THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Automatic Radiology Report Generator

Submitted by :
Anubha Kapur (102216090)
Madhav Galhotra(102216082)

Department of Computer Science

Lab Instructor: Ms.Priya Raina

A report submitted in partial fulfilment of the requirements for the degree of Bachelor of Engineering (B.E.) in *Computer Science at Thapar Institute of Engineering and Technology (TIET)*

November 21, 2025

Abstract

Automated radiology report generation is an important challenge in medical imaging, requiring models that can both understand visual patterns and communicate clinical findings in clear, structured language. While existing deep learning systems can classify abnormalities or retrieve similar cases, they often fall short in producing coherent, contextually accurate, and clinically meaningful full-length reports. Vision-only models struggle to translate pixel information into higher-level reasoning, and language models alone cannot generate trustworthy descriptions without medically relevant conditioning.

This project presents a **hybrid, retrieval-augmented generative AI framework** that combines visual encoding, clinical similarity search, and language modeling to generate radiologist-style chest X-ray reports. The system performs:

- **Image feature extraction** using a Vision Transformer (vit_base_patch16_224)
- **Nearest-neighbor retrieval** of similar radiographs via cosine similarity
- **Prompt generation** using MeSH terms and findings from retrieved cases
- **Report synthesis** using a fine-tuned DistilGPT-2 model
- **Performance evaluation** through BLEU, ROUGE, BERTScore, perplexity, and embedding visualizations

The workflow is trained and validated on the Indiana University Chest X-ray dataset, enabling the model to learn realistic clinical writing patterns from paired images and expert-authored reports. Unlike conventional classification pipelines, this approach treats report creation as an end-to-end **retrieval-augmented narrative generation** task, supporting more clinically aligned and context-aware outputs.

The system integrates:

- A **ViT-based visual encoder** producing 768-dimensional embeddings
- A **retrieval module** for identifying top-K=1 similar case
- A **prompt engineering pipeline** based on MeSH-derived cues
- A **fine-tuned DistilGPT-2 generator** for producing full diagnostic reports
- A **complete evaluation suite** and a user-friendly inference interface

The resulting application accepts a single chest X-ray and generates a detailed, coherent, and clinically relevant report grounded in both image similarity and learned medical language. By unifying retrieval mechanisms with transformer-based generative modeling, the project offers a scalable and interpretable solution that moves automated radiology reporting closer to practical clinical use.

Contents

Introduction

Radiology reports are essential for documenting observations from medical images and communicating them to clinicians. However, generating these reports manually is time-consuming and requires detailed visual analysis and precise medical language. With the availability of large public datasets that contain paired chest X-rays and clinical reports, it has become possible to explore automated approaches to assist in this process.

Recent progress in computer vision and natural language processing has enabled models to extract features from images and generate text, but most existing systems still focus on limited tasks such as classifying abnormalities or producing short captions. They typically do not generate complete, structured radiology reports similar to the ones written by experts. This gap mainly exists because medical report generation requires both accurate image understanding and familiarity with domain-specific terminology.

The repositories used in this project implement a practical solution built around a **hybrid pipeline**. Instead of relying only on a vision model or only on a language model, the system combines multiple components that work together. A Vision Transformer is used to extract deep visual features from chest X-rays. These features are compared using cosine similarity to retrieve the most similar images from the dataset. The clinical findings and MeSH terms associated with these retrieved cases help create a meaningful prompt. A GPT-2 model, fine-tuned on the Indiana University Chest X-ray reports, then uses this prompt to generate a complete radiology report.

This approach allows the system to remain grounded in real examples from the dataset while still using a language model to produce coherent, structured text. The repositories also include modules for training, evaluation using BLEU, ROUGE, BERTScore, and perplexity, and basic visualization such as t-SNE plots of embeddings. Together, these components form a straightforward, retrieval-guided radiology report generation system.

