Phase 2 Report

Manjit Ullal, Oliver Spohngellert, Saurabh Parkar

April 28, 2021

1 Motivation

Lung diseases are some of the most common medical conditions in the world. Tens of millions of people contract a lung disease every year in the U.S. alone. Altogether, lung diseases accounted for more than 1 million deaths in the U.S. in 2010, according to the NHLBI. The main goal of this project is to build a medical image classifier that detects Lung diseases. It can be challenging to diagnose these diseases, as it requires expertise and resources, which could be in the form of personnel or in the form of medical equipment. These resources are scarce, especially in developing countries. We aim to develop a low cost solution to diagnose these diseases using Deep Learning.

2 Phase 1 results

In phase 1 we aimed to be able to detect COVID-19 and Pneumonia using the COVIDx dataset. We achieved a best AUC score of 0.987, but we showed that this performance did not transfer well to our custom NIH test set. Due to this, we decided to move to using the NIH Chest X-Ray dataset.

3 Methods

3.1 Datasets

We used two datasets in our project: the NIH Chest X-Ray dataset for disease classification, and Pulmonary Chest X-Ray dataset for lung segmentation.

NIH Chest X-Ray

The NIH Chest X-Ray dataset contains 112,120 images from 30,805 patients. This dataset is labeled in a multi-task way, meaning each image can contain multiple (or no) diseases. In total there were 14 total disease labels, including pneumonia, emphyzema, effusion, atelectasis and more. See Table 3 for the number of images for each disease.

Pulmonary Chest X-Ray

This set contains 138 posterior-anterior x-rays, of which 80 x-rays are normal and 58 x-rays are abnormal with manifestations of tuberculosis. Overall, there are 318 image and mask pair in train set and 35 in test set.

3.2 COVID-Net Evaluation - *Low Risk*

As a leftover from phase 1, we decided to evaluate the COVID-Net model, which was created by the researchers who created the COVIDx dataset [5]. We used similar methods as in phase 1, evaluating its performance on the COVIDx test set, the NIH test set we created, and creating CNN feature visualizations.

3.3 Multi-Label Classification - *Medium Risk*

For Multi-Label Classification, our general methodology involved a train-test-val split of 80-10-10. We also used data augmentations such as random rotation, horizontal flip, brightness, contrast, and hue saturation. Using these augmentations allowed the model to learn features invariant to these conditions.

Baseline Models

As a baseline, we trained various common CNN architectures on the NIH dataset. This included VGG16, Efficient-net, Resnet-101, and Resnet-152. All of these except Resnet-101 were trained with Binary Cross Entropy Loss, and Resnet-101 was trained with Asymmetric Loss. Training was stopped when validation performance plateaued.

Resnet-101 with Asymmetric Loss

An Image contains on average few positive labels, and many negative ones. Positive-negative imbalance dominates the optimization process, and can lead to under-emphasizing gradients from positive labels during training, resulting in poor accuracy. The Asymmetric loss [4] enables to dynamically down-weights and hard-thresholds easy negative samples, while also discarding possibly mislabeled samples and helps improve model performance.

Attention models

After setting the baselines we selected the best performing model to construct 2 neural network architectures based on implicit and explicit attention mechanisms.

Implicit Attention (Attention Guided CNNs)

In this architecture (Figure 6), we train a Resnet model over the global xray image and then extract a patch from the original image where the excitation was the highest (using CAMs) [2]. Next, we train another Resnet model over these patches and concatenate the feature vectors from both these networks to train it over a feed forward neural network. This helps the model to focus on the disease specific region of the image and avoid noisy features.

Explicit Attention (Squeeze and Excitation)

The squeeze and excitation block (Figure 7) is a module that can be added to any neural network without changing the actual architecture [3]. This block downsamples and upsamples the feature vector from the last layer and computes a weighted feature vector based on the importance of different feature maps. This helps to include channel-wise feature dependencies and pay more attention to the discriminitive regions of the image.

3.4 Image Segmentation - *High Risk*

The main goal of the use of segmentation was to analyze the impact data quality has on classification performance. Successful segmentation will lead to removal of noise in the images, leading to better classification. After the segmentation we analyze the output of segments and investigate how model loss and the relevant labels of the images are impacted by the segment.

4 Results

4.1 COVID-Net

As can be seen in Figure 1, the COVID-Net performance closely follows those of the models we trained in phase 1. The model performs much better on COVIDx than on NIH. Further, based on Figure 2, the model sometimes seems to use relevant information but othertimes does not. Based on all of this, we can make the conclusion the COVID-Net model suffers similar deficiencies to the models we trained in phase 1.

4.2 Multi-Label Classification

4.2.1 Baseline Models

As can be seen in Table 1, each of the models gave promising performance with Resnet-152 performing the best. The models did not perform the same on all classes, see Figures 3, 4, 5 for examples.

Model	VGG16	Efficient-Net	Resnet-101	Resnet-152
AUC	0.85	0.84	0.89	*0.92*

Table 1: Performance of Baseline Models

There is a general trend among the models, where some of the classes are easier to classify compared to the rest. VGG and Efficient-net achieve reasonable results with mediocre performance on half the classes. Resnet-101 performs better with Asymmetric loss (Fig 5). It is likely that performance would improve if there were more images with multiple labels. Best performance is from Resnet-152.

