Classifying Respiratory Diseases using Chest X-Ray Imaging

Jade Benson, Egemen Pamukçu, Jacob Jameson

Specific Aims of the Project

This project aims to classify the NIH chest X-ray dataset through the use of a deep neural net architecture. We optimize our model through incremental steps. We first spend time exploring our dataset, then experiment with different architectures, and ultimately create our final model. The motivation behind this project is to replicate or improve upon the results as laid out in the following paper: *ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases*[1].

Our dataset was gathered by the National Institute of Health (NIH) and contains 112,120 anonymized frontal chest X-ray images from 30,805 patients. The data comes from a Natural Language Processing (NLP) analysis of radiology reports and may include areas of lower confidence in diagnoses. As a simplifying assumption, we assume that based on the size of the dataset, that the dataset is accurate in diagnosis.

Background research

Chest X-ray exams are one of the most frequent and cost-effective medical imaging examinations. However, clinical diagnosis of chest X-ray can be challenging, and sometimes believed to be harder than diagnosis via chest CT imaging. There has been promising work reported in the past, especially in recent deep learning work on Tuberculosis classification. Achieving clinically relevant computer-aided detection and diagnosis in real world medical sites on all data settings of chest X-rays is still very difficult, if not impossible, when only several thousands of images are employed for study. This is evident from *Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation*[2], where the performance of deep neural networks for thorax disease recognition is severely limited by the availability of only 4143 frontal view images[3] (Openi was the previous largest publicly available chest X-ray dataset to date). In 2017, the National Institute of Health (NIH) released one of the largest publicly available chest X-ray datasets to date, which includes demographic information and images from more than 30,000 patients.

---

[1] Wang, Xiaosong, 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." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097-2106. 2017.

[2] Hoo-chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, Ronald M. Summers, Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation, IEEE CVPR, pp. 2497-2506, 2016

[3] Open-i: An open access biomedical search engine. https: //openi.nlm.nih.gov

Many researchers have used this dataset, as well as others like it, including the Indiana University Chest X-ray Collection[4], the Japanese Society of Radiological Technology (JSRT) database[5], and the Shenzhen dataset[6], to create deep learning models capable of detecting and classifying various lung diseases. These have included instances where a single patient may present with more than one disease, as well as localizing these diseases within the image files.

While introducing the NIH dataset and its labelling technique, Wang et al. provide a DCNN based multi-label weakly supervised method for disease classification, setting benchmark results for future work on the dataset. Pre-trained networks on ImageNet17 including AlexNet, GoogLeNet, VGGNet-16, and ResNet-50 were used without the final classification layers. Some additional layers were added to the architecture to adapt it to the classification task, including a pooling layer, transition layer, and weighted Cross Entropy Loss layer. For localization, the pooling filters information to be passed down, and the weights of the prediction layer can then be used in spatial maps, producing weighted maps localizing active X-ray areas for each disease class.

Approach

I. Data Exploration

We began by exploring the data to better understand its structure and characteristics, in order to inform our classification approach. Table 1 displays the available information about patients' demographics and X-ray characteristics at baseline. Our sample included slightly more men than women (54% male) and the median age was 48 years old at baseline (first X-ray recorded). Although the dataset includes 112,120 images, the number of images per patient is right-skewed where the majority only have one X-ray (56.8%) and 92.8% have less than 10 X-ray images. The distribution of the number of X-rays per patient can be seen in Figure 1.

Table 1. Demographics and X-ray characteristics for the 30,802 patients with age information.

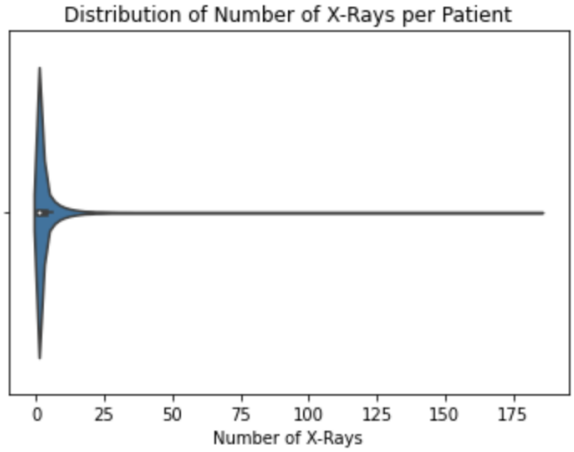
	Median (minimum, maximum) or Percent
Male	54%
Age	48 (1, 95)
Number of X-Rays	1 (1, 184)

[4] Demner-Fushman D, KohliMD, RosenmanMB, Shooshan SE, Rodriguez L, Antani Setal. Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association. 2016 Mar 1;23(2):304-310. Available from:DOI: 10.1093/jamia/ocv080.

