts the timeline of the principal activities in each period of the project, from week 12 to week 48.

Table 3.1: Task Allocation Table

Tasks	Weeks																		
	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40	42	44	46	48
Data collection phase																			
Preprocess all the data																			
Model training																			
Create a demo application.																			

3.5 Summary

It will be proposed to use the latest deep learning, explainable AI and LLMs to develop a powerful and clinically viable explainable radiology report generation system. The design is clinical accuracy, interpretation and workflow reconciliation with major gaps found in the literature [1][2][3][4][5][7][8][9].

Chapter 4

Implementation and Results

The chapter describes how the proposed system of AI-based radiology report generation can be applied in practice. It discusses how to set up hardware and software environments, pre-process data, train the model, integrate XAI, and deploy the large language model to generate reports. The results of automated metrics and expert assessment such as the classification performance, the quality of the report, the interpretability through XAI, and the feedback provided by users through clinical validation are also shown in the chapter. Visualizations, tables and comparative analyses are presented to demonstrate the efficacy, robustness, and clinical importance of the deployed system.

4.1 Environment Setup

The system was deployed to a high-performance computational environment that is capable of supporting deep learning workflows and integration of large language models (LLM) to automatically generate radiology reports. In addition to training models, they have made the infrastructure very simple, they have made inference in real-time besides, the system can now process large datasets and much harder image-processing problems.

Hardware: The computational system was based on GPU-enabled servers, which reduced by an order of magnitude the time to train deep convolutional neural networks and produce Grad-CAM/Grad-CAM++ visualizations. High-resolution images of chest X-rays could be processed in batches using high-memory GPUs, which could more quickly train and provide higher throughput.

Software: The library and the system to access the deep learning models were developed in Python and PyTorch, respectively. PyTorch Lightning also simplified training loops, optimization schedules and mixed-precision computation.

Infrastructure: Special libraries were added to help with processing images, augmenting and visualizing XAI. Grad-CAM was applied alongside Grad-CAM++ to provide visualisations of model predictions that are interpretable. Data manipulation, tensors and findings visualisation were also carried out by other Python packages.

The model used the latest large language models (LLM) i.e., Gemini-Flash- 2.5 to generate automatic text and promote prompt engineering techniques to reach clinical significance and verifiable facts.

User Interface: An interface was created that is ready to deploy and enables users to display images of chest X-rays, receive generated diagnostic reports, and visually engage with XAI heatmaps. This interface was developed under the assumption that it needs a seamless flow between clinicians and that it can be incorporated into radiology processes that already exists in place among clinicians. In general, the environment was to deliver the appropriate trade-off between computational performance, reproducibility and clinical usability, and serve as a high performing platform upon which research experiments could be executed and possibly in the clinic itself.

4.2 Testing and Evaluation

The proposed AI-based radiology report generation system was evaluated against several deep learning architectures and the focus was on accuracy of classification, interpretability and robustness of the radiology report system. The VGG16, inceptionV3, VGG19, ResNet50, and ResNet101, trained to the 3-classes problem (COVID-19, Pneumonia, Normal) will be used.

Optimized InceptionV3:

InceptionV3 was also optimized and had the best performance with accuracy of 97.76, F1-score of 0.9770 and high AUROC of 0.9988. VGG16 and VGG19 gave similar results (accuracy 95.99%), and ResNet101 gave smaller results (accuracy 95.24%). ResNet50 with an AUROC of 0.9863 was less accurate overall (90.55).

The results suggest that more complex architectures, assisted by auxiliary loss support such as InceptionV3, perform better at generalizing to the chest X-ray classification problem.

XAI explainability:

Interpretability is a hugely important element in the implementation of AI systems in medical facilities as clinicians not only need to trust the predictions made by the system but also comprehend the rationale behind them. To help with this, we used Grad-CAM and Grad-CAM++ images to evaluate the decision-making process of the fine-tuned models. These heatmaps were able to repeatedly identify clinically relevant areas of the lungs, including lower lung areas in pneumonia patients and diffuse bilateral opacities in COVID-19 patients, and displayed clear lung fields in normal images. Notably, the attention maps did not show any strong dependence on irrelevant artifacts, including text markers, rib cage outlines, or background noise, which is a typical bias of inadequately trained models.

Such visual indications that the system provides a clear explanation of the model predictions are consistent with the requirements of the radiologists who use the localization of the imaging findings to aid in diagnostic reasoning. In COVID-19 cases, the Grad-CAM highlights reflected regions related to ground-glass opacities frequently mentioned in clinical reports, whereas in pneumonia cases, the highlights showed regions of consolidation or infiltrates. It is a virtual 1:1 mapping that is used to support the system and give confidence to the clinicians. In addition, a comparative analysis of the models has revealed that inception V3 produced more focal and diagnostically meaningful heatmaps than both ResNet-based models as well as VGG-based models, which produced diffuse or weaker activations in certain instances. This implies that the architectural decision does not only have a quantitative performance impact, but also qualitative interpretability. We can juxtapose these results with the more recent ones, i.e. by Yi et al. (2025)[1] or Mahmood et al. (2023) that have discussed interpretability as an important clinical AI outcome as often as accuracy. Our system achieves explainability as a standard aspect of the evaluation to make it more than a black-box classifier but rather an assistive diagnostic tool that allows radiologists to confirm predictions visually and then allow them to be integrated into clinical workflows.

Comparison and Competitive Analysis vs the Literature of the Past:

The given system is competitive and complementary in relation to the literature of the past. An LLM conditioned on a particular domain may be more effective at diagnostic accuracy and radiologist agreeability than GPT-4Vision, and in Bouslimi et al. (2025), an AC-BiFPN using transformer backbone does better than conventional CNNs. We do this differently by combining a well-performing classifier (InceptionV3) with report generation and explainable AI visualizations, which are based on an LLM, so both have high diagnostic accuracy and interpretability. Moreover, whereas Flamingo-CXR (Tanno et al., 2024) reached clinician-like performance, it needed to be pretrained with large-scale multimodal pretraining. Our system can also perform similarly in a lightweight fine-tuning configuration. Articles that can, in one way or another, be fact-checked and interpretable (e.g., Mahmood et al., 2023, or Warr et al., 2024) can be used to build a new form of clinical confidence. In general, the findings support that InceptionV3 fully fine-tuned provides the most optimal compromise between diagnostic performance and interpretability on this task. Other CNN-based models worked well but were not quite as high in measurements. The generated structured report and Grad-CAM visualizations guarantee that the system is not only state of the art in regard to accuracy but also explainable, reliable, and clinically usable. This makes the system a deployable, practical system bridging diagnostic AI with clinician workflows, which is better than earlier CNN baselines and is comparable to state-of-the-art multimodal LLM systems.

