RADAC-X: Radiology And Data Augmented Classifier for X-ray

Wonyoung Kim
Ohio State University
Columbus, OH 43210
kim.9519@osu.edu

Doeun Lee
Ohio State University
Columbus, OH 43210
lee.11501@osu.edu

Jiwoo Park
Ohio State University
Columbus, OH 43210
park.3620@osu.edu

Abstract

We propose an anatomical segmentation masking procedure on chest X-ray images to provide additional visual prior when both interpreting and learning chest X-rays. We incorporate fine-tuning approaches for **BiomedCLIP**, a pre-trained medical vision-language model, aiming to achieve accurate disease label predictions with improved classification accuracy against chest X-ray images. Experimental results demonstrate that combining original images with anatomical segmentation masks enables the model to achieve higher accuracy in disease label predictions. Furthermore, the fine-tuned model produces predictions aligned with standard radiological findings, contributing to the advancement of automated chest X-ray understanding.

1 Introduction

This project is a self-defined task that fine-tunes **BiomedCLIP**[1], a pre-trained medical vision-language model, to classify diseases more accurately based on chest X-ray images augmented with anatomical segmentation masks. The pre-trained model used as both the baseline and for fine-tuning is **BiomedCLIP** [1], which was originally trained on PMC-15M for a variety of tasks and tested with PCam, LC25000, TCGA-TIL, RSNA for an image classification task. We conduct several experiments on fine-tuning either only the linear classifier head(with the frozen vision encoder) or both the vision encoder and the classification head using original chest X-ray images or those combined with anatomical mask segmentation(e.g., lung and heart boundaries) to enhance anatomical focus. The augmentation of masks aims to guide the model's attention toward clinically relevant regions. Our goal is to enable the model to produce interpretable outputs through classification, providing standardized diagnostic labels that reflect radiological findings. This approach enables explainable decision-making in medical images while remaining lightweight.

2 Motivation

The development of vision models suggested a variety of new capabilities in different areas. Although domain-specific applications of the models are relatively unexplored, they offer the potential to reduce the efforts of experts on time-consuming tasks. Especially, while the use of models in the medical field is in a primitive stage due to its complexity, it is one of the major domains that will benefit from the task-specific application of the models. Therefore, we plan to explore the interesting yet challenging task of addressing the capability of these models in the medical area to assist medical professionals in data association and clinical decision-making. This project uses **anatomical masking strategy** to assist clinicians by providing a tool capable of **diagnosis of pulmonary disease** based on the **explicitly segmented region**.

.

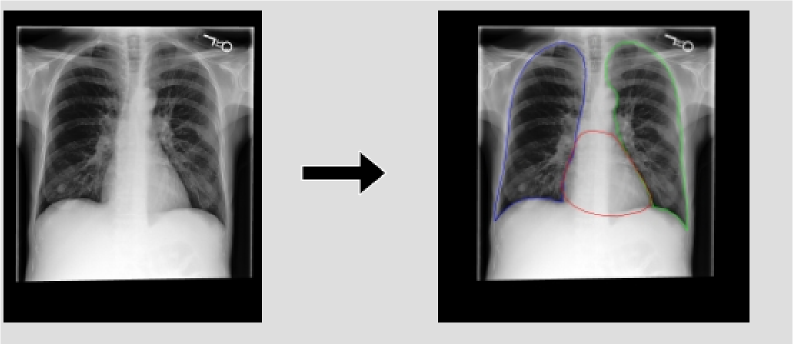

Figure 1: Original image (left) and masked image (right)

3 Approach

In this section, we describe the creation of masked images along with our baseline and advanced approaches for label classification of chest X-rays using a vision-language model. We begin by leveraging the pre-trained **BiomedCLIP** model as a baseline, followed by our fine-tuning strategies using images from **MIMIC-CXR-JPG** [2] incorporated with anatomical masks.

3.1 Data Masking

Chest X-rays from **MIMIC-CXR-JPG** [2] were paired with corresponding **CheXmask** [3] segmentation masks to create an overlay of borders in the anatomical structures as shown in Figure 1. We hypothesize that using masked images will improve classification accuracy and significantly enhance model interpretability. By combining segmentation-based anatomical boundary information with the original image, our fine-tuning approach allows the model to focus on clinically relevant regions and associate them with labels at higher precision.

3.2 Baseline Approach

We use a pre-trained **vision-language model(BiomedCLIP)** [1] as the baseline. The model consists of a vision encoder (**ViT**) and a language encoder (**PubMedBERT**). BiomedCLIP has demonstrated its ability to learn representation across a broad range of medical domains, beyond chest X-ray images. It is pre-trained on a large-scale biomedical dataset (PMC-15M) consisting of high-quality parallel image-text pairs, ensuring satisfactory performance on zero-shot classification and retrieval tasks.

