Disease Detection in Chest X-rays using CLIP Features with Interpretable ROI Analysis

Manasa Mangipudi
manasa.mangipudi@rutgers.edu
Rutgers University
New Brunswick, NJ, USA

ABSTRACT

Chest radiography is a widely used diagnostic tool for detecting thoracic diseases. However, accurately interpreting chest X-rays requires significant expertise and can be time-consuming. This study proposes a novel multi-modal deep learning approach for classifying chest X-rays as normal or diseased by leveraging both the radiographic images and the associated textual reports, which contain valuable information about the physician's motivation for recommending the X-ray. The proposed pipeline consists of a CLIP (Contrastive Language-Image Pre-training) model to learn joint visual-textual representations, followed by a simple classifier head for final predictions. The CLIP model, comprising a Vision Transformer (ViT) for image encoding and BERT for text encoding, is trained on a large dataset of chest X-rays and corresponding reports using a contrastive objective function. By incorporating the physician's reasoning from the textual reports, the model gains a more comprehensive understanding of the patient's condition. The learned multi-modal representations are then used to train a linear classifier to distinguish between normal and diseased cases. To provide insights into the model's decision-making process, Grad-CAM (Gradient-weighted Class Activation Mapping) visualizations are generated, highlighting the salient regions in the input chest X-rays that contribute to the predictions. Experimental results on a large-scale chest X-ray dataset demonstrate the effectiveness of the proposed approach, achieving high classification accuracy. The Grad-CAM visualizations further showcase the model's ability to focus on clinically relevant areas of the radiographs when making decisions. The proposed multi-modal framework, which incorporates both visual and textual information, offers a promising avenue for developing interpretable and explainable AI-assisted tools to aid radiologists in chest X-ray analysis, potentially improving diagnostic efficiency and patient outcomes.

KEYWORDS

Chest X-ray classification, Multi-modal learning, Contrastive Language-Image Pre-training (CLIP), Vision Transformer (ViT), BERT, Grad-CAM, Explainable AI

Multi-modal Machine Learning for Sensing Systems, Times,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Manasa Mangipudi. 2024. Disease Detection in Chest X-rays using CLIP Features with Interpretable ROI Analysis. In *Proceedings of (Multi-modal Machine Learning for Sensing Systems).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Chest radiography is one of the most commonly performed diagnostic imaging examinations, playing a vital role in the detection and management of various thoracic diseases . However, interpreting chest X-rays is a challenging task that requires significant expertise and experience. Radiologists need to carefully analyze the images, taking into account a wide range of potential abnormalities, such as pneumonia, lung nodules, and pleural effusions. This process can be time-consuming and is subject to inter-observer variability.

In recent years, deep learning techniques have shown promising results in automating the analysis of medical images, including chest X-rays. Convolutional Neural Networks (CNNs) have been widely used for this purpose, achieving high accuracy in classifying chest X-rays as normal or abnormal(diseased). However, most existing approaches rely solely on the visual information present in the radiographs, without considering the rich contextual information available in the associated textual reports.

Radiological reports contain valuable insights from radiologists, describing their observations, findings, and impressions based on the X-ray images. These reports often provide a more comprehensive understanding of the patient's condition and can aid in the accurate interpretation of the radiographs. Integrating the information from both the visual and textual modalities has the potential to enhance the performance and robustness of automated chest X-ray analysis systems.

In this study, we propose a novel multi-modal deep learning approach for classifying chest X-rays as normal or diseased by leveraging both the radiographic images and associated textual reports. Our pipeline consists of a CLIP (Contrastive Language-Image Pre-training) model [5] to learn joint visual-textual representations, followed by a simple classifier head for final predictions. The CLIP model, comprising a Vision Transformer (ViT) [6] for image encoding and BERT [7] for text encoding, is trained on a large dataset of chest X-rays and corresponding reports using a contrastive objective function.