Problem Statement

Radiology reports are essential for interpreting chest X-rays, but generating them requires expert knowledge and considerable time. Existing machine learning models mainly perform classification or abnormality detection, producing only labels rather than full diagnostic reports. These outputs lack the descriptive detail and structured clinical language found in real radiology reports.

Vision-only models can analyze image features but cannot translate them into meaningful medical text. Language models like GPT-2 can generate fluent reports, but without direct grounding in the X-ray image, their output may be inaccurate or irrelevant. This disconnect creates a major limitation: **there is currently no unified model that both understands the visual content of a chest X-ray and generates a complete, clinically relevant report.**

The problem addressed in this project is the gap between image interpretation and medical report generation. There is a clear need for a system that links visual information from the X-ray with domain-specific textual descriptions so that the resulting reports are coherent, medically meaningful, and grounded in actual image content.

Objectives

The primary objective of this project is to develop an automated system capable of generating clinically coherent and image-grounded radiology reports for chest X-rays. To achieve this, the project focuses on the following specific goals:

- **Generate complete, human-like radiology reports** that mimic the structure and writing style of expert radiologists.

- **Extract deep visual features using a Vision Transformer** to capture detailed image representations from chest X-rays.

- **Retrieve the most similar radiographs** from the dataset using cosine similarity to provide relevant clinical context.

- **Construct meaningful prompts** using MeSH terms and findings from retrieved cases to guide report generation.

- **Fine-tune a DistilGPT-2 language model** to produce structured radiology reports based on the prompt and retrieved information.

- **Evaluate the system’s performance** using BLEU, ROUGE, BERTScore, perplexity, and embedding visualizations.

- **Develop a simple inference interface** that allows users to input a chest X-ray and receive an automatically generated report.

Dataset Description

This project uses the **Indiana University Chest X-ray (Open-I) dataset**, a publicly available collection of chest radiographs paired with their corresponding diagnostic reports. This dataset is widely used for research on automated medical report generation because it provides both the **visual data** required for feature extraction and the **textual data** needed for language model training.

Dataset Size

The dataset consists of:

- **~7,470 chest X-ray images**
- **~3,955 expert-written radiology reports**
- **Metadata** describing:
    - Projection types (PA, lateral, etc.)
    - MeSH terms
    - Findings and impressions
    - Unique image identifiers and mappings

These metadata files help link each image to its corresponding report and provide clinically meaningful terms used during prompt construction.

Data Format

- **Image Folder**
  Contains individual **PNG** chest X-ray images, each stored with a unique filename representing its image ID.
- **CSV Metadata Files**
  Two main CSV files are used in the repositories (**indiana_reports.csv, indiana_projections.csv**)

These files include:

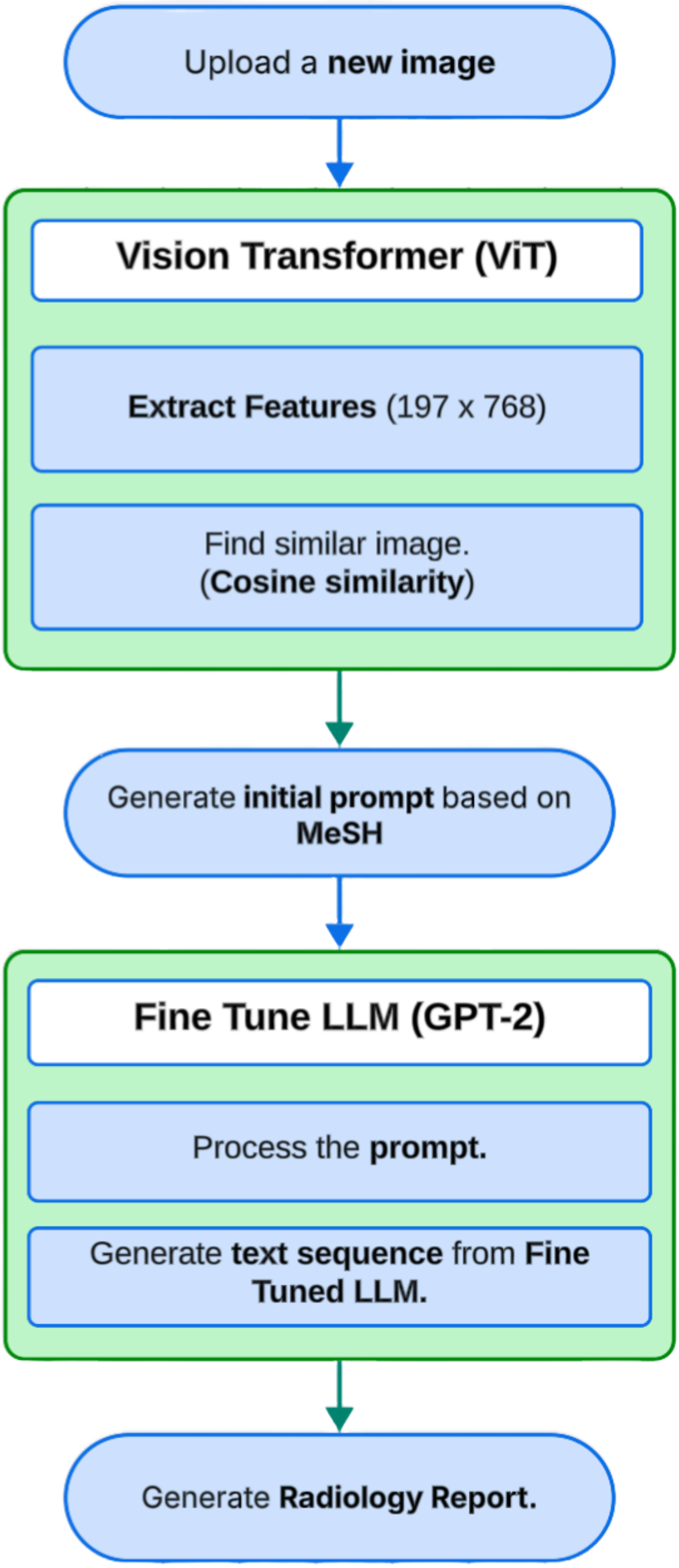
- Image IDs
- MeSH terms
- Full report text
- Findings and impression sections
- Projection information

Use in This Project

Within the project pipeline:

- The **ViT model** uses the chest X-ray images to extract **768-dimensional visual embeddings**.
- The **DistilGPT-2 model** uses the textual reports during fine-tuning to learn clinical language patterns.
- **MeSH terms** from the metadata are used to build meaningful prompts for guiding report generation.

System Architecture Diagram


Upload a **new image**
Vision Transformer (ViT)
Extract Features (197 x 768)
Find similar image.
(**Cosine similarity**)
Generate **initial prompt** based on **MeSH**
Fine Tune LLM (GPT-2)
Process the **prompt.**
Generate **text sequence** from **Fine Tuned LLM.**
Generate **Radiology Report.**

Methodology

Image Feature Extraction – ViT (vit_base_patch16_224)

The system utilises a Vision Transformer as it provides strong global feature representations for medical images. ViT processes an image as a sequence of patches, enabling it to capture long-range dependencies and subtle radiological patterns more effectively than traditional CNNs.

Input Preprocessing Steps:

Before extracting features, the images undergo the following preprocessing steps (implemented in extract_features.py):

- Resize the chest X-ray image to the ViT input size
- Normalize pixel intensities using ImageNet mean and standard deviation
- Convert the image into patches (16×16) internally via the ViT architecture
- Prepare the image tensor for inference (batch dimension, device placement)

768-Dimensional Embeddings

The ViT model produces a 768-dimensional embedding for each image, taken from the model's final pooling layer or CLS token. These embeddings are saved and later used for similarity-based retrieval.

