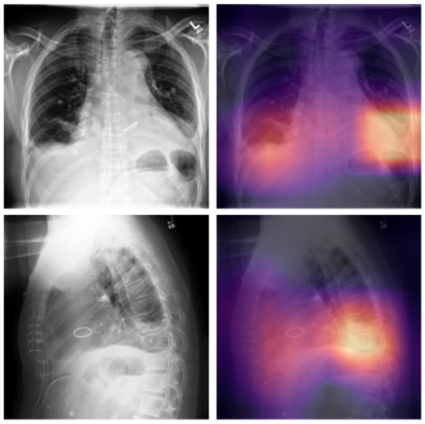
Modern Methods for Automation of Thoracic Radiography Interpretation in the Face of Uncertainty

Derek Levesque, Candidate for Masters in Analytics Degree
Georgia Institute of Technology, Atlanta, Georgia

https://github.gatech.edu/dlevesque3/CSE6250_final_project_code

https://youtu.be/cv-CD_DgYsM



Abstract

Thoracic radiography is the most common imaging examination globally, critical for screening, diagnosis, and management of many life threatening diseases, including COVID-19. Automation of Chest X-rays is considered a 'low-hanging fruit' in terms of the interest and motivation that exists towards this problem. Numerous approaches and studies have already been conducted for years and as such, there is not much room for an entry-level machine learning practitioner to reinvent the wheel and discover new approaches. In that vein, I decided to replicate the existing baseline approach of a convolutional neural network with the hope of then discovering possible new approaches in my next iteration. I will also incorporate Scala and Hadoop for the deep learning portion and will perform the data processing stage using Amazon AWS. I will focus on generalizability and test the model on different publicly available datasets and will also experiment with different methods of dealing with the uncertainty in the training labels which exists due to the automated labeling approach needed to deal such a large quantity of images in addition to the inherent uncertainty in any medical diagnosis.

Background

X-rays are the oldest and most common form of medical imaging in medical systems across the globe. The technology itself was discovered in 1897 and was originally used to detect military battlefield injuries. Imaging with x-rays involves directing ionizing radiation to a specific part of the body in order to produce images of the inside of the body which can then be used for further diagnosis. In other words it is a fundamental and critical tool for the most basic medical systems to the most advanced.

An x-ray of the chest in particular produces images of not only the thoracic cavity and ribs, but also the heart, lungs, esophagus and spinal column. Because of this wide variety of information on a wide range of different biological systems all from a non-invasive tool, the x-ray is critical in both the prevention of a variety of diseases and conditions as well as a tool for confirming a diagnosis.

Heart Disease is the #1 killer in the United States and many western nations and the reading of chest X-rays by lab technicians is one of the most fundamental tools to prevent and recognize a variety of serious disorders. By automating this process and having machine learning algorithms be able to read and recognize patterns in X-rays faster and more accurately than technicians will be a game changer.

In the midst of the current COVID-19 repository disease pandemic the significance of chest radiography to global health is clearer than ever before . While it may be difficult to pinpoint Covid-19 as a specific phenotype vs a more common strain of normal pneumonia, the underlying need for a rapid, cost-effective diagnostic tool the same. Since we have established the primacy of radiography and its importance in normal times and in a global pandemic it becomes clear the benefits that automation would provide.

Purpose - Benefits of Automation

Automation of the routine process of radiograph interpretation, a process subject to human error and normal human fallibility, would provide a substantial benefit for the large-scale screening required during a global health situation such as the current Covid-19 pandemic.

The cost of healthcare in the United States is astronomical compared to other nations with a national health care system. This pricing crisis affects the economy in multiple ways, the primary being the cost of health insurance is out of reach for millions of Americans. Technological advances such as Machine Learning and Artificial Intelligence have unearthed a great opportunity to automate expensive and labor-intensive human tasks with the dual goals of reducing costs while simultaneously allowing human mental capital to be put to more effective uses.

Literature Survey:

Since this dataset is highly publicized and is associated with Andew NG and also part of a public competition there is a large quanity of literature available on the various approaches. Many of them have strong results and are already as good or better than human radiologists.

The goal and guiding principal of almost every approach is the development of a deep learning model which is able to distinguish between a variety of chest/lung pathologies and identify one or multiple areas of concern while also determining the specific pixel region of the image that pinpoints the area in question.