To enhance the interpretability and explainability of our model, we employ Grad-CAM (Gradient-weighted Class Activation Mapping) [8] visualizations. Grad-CAM highlights the salient regions in the input chest X-rays that contribute most to the model's predictions, providing insights into the decision-making process. This is particularly important in the medical domain, where understanding

the reasoning behind the model's outputs is crucial for building trust and aiding clinical decision-making.

The main contributions of this study are as follows:

(1) We propose a novel multi-modal deep learning framework for chest X-ray classification that combines visual and textual information using a CLIP model.
(2) We demonstrate the effectiveness of our approach on a large-scale chest X-ray dataset, achieving high classification accuracy.
(3) We generate Grad-CAM visualizations to provide interpretability and explainability, highlighting the salient regions in the chest X-rays that influence the model's predictions.

The rest of the paper is organized as follows. Section 2 presents the related work on chest X-ray classification and multi-modal learning. Section 3 describes the proposed methodology, including the dataset, preprocessing steps, CLIP model architecture, and Grad-CAM visualization technique. Section 4 presents the experimental setup and results, along with a discussion of the findings. Finally, Section 5 contains results and concludes the paper and outlines future research directions.

2 RELATED WORK

Radiology report generation is an important task that aims to automatically produce textual descriptions of medical images. This has the potential to reduce radiologists' workload and improve the efficiency and quality of the radiology reporting process. In recent years, deep learning approaches have shown promise for this task.

Early work on radiology report generation adapted CNN-RNN architectures and attention mechanisms that were successful for general image captioning tasks. For example, Jang et al. [2] used a CNN-RNN with visual attention on pathology images. They introduced a hierarchical LSTM with co-attention on both visual and semantic features to generate chest x-ray reports on the Indiana University Chest X-ray (IU-Xray) dataset [9].

Subsequent work has explored additional strategies to improve report generation, especially for handling longer, multi-sentence reports. Hierarchical recurrent architectures decompose the problem into generating topic vectors and then sentences conditioned on the topics. Li et al. [3] leveraged information from both frontal and lateral x-ray views. Knowledge graphs and retrieval-based methods have also been studied.

More recently, transformer-based language models pre-trained on large text corpora, such as GPT-2, have been applied to report generation, yielding state-of-the-art results. The work of Alfarghaly et al. [1] proposes a novel approach called CDGPT2 that fine-tunes a pre-trained distilled GPT-2 model conditioned on both visual features from a CNN and semantic features based on tags extracted from the x-ray. This avoids the need for a specialized medical vocabulary. The authors also introduce semantic similarity metrics to more comprehensively evaluate the generated reports.

Another direction has been to explore methods for transferring representations learned from report generation to downstream tasks like disease classification, in order to improve performance and enable utilization of smaller radiology report datasets. Xue and Huang [4] developed a recurrent attention model for report generation on the IU-Xray dataset, and showed that transferring the encoder's visual features to classifiers trained on the larger ChestX-ray14 dataset [9] could boost classification accuracy.

In summary, radiology report generation is an active area of research within medical image analysis and understanding. Progress is being made by incorporating insights from image captioning, hierarchical text generation, large language models, and transfer learning. However, challenges remain in producing reports with sufficient factual completeness and clinical accuracy. Promising avenues for future work include improving clinical relevance evaluation metrics, leveraging multi-modal and background information, and optimizing architectures for long document generation.

3 MATERIALS AND METHODOLOGY

3.1 Dataset

The dataset used in this study is the IU Chest X-Ray dataset[9] consisting of chest radiographs and associated textual reports. The dataset contains 3,955 radiology reports and 7,470 associated chest x-rays from 3,955 unique patients. The reports contain an impression section summarizing the key findings and a findings section with a detailed description. Each report is associated with 1-2 frontal/lateral x-ray images.