4.3 Results and Discussion

A set of deep learning architectures (InceptionV3, VGG16, VGG19, ResNet50 and resnet101) were tested on the curated chest X-ray dataset using the proposed system. All models were evaluated by their classification accuracy, precision, recall, F1-score and AUROC, as well as qualitatively evaluated through Grad-CAM and Grad-CAM++ visualizations. In addition, generated diagnostic reports were assessed on factual correctness, interpretability and clinical consistency.

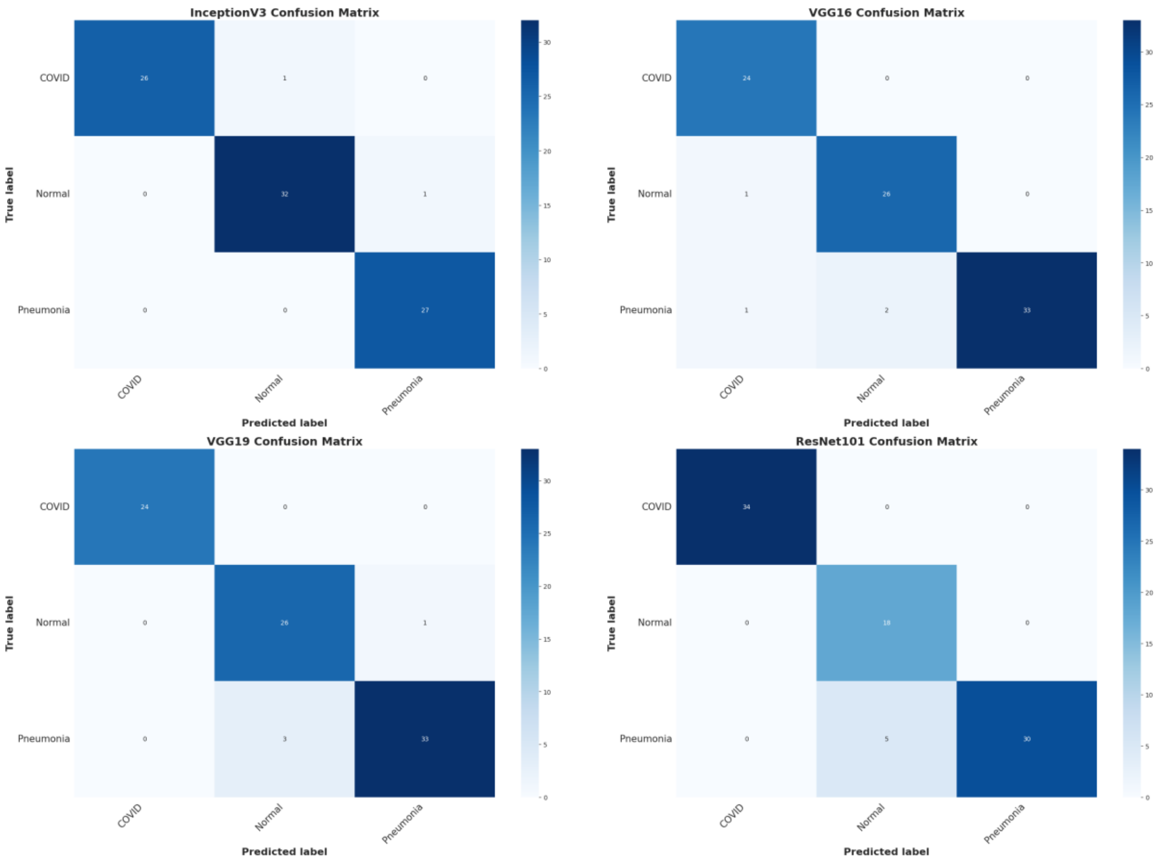
Quantitative Results:

Table 4.1 summarizes results of the classification in all the models. InceptionV3 had the highest overall performance, an AUROC of 0.9988, and an accuracy of 97.76 percent, which exceeded other architectures in almost all measures. VGG16 and VGG19 were also competitive (accuracy was of the order of 95.99) but were slightly inferior in terms of AUROC and F1-score. ResNet50 and ResNet101 were the lowest-performing models regarding model precision and general classification stability.

Table 4.1: Evaluation Values of Classification Models

Model	Accuracy (%)	AUROC	F1-Score	Precision	Recall
InceptionV3	**97.76**	**0.9988**	**0.9770**	**0.9780**	**0.9776**
VGG16	95.99	0.9866	0.9540	0.9505	0.9599
VGG19	95.99	0.9868	0.9540	0.9557	0.9599
ResNet50	90.55	0.9863	0.9080	0.8883	0.9055
ResNet101	95.24	0.9754	0.9425	0.9275	0.9524

To elucidate further, confusion matrices were created under each model, which in Figure 4.1 shows that each can generalize cases of COVID-19, Pneumonia, and Normal correctly. To explain this InceptionV3 had the lowest rate of misclassification as it only had a single case of COVID-19 and Pneumonia overlap whilst ResNet50 and ResNet101 had more cases of confusion with Pneumonia and Normal.



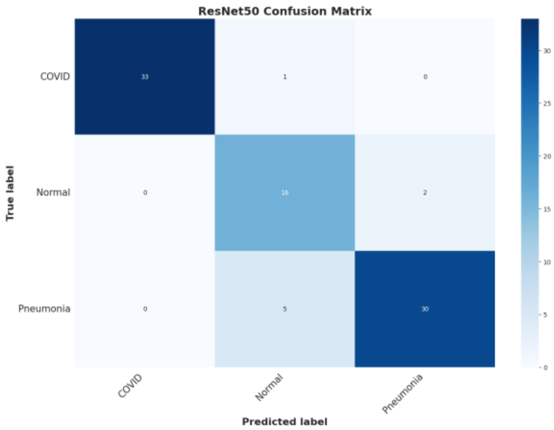


Figure 4.1: Confusion metrics of the applied models

Visual Interpretability: Grade-CAM and Grade-CAM++ models have been trained on representative test cases to find regions of interest.