4.2.2 Attention Models

Model	**Atel**	**Card**	**Effu**	**Infi**	**Mass**	**Nodu**	**Pne1**
Resnet-152	0.814	0.907	0.878	0.704	0.833	0.768	0.759
Resnet-152 with Implicit Attention	**0.824**	**0.916**	**0.885**	0.706	**0.844**	**0.773**	**0.767**
Resnet-152 with Explicit Attention	0.816	0.913	0.883	**0.712**	0.833	0.754	0.757
Model	**Pne2**	**Cons**	**Ede**	**Emp**	**Fibr**	**Pleu**	**Hern**
Resnet-152	0.863	0.793	0.890	0.924	**0.834**	0.779	0.881
Resnet-152 with Implicit Attention	0.867	**0.811**	0.891	0.914	0.823	**0.785**	**0.923**
Resnet-152 with Explicit Attention	**0.880**	0.804	**0.895**	**0.925**	0.826	0.776	0.890

Table 2: ROC AUC scores corresponding to the different conditions

Based on the above results, the model with attention mechanism almost always performs better than the Vanilla Resnet. Moreover, implicit attention works slightly better than explicit attention. We further visualized these 2 models to better interpret the results using class activation maps (Figure 8). The figures show that explicit attention picks up some signal from the non lung regions as noise which was the same problem we faced in phase 1. Whereas, the implicit model completely eliminates this problem. A benefit that the explicit attention can provide is by applying equal attention to all the regions of high excitation rather than just a single large region.

4.3 Image Segmentation

Unet model [1] was used for X-ray segmentation. The best performance being, AUC score of 0.97 (Fig 10). We can see that the segmentation model works excellent on the Pulmonary dataset which had the true masks (Fig 11). So we use this model to investigate the NIH dataset (Fig 12). Clearly the output is not the best. Since the pulmonary dataset has only 2 conditions, while NIH has 14, we needed to experiment with different thresholds for optimal performance, and settled on 0.1. Since NIH dataset has multiple conditions, the segmentation's may not be the best for classification, and using just that did not improve model performance on classification.

Quality of Images

In Fig 12, we have presented two sets of images. The Image on top is where the original image is clear and hence the segmentation is clear. However the segmentation suffers in the bottom image due to poor quality. This confirms our hunch that image quality is not standard across all image classes.

Quality of Information

Segmented output is a good proxy for the quality of the original image. In Fig 13 we capture the segmented area for all the images on the test set and rank them, and investigate them based on the percentage of segmented area. It's not surprising to see lower segmented area corresponds to the model not being able to capture any quality information. It is interesting to note that higher segmented area does not necessarily mean better information as we can see in the bottom image that it includes a lot more area not relevant for classification.

To understand the effect of image quality on the predictions we plotted the two on a graph (14). It was observed that, as the percent information in an xray increases, the hamming loss for that prediction decreases. Furthermore, the images with fewer conditions had a lower loss even for xrays with very low information (poor quality). Lastly, as the number of conditions in a given image increases, the model finds it difficult to predict all of them correctly increasing the hamming loss. These observations show that image quality plays an important role in the above tasks.

5 Conclusion

Over the course of the semester we have tried to lay a solid ground work on diagnosing lung diseases using deep learning models. We started with COVID-Net model. We can see that the COVID-Net model does not do as well on the NIH dataset as it did on the COVIDx test set. However, we can not fully conclude whether this is due to problems with the network, or that the NIH set is more difficult. Further, we achieved very good results in multi-label classification using standard models and using attention. Implicit attention achieved the best results by attenuating the signal from noisy regions in the xray images. Finally, image segmentation helped us understand how image quality impacts the ability for the model to make predictions. In conclusion, we were able to classify diseases on chest X-Rays using Deep Learning techniques.

6 Bibliography

References

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *CoRR* abs/1505.04597 (2015). arXiv: 1505.04597. URL: http://arxiv.org/abs/1505.04597.

[2] Qingji Guan et al. *Diagnose like a Radiologist: Attention Guided Convolutional Neural Network for Thorax Disease Classification*. 2018. arXiv: 1801.09927 [cs.CV].

[3] Jie Hu et al. *Squeeze-and-Excitation Networks*. 2019. arXiv: 1709.01507 [cs.CV].

[4] Emanuel Ben-Baruch et al. *Asymmetric Loss For Multi-Label Classification*. 2020. arXiv: 2009.14119 [cs.CV].

[5] Linda Wang, Zhong Qiu Lin, and Alexander Wong. "COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images". In: *Scientific Reports* 10.1 (Nov. 2020), p. 19549. ISSN: 2045-2322. DOI: 10.1038/s41598-020-76550-z. URL: https://doi.org/10.1038/s41598-020-76550-z.

7 Figures and Tables

	Atelectasis	Cardiomegaly	Consolidation	Edema	Effusion
Count	11559	2776	4667	2303	13317
	Emphyzema	**Fibrosis**	**Hernia**	**Infiltration**	**Mass**
Count	2516	1686	227	19894	5782
	Nodule	**Pleural Thickening**	**Pneumonia**	**Pneumothorax**	
Count	6331	3385	1431	5302	