Similarity Search

Cosine Similarity

After extracting embeddings for all images in the dataset, the system computes **cosine similarity** between the input X-ray and every precomputed image embedding. This is implemented in the retrieval scripts under the `src` directory.

Top-K=1 Image Retrieval

The top most similar image is selected based on similarity scores. This retrieved image serve two purposes:

- It provides **clinical grounding** from known, labeled examples
- It helps construct meaningful prompts for GPT-2

Using Neighbor MeSH/Findings for Grounding

For each retrieved image, the system extracts:

- **MeSH terms**
- **Findings / Impression text**

These pieces of text are used as context to guide the generation process.

Prompt Engineering

Prompt construction is performed during report generation using metadata and retrieved findings. The repositories build prompts using the following structure:

What Text is Used

From the retrieved radiographs, the system uses:

- MeSH terms
- Key findings and impression sections
- Short, clinically relevant text fragments

How Prompts Guide Distil GPT-2

The prompt serves as the **conditioning text** for GPT-2. It acts as a starting point that includes important medical terminology and relevant findings. GPT-2 continues generating from this prompt, producing a full diagnostic report grounded in:

- retrieved image similarity
- clinically relevant terminology
- dataset-specific writing style

This improves relevance and reduces hallucination.

DistilGPT-2 Fine-Tuning

DistilGPT-2 fine-tuning is implemented in train_gpt2.py.

Training Setup

- The original Indiana University radiology reports serve as training text.
- GPT-2 tokenizer is applied to all reports.
- Data is split into training and validation sets.
- A standard language model training loop is used via HuggingFace's Trainer API.

Loss Function

The model uses the **causal language modeling (CLM) loss**, i.e., cross-entropy loss over next-token prediction.

Hyperparameters

- Subset Size: 1000
- Learning rate = 5e-5
- Batch size: 2
- Number of epochs: 20
- Optimizer: AdamW

Checkpointing and logging are handled during training to track model performance.

Report Generation Pipeline

The final radiology report is generated through a multi-step process:

Step-by-Step Flow

1. **The user inputs a chest X-ray image** into the system.
2. **ViT extracts a 768-dimensional embedding** from the input image.
3. **Cosine similarity retrieval** identifies the most similar images from the dataset.
4. **MeSH terms and findings** from those similar cases are extracted.
5. A **prompt** is constructed using these retrieved textual elements.
6. The **fine-tuned GPT-2 model** takes the prompt and generates a full diagnostic report.
7. The generated report is **post-processed** (trimming, cleaning) and displayed to the user.

This pipeline is implemented across generate_report.py and supporting utilities in the repository.

