



**THE GEORGE  
WASHINGTON  
UNIVERSITY**  
WASHINGTON, DC

# **Medical Report Generation using Chest X Ray Images**

## **Final Project Report (Team 6)**

**for the course**

**DATS 6312 ‘Natural Language  
Processing’**

**Fall 2024**

**Abishek Chiffon Muttu Raja**

**Anand Raj**

**Saniya Shinde**

**Shanun Randev**

# INTRODUCTION

In modern healthcare, radiology plays a crucial role in diagnosing and managing a wide array of medical conditions. Chest X-rays, in particular, are among the most frequently used diagnostic tools for detecting abnormalities such as Pneumonia, Hernia and Cardiomegaly. The motivation behind this project is to create a tool that supports radiologists in their demanding roles by automating the generation of preliminary radiology reports. By leveraging advanced computer vision and large language models, the system aims to draft accurate and detailed reports from chest X-ray images, serving as a starting point for radiologists to review and refine. This collaboration between AI and radiologists enhances productivity, reduces delays, and minimizes the risk of errors stemming from workload fatigue.

The report begins with an Introduction that outlines the motivation and scope of the project, followed by a detailed description of the dataset, highlighting its structure and features. The Methodology section explains the technical approach, including preprocessing steps, models used, and training processes. The Results section presents the performance outcomes, supported by relevant metrics and analysis. The Conclusions summarize the key findings from the project. The Further Improvements section explores potential advancements, such as integrating additional datasets to improve accuracy and usability. Finally, the References provide a comprehensive list of all sources cited throughout the report.

# DESCRIPTION OF THE DATASET

The MIMIC-CXR is a publicly available chest X-ray dataset for chest radiography research. It comprises 15,000 chest X-ray images in dicom format and their associated radiology reports in xml format. The dataset has the following key features:

- Image File Path: Location or link to the corresponding chest X-ray image.
- Findings: A textual description of abnormalities or observations made by the radiologist.
- Impression: A concise summary of the radiologist's primary conclusions.

The dataset has 14 labels corresponding to common chest X-ray pathologies. The pathology labels include Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, Pleural Effusion, Pleural Other, Pneumonia, Pneumothorax, Support Devices, and No Finding.

# METHODOLOGY

The project was structured in several key stages, each critical to the successful implementation of the model:

- 1. Data Collection and Preprocessing:** The initial step involved converting DICOM images to PNG format, extracting relevant information, preparing datasets for further analysis, and performing text and image transformations.

- 2. Label Extraction:** The next stage utilized ChexBERT to generate multi-label classifications from the associated radiology reports, enabling the identification of relevant medical conditions in each image.
- 3. Model Architecture Design:** During this stage, different combinations of image encoders, alignment models, and language models were experimented with to determine the most effective architecture for processing medical image data and generating accurate reports.
- 4. Training and Evaluation:** Once the model architecture was defined, the models were trained using the preprocessed data. Performance was closely monitored and evaluated using key metrics such as ROUGE-L to assess the quality of the model's output.
- 5. Optimization:** To enhance model efficiency, Parameter-Efficient Fine-Tuning (PEFT) techniques were implemented.
- 6. Report Generation:** The final step involved generating medical reports from the trained model and evaluating the quality of these reports to ensure they met clinical standards and provided accurate diagnoses.

Below is an in depth explanation of each of these stages:

## **1. Data Collection and Preprocessing**

### **a. Data Extraction**

#### **i. DICOM to PNG Conversion**

Chest X-ray images were originally stored in DICOM format, a standard for handling, storing, and transmitting medical imaging information. To optimize storage and processing efficiency, these images were converted to PNG format using a custom script. The conversion significantly reduced file sizes without compromising image quality, facilitating faster data loading and manipulation during model training.

#### **ii. Script to Create CSV with Image\_ID, Image\_Path, Findings, and Impressions**

A dedicated script was developed to streamline data organization. Following fields were extracted to create the csv.

image\_ID: Extracted unique identifiers corresponding to each image.

image\_path: Consolidated file paths to each PNG image.

findings and impressions: Parsed the XML reports to extract relevant clinical findings and impressions.

This structured format facilitated efficient data handling and model training.

### **b. Data Pre-processing**

The following steps were implemented in data pre-processing.

- i.** Findings column was cleaned by expanding abbreviations (e.g. lat was expanded as lateral), removing special characters, and fixing spacing around punctuation.
- ii.** The dataset was then filtered to remove invalid or missing entries, and the findings were mapped to a list of labels indicating the presence of specific diseases.
- iii.** Data augmentation techniques, such as resizing to (224,224), random rotations, flips, noise addition, and normalization were applied to images for training.

### **c. Dataset Split**

The `get_dataloaders` function was designed to create and return PyTorch `DataLoader` objects for training and validation datasets. It took several parameters to customize the data loading process: `csv_with_image_paths` (the path to a CSV file containing image paths) and `csv_with_labels` (the path to a CSV file containing corresponding labels) were the main inputs. The `batch_size` parameter defined the number of samples per batch, with a default value of 8. The `train_split` parameter determined the fraction of the data to be used for training, with a default value of 0.85 (85% for training and 15% for validation). The `num_workers` parameter controlled the number of subprocesses to use for data loading, set to 4 by default, which helped speed up the process. The `seed` parameter ensured reproducibility of the data split by setting a fixed random seed (default was 42). Lastly, the `collate_fn` parameter allowed the user to specify a custom collate function (default was `custom_collate_fn`), which was responsible for merging samples in a batch, particularly useful for handling variable-length inputs like text.

