Interpretable Multimodal Deep Learning Model on MIMIC-CXR Dataset

Zhenghui Chen
Stanford University
450 Jane Stanford Way, Stanford, CA 94305
zhengh04@stanford.edu

Abstract

The advancement of deep learning techniques in both image and text processing has allowed for the creation of multi-modal models that make predictions using multiple forms of data. Seeing as the field of medicine contains a lot of image data and text data, it makes perfect sense to apply multi-modal models in this field with the hopes of improving model performance and results. However, it is not just results that matter, we need to be able to understand the reasoning behind these models' decisions in order to improve clinical trust and mass adoption. This project aims to tackle both these problems at once by applying multi-modal models for biomedical diagnosis of chest-related diseases from the MIMIC-CXR dataset containing chest images and radiology reports. Additionally, model interpretability is a main focus of this project as we will use techniques like Grad-CAM and SHAP values to identify influential features of the images and text that ultimately contribute to a decision. Through comparing unimodal models and their performance with a multimodal model's performance, we can see that the additional information provided through another medium allows for the multimodal model to perform better on the same classification task. Additionally, analyzing our Grad-CAM heatmap on images and SHAP values on radiology reports gives us a useful glimpse into why the model is making the decision that it is. With this project, I aim to demonstrate that interpretable multi-modal models not only improve diagnostic accuracy but can also be confidently interpreted to enhance the transparency and reliability of AI systems in clinical situations.

1. Introduction

The field of medical diagnosis has long relied on expert interpretation of both images and texts but current AI solutions aiming to tackle medical diagnoses only focus on using mainly images. These models have demonstrated great performance using only one stream of data but what would happen if we introduced additional data modalities to make it more life-like? Another problem with current models is the black box nature of them - we provide a medical image and it outputs a prediction instantly. In order to popularize the acceptance of AI diagnosis tools, models also need to be interpretable, there needs to be explanations and reasoning behind decisions made. Espeically with multiple data modalities being presented to a multi-modal model, this interpretability is more important than ever. This project aims to explore the performance of multi-modal models and interpret them. This will be done with the development of a multi-modal deep learning model that processes chest X-ray images and their associated radiology reports from the MIMIC-CXR dataset to make predictions of thoracic disease labels. To do this, our inputs will be images of chest radiographs along with the corresponding radiology report of the image(s). This input is passed into our multimodal model that extracts the features from the images and radiology report to output a multi-label classification vector indicating the presence or absence of 14 thoracic problems. This model aims to do two things. The first being how it compares to unimodal models, which will be found by comparing its performance with models trained on just the images or the texts. Model performance will be determined using the standard metrics of loss and accuracy. The second thing is what methods are the best to interpret these models. To do this, we will experiment with methods such as Grad-CAM for images and SHAP values for text to determine if their outputs provide any sensible explanations for why certain decisions were made by the model.

1.1. Literature Review

To understand more about multi-modal models, we can look at the article titled "A Survey on Deep Multimodal Learning for Computer Vision: Advances, Trends, Applications, and Datasets" by Khaled Bayoudh, which goes over different architecture types, application areas, and present challenges [2]. Deep multi-model models are categorized into early fusion, late fusion, and hybrid fusion models. Since we're working with different modalities of data, which usually require different models to examine

them, we need to combine the things we learn at some point in the model, which is what fusion refers to. Early fusion combines raw features from different modalities at the input level, late fusion combines separate predictions from unimodal models at the decision level, and hybrid fusion fuses data at multiple levels. However, just like all models, these still come with challenges. The main challenge is the difficulties of working with different data types. These include things like difficulties combining different dimensional data, as well as temporal and semantic alignment between different modalities. In addition to these problems, there is an area of research focused on interpreting the reasoning of multi-modal models, which makes them a blackbox still. There is also the problem of scalability, as training and inference on large multi-modal datasets require a lot of computational resources. This article provides us with information necessary to structure and deploy a multi-modal model, as well as explains the interpretability problem and why it is important to work on it.