Table 3: Counts of each disease in the NIH Chest X-Ray Dataset

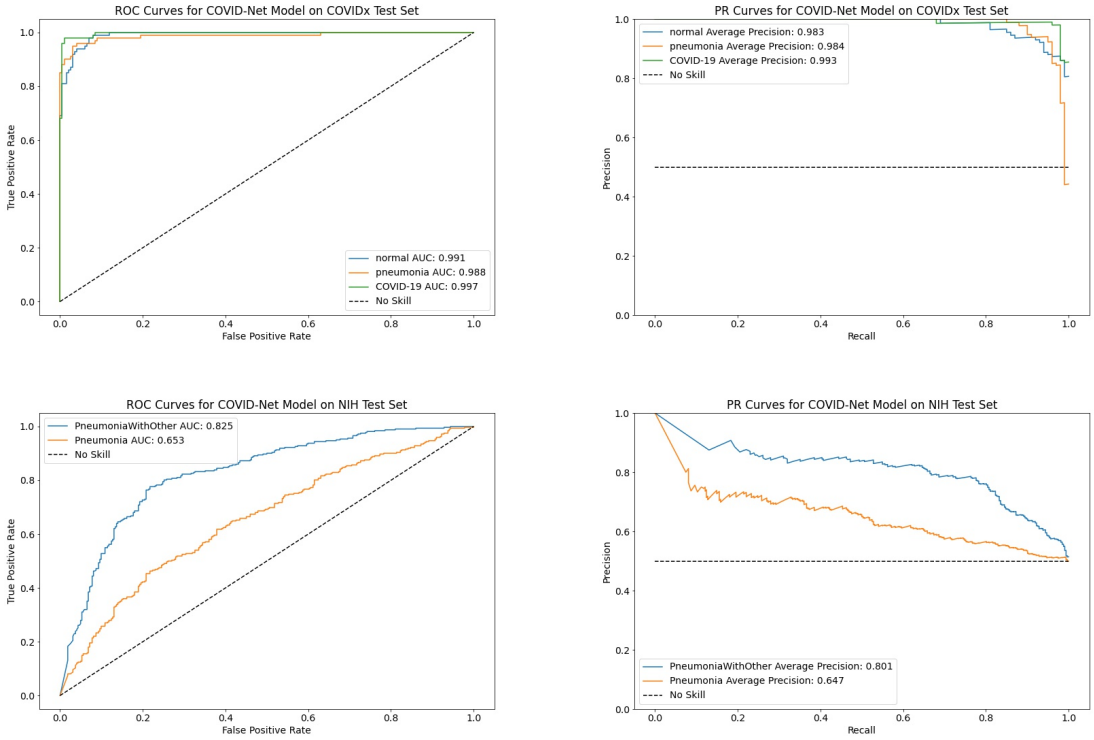
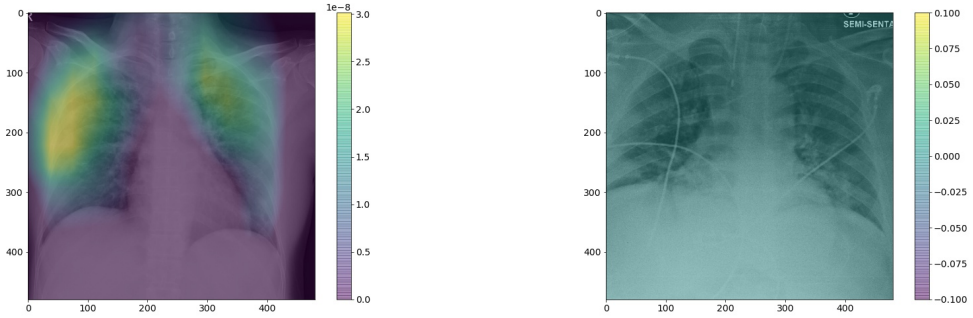
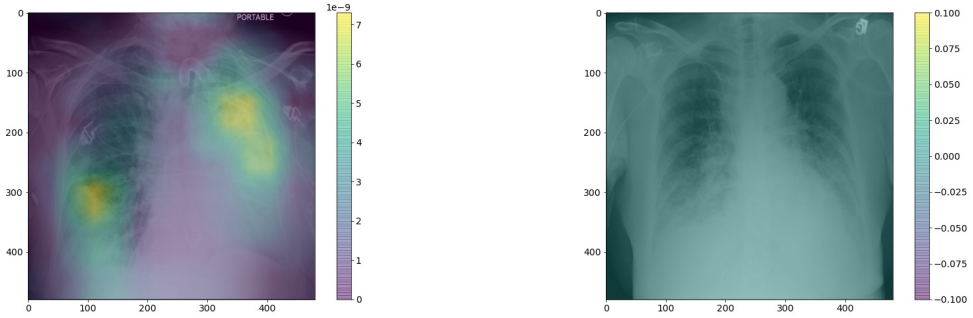


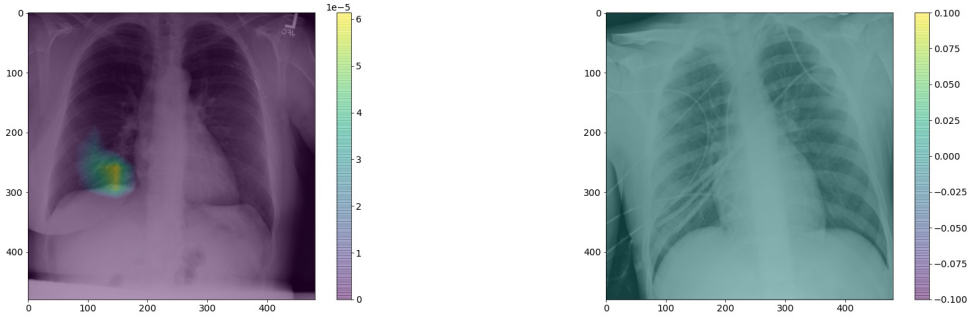
Figure 1: ROC and PR curves for COVID-Net Model