## **2. Extracting Labels using Chexbert**

ChexBert is a transformer-based model fine-tuned for medical text classification, specifically designed to extract multi-label classifications from chest X-ray radiology reports. It is built on the BERT (Bidirectional Encoder Representations from Transformers) architecture and offers superior performance in identifying the presence or absence of 14 thoracic conditions, providing a reliable method for generating ground truth labels from textual radiology reports.

### **i. Text Processing with ChexBERT**

The textual content of chest X-ray reports, particularly the "Findings" and "Impressions" sections, was extracted for ChexBERT. The extracted text was tokenized and formatted according to the input requirements of ChexBERT. The tokenized input was passed through the ChexBERT model, which produced high-dimensional contextual embeddings.

### **ii. Label Extraction**

After obtaining the contextual embeddings from ChexBERT, the next crucial step was label extraction. This process involved applying a classification layer that used the embeddings to predict the presence or absence of various clinical conditions. The predictions were made in the form of probabilities for each condition. To convert these probabilities into binary labels (0 or 1), a threshold of 0.5 was applied. The probabilities above this value indicate the presence of a condition (labelled as 1), and those below indicate its absence (labelled as 0).

### **iii. Dataset Preparation**

The final stage of the workflow involved the preparation of a dataset suitable for multi-label classification tasks. The binary labels obtained from the previous step were integrated into a CSV file, enriching the dataset with detailed annotations regarding the clinical conditions identified in the X-ray reports.

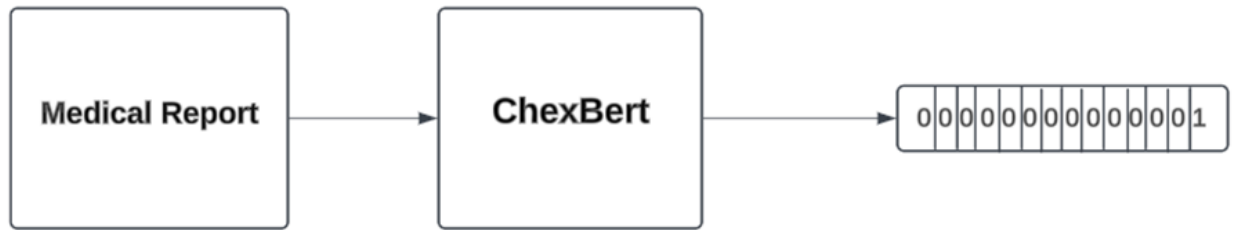


Figure 1: Output from ChexBert Model

### 3. ChexNet for Structural Findings Extraction

After experimenting with BioViLT, we found through several research papers that the model could be improved further by adding another layer of multi-label classification as input to the main ViLT models. Therefore, we decided to fine-tune ChexNet, a deep learning model based on the DenseNet-121 architecture, utilized for multi-label classification of chest X-ray images. It has been pretrained on chest X-rays, and its primary function was to identify structural abnormalities such as cardiomegaly, pneumonia, and atelectasis. The integration of ChexNet provided structured findings that enhanced the contextual understanding of the images, thereby improving the quality of generated medical reports.

The challenge was that the dataset was heavily imbalanced, so we had to use methods like a custom WeightedBCELoss class that applies different weights to positive and negative samples, uses WeightedRandomSampler to oversample minority classes, and implements inverse frequency weighting to give more importance to minority classes. We achieved an F1-micro score of 0.70.

#### Architecture:

- **Base Model:** DenseNet-121

#### Training Configuration:

- **Pretraining:** The model leverages pre-trained DenseNet-121 weights trained on ImageNet. Specific weights for ChexNet were also fine-tuned for chest X-ray analysis.
- **Layer Freezing:**
  - Initial layers (conv0, norm0, denseblock1, transition1, etc.) were frozen to retain the pre-trained features.
  - Training focused on the last two dense blocks and the classifier head to adapt the model to the chest X-ray dataset.
- **Custom Classifier:**
  - Input: 1024 features from DenseNet-121
  - Hidden Layer: 512 units with ReLU activation
  - Dropout: 0.3 for regularization
  - Output Layer: 14 sigmoid-activated nodes for multi-label classification.
- **Training Procedure**
  - Loss Function:** Binary Cross-Entropy (BCE) Loss for multi-label classification
  - Optimizer:** Adam with differential learning rates.
  - Scheduler:** ReduceLROnPlateau to adapt learning rates based on validation AUC.
  - Metrics:** Area Under the Curve (AUC) for each class; average AUC across classes for model evaluation.

## 4. Model Architecture

To identify the most effective model architecture for generating medical reports from chest X-ray images, four distinct combinations were experimented:

### i) **BioVilt + Alignment + BioGPT**

#### (1) **Model Used**

- (a) **BioViLT** : The BioViLT architecture serves as an advanced image encoder, specifically designed to extract high-level features from chest X-ray images for medical applications. We integrated a customized ResNet backbone (ResNetHIML) to enhance feature representation. The ResNetHIML backbone, based on ResNet-18 or ResNet-50, supports the extraction of intermediate layers for patch-wise embeddings and leverages pre-trained weights for transfer learning. An MLP Projector maps these features into a joint feature space of size 128 with batch normalization for stability.
- (b) **Alignment Model**: Bridges image embeddings with textual representations to ensure semantic compatibility.
- (c) **BioGPT**: BioGPT, developed by Microsoft Research, is a powerful natural language processing model designed specifically for biomedical applications. It is based on the GPT-2 architecture and features a vocabulary size of 42,384 tokens, with the base model comprising approximately 347 million parameters. This model is pre-trained on an extensive corpus of biomedical literature, enabling it to excel in tasks like summarizing complex biomedical content, assisting with medical writing, and extracting critical information from biomedical texts. Its specialized training makes BioGPT a standout performer in biomedical NLP tasks, surpassing other models in the field in terms of accuracy and applicability.