To better understand our dataset, let's look at an article that has used the dataset to perform a multilabel classification. In the article "NLP-Powered Healthcare Insights: A Comparative Analysis for Multi-Labeling Classification With MIMIC-CXR Dataset" by Ege Erberk Uslu, we see that the researchers decided to leverage NLP techniques to classify 14 distinct radiological findings from radiology reports of the MIMIC-CXR dataset [1]. To do this, the authors compared the performance of multiple transformer-based language models (BERT, BioBERT, ClinicalBERT, and CXR-BERT) and found that the model using CXR-BERT-GENERAL with the BERT classifier achieved the highest weighted F1-score of 0.8047. Even though this article only uses a unimodal model, it still gives us a few important takeaways. Firstly, this article validates text modality as a way to perform a multilabel medical diagnosis classification task. Secondly, it provides us the the best transformer model to work with, which is extremely useful when developing our multi-modal model since we know which text model works the best with the dataset. In another article, "Advancements in Chest Radiography Pneumonia Classification Through Fine-Tuning Using the MIMIC-CXR-JPG Dataset" by Yifan Zhang, Zhang uses the MIMIC-CXR dataset to fine-tune CNNs for the task of pneumonia classification [7]. The model is a CNN built using the FastAI library with a Huggingface pretrained backbone and has a baseline error rate of 0.7688. After fine-tuning with the MIMIC-CXR dataset, the error rate dropped to 0.3133, showing the effectiveness of the MIMIC-CXR dataset and also providing ideas of what architectures to use to analyze the X-ray images. From these articles and others on our dataset, we gain valuable insights into preprocessing techniques, model architectures, and training strategies that guide the development of a multimodal model built on this dataset.

Looking at an article that actually uses a multi-modal model, we can look at "Interpretability-Based Multimodal Convolutional Neural Networks for Skin Lesion Diagnosis" by Sutong Wang [3]. This article explores a multimodal deep learning model for skin lesion classification by combining dermoscopic images with structured clinical metadata. Their model architecture used a fusion method that concatenated CNN-based image features with a simple MLP on structured data before the classification head. This article also highly emphasizes interpretability as it uses Grad-CAM for visualizing important regions in the image and SHAP values for interpreting structured clinical input features. From this article, it is shown that their model outperformed unimodal baselines and the interpretation tools were useful in explaining why a particular prediction was made, which are both great signs for the use of multi-modal models. Additionally, this article provides a blueprint for constructing a multimodal model, specifically the architectural choice of late fusion which is when features are extracted from the image and text and concatenated before the MLP with the idea being that late fusion allows the model to integrate visual and contextual information right before classification to improve predictive performance and maintain interpretability.

From these articles and others attached in our references, we are able to see how others have tackled this problem before us whether it was a unimodal model or a multimodal model. We were able to get an insight into what algorithms/architectures are currently state-of-the-art, such as pretrained BERT models for text-based classification or different CNN models for image-based classification, and also clever approaches such as the late fusion method for a multimodal model, as well as Grad-CAM heatmaps and SHAP values for model interpretability. This project builds upon the ideas of these articles by leveraging proven unimodal models, integrating them into a multimodal model, and using various interpretability methods to make a step in the right direction in the development of transparent AI systems for biomedical diagnoses.

2. Dataset

We are using the MIMIC-CXR v2.0.0 dataset along with the MIMIC-CXR-JPG v2.1.0, which are multimodal datasets containing 377,000 chest X-ray images along with associated radiology reports from over 65,000 patients. The difference between these two datasets is that MIMIC-CXR-JPG contains preprocessed JPG images while MIMIC-CXR contains the images in DICOM form, which is less ideal for machine learning due to its higher difficulty to work with. This dataset was collected at Beth Israel Deaconess Medical Center between 2011 and 2016 and is one of the most popular clinical datasets containing high-resolution images