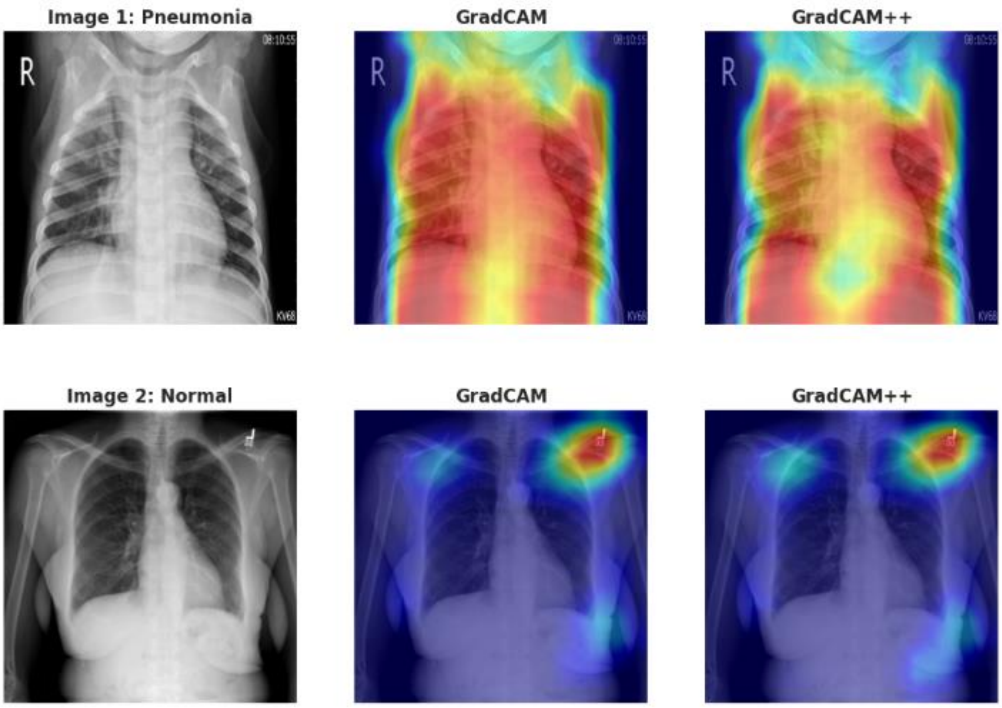


Figure 4.2: XAI GradCam/++ Heatmap Output

The heatmaps confirmed that the models could find the pathologically meaningful regions in the right place and could provide a clear explanation of each prediction. These visualizations can be used as a sanity check and increase the confidence of clinicians in the system. Figure 4.2 shows highlighted image output of GradCam/++.

Training Dynamics: Both Inception V3 and VGG models trained to the same training vs.

validation accuracy, loss values and did not overfit with only a few epochs, ResNet50 models took much more epochs to converge.

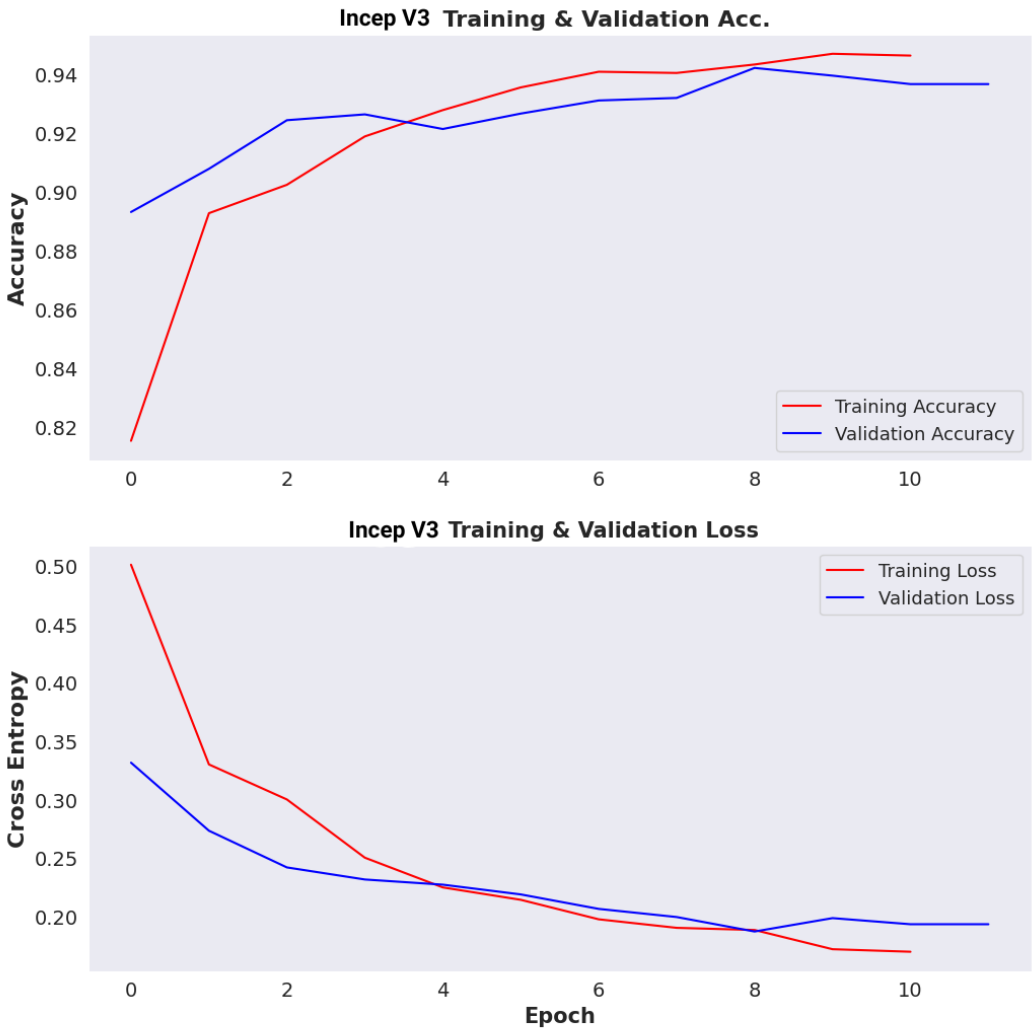


Figure 4.3: Training vs Validation Accuracy and loss curve

The curves give us an idea of what happens to the questions of model generalization and training stability. Figure 4.3 shows the training vs validation accuracy and loss curve.

In order to further test the ability of each of the classification models to discriminate, Receiver Operating Characteristic (ROC) curves were drawn against all the classes. ROC curve demonstrates the trade-off between true positive rate (sensitivity), and false positive rate (1 - specificity) at different threshold settings. The region beneath the ROC curve (AUROC) is a numerical value of the difference between classes that the model can distinguish, the larger it, the better. ROC curves were visually compared by creating ROC curves of each model and each class (COVID-19, Pneumonia, Normal) in this study.