#### (2) **Configuration:**

##### 1) **BioVilt:**

- a) **Architecture**: ResNet-50 backbone.
- b) **Output Dimensions**: Global embedding of 512 dimensions

##### 2) **Alignment Module:**

- **Text Encoder**: Microsoft BioGPT.
- **Projection Layers**: Linear layers mapping image embeddings to BioGPT's hidden size (768 dimensions).
- **Loss Function**: Contrastive Loss.

### 3) BioGPT:

#### PEFT Configuration:

Parameter-Efficient Fine-Tuning (PEFT) is a strategy to fine-tune large language models like BioGPT using fewer trainable parameters. The specific configuration used here is LoRA (Low-Rank Adaptation). LoRA introduces additional trainable rank-reduced matrices during fine-tuning without modifying the original model weights directly.

#### Key parameters:

##### **Rank: 16**

This determines the rank of the low-rank adaptation matrices added during fine-tuning. A higher rank increases the capacity of the adaptation but also increases the computational cost.

##### **Alpha: 32**

Alpha is a scaling factor that balances the contribution of the LoRA parameters with the pre-trained weights. Larger alpha values give more weight to the pre-trained parameters, making fine-tuning less aggressive.

##### **Dropout: 0.1**

Dropout is applied to the LoRA layers to prevent overfitting during fine-tuning.

#### **Generation Parameters: These parameters control the text generation process in BioGPT:**

- **max\_length: 150**

Specifies the maximum number of tokens that can be generated in the output text.

- **temperature: 0.8**

A lower temperature produces more deterministic outputs, while a higher value introduces randomness. A temperature of 0.8 ensures some diversity in generated reports while maintaining coherence.

- **top\_k: 50**

Limits sampling to the top 50 tokens with the highest probability at each generation step. This avoids low-probability tokens while allowing some variety.

- **top\_p: 0.85**

Enables nucleus sampling, where tokens are chosen from the smallest cumulative probability distribution above 85%. This ensures a balance between diversity and coherence.

### 3) Integration and Embedding Flow

- I. **Image Preprocessing:** PNG images are resized and augmented.
- II. **Image Encoding:** BioVilt extracts features, producing embeddings.
- III. **Alignment:** Image embeddings are projected to align with BioGPT's text embeddings.
- IV. **Report Generation:** Aligned embeddings are fed into BioGPT to generate natural language reports.

## ii) BioVilt + ChexNet + Alignment + BioGPT

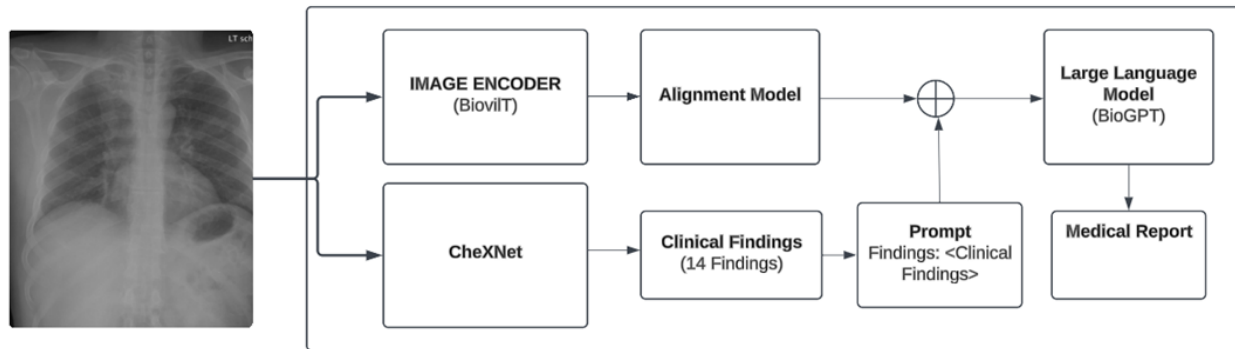


Figure 2: Architecture of Biovilt + ChexNet + Alignment + BioGPT

### 1) Model Used

- **BioVilt:** ResNet-50 based image encoder.
- **ChexNet:** Multi-label classifier for structural findings.
- **Alignment Model:** Integrates image and label embeddings with text embeddings.
- **BioGPT:** Fine-tuned for biomedical report generation.

### 2) Configuration

#### 1. BioViT:

1. **Architecture:** ResNet-50 backbone.
2. **Output Dimensions:** Global embedding of 512 dimensions.

#### 2. ChexNet

1. **Architecture:** DenseNet-121 backbone.
2. **Output Dimensions:** Multi-label predictions for 14 clinical findings.

#### 3. Alignment Module

1. **Text Encoder:** Microsoft BioGPT.
2. **Projection Layers:** Linear layers to project image embeddings to 768 dimensions. A separate linear layer to project the text in the ground truth report as well.
3. **Loss Function:** Contrastive Loss.

#### 4. BioGPT

##### 1. PEFT Configuration: LoRA with

- rank 16
- alpha 32
- dropout 0.1.