along with clinical text. For this project, we used a random subsample of 10850 datapoints with a 70/20/10 split for our data into the training, validation, and test sets. A single datapoint in our dataset consisted of a study_id, dicom_id, image_path, radiology_report, which split it belonged to, and its labels. Since this project is trying to tackle multilabel classification, each sample will also have fourteen associated labels from the CheXpert scheme (atelectasis, cardiomegaly, consolidation, edema, enlarged cardiomediastinum, fracture, lung lesion, lung opacity, pleural effusion, pleural other, pneumonia, pneumothorax, support devices, and no finding) where 1 indicates the presence of a label, 0 indicates the absence of a label, -1 indicates uncertainity about the label, and -2 indicating no mention of the label. To preprocess the image data, we resized all X-ray images to 224 x 224 pixels as well as normalized pixel values to [0, 1] to make it compatible with a CNN. No data augmentation was done to the images. To preprocess the text data, we extracted the text from the "Findings" and "Impression" fields (due to these being the most important fields), lowercased characters, and removed punctuation as well as special characters. Below is an example of an unprocessed and preprocessed image and text data that is inputted into our multimodal model.

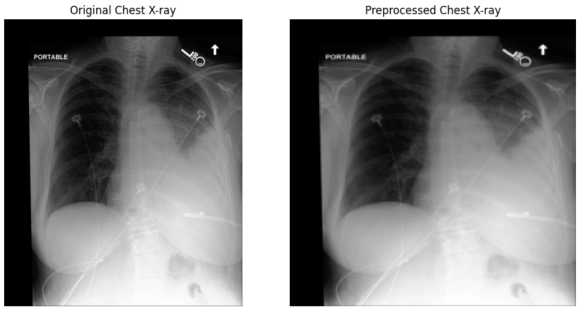

Figure 1. Image Input

```
=== Original Radiology Report ===
                                FINAL REPORT
 EXAMINATION:  CHEST (PORTABLE AP)

 INDICATION:  ___ year old woman with pleural effusion, s/p chest tube placement
 ___  // chest tube interval monitoring: PLEASE DO AT ___      chest tube
 interval monitoring: PLEASE DO AT ___

 IMPRESSION:

 Comparison to ___.  Unchanged position of the left pleural
 pigtail.  Minimal increase in extent of the pre-existing pleural effusion.
 The left perihilar mass and the surrounding parenchymal opacity with air
 bronchograms is constant.  Unchanged normal appearance of the right lung.


=== Cleaned Radiology Report ===
Comparison to ___. Unchanged position of the left pleural pigtail. Minimal increase in extent of the pre-existing p
leural effusion. The left perihilar mass and the surrounding parenchymal opacity with air bronchograms is constant.
Unchanged normal appearance of the right lung.
```

Figure 2. Text Input

3. Methods

Reiterating the problem, we want to input image and text data into a multimodal model to output a vector containing probabilities for 14 disease labels. To do this, I first decided to create unimodal baseline models to provide baseline numbers and also act as building blocks for our multimodal model. For our text-only model to read radiology reports, I decided to use a pretrained BiomedVLP CXR-BERT model, which is a vision language transformer from Microsoft, designed specifically for chest radiology and trained on radiology reports of the MIMIC-CXR dataset. BERT (Bidirectional Encoder Representations from Transformers) is a deep learning model that utilizes bidirectional attention to look at all words in a sentence at once to capture the full context of a text. For classification tasks, such as classifying a radiology report, BERT adds a special [CLS] (classification token) to the beginning of the input text after processing the entire text, which can be treated as a summary of the entire text and passed into a classifier. For our project, each radiology report is tokenized using the BERT tokenizer associated with the pretrained model to output the CLS token, which is then passed to a multi-layer perceptron classifier with a final sigmoid output for each of the 14 diagnostic labels. Since we are using the model as is and we are not training it, we did not use a loss function for this text-only model. I decided on using the BiomedVLP CXR-BERT model since it's already pretrained on domain-specific data, which avoids the need to train a BERT model from scratch, and it also outperforms general-domain BERTs on radiology-specific tasks. For our image-only model to analyze chest X-ray images, I decided to fine-tune a ResNet-18 CNN, which is a well-established CNN that uses residual connections to enable stable training of deeper models and is pretrained on ImageNet. This model contains 1 input convolutional layer, 4 residual blocks which each contains 2 convolutional layers, and 1 fully connected output layer. Additionally, we made other modifications to the model such as adjusting the input layer to just take in grayscale images as well as modifying the classifier head to be compatible with our task. To do this, we modified the first convolutional layer by changing "Conv2d(3, 64, kernel_size=7, stride=2, padding=3)" to "Conv2d(1, 64, kernel_size=7, stride=2, padding=3)" which allows the model to directly process grayscale inputs without channel duplication. For the classifier head, we replaced the original fully connected layer, which had 1000 logits, with a custom fully connected layer that extracted the 512-dimensional image feature vector and projected it to 14 outputs using a linear layer followed by a sigmoid activation to get our probabilities for each label. Because this is a multi-label classification problem, we apply binary cross-entropy loss independently to each label as our loss function. This model is then fine-tuned by training on our dataset using an Adam optimizer, where layers up to and including layer 2 are frozen so we just train layer 3, layer 4, and the final classification head.