However, the baseline model is pre-trained primarily on general medical images-text pairs rather than explicitly on diagnostic tasks. As a result, it exhibits the limited ability to capture subtle diagnostic differences in the medical images that are critical for distinguishing between specific diseases. Consequently, it is expected to struggle in focusing on anatomically relevant regions or nuanced clinical cues, leading to the performance deficiency in disease classification task and reducing clinical interpretability.

3.3 Advanced approach

To address the limitations of BiomedCLIP [1], we applied two approaches for fine-tuning. For each approach, we use two separate train sets (original images and masked images) separately to observe the effect of masking, resulting in four fine-tuned models in total. To prioritize the assessment of the vision encoder in this classification task, we use the vision encoder from BiomedCLIP and define a linear classifier for fine-tuning and evaluation.

3.3.1 Linear Classifier

We focus on the effect of masks on the interpretation of visual cues. We adopt a fine-tuning strategy where the vision encoder is frozen and a randomly initialized linear classifier head is used for fine-tuning using cross-entropy loss. As the semantics of the masked area cannot be recognized by the frozen vision encoder, we rely on the ability of the linear classifier to interpret the extracted features under this fine-tuning framework. With the original images, the linear classifier is expected to interpret the cues from the pre-trained model that were not easily separable by the baseline. This setup was inspired by **MedCLIP**'s fine-tuning method. [4]

3.3.2 Linear Classifier and Vision Encoder

In this approach, we fine-tune both the vision encoder and a randomly initialized linear classifier with cross-entropy loss for fine-tuning. Allowing the vision encoder to adapt its feature extraction to masked images enables the model to learn task-specific features localized to specific areas in the chest X-rays. Therefore, this fine-tuned model will assess the capability of knowledge extraction by region segmentation along with their interpretation. Fine-tuning on the original images will further refine the pre-trained representations of chest X-rays by associating them with the target classes.

3.4 Library, computational resource

- **Programming Language**: Python
- **Libraries & Frameworks**:
  - **Deep Learning**: PyTorch, Hugging Face Transformers
  - **Data Processing**: pandas, NumPy, scikit-learn
- **Computational resource**: Ascend NVIDIA A100 GPU

4 Experiments/Evaluation/Validation

4.1 Data

4.1.1 Data Sources

- **MIMIC-CXR-JPG** [2]: A large scale dataset consisting of frontal-view chest X-ray images paired with multi-label diagnoses.
- **CheXmask** [3]: Anatomical segmentation masks corresponding to MIMIC-CXR-JPG images, including heart, left lung, and right lung.

4.1.2 Data processing

We apply several data preprocessing steps for model fine-tuning and evaluation. First, anatomical segmentation masks that highlight anatomical structure boundaries are overlaid on the original chest X-ray images to improve anatomical focus. Next, all images are padded and resized to 224×224 pixels to ensure compatibility with CLIP-based vision encoders. Furthermore, we apply filtering rules to build a clear and consistent dataset. Specifically, we only train the frontal-view images and choose cases that have at least one positive label among the five target disease categories, which are Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion.

4.1.3 Data split

After filtering, the datasets were divided as follows:

- **Train set**: 87,574
- **Validation set**: 17,514
- **Test set**: 2,045

4.2 Metric

The primary evaluation metric is **accuracy**. We frame the task as a multi-class classification over five pulmonary disease categories. Model performance is evaluated separately on masked and original images during both training and testing, aiming for higher accuracy than the baseline approach.

4.2.1 Test setup

As a baseline, we perform zero-shot classification using the pre-trained BiomedCLIP [1] model without any fine-tuning. We then explore diverse fine-tuning strategies: (1) fine-tuning only the linear classifier head while keeping the vision encoder frozen, and (2) fine-tuning both the classifier head and the vision encoder. We perform separate fine-tuning for distinct datasets (original and masked images). For evaluation, we compare the model's performance when testing on original chest X-ray images versus anatomically masked images. During both training and evaluation for classification, we exclusively use the vision encoder for the the scope of this project.

4.3 Results and/or Further Analysis

Fine-tune Method	**Train Image**	**Accuracy (Original)**	**Accuracy (Masked)**
Baseline (Zero-shot)	–	0.2367	**0.3467**
V1: Linear classifier head	Original	0.4460	**0.4509**
V2: Linear classifier head	Masked	0.4205	**0.4215**
V3: Classifier + vision encoder	Original	0.4005	**0.4039**
V4: Classifier + vision encoder	Masked	0.3484	**0.4318**

Table 1: Comparison of fine-tuning methods with different training images on test sets.