##### 2. Generation Parameters:

- max\_length=150
- temperature=0.8
- top\_k=50
- top\_p=0.85



### 3) Integration and Embedding Flow

- i. **Image Preprocessing:** PNG images are resized and augmented.
- ii. **Image Encoding:** BioVilt extracts features, producing image embeddings.
- iii. **ChexNet Classification:** Identifies and extracts structural findings, generating binary labels.
- iv. **Alignment:** Combined embeddings are projected to align with BioGPT's text embeddings.
- v. **Concatenation of Image Embeddings and Prompt:** The Image embeddings and the prompt text embeddings are concatenated having a separator (<SEP>) token in between them and passed to BioGPT.
- vi. **Report Generation:** Concatenated embeddings are fed into BioGPT to generate natural language reports.

#### iii) BioMed + ChexNet + Alignment + BioGPT

After implementing models like BioViLT along with ChexNet, we decided to try out Biomed with ChexNet and BioGPT because Biomed is adapted for tasks requiring high generalization across unseen medical data. We initially used Biomed as a text embedding extractor, and then we wanted to determine whether training the entire model—along with image and text encoder with BioGPT—for report generation would improve performance, unlike BioViLT where we train only the alignment model and BioGPT. Following this analysis, we proceeded to train the complete model as well.

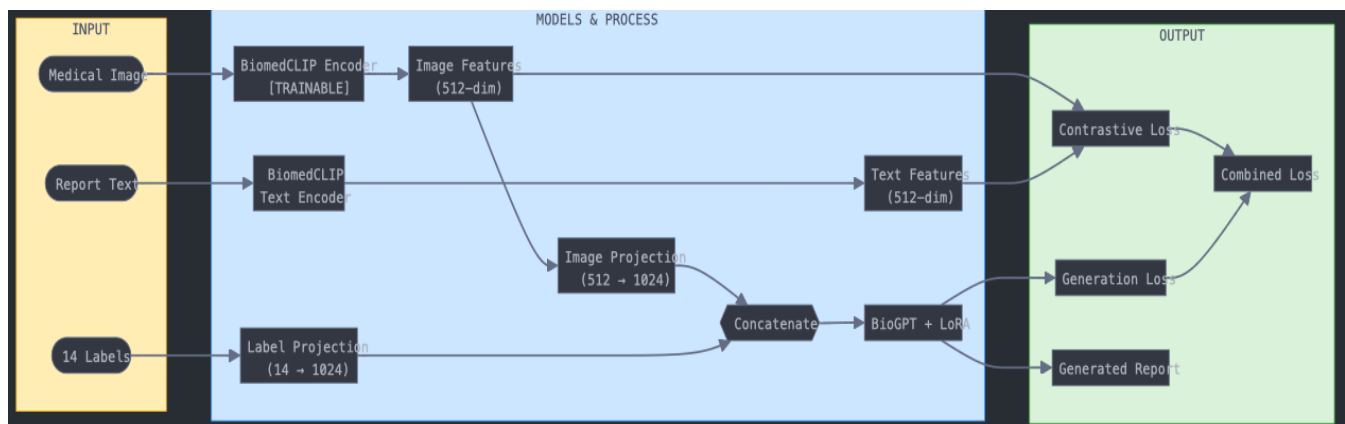


Figure 3: Architecture of BioMed + ChexNet + Alignment + BioGPT

#### 1. Model Used

##### i) BioMed Vision Transformer (ViT-Base/16) for image encoding:

BioMed is a transformer-based architecture used in biomedical image processing. The core architecture comprises a Vision Encoder and a Text Encoder. The Vision Encoder utilizes ViT-Base/16, a Vision Transformer that processes medical images by dividing them into  $16 \times 16 \times 16$  times  $1616 \times 16$  patches,

encoding them through transformer layers with self-attention mechanisms to capture spatial and contextual relationships, and producing a final output embedding of dimension 512. Complementing this, the Text Encoder leverages PubMedBERT-base, a specialized BERT model pre-trained on extensive biomedical literature. This encoder processes medical text to generate context-rich embeddings, also of dimension 512, ensuring alignment with the Vision Encoder for seamless integration in tasks such as image-text retrieval, report generation, and disease diagnosis.

ii) **ChexNet:** Multi-label classifier for structural findings.

iii) **Alignment Model:** Integrates image and label embeddings with text embeddings.

iv) **BioGPT:** Fine-tuned for biomedical report generation.

## 2. Configuration

### BioMed:

- **Architecture:** ViT-Base/16 Transformer layers.
- **Output Dimensions:** 512 dimensions.

### ChexNet

- **Architecture:** DenseNet-121 backbone.
- **Output Dimensions:** Multi-label predictions for 14 clinical findings.

### Alignment Module

- **Text Encoder:** Microsoft BioGPT.
- **Projection Layers:** Linear layers mapping combined image and label embeddings to 768 dimensions.
- **Loss Function:** Contrastive Loss.

### Training Setup:

- **Batch size:** 8
- **Number of epochs:** 18
- **Training/validation split:** 80/20
- **For BiomedCLIP model:** AdamW with learning rate of 1e-5
- Uses Reduce LR on Plateau scheduler for both optimizers

### Training Specifics:

- I. **Loss function:** Combines contrastive loss and generation loss
  - A. 30% weight for CLIP contrastive loss
  - B. 70% weight for generation loss
- II. **Early stopping patience:** 5 epochs
- III. **Uses ROUGE-L score for model selection**

### BioGPT Configuration

#### 1. PEFT Configuration: LoRA with

- rank 16

- alpha 32
- dropout 0.1.