After these unimodal models were complete, I moved on to creating the multimodal model. For the model to take in both imaging data and textual information, I planned to use

$$L = \frac{1}{14}\sum_{i=1}^{14} \text{BCE}(y_i, \hat{y}_i) = \frac{1}{14}\sum_{i=1}^{14} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

Figure 3. Binary Cross-Entropy Loss Equation

the two unimodal models as building blocks and combine them to form our multimodal model. The fusion method that I decided to go with was the late fusion paradigm, where each modality is processed independently by its own encoder, and the learned representations are then fused at the decision level. From our image encoder, we can see that the fine-tuned ResNet-18 model yields a 512-dimensional feature vector, and our text encoder, the pretrained BiomedVLP CXR_BERT, outputs a 768-dimensional feature vector, which when concatenated gives us a 1280-dimensional feature vector. This feature vector retains the modality-specific features, which are then input into a 2-layer feedforward network to output the final 14 sigmoid-activated predictions for our multi-label classification task. This model also uses the binary cross-entropy loss that is used in the unimodal image model to allow for comparability between unimodal and multimodal results. This model was then trained using the Adam optimizer under similar conditions to the ResNet-18 model, except no layers were frozen. Late fusion was chosen because of its modular design, allowing for the reuse of pretrained models and easy use of interpretability methods on each individual encoder, and also because it requires minimal preprocessing compared to early or joint fusion methods, which may require aligning modalities in space.

After building our multimodal model, we now want to interpret it. To do this, we use Grad-CAM (Gradient-weighted Class Activation mapping) on image interpretability and SHAP values for text interpretability. The high-level way Grad-CAM works is by highlighting the spatial regions that the model "looks at" when predicting a particular condition. Going deeper, we calculate how important each filter in the last convolutional layer is for the class we're interested in. This is done by looking at the gradient output score with respect to that filter's activation map and averaging it over the whole spatial area. Then, each filter's activation map is weighted by its importance and added altogether, which creates a single map that shows where the model was looking to make its decision. The ReLU function is then applied to keep only the parts that positively contribute to the prediction, since we want to know what spatial regions support the model's decision. This map is then resized to match the original image size and overlaid as a heatmap on top of the image to provide a visual explanation of where the model focused its attention for that specific prediction. The high-level way SHAP values work is by assigning each word of a text a Shapley value to indicate its contribution to the model's prediction. Going deeper, SHAP determines this by masking each word and observing how the model's prediction changes. If removing a word significantly changes the prediction, the word is likely to be important. SHAP also evaluates many different permutations of words to see how the model behaves, since the impact of the word can also depend on the context. Therefore, for each word, SHAP calculates an average effect on the prediction across these permutations, which is our SHAP value. This value tells us how much that specific word contributed to the final prediction positively or negatively. This is then used to generate color-coded visualizations where red words are associated with pushing the model towards the prediction and blue words are associated with pushing it away the prediction.