The most common approaches rely on Depp Learning and more specifically Convolutional Nerual Networks (CNN)

Conduct literature search to understand the state of arts and the gap for solving the problem. Formulate the data science problem in details (e.g., classification vs. predictive modeling vs. clustering problem).

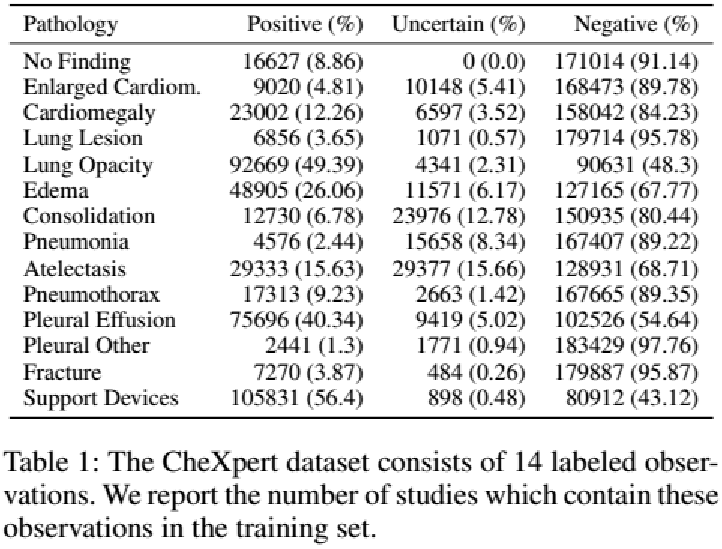
Data

There are several public datasets of chest X-ray information each with their own quirks and pros and cons. My goal is to focus primarily on the CheXpert dataset summarized below:

> **Chexpert dataset from Standford University** The CheXpert dataset consists of Chest radiographic examinations and associated radiology reports from Stanford Hospital's inpatient and outpatient centers performed between October 2002 and July 2017.[1]
>
> - 224,316 chest radiographs of 65,240 patients
> - Available in full resolution (500GB) or downscaled resolution of 320 x 320 pixels (11GB) - Labels of 14 diagnostic criteria from associated radiograph reports

The cheXpert dataset is the optimal dataset because it has a highly reliable ground truth, verified by multiple human radiologists. This is critical in evaluating and comparing deep learning models with medical imaging and provides the foundation for comparison of model results vs this benchmark.

The images each have an associated radiological report (not made public) in which labels for the 14 conditions were generated. There was a 3 part pipeline in which an NLP-based tool extracted text from the reports that matched the



Pathology	Positive (%)	Uncertain (%)	Negative (%)
No Finding	16627 (8.86)	0 (0.0)	171014 (91.14)
Enlarged Cardiom.	9020 (4.81)	10148 (5.41)	168473 (89.78)
Cardiomegaly	23002 (12.26)	6597 (3.52)	158042 (84.23)
Lung Lesion	6856 (3.65)	1071 (0.57)	179714 (95.78)
Lung Opacity	92669 (49.39)	4341 (2.31)	90631 (48.3)
Edema	48905 (26.06)	11571 (6.17)	127165 (67.77)
Consolidation	12730 (6.78)	23976 (12.78)	150935 (80.44)
Pneumonia	4576 (2.44)	15658 (8.34)	167407 (89.22)
Atelectasis	29333 (15.63)	29377 (15.66)	128931 (68.71)
Pneumothorax	17313 (9.23)	2663 (1.42)	167665 (89.35)
Pleural Effusion	75696 (40.34)	9419 (5.02)	102526 (54.64)
Pleural Other	2441 (1.3)	1771 (0.94)	183429 (97.76)
Fracture	7270 (3.87)	484 (0.26)	179887 (95.87)
Support Devices	105831 (56.4)	898 (0.48)	80912 (43.12)

Table 1: The CheXpert dataset consists of 14 labeled observations. We report the number of studies which contain these observations in the training set.