## 2. Generation Parameters:

- **max\_length**=150
- **temperature**=0.3
- **top\_k**=50
- **top\_p**=0.5

After experimenting with these parameters, we found that these specific configurations performed well. Since Biomed was already trained on biological images, we wanted BioGPT to be more focused, so we kept the temperature, top\_p, and top\_k values small.

## 3. Integration and Embedding Flow

- **Image Preprocessing:** PNG images are resized and augmented.
- **Image Encoding:** BioMed (ViT) extracts features, producing embeddings.
- **ChexNet Classification:** Identifies and extracts structural findings, generating binary labels.
- **Feature Concatenation:** Image embeddings are concatenated with ChexNet labels to form a comprehensive feature vector.
- **Alignment:** Combined embeddings are projected to align with BioGPT's text embeddings.
- **Report Generation:** Aligned embeddings are fed into BioGPT to generate natural language reports.

## 4) BLIP2 + ChexNet + BioGPT

From our experience with BioViLT and BiomeD, we learned that training only alignment models, using BiomeD as an image extractor, or even training the entire BiomeD CLIP model were not sufficiently effective approaches. Therefore, we wanted to experiment with a larger model using an innovative approach. Through our research of academic papers, we discovered BLIP, a well-generalized model that uses an innovative QFormer which can be trained while keeping the image and language models frozen. However, we ultimately used it only as a text embedding extractor since we were unable to train it due to computation power limitations. In the end, we trained only the ChexNet + BioGPT component.

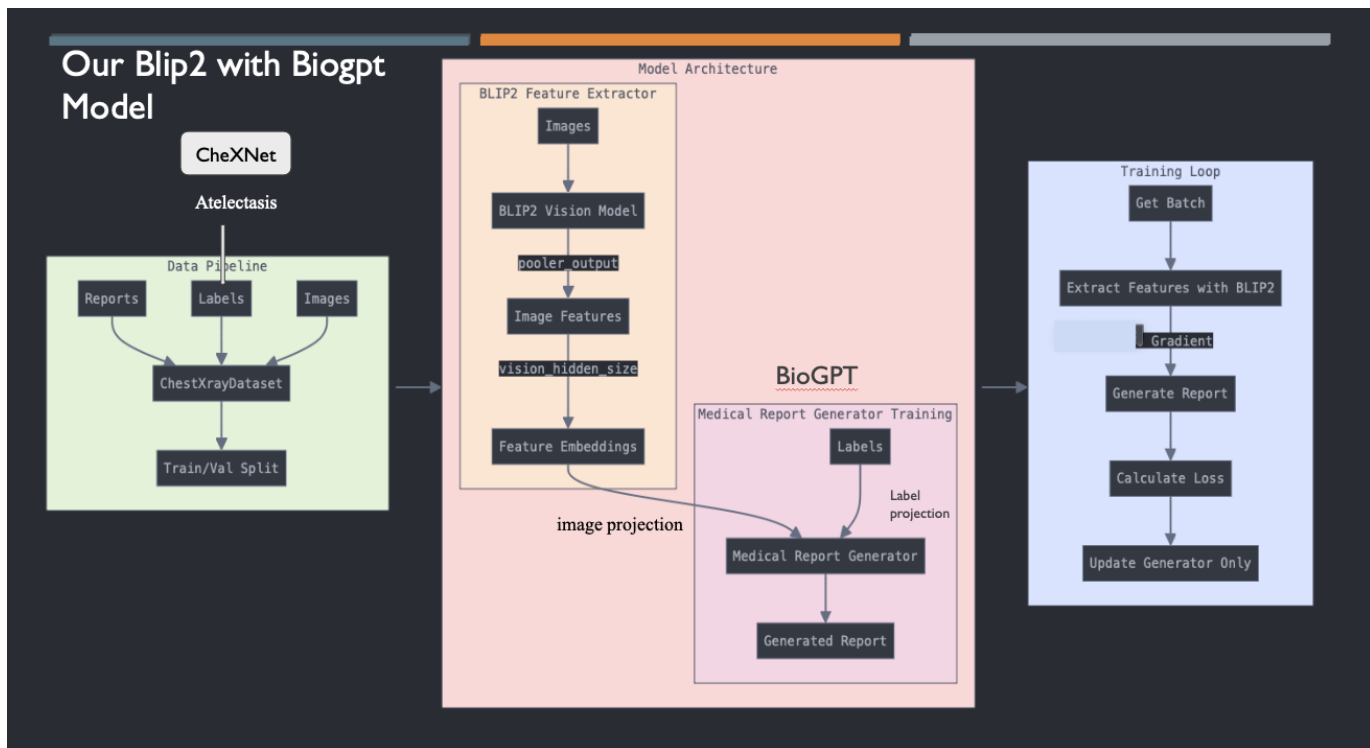


Figure 4: Architecture of BLIP2 + BioGPT

## 1) Model Used

- a. **BLIP2:** BLIP2 (Bootstrapped Language-Image Pretraining 2) by Salesforce is an advanced vision-language model designed to bridge visual and textual understanding. It introduces a novel two-stage pretraining strategy, enabling efficient alignment of vision and language models without requiring computationally expensive end-to-end optimization. The model leverages a lightweight Querying Transformer (Q-Former) to extract informative features from vision encoders, like ViT (Vision Transformer), and aligns them with text representations from large language models (LLMs) such as OPT or GPT.
- b. **ChexNet:** Multi-label classifier for structural findings.
- c. **BioGPT:** Fine-tuned for biomedical report generation.