Evaluation Metrics

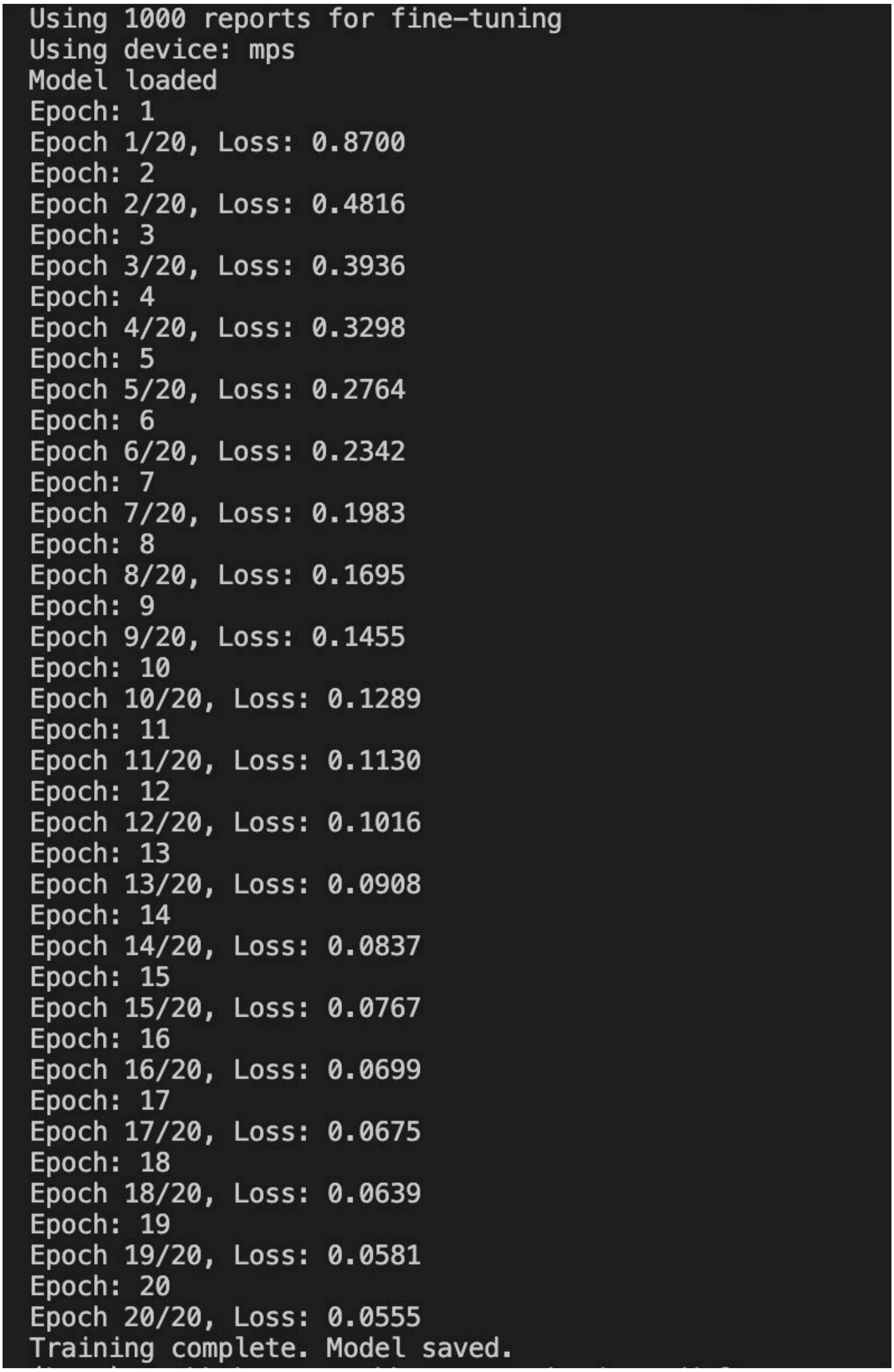
The repositories include an evaluation script (evaluate.py) that applies a variety of NLP metrics to compare generated reports with ground-truth reports.

- **BLEU Score** – measures n-gram overlap between generated and reference reports
- **ROUGE Score** – evaluates recall-based overlap, especially useful for longer text
- **BERTScore** – measures semantic similarity using contextual embeddings
- **Perplexity** – evaluates how well the language model predicts text
- **t-SNE Visualization** – visualizes high-dimensional image embeddings to show clustering of similar cases

Together, these metrics provide both quantitative and qualitative insights into the model’s performance.