(a) COVID-19 Visualizations



(b) Pneumonia Visualizations



(c) Normal Visualizations

Figure 2: CNN Visualizations for COVID-Net Model

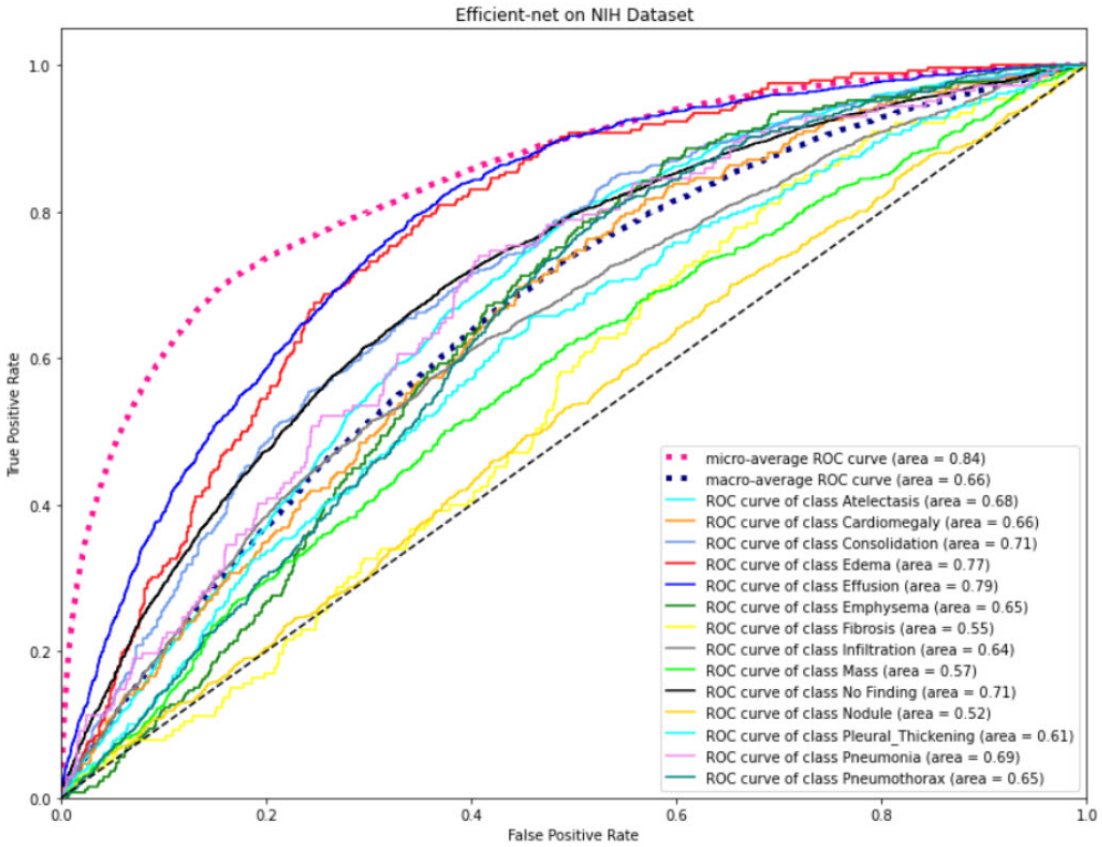


Figure 3: ROC for Efficient-net Model

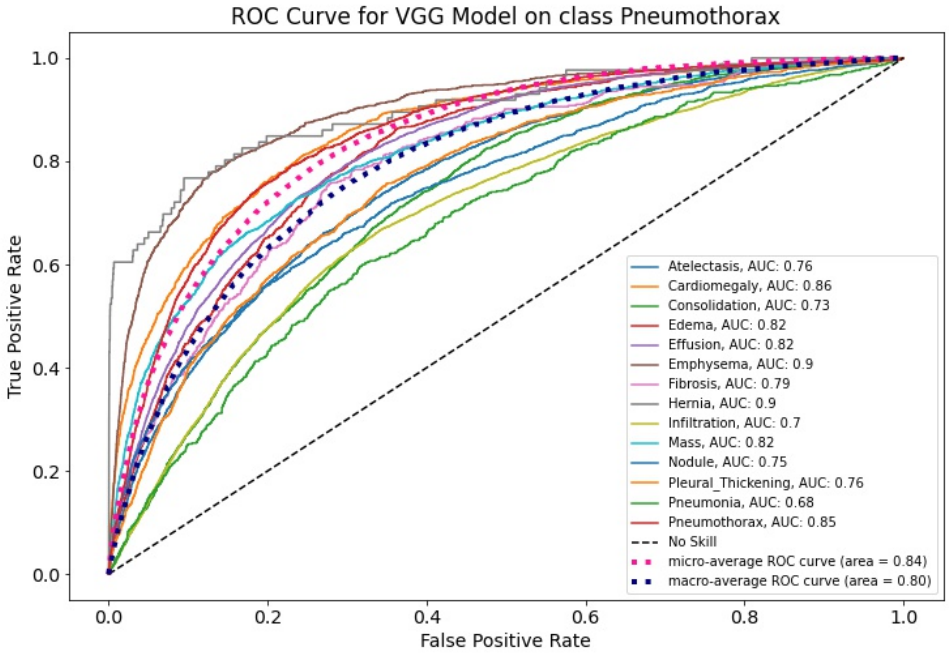


Figure 4: ROC for VGG16 Model

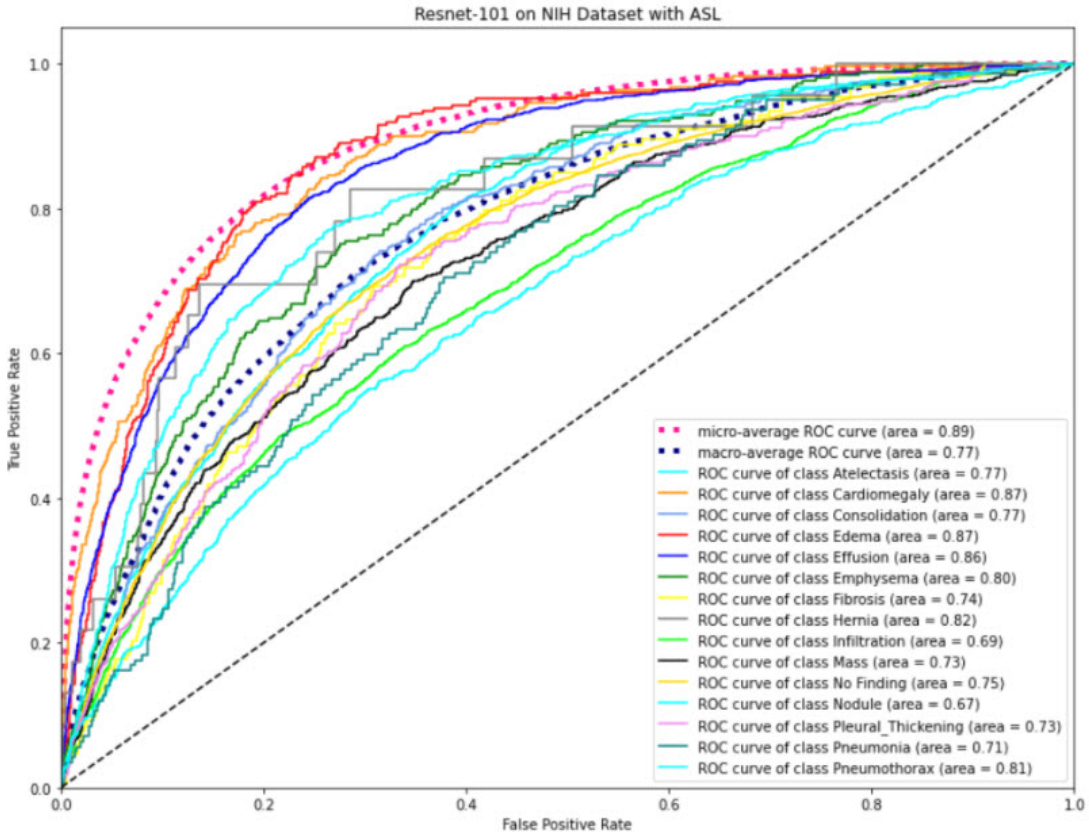