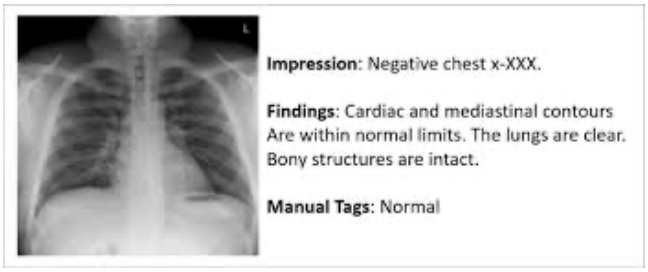

Figure 1: One sample from dataset

The radiological reports were preprocessed by extracting the indication and findings sections, concatenating them into a single text field, and applying text cleaning steps including lowercasing, removing special characters, stop word removal (using NLTK), and removing mentions of 'x' or 'year old'. The reports were then labeled as 'normal' or 'diseased' based on the 'Problems' field, and converted into numeric classes (0 for normal, 1 for diseased). After data cleaning, the dataset is split into training (2,725 patients), validation (584 patients) and test (585 patients) sets.

The associated chest x-ray images were normalized and resized to a consistent 224x224 resolution. During the dataset creation, for each report, all associated x-ray images were included, leveraging the multiple views available. This resulted in a final dataset with each example consisting of an x-ray image, the associated report text, and the normal/diseased label.

3.2 CLIP (Contrastive Language-Image Pre-training)

The core of the proposed methodology is the CLIP model, which learns to align the visual and textual representations into a shared multi-modal embedding space. CLIP, introduced by [5], has shown impressive performance on a wide range of vision-language tasks by learning from large-scale image-text pairs in an unsupervised manner.

The key idea behind CLIP is to train two encoder networks - one for images and one for text - to maximize the cosine similarity between the embeddings of matched image-text pairs, while minimizing the similarity between unmatched pairs. This contrastive objective encourages the model to learn meaningful and semantically aligned representations across modalities.

For the image encoder, we use a Vision Transformer (ViT) architecture, specifically the ViT-B/16 model. ViT, proposed by [6], applies the Transformer architecture directly to image patches, treating them as a sequence of tokens. This allows the model to capture long-range dependencies and learn more expressive image representations compared to traditional convolutional neural networks (CNNs). We initialize the ViT with weights pre-trained on the ImageNet dataset, which provides a strong starting point for fine-tuning on our specific task.

The text encoder is a BERT (Bidirectional Encoder Representations from Transformers) model, which has become the standard for natural language processing tasks. BERT, introduced by [7], is a deep bidirectional Transformer that learns contextualized word representations by training on large-scale text corpora. We use the BERT-base model and initialize it with pre-trained weights to leverage the vast amounts of linguistic knowledge captured during pre-training.

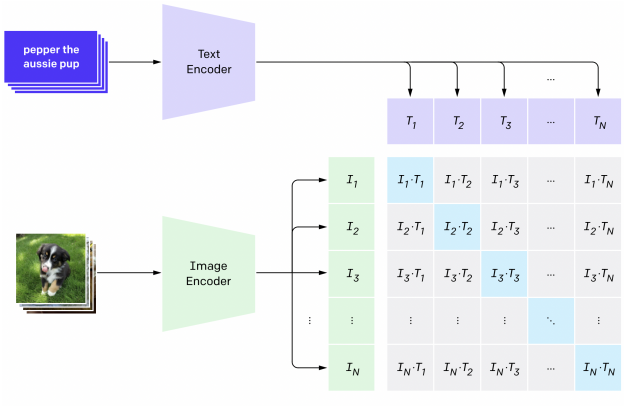

Figure 2: CLIP

The CLIP model is trained using a contrastive loss, specifically the InfoNCE loss, which maximizes the cosine similarity between matched image-text pairs while minimizing the similarity between unmatched pairs. Given a batch of N image-text pairs, the InfoNCE loss for a single pair (i, t) is defined as:

$$\mathcal{L}_{i,t} = -\log \frac{\exp(\text{sim}(i,t)/\tau)}{\sum_{t'} \exp(\text{sim}(i,t')/\tau)} \tag{1}$$

where $\text{sim}(i, t)$ is the cosine similarity between the image embedding i and text embedding t, τ is a temperature parameter that controls the softness of the distribution, and t' ranges over all text embeddings in the batch.