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right)$$

Figure 4. Grad-CAM Importance Weight and Applied ReLU

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} [f(S \cup \{i\}) - f(S)]$$

Figure 5. Shapley Value ϕ_i for i-th Token in Text

4. Results

As mentioned above, we trained three models, the ResNet-18 CNN (image-only model), a BiomedVLP CXR-BERT (text-only model), and the late fusion multimodal model. These models were all trained on our 10850 datapoints which included grayscale X-ray images, their corresponding radiology reports, and the labels for the 14 chest conditions. As mentioned in our methods, the BiomedVLP CXR-BERT model was pretrained and used as is so there weren't any hyperparameters associated with it or any training involved. For our CNN model and multimodal model, it was trained using the AdamW optimizer which decouples weight decay from the gradient update step which improves generalization compared to the standard Adam optimizer. For our learning rates, we chose to use a learning rate of 1 * 10e-5 since it allowed for stable and efficient training and led the model to convergence consistently. For our batch size, we went with a batch size of 32 to balance speed and GPU constraints. Both models were trained for 10 epochs since anything more or less led to undertraining and overfitting respectively. Both models also used a dropout of 0.5 to reduce overfitting in the models.

For our evaluation metrics, we mainly used accuracy and loss to evaluate model performance. However, we

used a custom formula for accuracy and loss as we chose to ignore missing labels, which we filled with -2. Firstly, we computed the binary cross-entropy loss only on valid labels, ignoring labels with -2. This way, the model is not penalized for predictions on missing labels.

```
def masked_bce_loss(preds, targets, ignore_val=-2):
    mask = (targets != ignore_val).float()
    targets_clamped = torch.clamp(targets, 0, 1)
    loss = F.binary_cross_entropy(preds, \
    targets_clamped, reduction='none')
    masked_loss = (loss * mask).sum() / mask.sum()
    return masked_loss
```

We also modified the accuracy function to compute the mean per-label accuracy for a multilabel classification task while excluding labels with -2. Additionally, since the models are outputing probabilties, we consider a prediction for a label to be correct if the sigmoid output exceeds a treshold of 0.5 and matches the ground truth labels.

```
def masked_accuracy(preds, targets, ignore_val=-2,\
threshold=0.5):
    mask = targets != ignore_val
    preds_bin = (preds > threshold).float()
    accs = []

    for i in range(targets.shape[1]):
        col_mask = mask[:, i]
        if col_mask.sum() < 2:
            continue
        acc = accuracy_score(targets[col_mask, i]\
        .cpu(), preds_bin[col_mask, i].cpu())
        accs.append(acc)

    return float(np.mean(accs)) if accs else 0.0
```

Using these evaluation metrics, we can see that our fine-tuned ResNet-18 model acheived a training loss of 0.2877 and a training accuracy of 0.7648 along with a validation loss of 0.4436 and a validation accuracy of 0.7047. Our multimodal model achieved a training loss of 0.1302 and a training accuracy of 0.8296 along with a validation loss of 0.2707 and a validation accuracy of 0.7916. On the test set, our multimodal model acheieved a test loss of 0.2592 and a test accuracy of 0.7950. From this, we can see that our multimodal model had a 10% better performance than the unimodal model suggesting multimodal models and data allow for better performance than unimodal models and data. We measured accuracy differently for our pretrained Biomed-VLP CXR-BERT as we computed the per-label accuracy by comparing predicted and ground truth binary labels for each disease label rather than the entire vector of labels as we did with the CNN and mutlimodal model.

In addition to just models, we utilized interpretability methods of Grad-CAM and SHAP values to see if they would provide any useful information on decisions that the models made. For Grad-CAM, we only created heatmaps for the top-3 highest scoring labels and their predicted probabilties since creating 14 for each image would be too much. In an image that we look at, we can see that our code outputs an X-ray image along with the three Grad-CAM heatmaps for the top three labels (no finding, cardiomegaly, and atelectasis). In these images, the attention is focused on the central thoracic cavity which aligns well with clinically relevant regions for the conditions of the labels. This visualization shows that the model is not only accurate but also provides interpretable reasoning for why it is making the decisions that it does. For the SHAP values, we the same approach of only calculating SHAP values for the top-3 highest scoring labels. The output of the SHAP values are visualized with a color-coded attribution map where red tokens are tokens that push the model towards predicting a label and blue tokens push it away from predicting that label. In our plot, we see that our model finds the SHAP values for the three classes of no finding, pleural effusion, and fracture. From this text "Severe cardiomegaly is unchanged as well as bilateral pleural effusions. There is no pneumothorax. Mild vascular congestion is re- demonstrated with no substantial change since the prior study", we can see that the word "no" is important in pushing the model to classify the text as no finding. Nextly, the words "pleural effusion" are pushing the model to classifying it as such. From this, we can see how these SHAP values allow us to understand which words are important in making a decision as well as not making a decision. Therefore, from both of these interpretability values, we can see that the model is focusing on clinically relevant features in both modalities which allows us to put a lot more trust into the diagnostic predictions.