Figure 5: ROC for Resnet-101 with ASL Model

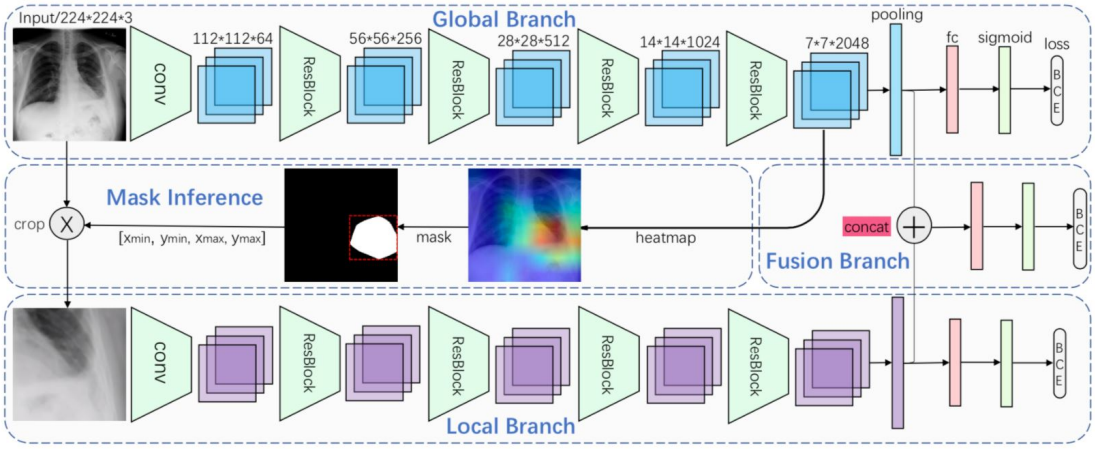


Figure 6: Attention Guided Convolutional Neural Network

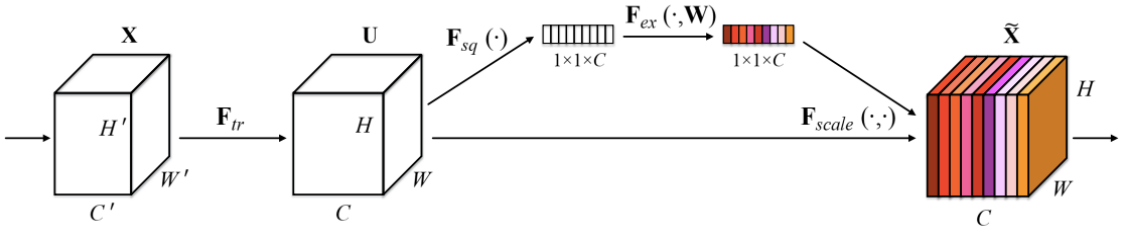


Figure 7: Squeeze and Excitation Networks

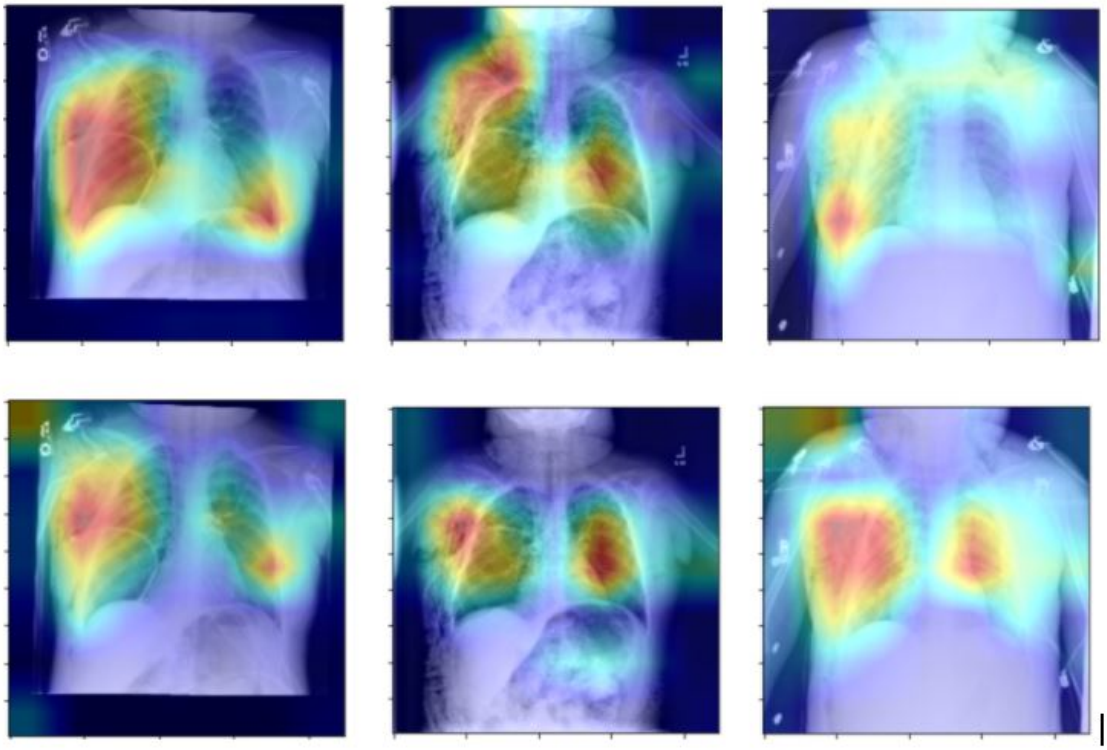


Figure 8: Class Activation Maps from Implicit Attention (Top Row) and Explicit Attention (Bottom Row)