The final loss is the average of the InfoNCE losses over all positive pairs in the batch:

$$\mathcal{L} = \frac{1}{N} \sum_{(i,t)} \mathcal{L}_{i,t} \tag{2}$$

The model is trained on dataset of chest X-ray images and associated reports, learning to align the visual and textual features in a shared embedding space. By minimizing the contrastive loss, CLIP learns to produce similar embeddings for matched image-text pairs and dissimilar embeddings for unmatched pairs, effectively capturing the semantic alignment between the visual and textual modalities.

3.3 Classifier Head

Once the CLIP model is trained, we use it as a feature extractor to obtain multi-modal embeddings for each image-text pair. The learned 512-dimensional image and text embeddings are concatenated to form a 1024-dimensional feature vector.

We then train a simple classifier head on top of these frozen multi-modal features to predict the normal/diseased label. The classifier head consists of a multi-layer perceptron (MLP) with two hidden layers of size 512, followed by a softmax output layer for binary classification.

This approach, known as linear probing, allows us to assess the quality of the learned CLIP representations for the specific task of chest X-ray classification. By keeping the CLIP weights frozen, we can evaluate how well the learned multi-modal embeddings capture the relevant information for discriminating between normal and diseased cases.

3.4 Pipeline

The proposed pipeline for classifying chest X-rays into normal and diseased categories consists of three main components: data preprocessing, CLIP (Contrastive Language-Image Pre-training) model for learning joint image-text representations, and a classifier head for predicting the final labels.

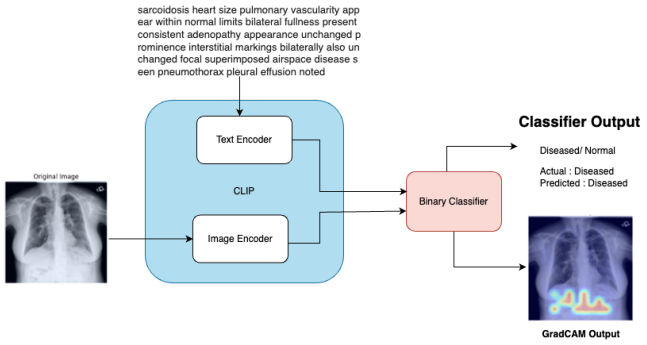

Figure 3: Complete Pipeline

The IU Chest X-Ray dataset, containing chest radiographs and associated textual reports, is used for this study. The reports are preprocessed by extracting the indication and findings sections, concatenating them, and applying text cleaning techniques. The images are normalized and resized to a consistent resolution.

The core of the pipeline is the CLIP model, which learns to align visual and textual representations into a shared embedding space. It consists of a Vision Transformer (ViT) for encoding images and a BERT model for encoding text. The ViT and BERT are initialized with pre-trained weights and fine-tuned during training. The model is trained using a contrastive objective, the InfoNCE loss, which maximizes the cosine similarity between matched image-text pairs while minimizing the similarity between unmatched pairs. After training, the CLIP model is used as a feature extractor to obtain multi-modal embeddings for each image-text pair. These embeddings are then passed through a simple classifier head, consisting of an MLP with two hidden layers, to predict the normal/diseased label. The classifier is trained using standard supervised learning techniques while keeping the CLIP weights frozen, a process known as linear probing.

To interpret the model's decisions and provide explainability, Grad-CAM visualizations are generated. Grad-CAM highlights the regions of the input image that contribute most to the predicted class, offering insights into the model's decision-making process. This is particularly important in medical settings, where understanding the reasoning behind predictions is crucial for building trust and aiding clinical decision-making. The proposed pipeline combines the power of CLIP for learning aligned image-text representations with a simple linear classifier for the task of chest X-ray classification. By leveraging pre-trained models and fine-tuning them with a contrastive objective, it obtains strong multi-modal representations that capture relevant information from both visual and textual modalities. The Grad-CAM visualizations further enhance the interpretability and explainability of the model's predictions.