## 2) Configuration

- a. **Blip2:**
  - i. **Model:** "Salesforce/blip2-opt-2.7b"
  - ii. **Function:** Purely serves as a feature extractor without the need for an alignment model
- b. **ChexNet**
  - i. **Architecture:** DenseNet-121 backbone.
  - ii. **Output Dimensions:** Multi-label predictions for 14 clinical findings.
- c. **Alignment Module**
  - i. **Text Encoder:** Microsoft BioGPT.
  - ii. **Projection Layers:** Linear layers mapping combined image and label embeddings to 768 dimensions.
  - iii. **Loss Function:** Cosine Embedding Loss.
- d. **BioGPT**
  - i. **PEFT Configuration:** LoRA with

1. rank 16
2. alpha 32
3. dropout 0.1.

**ii. Generation Parameters:**

1. max\_length=150
2. temperature=0.7
3. top\_k=50
4. top\_p=0.9

**3) Integration and Embedding Flow**

- a. Image Preprocessing:** PNG images are resized and augmented.
- b. Image Encoding:** BLIP2 extracts features, producing embeddings.
- c. ChexNet Classification:** Identifies and extracts structural findings, generating binary labels.
- d. Feature Concatenation:** Image embeddings are concatenated with ChexNet labels to form a comprehensive feature vector.
- e. Report Generation:** Combined embeddings are directly fed into BioGPT without an additional alignment model, leveraging BLIP2's inherent capabilities to bridge image and text modalities.

## **5. Model Architecture - CXR- MATE**

After Implementing Heavy models like Blip2 and BioMed(Specialized CLIP-based model). We decided to develop and fine-tune a model which is specifically used in biomedical engineering.

- 1) **Model Used** - CXR MATE is an advanced DL model designed for chest Xray analysis and medical image interpretation.



Figure 5: Architecture of CXR Mate

In the CXR pipeline, filtering of images is done based on valid medical findings. Then we apply some image augmentations and transformations.

The module supports 14 condition labels which we predict using ChexNet model. In the training loop we load CXR Mate model and tokenizer and the medical report generator. The training loop implements image feature extraction, report generation and checkpoint saving for inference.

**Configuration:** Epochs - 20, Batch size - 8, Learning rate is different for different model components.

**PEFT Configuration:** LoRA with Rank - 16, alpha - 32, dropout - 0.1

## 4. RESULTS

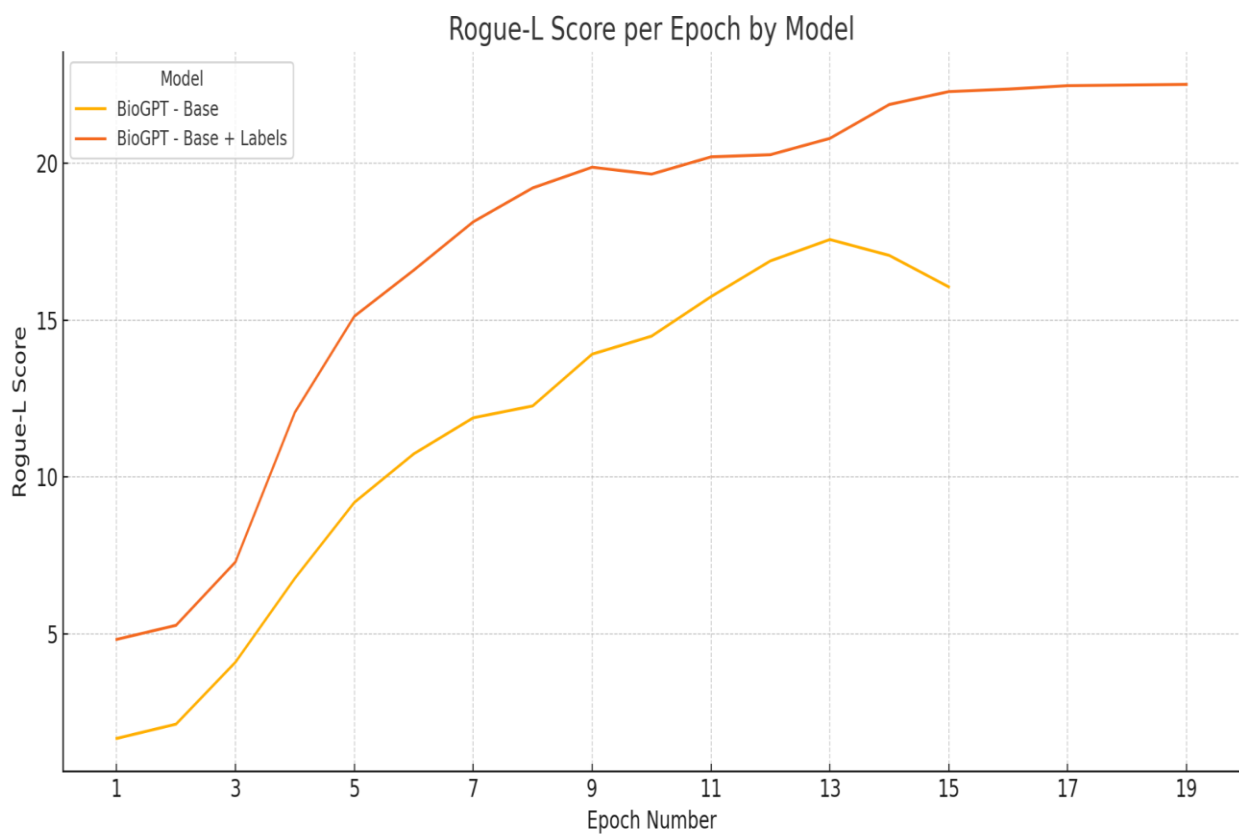
In this analysis, a comprehensive comparison is conducted between 4 distinct models. ROUGE (Recall Oriented Understudy for Gisting Evaluation) metric is used as the primary evaluation metric. ROUGE measures the overlap of predicted text against reference text across several dimensions, including recall, precision, and f1-score to evaluate the quality of generated summaries.