[5] JSRT Database, Japanese Society of Radiological Technology. Available from: http://db.jsrt.or.jp/eng.php.

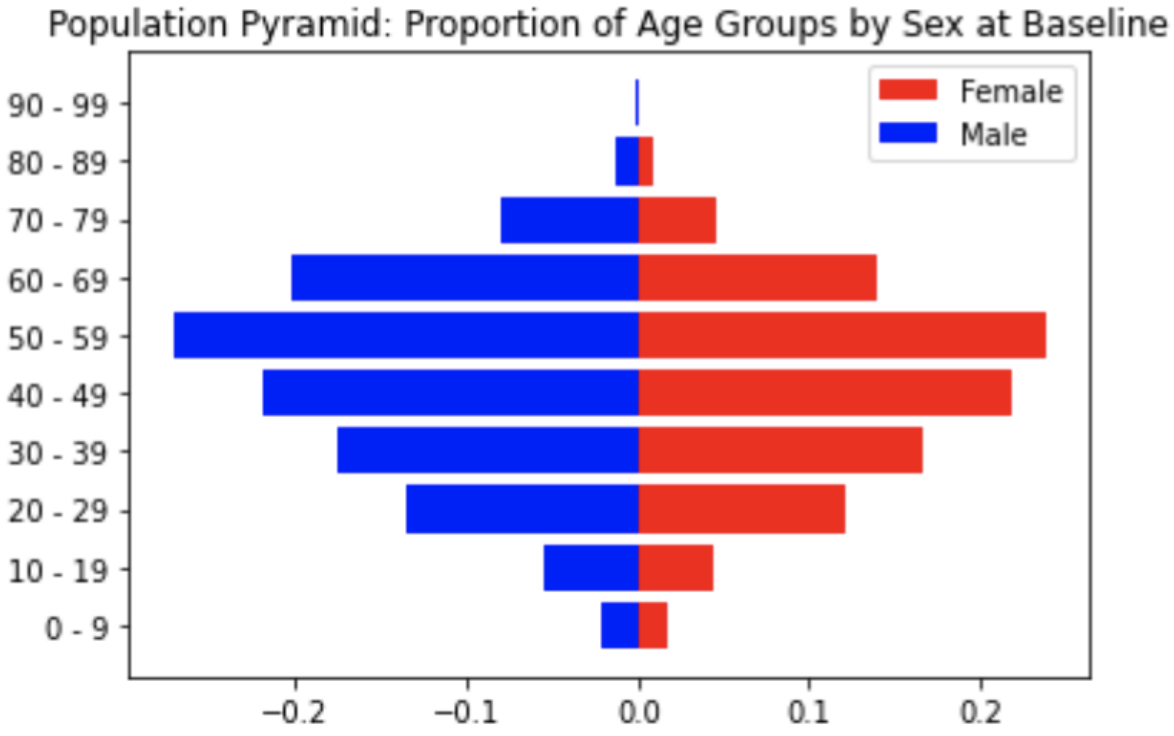
[6] datasets for computer-aided screening of pulmonary diseases. Quantitative Imaging in Medicine and Surgery. 2014; 4(6):475-477. DOI:10.3978/j.issn.2223-4292.2014.11.20.

Figure 1. Distribution of the number of X-ray images per patient.



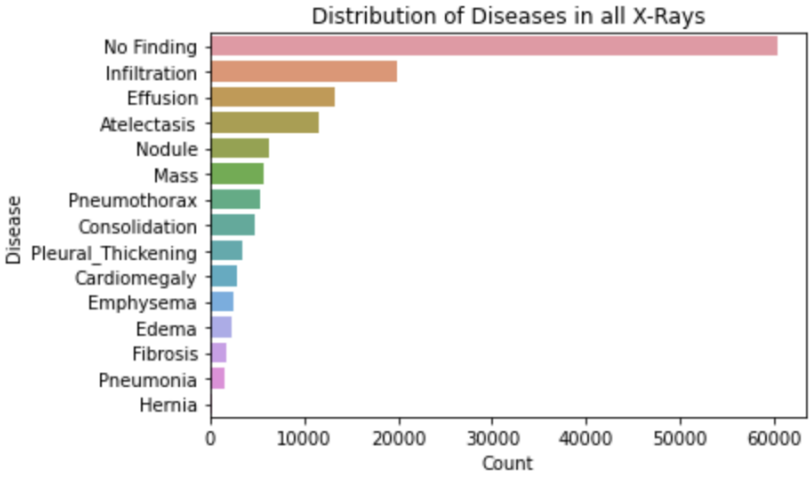
The dataset was primarily composed of middle-aged adults. Adults aged 50 - 59 made up 23.5% of the sample and another 20.1% were adults aged 40 - 49. There were comparable proportions of men and women in each age category. The distribution of age groups by sex is displayed in Figure 2.