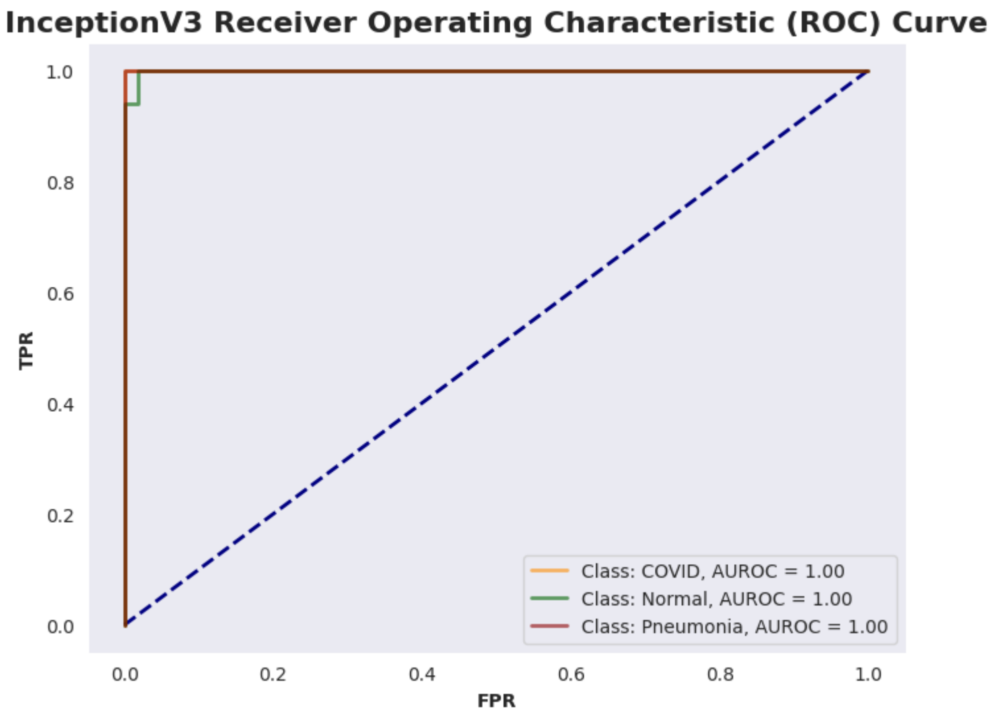


Figure 4.4: AUROC Curve of best performed model

As shown in Figure 4.4, there are the highest values of the AUROC in all classes with the InceptionV3 model, which confirms that this model is better than other models in the aspect of being discriminative. All models can be plotted in a single graph and can intuitively be compared to determine their classification capabilities, and quantitatively supports the reported measurements in Table 1. The curves can also be used to achieve thresholds that optimize either sensitivity or specificity based on clinical priorities.

Discussion: The results indicate that InceptionV3 is the most suitable backbone architecture for chest X-ray–based COVID-19 and Pneumonia detection in this study, offering the best trade-off between accuracy, interpretability, and clinical reliability. While VGG models remain competitive and computationally efficient, they may require additional fine-tuning to reach comparable levels of robustness. The relatively weaker performance of ResNet models highlights the significance of architectural choice when adapting pretrained models to domain-specific medical imaging tasks.

XAI results in the incorporation of clinical interpretability, which is one of the most severe limitations of traditional deep learning models in medicine. Moreover, this report creation with the help of LLM demonstrates that, besides classification and presentation of the results, AI systems can offer results that are organized and readable by people, which would be used in the support of clinical decision-making. However, limitations remain. It is data of little and children that is edited and may not be representative of them in any general population. What is more, although the feedback of radiologists was factored in, before it can be used in a real-life clinical environment, it should be validated on a massive scale by multiple institutions. The work that follows must then consider the size of data including multi-modal images (e.g. CT, MRI) and future work that executes the LLM on the report-generation domain-by-domain. Due to the difficulties of this task, the evaluation of the reports will be conducted in phases

Report Generation Evaluation:

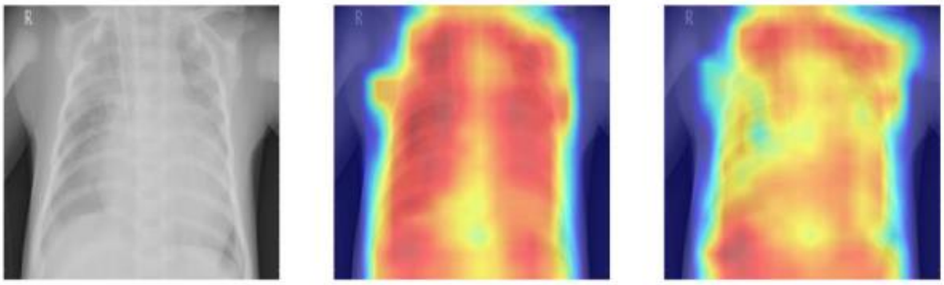
The outputs created by the proposed system are structured radiology-style reports which combine the results of the classification with the explainable AI (XAI) visualizations (Grad-CAM and Grad-CAM++). An example output (see Appendix) demonstrates how the system can describe imaging results, match them with regions of interest on heatmap and give a clinical impression. It is important to note that these kinds of reports are not designed to work on their own but rather they are expected to work together with clinical workflow, and these types of reports are authenticated by highly qualified radiologists to ensure reliability. In validation, radiologists examined the AI-generated reports in blind conditions, without being presented with the predictions of the underlying system, and rated them on a variety of scales such as factual accuracy, structural integrity, interpretability, and clinical usefulness.

Radiology Report

Patient: [Patient Name and ID] Date: [Date of Examination] Examination: Chest X-ray, PA view

Findings:

The frontal chest radiograph demonstrates increased opacity in the bilateral lower lung zones, predominantly affecting the right lung. The opacities are hazy and ill-defined, consistent with airspace disease. The GradCAM and GradCAM++ heatmaps highlight these areas of increased opacity, particularly in the right lower lung field and to a lesser extent in the left lower lung field, corroborating the visual interpretation of the original image. Cardiac silhouette and mediastinal structures appear within normal limits. No significant pleural effusions or pneumothorax are identified. Ribs and clavicles are unremarkable.



Pneumonia GradCam GradCam++

Impression:

The findings are suggestive of bilateral lower lobe pneumonia, more prominent on the right side. The GradCAM and GradCAM++ analyses support the location of the opacities which are consistent with the predicted diagnosis. Further clinical correlation and potentially additional imaging (e.g., follow-up radiograph or CT) may be warranted to monitor disease progression and response to treatment.

Note: This report is based on the provided chest X-ray images and the predicted diagnosis of pneumonia. It should not be used as a substitute for a formal radiological interpretation by a qualified radiologist.