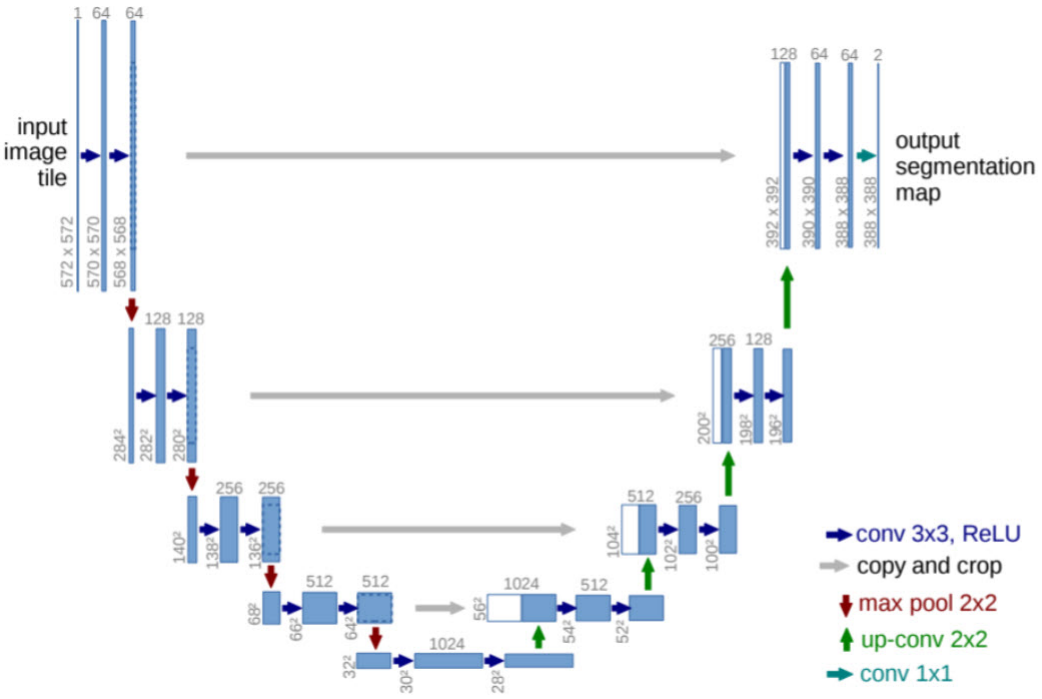


Figure 9: Unet Architecture

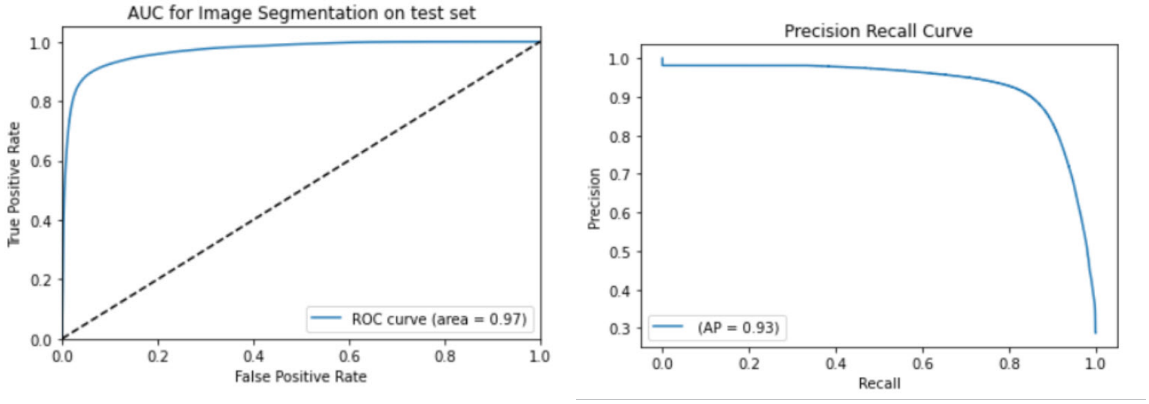


Figure 10: Unet Performance

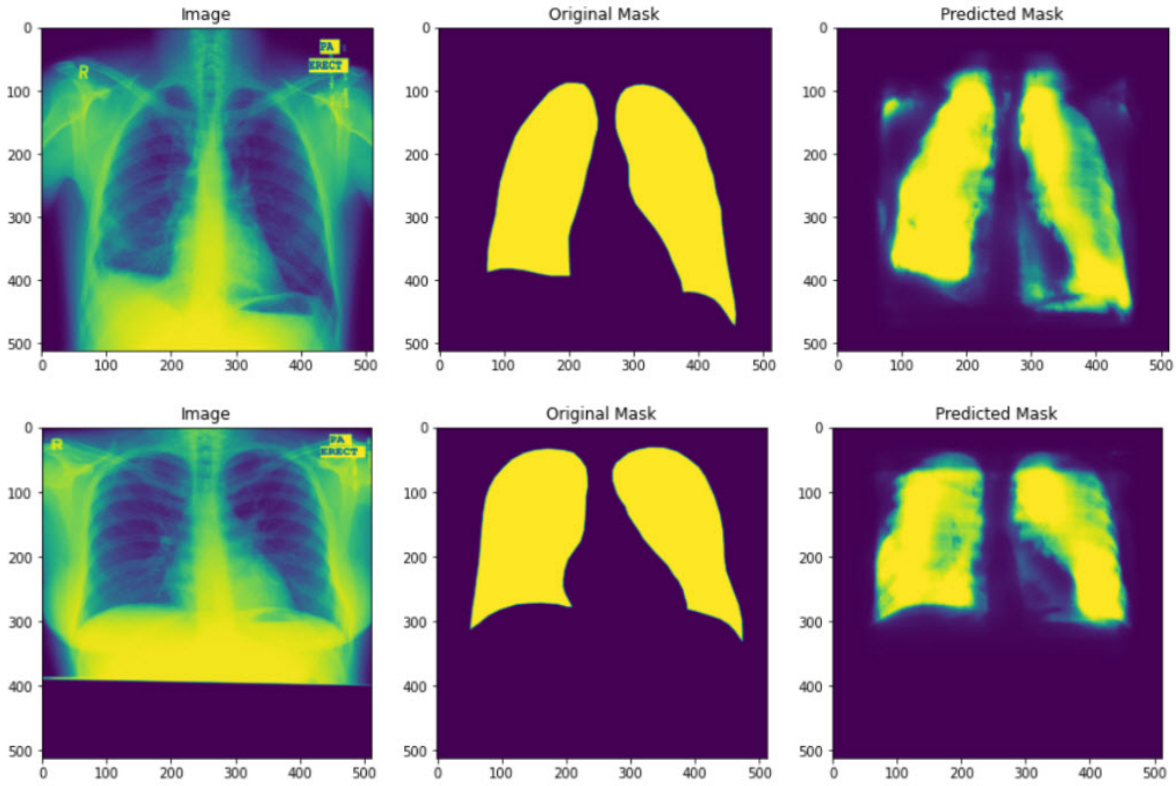