Figure 1: 14 Labeled Diagnostic Observations

list of the 14 medical conditions. Just because a medical term was mentioned does not provide enough information to confidently apply the correct label so to circumvent this challenge, radiology report search engines were used to detect the context of findings in reports—positive, negated, or uncertain findings.[2]

These automated uncertainty labels were then verified independently by multiple human radiologists, for the validation and test sets. The automated approch is possible because of the standardized format that radiography reports follow regardless of the lab. The rule-based labeling tool specifically focused on the ”Impressions” section of the reports which is similar to a conclusion but contains a bit more uncertainty inherent in any medical diagnosis. Impressions are an excellent gauge of the common sense and clinical judgment of the radiologist.[3]

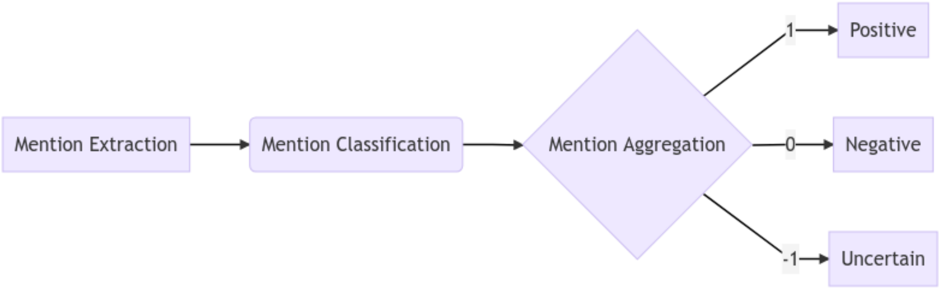


Figure 2: Automated Rule-based Labeling Process

”Mention Extraction” - extract text from Impressions Section that matches the list of observations

”Mention Classification” - classify the extracted mentions as negative, uncertain, or positive

”Mention Aggregation” - Label for Each Observation (0, -1, 1)

Technology Stack

Due to the hardware resources required to analyze a 500GB dataset of high definition radiographic images. I will attempt to leverage the public BigData tools offered by Amazon Webservices. I do not have personal experience with these tools and will use the credits offered to students at Georgia Tech.

Amazon Sagemaker is a solution that allows for building, training and deployment of models provides an Apache Spark library for data processing.

Databricks is an alternative platform for Data Scientists that also uses Amazon Web Services to train and deploy machine learning models. It is a web-based platform for working with Spark that provides automated cluster management and Juyter/Zeppelin style notebooks.

Both of these tools can use a Spark framework which I prefer as it seems to be the most efficient architecture because it facilitates both pre-processing and implementation of machine learning algorithms in its embedded libraries. This will take advantage of the Spark frameworks ability to distrubute the tasks onto multiple parallel worker nodes.

BigDL is a deep learning framework for Apache Spark which is perfect for the requirement to use big data tools learned in this course. This library is meant for distributed environments and can be used to help preprocess the image data.

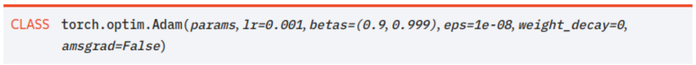
Experimental Setup:

My baseline intent is to train a model that takes a 320 x 320 pixel chest x-ray image as input and then calculates and outputs the multi-class probability that the patient should be diagnosed with each of the 14 observed condition.

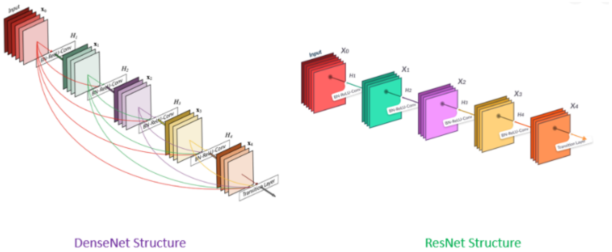
We can then compare the model output vs the ground-truth labels supplied by eight board-certified radiologist who individually annotated each of the studies in the test set and based on their expert opinion classified them individually as having been diagnosed vs not present.

The images were split into a training and validation set but I intend to use the provided validation set as the final test set and develop my own validation set that will serve the needs of my technology stack and allow for faster iteration and tweaking of the model.

We trained using two different object recognition architectures ResNet-152, and DenseNet-161 which are both considered standard networks and at one point 'state-of-the-art' architectures for image recognition. The output layer was set to 14 neurons to match the classes of disease and we used a sigmoid activation which can be interpreted as the probability of the label. We trained using the Adam Optimization Algorithm which is computationally more efficient and used a learning rate, η, was 10^{-4}



```
CLASS  torch.optim.Adam(params, lr=0.001, betas=(0.9, 0.999), eps=1e-08, weight_decay=0,
       amsgrad=False)
```



DenseNet Structure

ResNet Structure

Figure 3: Artificial Neural Networks

We ran 24 experiments based on the two policy choices regarding the implicit negatives, three choices for uncertainty labels and the two possible neural net architectures. We ran for 10 epochs because we determined that the model was not overfitting based on the learning curves.