Table 1 summarizes the performance of models fine-tuned on different training sets and evaluated on different testing sets.

We observe that our approach outperforms the baseline approach in every model-data pair. Additionally, our experimental results reveal several important observations regarding the effects of anatomical segmentation masks on fine-tuning and inference performance.

First, incorporating masks during the fine-tuning stage did not consistently improve model accuracy compared to training on original, unmasked images. Specifically, models fine-tuned on masked datasets (V2, V4) exhibited lower or comparable training and testing accuracies relative to those trained on original images (V1, V3). This suggests that introducing masks during training might disrupt the model's ability to leverage broader contextual information available in full chest X-rays. Possible explanations include (1) the loss of contextual cues necessary for accurate disease classification and (2) the introduction of artifacts through the masking process, which may inadvertently confuse the model during representation learning.

However, an important phenomenon was observed during inference (testing): Evaluating on masked images generally led to improved classification performance across most fine-tuning strategies. For instance, models trained on original images and tested on masked images (V1) achieved slightly higher accuracy (0.4509) compared to testing on original images (0.4460). Similarly, models fine-tuned on masked images also showed better performance when evaluated on masked test images.

These results suggest that masks act as lightweight visual priors during inference. By highlighting anatomical regions of clinical relevance (e.g., lung boundaries, heart contour), the masks guide the model's attention towards the most important structures for diagnosis. Rather than enhancing feature extraction during training, the masks appear to unlock spatial knowledge already embedded within the pretrained vision encoder, allowing the model to focus on diagnostically significant areas without explicit retraining on those regions.

5 Conclusion and discussion

In this study, we explore the effect of chest X-ray images with anatomical masking on pulmonary disease classification. Through diverse fine-tuning strategies and evaluation settings, we assessed whether masking could improve the performance of the model.

5.1 Insights

Overall, using masks during fine-tuning did not improve performance compared to training original images. Training models on masked images sometimes results in lower accuracy compared to training on original images. This suggests that the reason for the lower accuracy may be the loss of useful context information necessary for accurate diagnosis or the introduction of artifacts caused by the masking process. However, at test time, the model evaluated on masked images shows higher accuracy compared to original images. We suggest that masks act as visual priors, helping the model focus more effectively on relevant regions. This effect is particularly evident in the model fine-tuned on both the classifier head and the vision encoder, where testing on masked images led to better performance than on original images. Although masks did not enhance training performance, they seem to guide the model's attention during inference, acting as lightweight priors that reveal pretrained spatial knowledge. Therefore, we accept our initial hypothesis that masked images improve classification accuracy and enhance model interpretability.

5.2 Workload (every member)

- Reading relevant research papers(**BiomedCLIP** [1], **MedCLIP** [4], etc.)
- Get familiar with PyTorch, HuggingFace, and explainability techniques
- Data preprocessing: Applying anatomical masks, formatting input, and filtering datasets
- Model fine-tuning: Create four model variants across different datasets
- Optimization: Tuning hyperparameters and conducting evaluations in a total eight experiment settings.

5.3 Key challenges you face

- Low accuracy when fine-tuning with image-report pairs: Fine-tuning the model on image-report pairs results in suboptimal performance. It suggests that original pre-trained text generation might not align accurately with classification tasks, requiring careful adaptation.
- Using **MedCLIP**(original target model) [4] was tricky: We initially adopt **MedCLIP** as our baseline approach. However, despite fine-tuning, it does not achieve competitive accuracy. Due to the challenges in adapting **MedCLIP** for our classification objective, we decided to switch to using **BiomedCLIP** [1], which is better suited for our downstream task.
- Task exploration: Identifying an appropriate downstream task, such as classification, retrieval, and other alternative formulations, requires substantial exploration and experimentation to align with the project's objective.
- Building subsamples on large datasets: handling and filtering a large-scale dataset like **MIMIC-CXR-JPG** [2] and **CheXmask** [3] requires significant preprocessing effort, including filtering rules and masking procedures.

References

[1] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. Biomedclip: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2025.

[2] Alistair Johnson, Matthew Lungren, Yifan Peng, Zhiyong Lu, Roger Mark, Seth Berkowitz, and Steven Horng. Mimic-cxr-jpg - chest radiographs with structured labels (version 2.1.0). *PhysioNet*, 2024.

[3] Nicolas Gaggion, Candelaria Mosquera, Martina Aineseder, Lucas Mansilla, Diego Milone, , and Enzo Ferrante. Chexmask database: a large-scale dataset of anatomical segmentation masks for chest x-ray images (version 1.0.0). *PhysioNet*, 2025.

[4] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, 2022:3876–3887, 2022.