- I. **ROUGE-L (Longest Common Subsequence)** - Evaluates the longest common subsequence between generated and reference text. The metric considers the sequence structure and how well the order of words in the generated summary matches the reference summary. Rouge L gives credit for correctly ordered content even if the content is spread across the summary.

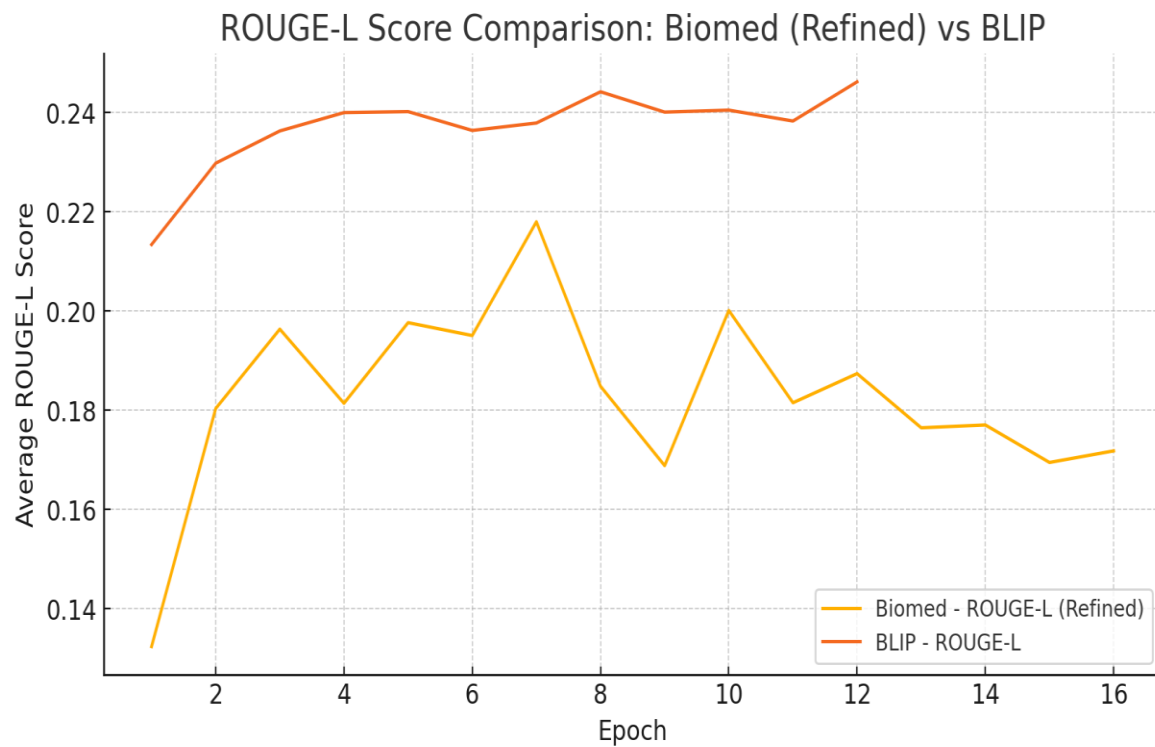
$$\text{ROUGE-L} = \frac{\text{LCS length}}{\text{Length of reference summary}}$$

- II. (Latex snippet taken from ChatGPT)

- III. **Graph snippets for (BioGPT + Image encoder) and (BioGPT + image encoder + chexNet Labels)**



**Deduction - Generated text coming from BioGPT + Base + ChexNet labels overlap more with the reference text than the base model.**



**Deduction:** BioMed (a CLIP-based) model has a lower average ROUGE L score as compared to the Blip2 model.

#### 4b. Sample outputs:

##### BioVilt + ChexNet + Alignment + BioGPT:


Chest X Ray Image	Ground Truth	Prediction
	<p>Clinical Findings: Opacity</p> <p>Report: there is a <b>large right hilar opacity</b>, xxxx in the posterior segment of the lung. <b>the heart size and mediastinal contour are normal. no pneumothorax or pleural effusions.</b> this appears to be hyperinflated no focal airspace consolidations.</p>	<p>Clinical Findings: Opacity</p> <p>Report: there is a <b>rounded opacity in the right lower zone</b> measuring 2.0 cm which is xxxx to be in the <b>posterobasal</b> segment. there is of uncertain etiology but would benefit from <u>followup</u> at xxxx some concern for neoplasm. a xxxx is recommended. <b>no airspace disease, effusion</b> or cavitory nodule. <b>normal heart size and mediastinum.</b> visualized xxxx of the chest xxxx are within normal limits.</p>