Preproccesing the Data:

Due to the size of the image data we used big data tools learned in the course to do the initial preprocessing. We used the docker image along with Hadoops HDFS for data storage and Spark for in-memory processing. By using the Spark Resilient Distributed Dataset (RDD) and 4 worker nodes allow this pre-proccesing to be done more effectively.

In order to quickly be able to assess the CNN model as it is tweaked and iterated a validation set is needed to evaluate performance. The goal is generalization to future xray images and we want to avoid overfitting on the training data. Since chest xrays are generally standardized across the globe and the training set is so large we are not too worried since it is highly likely that any chest xray will be quite similar to once already in the training set.

We needed to randomly divide the available training images into a training set and hold a portion out as a validation set. We will use this hold-out set to further tune the parameters of the model to get a more unbiased measure of performance

We took care to ensure that no single patient had images in both the training and hold-out set to help reduce bias further. To make this split we used Scala and the ’randomsplit’ method which uses a Bernouilli Distribution to split.

```
train_split, test_split = df.randomSplit(weights = [0.90, 0.10], seed = 11)
```

Neural networks work best when all the features are on the same scale. Image data is no exception. The gradient descent optimization technique is also more effective when the features are standard-scaled to a normal distribution (centered at mean zero with a standard deviation of one). This helps the learning process to be faster and more stable since the gradient calculation used in backpropagation is uniform for each RGB channel.

The previously mentioned BigDL package provides an image preprocessing library that contains all the common and standard transformation and augmetntation operations commonly used in big data pipelines. By using the ChannelNormalize method we create a distributed data structure of the pixel data.

We also used Pytorch to preprocess some training images that had text written on them and other color issues.

Experiment Design:

We used Amazon AWS and an EC2 p2.xlarge cluster with an Ubuntu AMI w CUDA installed and 16GB GPU. We ran 5 epochs with batch sizes of 48/64 for training and 512 for parameter tuning.

Due to the class imbalance as previously mentioned, the choices we make on how to deal with this will affect the models results. The class imbalance is not the same across all diseases so

In our data-set many of the labels are ”implicitly” classified as negatives. In these cases the radiologist did not specifically mention the presence of a disease nor did they confirm the condition was not present. Common sense will tell us there is a strong likelihood that lack of mention does in fact correspond to a true negative. However we also wanted to consider different priors such as .33 and .66 to consider scenarios with more uncertainty regarding these labels.

Metrics Used:

We used binary cross entropy loss a.k.a log loss as our cost function which is pretty standard for classification.[4] Since it is a sigmoid activation function we are not constrained on the input which allows flexibility in trying different things including label smoothing due to uncertainty inherent in some of the labels.

$$\textbf{Loss} = -\frac{1}{\text{outputsize}} \sum_{i=1}^{(\text{outputsize})} y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log(1 - \hat{y}_i t)$$

Our data suffers the *class imbalance problem* when the class distributions are highly imbalanced. Some labels have as many uncertain cases as negative cases and most labels have very few positive cases. This causes issues and low accuracy for detecting pathologies.[5] The worst example of this issue is for ”Pleural Other”, in which just 1.3% of the images contain this diagnosis. The **BigDL** framework has several data augmentation methods that can help try and ameliorate this issue.

We use AUC (Area Under Operating Characteristic) as our primary metric to compare our results to other studies and also on our hold-out set. AUC provides a quick summary in a single metric but allow us to look into the True Positive and False Positive rates if necessary to improve the model.

[6–10]

Results:

The table below is a summary of the best scoring results

In Figure 5 you will find the learning curves for our best performing model which makes clear that more epochs would start to overfit and there isn’t much improvement to be had.