Result & Analysis

Fine-Tuning Performance (Loss)



```
Using 1000 reports for fine-tuning
Using device: mps
Model loaded
Epoch: 1
Epoch 1/20, Loss: 0.8700
Epoch: 2
Epoch 2/20, Loss: 0.4816
Epoch: 3
Epoch 3/20, Loss: 0.3936
Epoch: 4
Epoch 4/20, Loss: 0.3298
Epoch: 5
Epoch 5/20, Loss: 0.2764
Epoch: 6
Epoch 6/20, Loss: 0.2342
Epoch: 7
Epoch 7/20, Loss: 0.1983
Epoch: 8
Epoch 8/20, Loss: 0.1695
Epoch: 9
Epoch 9/20, Loss: 0.1455
Epoch: 10
Epoch 10/20, Loss: 0.1289
Epoch: 11
Epoch 11/20, Loss: 0.1130
Epoch: 12
Epoch 12/20, Loss: 0.1016
Epoch: 13
Epoch 13/20, Loss: 0.0908
Epoch: 14
Epoch 14/20, Loss: 0.0837
Epoch: 15
Epoch 15/20, Loss: 0.0767
Epoch: 16
Epoch 16/20, Loss: 0.0699
Epoch: 17
Epoch 17/20, Loss: 0.0675
Epoch: 18
Epoch 18/20, Loss: 0.0639
Epoch: 19
Epoch 19/20, Loss: 0.0581
Epoch: 20
Epoch 20/20, Loss: 0.0555
Training complete. Model saved.
```

The image shows the training loss for DistilGPT-2 fine-tuned on a subset of 1000 reports for 20 epochs.

During fine-tuning of the DistilGPT-2 model on a subset of **1000 radiology reports** for **20 epochs**, a consistent reduction in training loss was observed.
The loss decreased from **0.8700 → 0.0555**, indicating stable convergence and effective learning of report structure and terminology.

Evaluation Metrics

BLEU Score: 0.2698

ROUGE Scores:

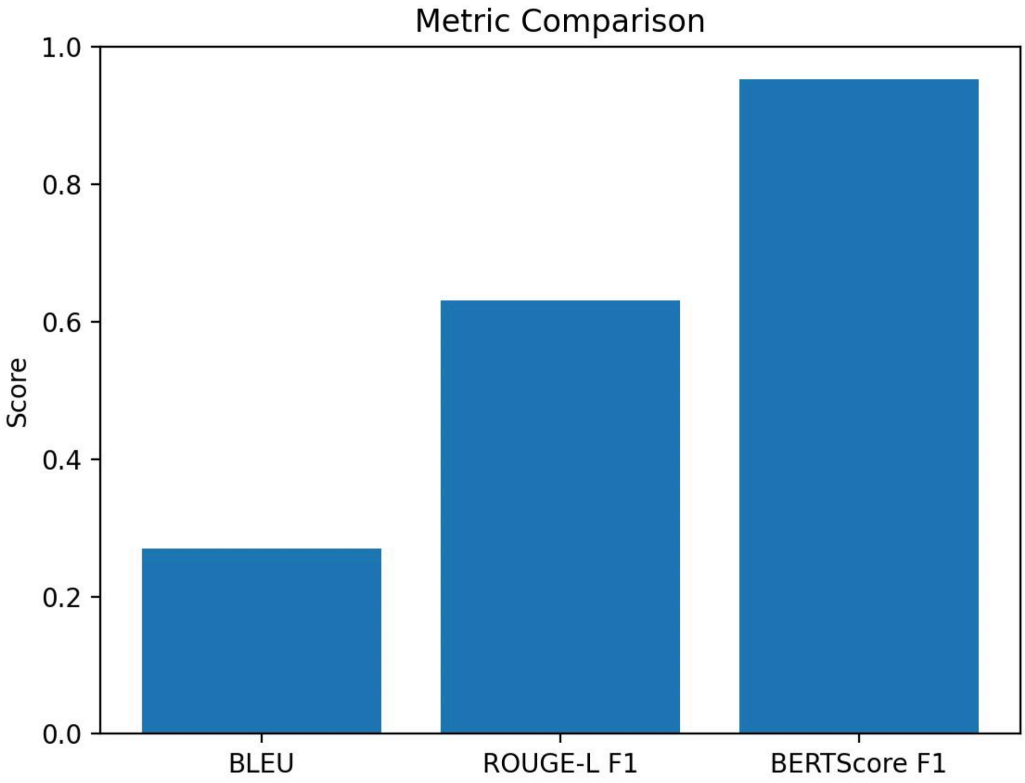
Metric	**Score**
ROUGE-1 F1	0.7368
ROUGE-2 F1	0.3529
ROUGE-L F1	0.6315