3.5 GradCAM - Visualization and Interpretability

To gain insights into the decision-making process of the trained model, we employ Gradient-weighted Class Activation Mapping (Grad-CAM), a popular technique for visualizing the regions of an input image that are most important for a model's prediction.

Grad-CAM, introduced by [8], works by computing the gradient of the target class score with respect to the feature maps of a convolutional layer in the model. These gradients are then globally average-pooled to obtain importance weights, which are used to generate a coarse localization map highlighting the important regions in the image.

In our proposed methodology, we apply Grad-CAM to the Vision Transformer (ViT) image encoder of the trained CLIP model. Specifically, we compute the gradients of the target class score with respect to the self-attention maps in the last layer of the ViT. These self-attention maps capture the relationships between different image patches and provide a rich representation of the image's spatial structure.

To generate the Grad-CAM visualization, we first forward pass the image through the ViT and obtain the self-attention maps from the last layer. We then compute the gradients of the target class score with respect to these self-attention maps using backward propagation. The gradients are globally average-pooled along the patch dimension to obtain importance weights for each attention head.

Next, we perform a weighted combination of the self-attention maps using the importance weights, resulting in a single attention map that highlights the most relevant regions for the target class. This attention map is then resized to the original image size using bilinear interpolation and overlaid on the input image using a heatmap colormap.

The resulting Grad-CAM visualization provides a transparent and interpretable view of the model's decision-making process, highlighting the regions of the chest X-ray image that contribute most to the predicted class (normal or diseased). This can be valuable for understanding the model's behavior, identifying potential biases, and building trust in the model's predictions. Furthermore, the Grad-CAM visualizations can be used in conjunction with the associated radiology reports to gain deeper insights into the relationship between the visual and textual information. By analyzing the overlap between the highlighted regions in the image and the relevant findings mentioned in the report, we can assess the alignment between the model's predictions and the radiologist's observations.

In summary, incorporating Grad-CAM visualizations into our proposed methodology provides a powerful tool for interpreting the model's decisions and understanding the interplay between the visual and textual modalities. By highlighting the most important regions in the chest X-ray images, Grad-CAM offers a transparent view of the model's decision-making process, enabling clinicians and researchers to better understand and trust the model's predictions. This interpretability is crucial for the adoption and deployment of AI-assisted diagnostic tools in real-world clinical settings.

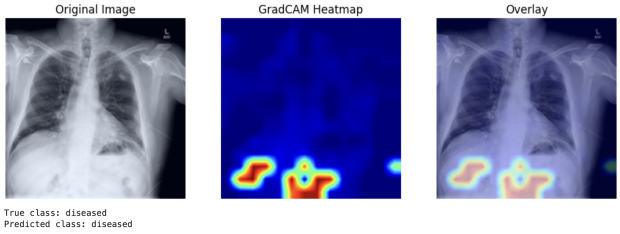

Figure 4: Sample Input and Output

4 EXPERIMENTS

We initially explored ResNet50 and DenseNet121 architectures for the image encoder in CLIP but found that they struggled to generalize to unseen data. This led us to investigate the Vision Transformer (ViT) architecture, utilizing a model with pre-trained weights on

the patch-16-224 dataset. For the text encoder, we employed a BERT model with uncased weights.

Extensive hyperparameter tuning was conducted, exploring batch sizes (32 and 64), learning rates (1e-4 to 1e-5), optimizers (AdamW and Adam) and learning rate schedulers (RandomScheduleron-Plateau and StepLR). The temperature parameter in the contrastive loss function was also tuned, with a value of 0.1 yielding the best results.

The final CLIP configuration achieving the best performance utilized a ViT image encoder with patch-16-224 pre-trained weights, a BERT text encoder with uncased weights, an AdamW optimizer with a learning rate of 0.00001, and a temperature parameter of 0.1. This combination allowed the model to learn meaningful representations from both visual and textual modalities, leading to high classification accuracy.

To adapt CLIP for classifying chest X-rays as normal or diseased, a classifier head consisting of linear layers was added on top of the learned embeddings. This classifier is also trained on similar parameters, a learning rate of 1e-5, with AdamW optimizer, with a batch size of 32 for 25 epochs. All experiments were conducted using Kaggle's P100 GPU.