In figure 6 you can see the confusion matrices for the pathologies, *edema* and *pleural effusion* which had the best performance using our validation set. This provides insight into the Type-1 and Type-2 errors which for data sets suffereing from class imbalance quickly eliminates the misleading nature of accuracy metric that exists in unbalanced classes.[11]

	ResNet-152				DenseNet-161			
	I-Zeros	I-0.33			I-Zeros			I-0.33
	U-0.66	U-0.66	U-0.66	U-Ones	U-0.66	U-0.66	U-Zeros	U-0.66
	scratch	scratch	pretrained	scratch	pretrained	scratch	scratch	pretrained
	5 epochs	10 epochs	10 epochs	10 epochs	5 epochs	5 epochs	5 epochs	10 epochs
No Finding	0.8291	0.8066	0.8201	0.8123	**0.8291**	0.7862	0.7891	0.8232
Enlarged Cardiomediastinum	0.4601	0.8291	0.8112	**0.8376**	0.5111	0.3342	0.5262	0.8259
Cardiomegaly	0.7302	**0.8321**	0.7369	0.8171	0.7332	0.7572	0.8031	0.7469
Lung Opacity	**0.8702**	0.8512	0.8192	0.8562	0.8232	0.8432	0.8594	0.8282
Lung Lesion	0.0042	0.0042	0.0222	0.0001	**0.0601**	0.0122	0.0154	0.0092
Edema	0.8152	0.7998	0.8363	0.7902	0.8151	0.7954	0.7912	**0.8462**
Consolidation	0.7921	0.8292	0.8110	0.7964	0.8170	0.7592	0.7522	**0.8392**
Pneumonia	0.3912	0.4444	0.6081	0.2868	0.5802	0.2588	0.2289	**0.7252**
Atelectasis	0.7532	**0.7934**	0.7613	0.6501	0.7622	0.7777	0.7354	0.7464
Pneumothorax	0.4982	0.4456	**0.6778**	0.4522	0.6362	0.4479	0.4778	0.6332
Pleural Effusion	0.8602	**0.8792**	0.7982	0.8640	0.8202	0.8584	0.8552	0.8262
Pleural Other	0.1090	0.0667	0.2312	0.1252	0.2782	0.0782	0.0872	**0.3322**
Fracture	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
Support Devices	0.7410	**0.8332**	0.7401	0.8202	0.7214	0.7532	0.7943	0.7240
Overall Mean AUC	0.5605	0.6005	0.6115	0.5792	0.5899	0.5229	0.5510	**0.6357**
5-observation focus Mean ***AUC***	0.7890	**0.8257**	0.7890	0.7829	0.7888	0.7892	0.7869	0.8011