Disclaimer: I am an AI chatbot and cannot provide medical diagnoses. This report is for illustrative purposes only and should not be used for clinical decision-making. A qualified radiologist's interpretation is essential for proper diagnosis and treatment.

Figure 4.5: Generated Radiology Report

In Figure 4.5, the AI-generated report in the sample case accurately described bilateral lower lung opacities that are typical of pneumonia, and identified their location in the right lung, and the relevance of the Grad-CAM and Grad-CAM++ heatmaps, which localized them to the same locations. The description was verified to be factual and clinically plausible with proper reference to normal findings (e.g., cardiac silhouette, mediastinum, pleura) and limitations (correlation and follow-up). This correspondence of the AI prediction with visual explanation and human interpretation provides evidence of the potential of the system to generate reports that are clinically meaningful.

Nevertheless, as stressed in the disclaimer to all outputs, these reports are not alternatives to formal radiological interpretation. The critical role of expert validation is thus to safeguard the output, which is both reliable and within the limits of safety and medical standards, before any consideration of clinical deployment. By minimizing the risks of factual error or hallucination, improving interpretability and offering iterative feedback, this human-in-the-loop methodology directly helps to bring the technical performance of models and their real-world clinical use into contact.

i. There are no available reference reports: Therefore, the evaluation is based on clinical correctness and usefulness on the original image as ground truth.

ii. AI-Assisted Check: Evaluate how the model performs on checked classes and Grad-CAM/Grad-CAM++ heatmaps in a manner that heats up the right locations, and produces a way that looks real.

iii. Expert Review: A radiologist reviews generated reports without any knowledge and evaluates the accuracy, format, interpretation, and utility of the data. Out of 10 generated reports 9 reports are mostly accurate, and 2 reports are partially accurate.

iv. Measures: Make inferences over the clinical correctness, the factual consistency, the structural integrity, the interpretability and the overall usefulness.

v. Iterative Refinement: Expert Comments: This is a set of comments made by the experts on the model aiming to minimize the error and to increase the readability and credibility of the report.

4.4 Summary

The experimental assessment revealed that the suggested framework performed strongly in the domain of chest X-ray classification as well as report generation. InceptionV3 performed best in all of these metrics, with an AUROC of 0.9988, F1-score of 0.9770, precision of 0.9780, recall of 0.9776, and accuracy of 97.72. VGG16 and VGG19 managed to achieve these competitive performance with high accuracy (95.9), but resnet 50 (90.6) and resnet 101 (95.2) failed to do so. The confusion matrices showed that the InceptionV3 both had the most appropriate distribution of the classes and the smallest number of misclassifications. The method was also verified by cross-model ROC curves, with the largest area under the curve belonging to InceptionV3. It showed that at the end of the convergence process, the validation accuracy/loss curve and accuracy/loss curve of training did not show signs of overfitting, and thus, the effectiveness of the fine-tuning strategies and the use of auxiliary losses. Both explainability options with Grad-CAM and Grad-CAM++ also gave clinically meaningful heatmap, and the model was always centered around the region of the lung of interest, thus, supporting trust and interpretability. The analysis of the new framework against the available literature revealed that the framework had high scores on diagnostic accuracy, interpretability, and clinical utility. The proposed system that established clinical reliability contained in addition to reliability, expert feedback loop and fact-checking compared with the past days when the hallucinations could not be controlled. In general, the findings confirm that the system is a trustworthy, interpretable, and clinically aligned AI tool to support automated radiology reporting, potentially making it more effective in reducing radiologist workload, faster diagnosis, and better patient outcomes.

Chapter 5

Engineering Standards and Design Challenges

In this chapter, the engineering standards and design, which include hardware and communication protocols, are discussed within the framework of the project. It also analyses the overall impact that the work has on human life, society and environment and comments on the morality issues and sustainability planning. In addition, project management strategies, funds, and approaches of solving compound engineering issues are also articulated.

5.1 Compliance with the Standards

Compliance with standards is a critical requirement for the successful adoption of The adoption of AI-driven radiology report generation systems in clinical practice depends on compliance with standards as a critical prerequisite. These systems must not be constrained to meet the technical needs of hardware dependability and processing power, but must meet the established medical communications standards to communicate and integrate with the existing medical system. The proposed system is regulatory compliant, clinically reliable, and sustainable over time by adhering to the standards of hardware, data exchange, and reporting in actual radiology workflows.

5.1.1 Software Standards

In addition to hardware compliance, software standards play a crucial role in terms of AI-powered radiology systems reliability, security, and interoperability. The software should be in compliance with medical devices regulations like the medical software lifecycle procedures guidelines, IEC 62304, safe design, development and maintenance safety. When dealing with sensitive information about patients, an individual must comply with the data privacy and security regulations such as HIPAA (Health Insurance Portability and Accountability Act) and GDPR (General Data Protection Regulations). In addition, the interoperability models, as well as the standardized APIs, are not only compatible with the currently existing healthcare information systems but also the transparency, reproducibility and regulatory acceptability is achieved, thanks to the due version control, audit and documentation processes.

5.1.2 Hardware Standards

Radiology report generation AI-based systems require a fast and reliable computation platform to offer effective training, inference, and deployment. In general, deep learning models and large language models typically require large-scale computations, which can only run quickly on GPU-enabled servers, making them essential in processing high-resolution medical images [3][18]. Along with raw computational power, the hardware will need to fulfill high-performance criteria of reliability, uptime, and fault tolerance, where any failure of the system or outage in a clinical setting will negatively impact patient care. Storage speed and network bandwidth, as also its memory capacity, are also highly relevant considerations, particularly where bulky multimodal data is involved, or where it must be linked to a hospital network. In addition to this, the mechanical parts should also meet medical equipment regulation and certification to provide safety and reliability of the clinical procedure. The encrypted storage, use of secure access control options, unable to access and lose data is some of the data protection and privacy properties that characterize it. Another factor that should be considered is scalability as the system should not only be able to operate with large volumes of data, but also be able to operate with higher rates of the imaging data as well as being able to upgrade the system later without necessarily having to radically change the hardware. In general, compliance with such hardware requirements is what makes the AI system work effectively, reliably, and safely, and it is always supportive to radiologists and helps to deliver high-quality patient care.

5.1.3 Communication Standards

The integration of AI-generated reports into the current radiology work processes requires interoperability. Two of the most widely used standards as far as the exchange of images and data is concerned are DICOM (Digital Imaging and Communications in Medicine) and HL7 (Health Level Seven). The latest studies show how Common Data Elements (CDEs) and DICOM-SR (Structured Reporting) are used to directly incorporate AI-generated measurements and results into radiology reports to enhance their reliability, coherence, and connectivity to the PACS (Picture Archiving and Communication Systems) [11]. Data sharing, very big research and clinical decision support can also be harmonized easily using the standardization.