Figure 11: Image Segmentation on Pulmonary X-ray dataset

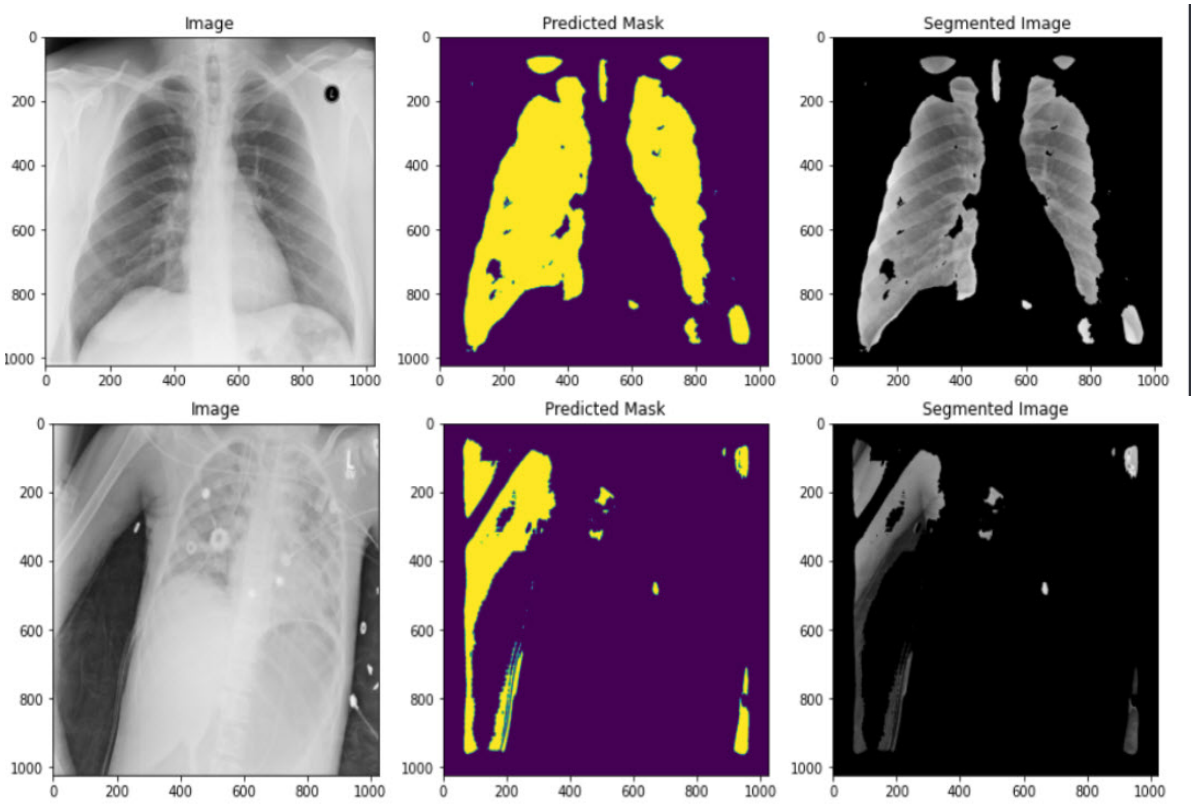


Figure 12: Image Segmentation on NIH X-ray dataset