Figure 4: AUC results for the Hold-Out Validation Set

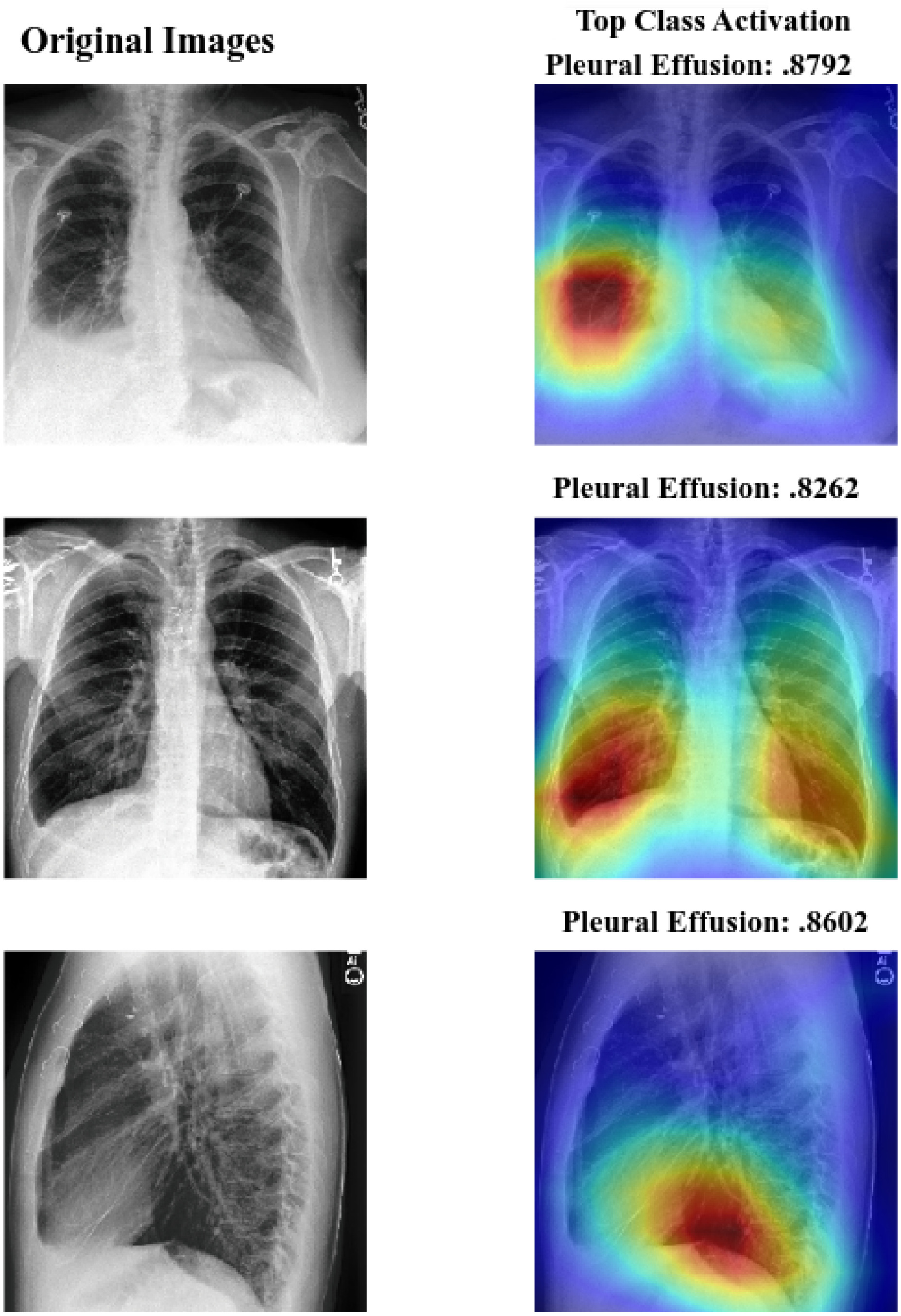


Figure 5: Class Activation Mappings using DenseNet161

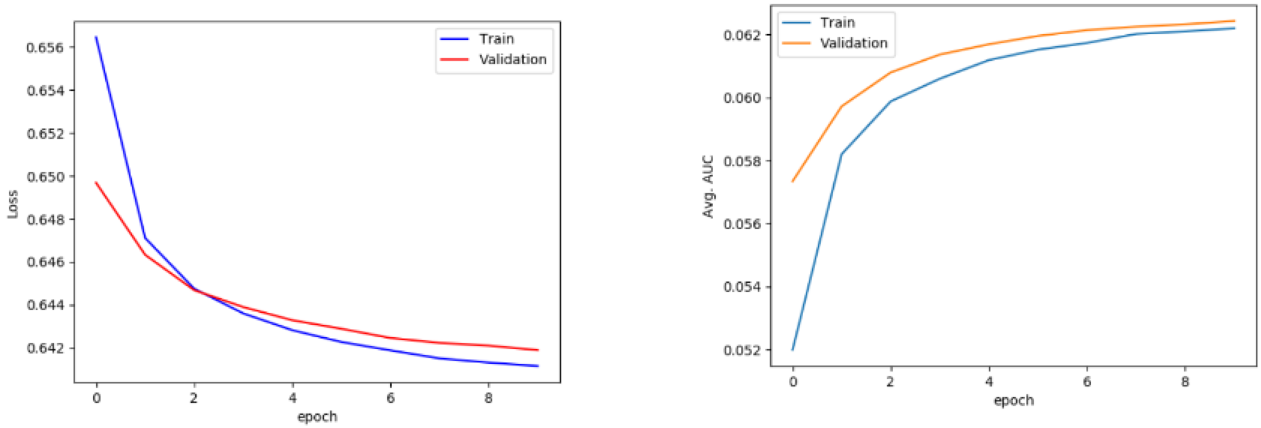
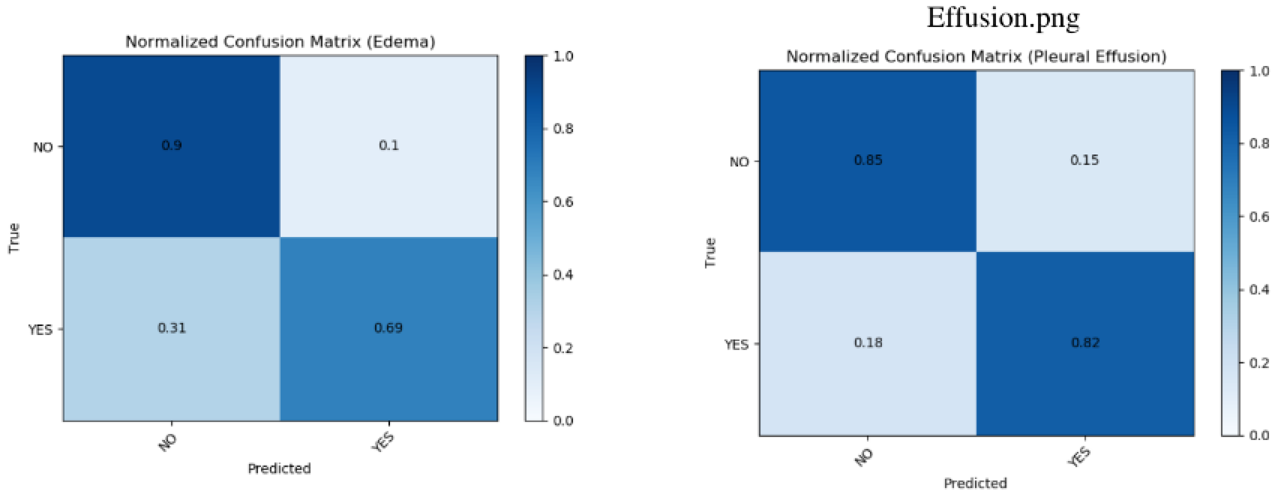