Figure 2. Proportion of patients' age groups by sex at baseline.



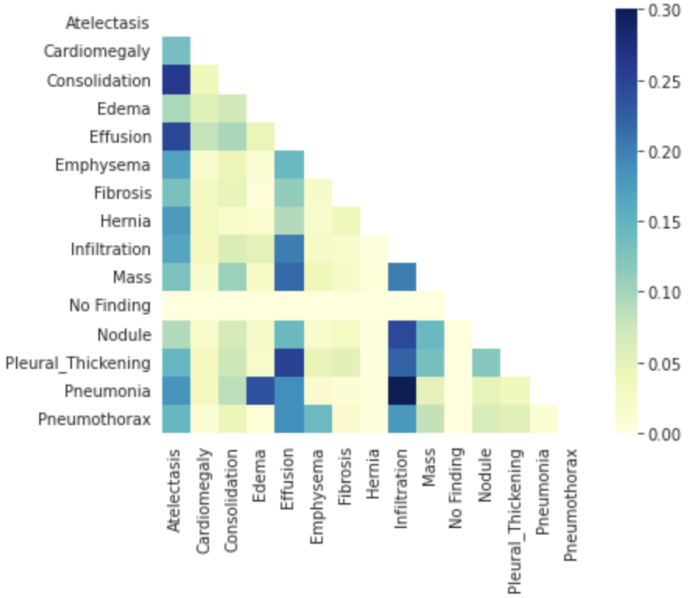
Each X-ray is labeled with either “No Finding” or 14 possible disease labels that can co-occur (Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural_Thickening, Pneumonia, & Pneumothorax). There were 60,361 X-rays (53.8%) that were normal (“No Finding”). Including multiple disease diagnoses for a single image, the disease category that occurs most often was infiltration (19,894) and the disease that occurs the least was hernia (227). The distribution of label frequency in all X-rays can be found in Figure 3.

Figure 3. Distribution of label frequency in all X-ray images (includes co-occurrences).