Figure 6: BioVilt + ChexNet + Alignment + BioGPT



### BioVilt + Alignment + BioGPT:

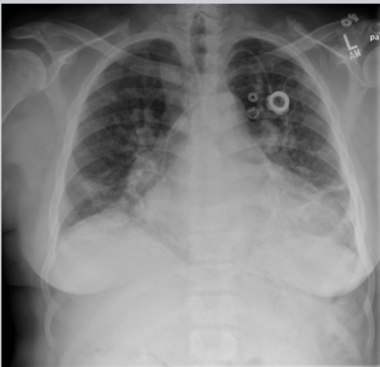
Chest X Ray Image	Ground Truth	Prediction
	<p>Report:</p> <p>Bilateral patchy pulmonary opacities noted. Interval improvement in left base consolidative opacity. Pulmonary vascular congestion again noted. Stable enlarged cardio mediastinal silhouette. Stable left XXXX. No evidence of pneumothorax. No large pleural effusions.</p>	<p>Report:</p> <p>Sternotomy and repair of a right-sided chest wall pericardiac loop in association with congestive heart failure is unchanged, but on the opposite side of this. The lungs are well expanded without evidence for acute or chronic consolidation. No pleural or pneumothorax found. No new areas of focal airspace disease. There has been interval resolution of left basilar atelectasis.</p>

Figure 7: BioVilt + Alignment + BioGPT

### Biomed + ChexNet + BioGPT:

**Generated:** The heart is normal in size and contour, mediastinal contours are within normal limits. There is no focal airspace consolidation or pleural effusion. No suspicious pulmonary nodules are identified. There are mild degenerative changes of the spine.

Target: XXXX XXXX and lateral chest examination was obtained. The heart silhouette is normal in size and contour. Aortic XXXX appear unremarkable. Lungs demonstrate no acute findings. There is no effusion or pneumothorax. There is degenerative changes of the skeletal structures

### Blip + ChexNet + BioGPT:

**Generated:** CONCLUSION: PA and lateral views of the chest were obtained. There is mild left basilar atelectasis. No focal consolidation, pleural effusion or pneumothorax is seen. The cardiomeastinal silhouette is normal. Imaged osseous structures are intact.

**Original:** Patchy left base opacity may be due to atelectasis versus subtle pneumonia.No pleural effusion or pneumothorax is seen. The cardiac and mediastinal silhouettes are unremarkable.

### CXR Mate:

#### Results:

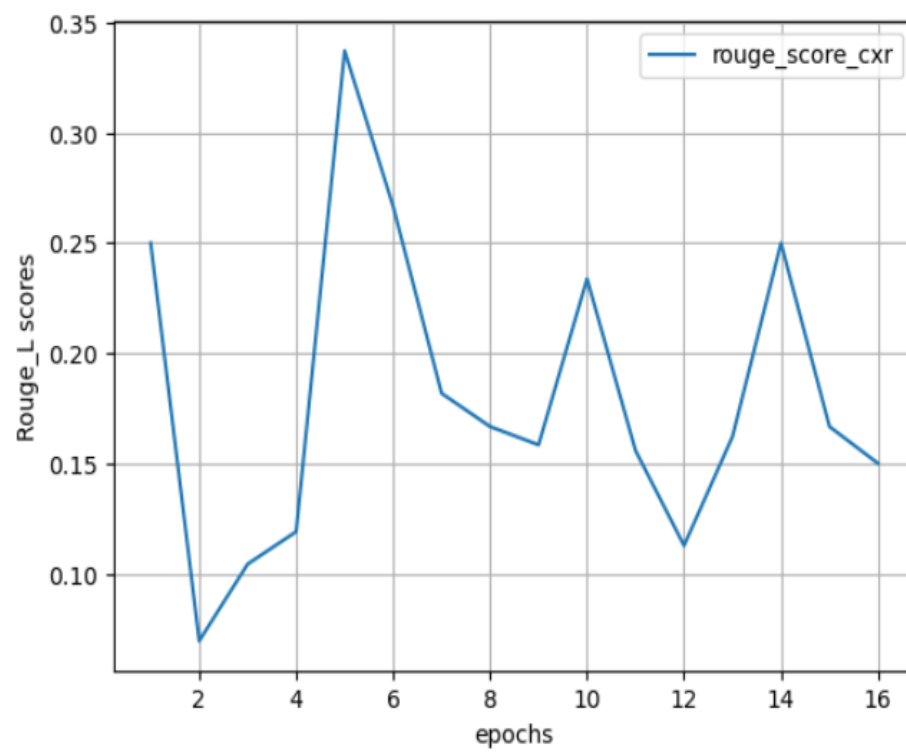


Figure 8: Performance of CXR Mate



Figure 9: Streamlit app

## CONCLUSIONS

Using ChexNet to add labels for the images as input to the models yielded significant benefits and improved overall performance. Training BioViLT + ChexNet + BioGPT models produced performance equivalent to the larger BLIP + ChexNet + BioGPT configuration, where we did not train BLIP. Although we trained the entire BiomeD model, our best ROUGE score was still lower than that achieved with the BLIP model when used as a text embedding extractor. Training bigger models like Blip-2 with a good amount of data will significantly improve the ROUGE L score.

## FURTHER IMPROVEMENTS

Training the Q-former (Alignment module) of Blip 2 model to enable cross attention so that image embeddings and text embeddings are aligned.

Using an ensemble of a custom CLIP and BLIP model could also improve the performance.

Using Multiple Images would help capture more features from the Images.

## REFERENCES

- [1]. [ChexBERT](#)
- [2]. [ChexNet](#)
- [3]. [BioVilt](#)
- [4]. BioGPT
- [5]. RaDialog-RG
- [6]. METransformer
- [7]. Blip2
- [8]. Clip
- [9]. CXR Mate