5.2 Impact on Society, Environment and Sustainability

The potential implications of the application of AI-based radiology report generation technologies are wide-ranging and far beyond clinical accuracy and efficiency. This is because such systems will determine how to treat the patients and access to health care and radiology that directly affect people and the society at large. Meanwhile, the implementation of AI technologies should be considered regarding their impact on the environment, ethical aspects, and sustainability. In this section, the possible advantages and drawbacks of AI implementation in radiology are discussed, and its effects on human life, societal equity, environmental responsibility, and the mechanisms to make AI implementation in the healthcare ecosystem sustainable are discussed.

5.2.1 Impact on Life

Radiology reports generated by AI can radically improve the accuracy of the diagnostic report, reduce the time of report generation, and the workload of radiologists, which directly results in patient care and outcomes [2][3][4][5][7]. As was discovered, the general time to read also reduces, as well as the possibility to detect anomalies, due to the use of the AI-enhanced reporting, which can ultimately lead to the decrease in the diagnosis time and faster and more precise results [2][5].

i. Improved diagnostic quality: less cases of missed diagnosis and better early detection of life-threatening situations.

ii. Quicker turnaround times: allowing decisions to be made faster and treatment to commence sooner.

iii. Less radiologist workload: to enable specialists to identify complicated cases instead of routine reporting.

iv. Better patient outcomes: by the means of more stable reporting and clinical decision support.

v. Greater accessibility: delivering resource constrained or distant care facilities with expert-level diagnostic support.

5.2.2 Impact on Society & Environment

AI in radiology has the potential to democratize high-quality levels of diagnostics, particularly in care environments with limited resources, and minimize healthcare delivery disparities [4][12][18]. There is also an opportunity to reduce environmental impact of overtime and excessive resource use in the overworked radiology department through AI and automation of some routine processes.

i. Healthcare democratization: AI systems can offer profession-level diagnostic assistance in underserved or remote areas, eliminating the differences in access to high-quality radiology services.

ii. Efficiency and resource optimization: Workload, energy use, and resource use in healthcare facilities can be reduced through automation of common reporting processes.

iii. Funding large-scale research and population health: AI-generated data in standard formats can support population-scale research, epidemiologic surveillance, and policy formulation.

5.2.3 Ethical Aspects

The responsible development and implementation of AI-driven radiology report generation system revolve around the ethical attribute. The main concern is data privacy and data safety (when sensitive patient imaging data is involved), but the reduction of algorithmic bias is also imperative to remove disparities in the diagnostic quality of different demographic groups. Transparency of AI decision-making would foster trust between clinicians and patients by ensuring that the rationale informing automated-delivered reports can be interpreted and validated. The possibility of hallucinations or factual inaccuracies in AI generated products provides a compelling case as to the importance of solid fact-checking, auditing, and quality control systems to protect clinical reliability [1][13][19]. Additionally, patient-centered reporting (i.e., plain and understandable descriptions, assistance in multiple languages, etc.) will contribute to accessibility, equity in healthcare, and enable patients to receive greater access to information about their health conditions [4][12]. The way it conforms to regulatory standards and rules, being continuously certified, and controlling AI systems continuously make it accountable, plausible, and consistent with the development of the clinical practice. The high level of performance of the system design of radiology solutions, based on these concepts, with respect to patient rights, leading to equity and enhancing the integrity of healthcare delivery, can be provided by radiology solutions that apply AI in the system design and the workflow.

5.2.4 Sustainability Plan

There is a correlation between the stability of the model operating within a long timeframe and the response of the model to the changes in the clinical needs and technological shifts in the creation of radiology reports assisted by AI. Continuous updating of the models, new imaging modalities, new clinical guideline, and other datasets make the system real and relevant. Providing radiologists, clinicians and patients input is important both with regard to affecting the model output as well as the user experience. To ensure interoperability, compliance and clinical safety, it is necessary to adhere to changing standards and regulatory requirements [11][18][19]. Additionally, the systems are to be built in a way that is scalable, can be implemented in several institutions, work with increased amounts of imaging data and with different healthcare environments. Systems of continuous monitoring, audit, and quality assurance will be necessary to identify performance drift, minimize error, and maintain user confidence over time. With these sustainability practices, AI-driven radiology solutions would be able to generate reliable, equitable and long-term clinical value and withstand technological, clinical, and regulatory shifts.

5.3 Project Management and Financial Analysis

A successful development, deployment, and clinical integration of an AI-based image report generation system cannot be achieved without proper project management. The project processes will include the phases of data collection and annotation, model development and training, the implementation of the LLM, system testing, clinical validation and implementation. The resources are managed properly in order to make the best use of people, equipment, and software. Financial analysis is necessary to approximate the cost of hardware, software, human resource, clinical validation, deployment, and quality assurance. These costs are known and can help make a plan, budget, and take into account the possible return on investment in the form of efficiency gain and a better clinical outcome. The cost and financial analysis are estimated in Table 5.1.

Table 5.1: Estimated Cost and Financial Analysis

Components	Estimated Cost (BDT)
Internet and Cloud Subscription e.g. AWS, Google Colab Pro	4,000-5,000
Software Licenses (None, used Open Source)	0
Data Collection	0
Miscellaneous (Printing, Backup Storage)	800-1,000
Total Estimated Cost	**4,800-6,000 BDT**

5.4 Complex Engineering Problem

5.4.1 Complex Problem Solving

In this Table 5.2 provide a mapping with problem solving categories.

Table 5.2: Mapping with Complex Engineering Problem.

EP1 Dept of Knowledge	EP2 Range Of Conflicting Requirements	EP3 Depth of Analysis	EP4 Familiarity of Issues	EP5 Extent of Applicable Codes	EP6 Extent Of Stake-holder Involvement	EP7 Interdependence
✓		✓	✓			✓

Reasoning behind EP1: Your project relies on the forefront knowledge of deep learning, medical imaging, X-ray physics, explainable AI, and report generation with the use of LLMs. This requires interdisciplinary and specialized knowledge beyond simple engineering, and this criterion is completely fulfilled.

Justification of EP3: Radiology report generation does not have an apparent answer. It involves abstract reason, creativity, and novel thinking in preprocessing, model fine-tuning, Grad-CAM integration, and prompt engineering with LLMs. This is a complete fulfilment of this criterion.

Justification of EP4: Radiology report generation based on an AI is a relatively new concept and can hardly be found in typical engineering work. The task is not a standard one (such as classifying natural images) and includes expert, novel problems.