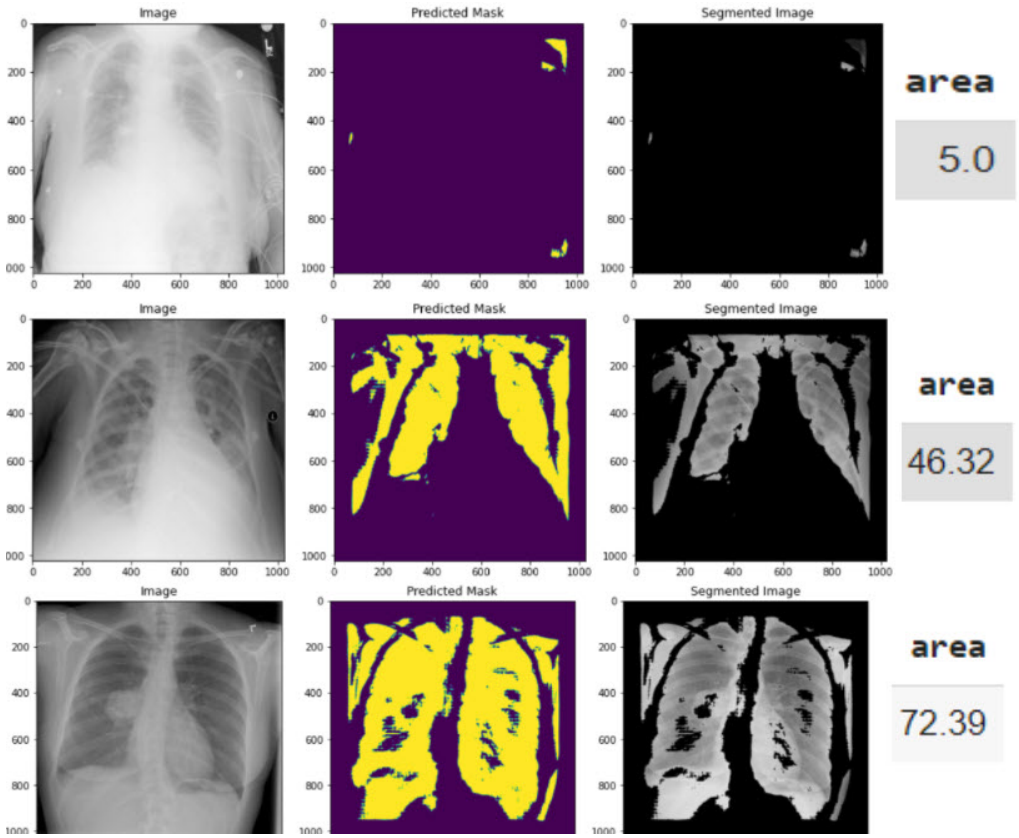


Figure 13: Segmentation Area Comparison

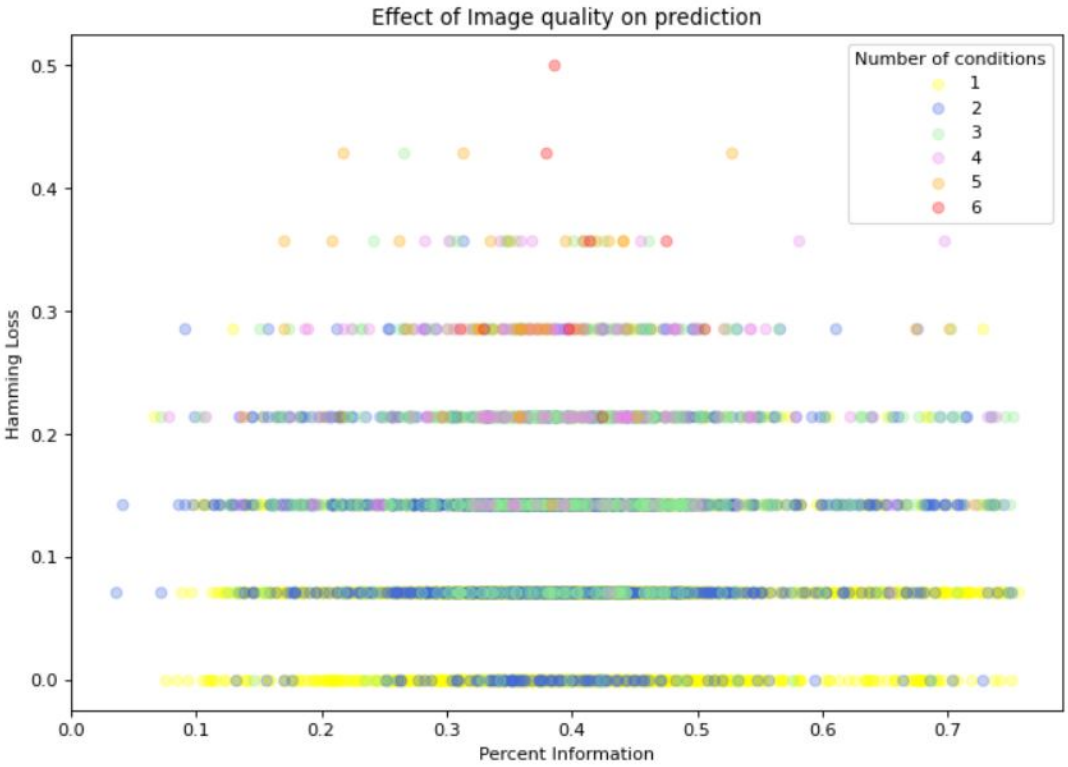


Figure 14: Effect of Image Quality on Predictions