5 RESULTS AND EVALUATION

5.1 Evaluation Metrics Selection and CLIP Assessment

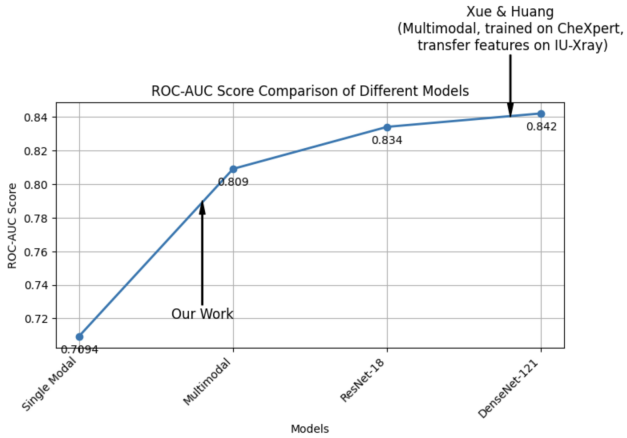

Figure 5: Our model vs SOTA Results

Our choice of evaluation metrics provides a comprehensive assessment of the model's performance across multiple dimensions. We achieved an accuracy of 83.17% on test dataset. The ROC-AUC score (0.8093) is particularly relevant for medical diagnosis tasks as it measures the model's ability to discriminate between classes across different classification thresholds, making it threshold-invariant. This is crucial in medical applications where the trade-off between sensitivity and specificity needs careful consideration.

The confusion matrix-derived metrics (precision, recall, and F1-score) offer complementary insights into the model's performance. In the medical context, recall (sensitivity) of 0.85 for the diseased class is particularly important as it indicates the model's ability to identify positive cases, while precision of 0.89 suggests high confidence in positive predictions. The F1-score (0.87 for diseased class) provides a balanced measure of both precision and recall, which is essential in clinical settings where both false positives and false negatives carry significant consequences.

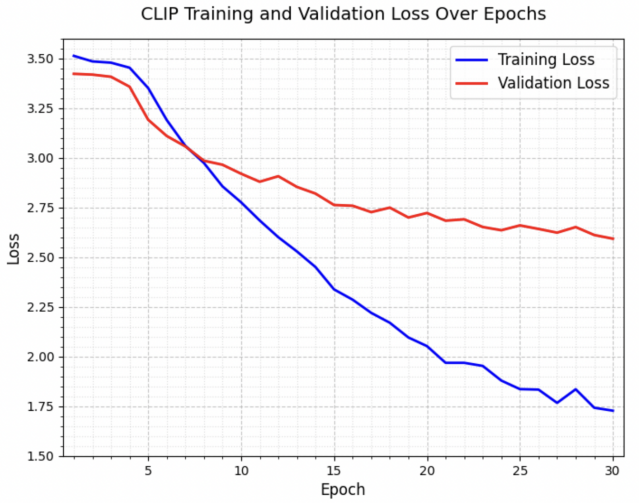

Figure 6: CLIP Training and Validation Loss

While CLIP itself is not directly evaluated using traditional classification metrics, its effectiveness can be assessed through the performance of the downstream classification task. The substantial improvement from single-modal (ROC-AUC: 0.7094) to multimodal (ROC-AUC: 0.8093) performance indicates that CLIP successfully learns meaningful joint representations of images and text. The CLIP training loss curve [6] demonstrates consistent convergence, decreasing from 3.50 to 1.75 over 30 epochs, with the validation loss maintaining a reasonable gap. This convergence pattern, combined with the strong performance of the subsequent classifier, suggests that CLIP effectively captures the semantic relationships between radiographic images and their corresponding medical reports.

Furthermore, the class-specific accuracies (80.11% for normal cases and 84.72% for diseased cases) and the overall accuracy of 83.14% validate CLIP's ability to learn discriminative features that are valuable for the classification task. The performance comparison with established architectures like ResNet-18 and DenseNet-121 which were trained IU X-Ray as transfer learning on top of CheXpert weights, they have achieved an ROC-AUC of 83.4% and 84/2% [4]. It can be seen that our results are pretty close to the SOTA results.