As explained, the model does begin to overfit when we train on more than 10 epochs but it does seem that 10 is the number where we maximize the validation and test accuracy before overfitting begins. This could be because there is not enough training data to allow for improved performance as the model is not seeing enough information from the 10850 images and texts. A remedy to this will be elaborated in our conclusion.

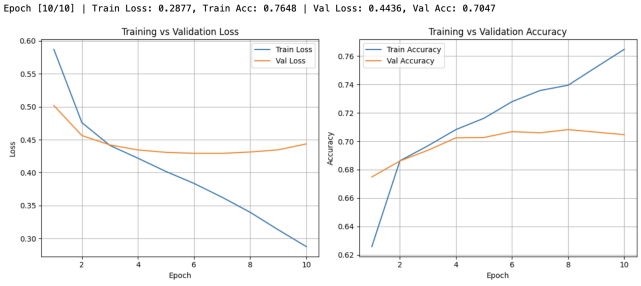

Figure 6. Loss and Accuracy Curves for Image-Only Model

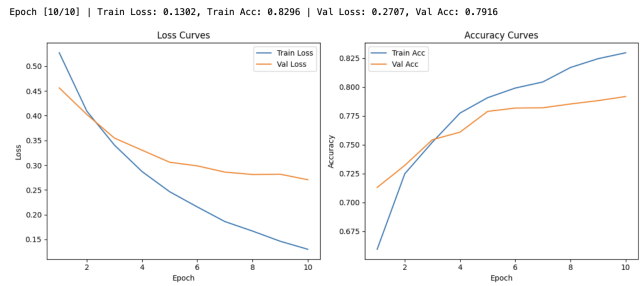

Figure 7. Loss and Accuracy Curves for Multimodal Model

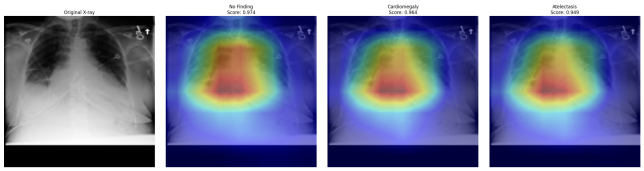

Figure 8. Original X-ray Image and GradCAM Heatmaps

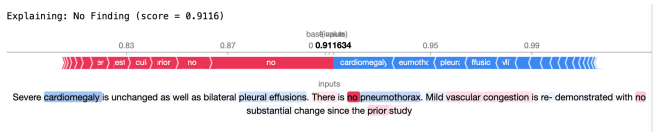

Figure 9. SHAP Values for No Finding

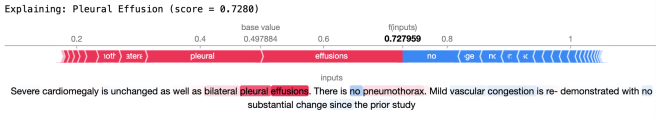

Figure 10. SHAP Values for Pleural Effusion

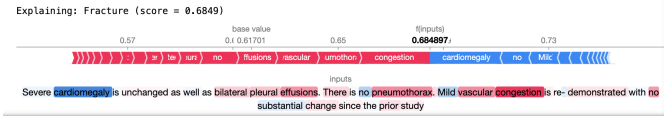

Figure 11. SHAP Values for Fracture

5. Conclusion

In this project, we developed an interpretable deep learning model for multilabel classification of chest X-ray images and radiology reports from the MIMIC-CXR dataset. To tackle this problem, we implemented three models: a text-only model using a pretrained BiomedVLP CXR-BERT encoder, an image-only model using a fine-tuned ResNet-18 model, and a multimodal model that fused both image and text embeddings using a late fusion method. Additionally, using Grad-CAM and SHAP value methods, we were able to get a glimpse into why our multimodal model made the decisions it did, making it more interpretable.