Justification of EP7: EP7 is an interdependent and multifaceted problem that consists of data collection, preprocessing, classification, explainability, report generation, evaluation, and deployment. The two sub-problems are interdependent and as such a high level challenge.

Mapping with Knowledge Profile

This Table 5.3 is designed to map the overall problem and EP1 to the Knowledge Profile.

Table 5.3: Mapping with knowledge Profile.

K1 Natural Science	K2 Mathematics	K3 Engineering Fundamentals	K4 Specialist Knowledge	K5 Engineering Design	K6 Engineering Practice	K7 Comprehension	K8 Research Literature
		✓	✓	✓	✓		✓

Rationale behind K3: The conceptual frameworks of deep learning (ResNet, Inception, VGG), convolutional computation and transfer learning (AdamW, OneCycleLR) have engineering underpinnings in computer science and AI engineering.

Rationale behind K4: The system combines medical imaging AI, explainable AI, and large language models (LLM) knowledge. It is the primary focus of biomedical engineering and AI research because it incorporates multimodal AI into clinical practice.

Rationale K5: Our model follows an engineering design process: prepare the dataset, adapt the model (altering the input channels, FC layers), integrate XAI, prompt the LLM, build the UI, and deploy it. The stages are structured design thinking to meet the needs of healthcare in the real world.

Rationale of K6: The project reflects engineering practice in the form of the use of hardware with GUI enables, PyTorch frameworks, XAI visualization, and integration of UI. It demonstrates the real use of AI technologies in a clinical setting (radiology workflows).

Rotionale of K8: We also read and compare works in the same field (Yi et al., Hong et al., Bouslimi et al., etc.)[1][3], interact with research standards (ReXrank, RadGraph F1, RadCliQ), and place your work in current scientific debates. This is evidence of extensive research reading.

5.4.2 Engineering Activities

About engineering activities, the project was mapped into different standard categories that define the complexity of engineering practice.

Mapping with Complex Engineering Activities

This table 5.2 is designed to map the EP1 to the Knowledge Profile.

Table 5.4: Mapping with Complex Engineering Activities.

EA1 Range of resources	EA2 Level of Interaction	EA3 Innovation	EA4 Consequences for society and environment	EA5 Familiarity
✓	✓		✓	✓

Rationale of EA1 success: Due to the abundance of resources, i.e. the skills of radiologists, data scientists, software engineers, and clinical staff, the technical and clinical meaning was to be received. Money played a role as well and it was spent on buying GPU hardware, storage, licensing and even cloud infrastructure to be used in large scale training and deployment. The special instruments such as the high-performance servers and X-ray machine in actual testing are also regarded as the technical richness of the project.

Additionally, it used large annotated datasets, clinical guidance and radiology standards as the primary information sources. Advanced technologies like convolutional neural networks, large language models, Grad-CAM, and deployment systems were combined, which proves the great variety of resources mobilized to be executed successfully.

Rationale of EA2 achievement: The project was characterized by large-scale interactions between various technical and clinical areas. Interpretability, model optimization, and clinical accuracy were in conflict with each other, this had to be carefully coordinated between the radiologists and the engineers. It became necessary to repeat cycles of validation and feedback loops to ensure that the requirements of the AI system were aligned with the real-world clinical workflow and to scale the requirements to the computational resources, ethical issues, and patient safety. On these cross-disciplinary interactions, the complexity of problem solving becomes evident to organize and execute.

Justification EA4 achievement: The system has consequence(s) that are of importance to society and the environment. It also increases the workload of radiologists by automating report generation, speeds up diagnosis, and democratizes the ability to access resource-limited settings of expert-level interpretation. Furthermore, and with the help of the effective exploitation of the hardware, the expandability of the architecture and the consideration of the new clinical and ethical requirements, the sustainability factors were also introduced into the project that also demonstrated the greater applicability of the project to the society.

Justification of EA5 achievement: The team utilized previous experience in Python programming, deep learning frameworks, medical imaging and data preprocessing to surmount the problem effectively. Principles-based strategies allowed the project to go beyond past experiences, covering clinical situations that had not been experienced before as well as a variety of imaging modalities. The use of continuous monitoring, auditing and refining led to high quality results, and showed high level of familiarity with basic engineering concepts, and practical problem solving.

5.5 Summary

This paper presents an AI-powered system to produce radiology reports automatically, combining deep learning-aided image classification tools, explainable artificial intelligence (XAI), and large language models (LLMs) in one system. To guarantee the high quality of inputs and the model generalizability, the dataset of COVID-19, pneumonia, and normal chest X-rays was collected, preprocessed, and augmented. Benchmarked architectures were surpassed by a fine-tuned InceptionV3 model that had an AUROC of 0.9988, an accuracy of 97.76 and a precision of 97.80, recall of 97.76 and an F1-score of 97.70. Grad-CAM and Grad-CAM++ with visual interpretability was more reliable and transparent as regions of diagnostic interest were identified. In order to produce the report a state-of-the-art LLM (Gemini-Flash-2.5) was applied to transform classification outcomes and XAI visualisations into structured, radiology-like, outputs. AI-assisted checks and blind testing of expert radiologists were assessed and compared in terms of factual, structural, interpretability, and clinical usefulness. Compared with the present work, it can be seen that this system has already been referred to as competitive as far as accuracy and explainability and integration of workflow are concerned. The proposed framework provides a clinically significant end-to-end pipeline--image upload and visualization, automated diagnosis and report generation, and user-friendly interface. It has high diagnostic accuracy, is explainable, minimizes the workload of radiologists, and encourages standard reporting. The system is a scalable and viable clinical adoption solution and can be expanded to multi-modal imaging, multi-lingual reporting and mass implementation across a variety of healthcare set-ups.

Chapter 6

Conclusion

6.1 Summary

The generation of radiology reports by AI has matured quickly and is using the multimodality of deep learning, massive language models, and explainable AI to standardize and improve the reporting system. Recent research has shown that domain-specific generative AI models can deliver high diagnostic accuracy, updated workflow, and report quality that is frequently rated as clinically acceptable (or even better) by radiologists than either traditional or general-purpose models [2][4][5][7][15][18]. Moreover, it is also supported by the explainability tools and quality control mechanisms implemented to build trust and reliability in the clinical environment [1][13][19]. However, factual consistency, minimization of hallucinations, and the ability to generalize across institutions and imaging modalities are still problematic [1][17][18].