To comprehensively assess the impact of multimodal learning, we developed a single-modal classifier utilizing a Vision Transformer (ViT) initialized with pre-trained weights. Upon evaluation on the test dataset, this unimodal approach yielded an ROC-AUC of 70%. This substantial difference in performance demonstrates the significant advantage of multimodal machine learning, where

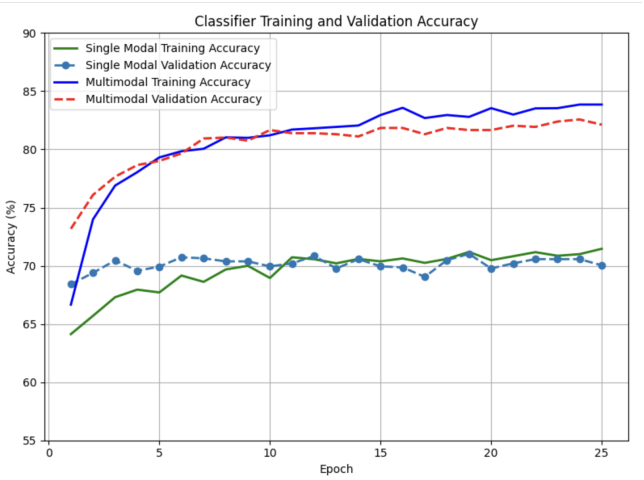

Figure 7: Classifier Accuracy Only Image vs Multimodal

the ROC-AUC is of 80%. By incorporating multiple modalities, our approach not only enhances feature representation but also substantially improves the model's discriminative capabilities, highlighting the potential of integrated learning strategies in medical image classification.

6 CONCLUSION AND FUTURE DIRECTIONS

While our proposed multimodal approach achieved moderate performance with 83.17% accuracy, its key strength lies in successfully demonstrating the value of integrating both visual and textual information in medical image analysis. The model shows that combining radiological images with clinical reports can provide complementary insights, even though there is substantial room for improvement in performance metrics. Future work could explore alternative architectures and fusion techniques, investigate other CLIP-based approaches such as zero-shot classification, and develop models that can generate detailed radiological findings from images. The framework's ability to process both images and text opens up possibilities beyond simple classification, including automated reporting and semantic image retrieval. While our current implementation has limitations, it serves as a proof of concept for multimodal approaches in medical imaging, with potential applications in automated reporting, anomaly detection, and clinical decision support.

7 RELATED PAPERS

(1) O. Alfarghaly, R. Khaled, A. Elkorany, M. Helal, and A. Fahmy, "Automated radiology report generation using conditioned transformers," *Informatics in Medicine Unlocked*, vol. 24, p. 100557, 2021.
(2) J. Jang, D. Kyung, S. H. Kim, H. Lee, K. Bae, and E. Choi, "Significantly improving zero-shot X-ray pathology classification via fine-tuning pre-trained image-text encoders," *Scientific Reports*, vol. 14, no. 1, p. 23199, 2024.
(3) M. Li, B. Lin, Z. Chen, H. Lin, X. Liang, and X. Chang, "Dynamic Graph Enhanced Contrastive Learning for Chest X-ray Report Generation," *arXiv preprint arXiv:2303.10323*, 2023.
(4) Y. Xue and X. Huang, "Improved Disease Classification in Chest X-rays with Transferred Features from Report Generation," *arXiv preprint arXiv:2407.14474*, 2024.
(5) A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 8748-8763, 2021.f
(6) A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *International Conference on Learning Representations (ICLR)*, 2021.
(7) J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4171-4186, 2019.
(8) R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618-626, 2017.
(9) P. Rajpurkar, I. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta, B. Yang, K. Zhu, D. Lungren, and A. Ng, "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," *arXiv preprint arXiv:1711.05225*, 2017.