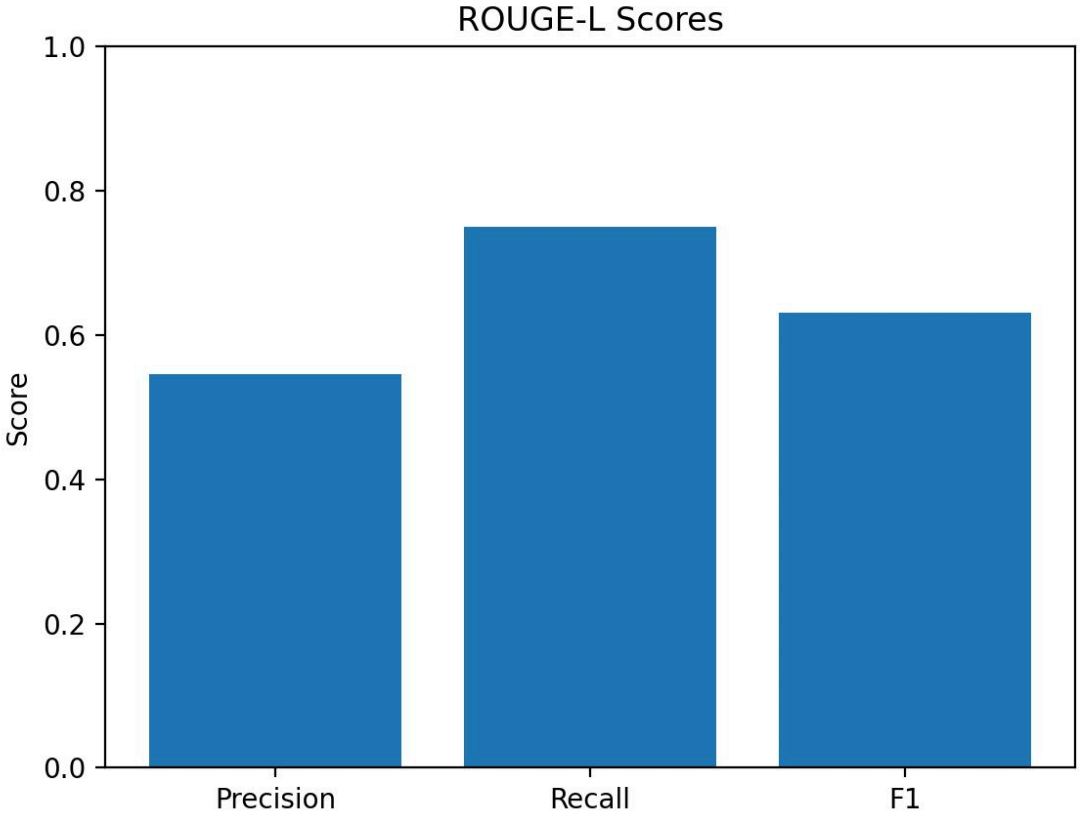
BERTScore:

Metric	**Score**
Precision	0.943
Recall	0.963
F1	0.953

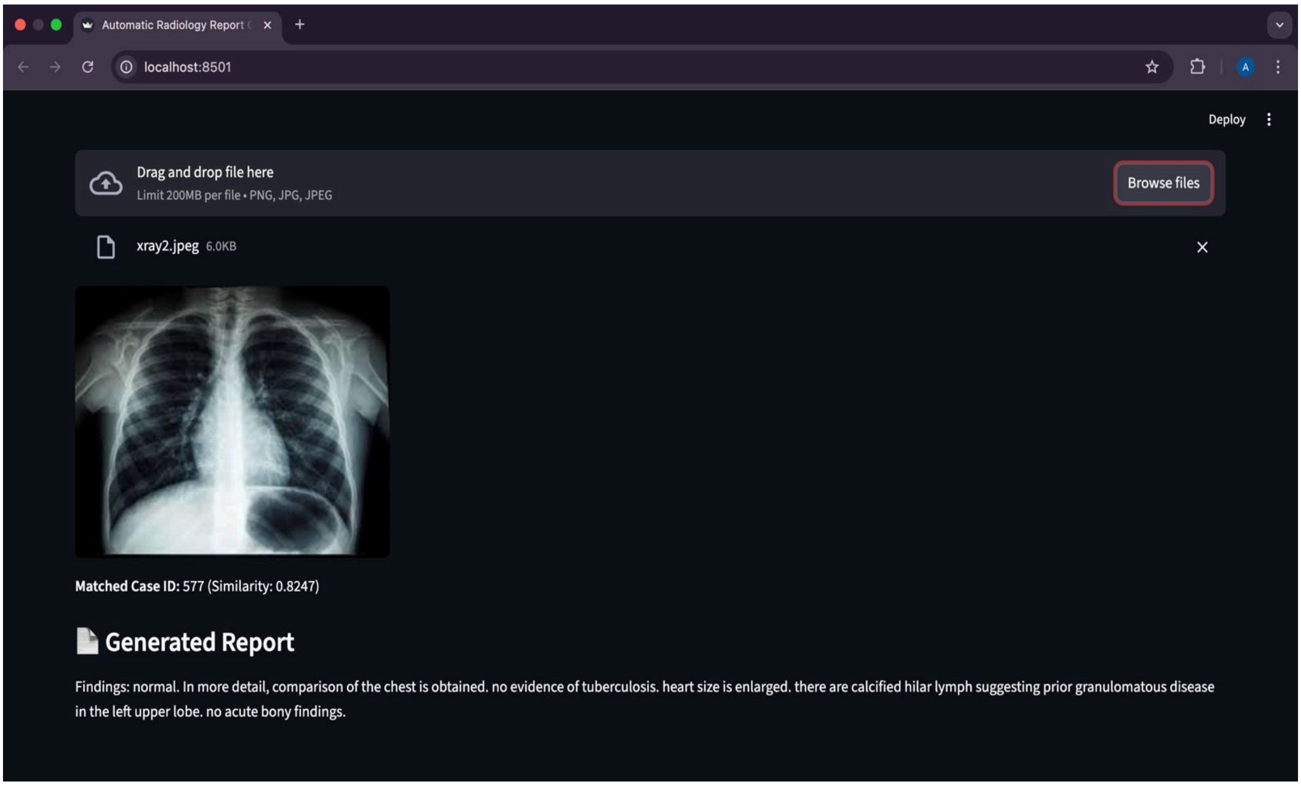
The evaluation metrics indicate that the generated reports have strong lexical overlap (BLEU/ROUGE) and high semantic similarity (BERTScore) with ground-truth reports from the dataset.

A BERTScore F1 of **0.953** suggests the generated reports capture clinically important meaning even when textual phrasing differs.


Metric Comparison
Score
1.0
0.8
0.6
0.4
0.2
0.0
BLEU
ROUGE-L F1
BERTScore F1


ROUGE-L Scores
Score
1.0
0.8
0.6
0.4
0.2
0.0
Precision
Recall
F1

Application Screenshot



Matched Case ID: 577 (Similarity: 0.8247)

Generated Report

Findings: normal. In more detail, comparison of the chest is obtained. no evidence of tuberculosis. heart size is enlarged. there are calcified hilar lymph suggesting prior granulomatous disease in the left upper lobe. no acute bony findings.

Future Scope

- Fine-tune the model on larger and more diverse radiology datasets to further improve generalization and clinical accuracy.
- Extend multimodality beyond X-ray + text, incorporating ECG, CT/MRI for richer patient-level understanding.
- Implement lightweight continual-learning pipelines so the system can adapt to new hospital data without full retraining.

Source Code Repository

https://github.com/anubhakapur/Automatic-Radiology-Report-Generator