From the three architectures, the multimodal model achieved the best performance, with an accuracy of 79.5% compared to the CNN model's performance of 70.47%. This performance improvement highlights the usefulness of multimodal data and multimodal models and highlights the complementary nature of radiology reports and imaging data as X-rays provide spatial evidence while textual reports often contain nuanced clinical summaries which work together to provide more context for our multimodal model that unimodal models might not catch.

For future works, I would attempt to improve the model by trying different fusion strategies such as early and hybrid methods to test which strategies are best in improving model performance. Additionally, I would train on the full dataset of 377,000 data points rather than our 10850 which is a very small subsample. This would hopefully improve model generalization and allow the network to better capture less common patterns present in underrepresented classes. Outside of model improvement, I would want to use clincian/radiologist feedback to validate interpretability outputs to make sure that they hold clinical importance. With these next steps on this project, we can go one step further in making transparent AI-assisted medical diagnosis tools a reality.

References

[1] E. E. Uslu, E. Sezer, and Z. A. Guven, *NLP-Powered Healthcare Insights: A Comparative Analysis for Multi-Labeling Classification With MIMIC-CXR Dataset*, IEEE Access, vol. 12, pp. 67314–67324, 2024. doi: 10.1109/ACCESS.2024.3400007.

[2] K. Bayoudh, R. Knani, F. Hamdaoui, and A. M. Alimi, *A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets*, Visual Computer, vol. 38, pp. 2939–2970, Aug. 2022. doi: 10.1007/s00371-021-02166-7.

[3] S. Wang, Y. Yin, D. Wang, Y. Wang, and Y. Jin, *Interpretability-Based Multimodal Convolutional Neural Networks for Skin Lesion Diagnosis*, IEEE Transactions on Cybernetics, vol. 52, no. 12, pp. 12623–12637, Dec. 2022. doi: 10.1109/TCYB.2021.3069920.

[4] H. N. Saleem, U. U. Sheikh, and S. A. Khalid, *Classification of Chest Diseases from X-ray Images on the CheXpert Dataset*, in *Innovations in Electrical and Electronic Engineering*, S. Mekhilef, M. Favorskaya, R. K. Pandey, and R. N. Shaw, Eds., Lecture Notes in Electrical Engineering, vol. 756, Springer, Singapore, 2021. doi: 10.1007/978-981-16-0749-3_64.

[5] W. Chai and G. Wang, *Deep Vision Multimodal Learning: Methodology, Benchmark, and Trend*, Ap-

plied Sciences, vol. 12, no. 13, p. 6588, 2022. doi: 10.3390/app12136588.

[6] D. Huang, C. Yan, Q. Li, and X. Peng, *From Large Language Models to Large Multimodal Models: A Literature Review*, Applied Sciences, vol. 14, no. 12, p. 5068, 2024. doi: 10.3390/app14125068.

[7] Y. Zhang, *Multimodal Machine Learning for Automated Medical Image Diagnosis*, Technical Report, Department of Computer Science, Northwestern University, 2022. [Online]. Available: https://www.mccormick.northwestern.edu/computer-science/documents/yifan-zhang-tr.pdf

[8] K. He, X. Zhang, S. Ren, and J. Sun, *Deep Residual Learning for Image Recognition*, arXiv preprint arXiv:1512.03385, 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[9] A. Mahbod, G. Schaefer, I. Ellinger, R. Ecker, A. Pitiot, and C. Wang, *Fusing fine-tuned deep features for skin lesion classification*, Computerized Medical Imaging and Graphics, vol. 71, pp. 19–29, Jan. 2019. doi: 10.1016/j.compmedimag.2018.10.007.

[10] S. Lundberg and S.-I. Lee, *A Unified Approach to Interpreting Model Predictions*, arXiv preprint arXiv:1705.07874, 2017. [Online]. Available: https://arxiv.org/abs/1705.07874

[11] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-Y. Deng, R. G. Mark, and S. Horng, *MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports*, Scientific Data, vol. 6, p. 317, Dec. 2019. doi: 10.1038/s41597-019-0322-0.