Figure 6: Learning Curves



(a) Confusion matrix (Edema)

(b) Confusion matrix (Pleural Effusion)

Figure 7: Confusion Matrix for Best Performing Models

Summary:

By using convolutional neural networks to understand image pixel data on a localized basis allows for neural networks to correctly diagnose various lung pathologies with accuracy on par with human radiologists. This is a dream come true and demonstrates just how data science and machine learning can benefit humanity in meaningful ways. The approach taken in this project can be applied to other domains as well and it will be exciting to see what else develops over the next decade. Since we are new to this field and still learning this project was designed to replicate the process that others had previously developed so this is not a true research paper in that sense. In the future we would have more time to clean up some anomolies in the images or perhaps use different training data to try and replicate the results.

References

1. Chexpert: A large dataset of chest x-rays and competition for automated chest x-ray interpretation. *CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison*.

2. Andrew S. Wu, Bao H. Do, Jinsuh Kim, and Daniel L. Rubin. Evaluation of negation and uncertainty detection and its impact on precision and recall in search. *Journal of Digital Imaging*, 24(2):234–242, Apr 2011.

3. Ferris M. Hall. Language of the radiology report. *American Journal of Roentgenology*, 175(5):1239–1242, Nov 2000.

4. Li Yao, Eric Poblenz, Dmitry Dagunts, Ben Covington, Devon Bernard, and Kevin Lyman. Learning to diagnose from scratch by exploiting dependencies among labels. 2018.

5. Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

6. Pranav Rajpurkar, Jeremy Irvin, and Andrew Y. Ng et. al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *CoRR*, abs/1711.05225, 2017.

7. Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M. Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. 2018.

8. Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Li Fei-Fei. Thoracic disease identification and localization with limited supervision. 2018.

9. Ikezoe J et al. Shiraishi J, Katsuragawa S. Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *AMERICAN JOURNAL OF ROENTGENOLOGY 174 (1)*, Jan 2000.

10. Mohammad Farukh Hashmi, Satyarth Katiyar, Avinash G Keskar, Neeraj Dhanraj Bokde, and Zong Woo Geem. Efficient pneumonia detection in chest xray images using deep transfer learning. *Diagnostics*, 10(6), 2020.

11. Jason Jinquan Dai, Yiheng Wang, Xin Qiu, Ding Ding, Yao Zhang, Yanzhang Wang, Xianyan Jia, Cherry Li Zhang, Yan Wan, Zhichao Li, and et al. Bigdl. *Proceedings of the ACM Symposium on Cloud Computing*, Nov 2019.