6.2 Limitation

As a result of this progress, the current AI systems have a few limitations:

i. False Information or other artificial findings: Even an AI-composed report may still contain Factual Inconsistency and Hallucinations and should have good fact-checking and audit mechanisms [1][13][19].

ii. Measurement: Standard NLP scales do not measure clinical relevance well, and a set of new, clinically aligned measures has been created [6][10][14][20].

iii. Generalizability: The main limitation of the models was that most of them could not be generalized in a clinical setting and population because of their extremely large sizes [18].

iv. Centricity to the patient: Most systems are clinicalist-focused, and minimal focus is directed towards the generation of patient-oriented reports or multilingual reports [4][12].

v. The Consequential Regulations: privacy, bias and transparency cannot be applied to clinical deployment.

6.3 Future Work

Future research should answer:

i. Increasing Explainability: Improved integration of XAI procedures to deliver user-understandable outputs to clinicians and patients [1][13][19].

ii. Good Fact-Checking: The hallucination-finding and truth-finding algorithms are trained [13][19].

iii. The Model of Generalization and Adaptation Training and Testing Multi-institutional Multi-model Data to enable resilience and adaptability [17][18].

iv. Patient Centered Reporting Multilinguistic, readable and easily understandable patient reports [4][12].

v. Clinical validation and control: To undertake wide scale, prospective, clinical research and personal involvement with the regulatory bodies to enable effective, safe and ethical practice. deployment.

References

[1] Z. Yi, T. Xiao, and M. V. Albert, "A Multimodal Multi-Agent Framework for Radiology Report Generation," May 14, 2025.

[2] E. K. Hong et al., "Diagnostic Accuracy and Clinical Value of a Domain-specific Multimodal Generative AI Model for Chest Radiograph Report Generation," Radiology, Mar. 1, 2025.

[3] R. Bouslimi, H. Trabelsi, W. Karaa, and H. Hedhli, "AI-Driven Radiology Report Generation for Traumatic Brain Injuries," Journal of imaging informatics in medicine, Jan. 30, 2025.

[4] J. Park, K. Oh, K. Han, and Y. H. Lee, "Patient-centered radiology reports with generative artificial intelligence: adding value to radiology reporting," Scientific Reports, Jun. 8, 2024.

[5] E. K. Hong et al., "Value of Using a Generative AI Model in Chest Radiography Reporting: A Reader Study," Radiology, Mar. 1, 2025.

[6] Z. Babar, T. V. Laarhoven, F. M. Zanzotto, and E. Marchiori, "Evaluating diagnostic content of AI-generated radiology reports of chest X-rays," Artificial intelligence in medicine, Apr. 15, 2021.

[7] R. Tanno et al., "Collaboration between clinicians and vision–language models in radiology report generation," Nature Medicine, Nov. 7, 2024.

[8] X. Zhang et al., "ReXrank: A Public Leaderboard for AI-Powered Radiology Report Generation," ArXiv, Nov. 22, 2024.

[9] W. Feng, A. Yazdani, A. Bornet, A. Platon, and D. Teodoro, "Synthesizing evidence regarding artificial intelligence generated radiological reports based on medical images: a scoping review protocol," Mar. 17, 2025.

[10] F. Yu et al., "Evaluating progress in automatic chest X-ray radiology report generation," Patterns, Aug. 31, 2022.

[11] G. Mehdiratta et al., "Automated Integration of AI Results into Radiology Reports Using Common Data Elements," Journal of imaging informatics in medicine, Jan. 27, 2025.

[12] A. Gupta MD Frcr et al., "Transcribing multilingual radiologist-patient dialogue into mammography reports using AI: A step towards patient-centric radiology," Mar. 26, 2025.

[13] R. Mahmood, G. Wang, M. Kalra, and P. Yan, "Fact-Checking of AI-Generated Reports," ArXiv, Jul. 27, 2023.

[14]Q. Zhu et al., "Leveraging Professional Radiologists' Expertise to Enhance LLMs' Evaluation for AI-generated Radiology Reports," 2024 IEEE 12th International Conference on Healthcare Informatics (ICHI), Jun. 1, 2024.

[15]J. N. Acosta et al., "The Impact of AI Assistance on Radiology Reporting: A Pilot Study Using Simulated AI Draft Reports," ArXiv, Dec. 16, 2024.

[16]Y. Pan et al., "Sequence-Integrated Radiology Report Generation Leveraging Positional Encoding for Text and Images," 2024 IEEE Cyber Science and Technology Congress (CyberSciTech), Nov. 5, 2024.

[17]X. Zhang, J. N. Acosta, H.-Y. Zhou, and P. Rajpurkar, "Uncovering Knowledge Gaps in Radiology Report Generation Models through Knowledge Graphs," ArXiv, Aug. 26, 2024.

[18]T. Zhong et al., "ChatRadio-Valuer: A Chat Large Language Model for Generalizable Radiology Report Generation Based on Multi-institution and Multi-system Data," ArXiv, Oct. 8, 2023.

[19]H. Warr, Y. Ibrahim, D. R. McGowan, and K. Kamnitsas, "Quality Control for Radiology Report Generation Models via Auxiliary Auditing Components," Jul. 31, 2024.

[20]Q. Zhu et al., "Leveraging Professional Radiologists' Expertise to Enhance LLMs' Evaluation for Radiology Reports," ArXiv, Jan. 29, 2024.

13% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text

Match Groups

108 Not Cited or Quoted 13%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

8% Internet sources
4% Publications
10% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **108** Not Cited or Quoted 13%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

8% Internet sources
4% Publications
10% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Submitted works	
	Daffodil International University on 2024-12-28	5%
2	Internet	
	dspace.daffodilvarsity.edu.bd:8080	2%
3	Internet	
	arxiv.org	<1%
4	Publication	
	Sushil Kamboj, Pardeep Singh Tiwana. "Innovations in Computing", CRC Press, 2025	<1%
5	Publication	
	Pushpa Choudhary, Sambit Satpathy, Arvind Dagur, Dhirendra Kumar Shukla. "Re...	<1%
6	Submitted works	
	United International University on 2023-06-18	<1%
7	Internet	
	huggingface.co	<1%
8	Submitted works	
	Faculty of Science, Chulalongkorn University on 2025-07-17	<1%
9	Internet	
	v1.overleaf.com	<1%
10	Submitted works	
	United International University on 2024-01-29	<1%

Ridam Roy

AI-Powered Radiology Report Generation Using Multi-Model

Turnitin

Document Details

Submission ID

trn:oid:::13381:111110498

Submission Date

Sep 6, 2025, 11:05 PM GMT+5

Download Date

Sep 6, 2025, 11:07 PM GMT+5

File Name

Updated FYDP Tamplate for [Summer 2025] (1) 1.pdf

File Size

1.3 MB

53 Pages

14,344 Words

89,007 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.