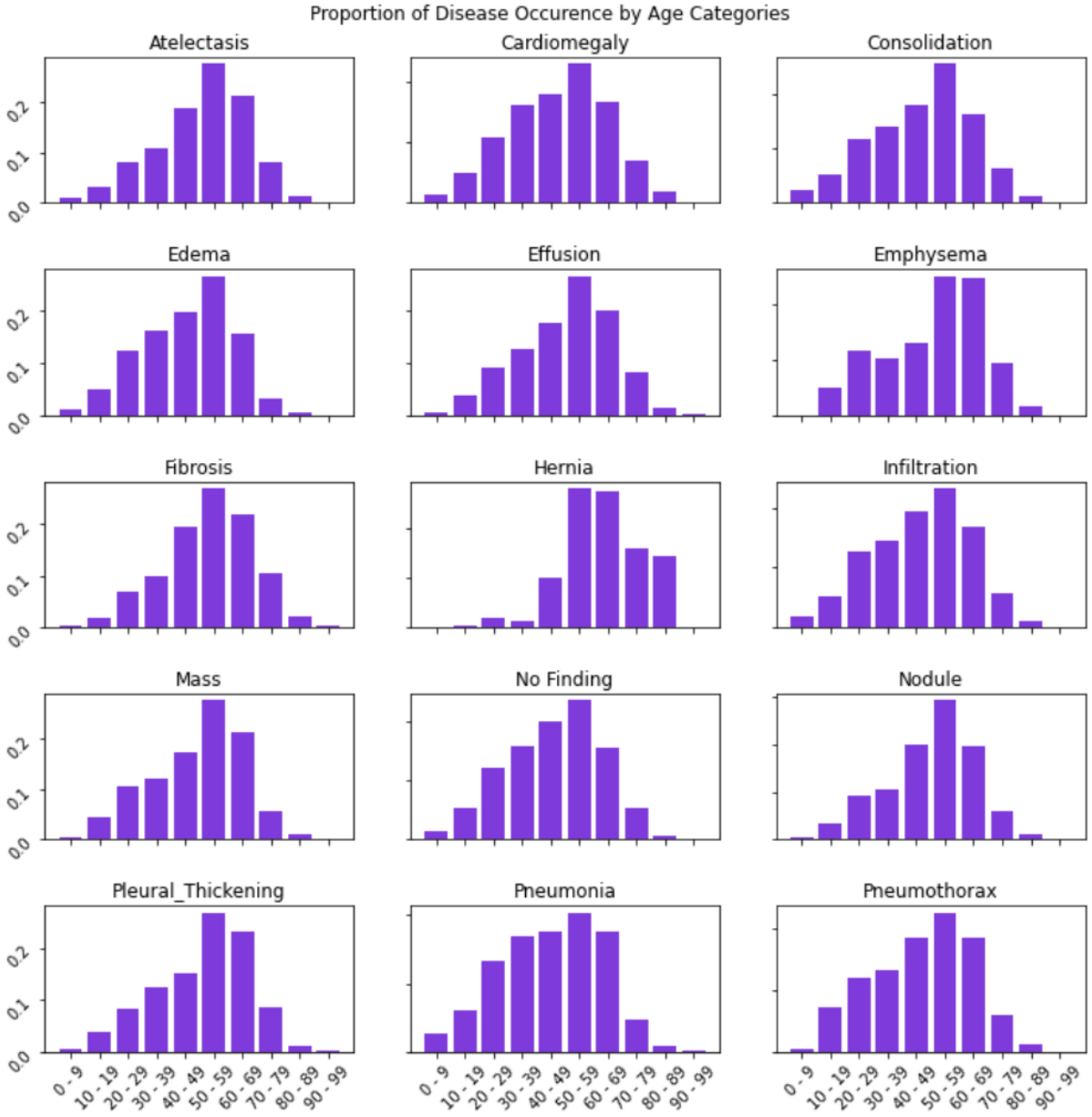
X-ray images that were identified as having some disease present, often had many co-occurring diseases. The diseases that most often appeared together were pneumonia and infiltration - 42% of pneumonia cases presented this way. Nodules and infiltration co-occurred in 24% of nodule cases. Figure 4 displays a heat map of the proportion of label co-occurrences.

Figure 4. Heatmap of the proportion of disease label co-occurrence in all X-ray images.



There were also different proportions of diseases by age category as illustrated in Figure 5. Emphysema primarily occurred in older adults, while pneumonia occurred throughout the age range. Younger adults and middle aged adults made up most of the normal findings. It is important to note that the sample is primarily composed of middle-aged adults, which we can see in the patterns of disease proportions across all labels.

Figure 5. Proportion of disease occurrence by age category.

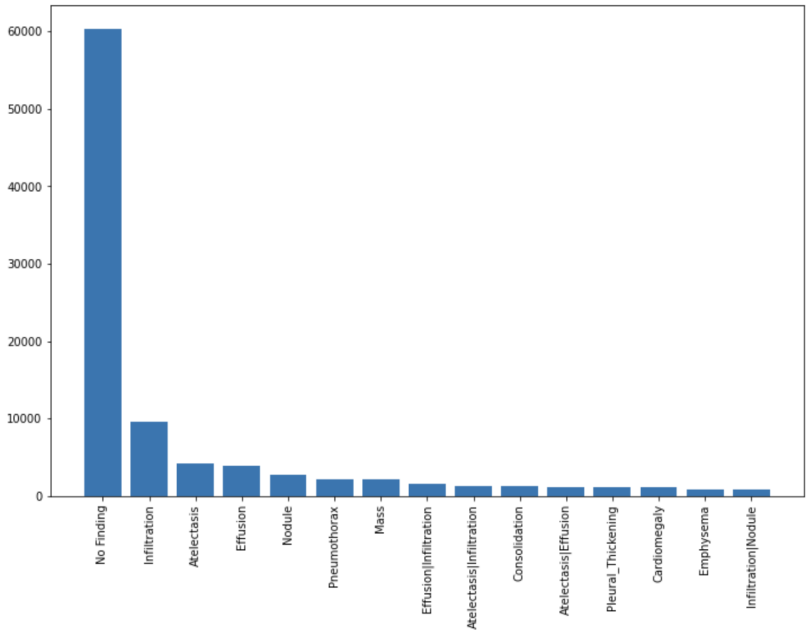


Additionally, we checked whether X-ray position may be associated with disease status. As we previously discussed in class, prior COVID-19 X-ray classification techniques had primarily identified the patient's position and the site location where they received the X-ray. Both of these factors were associated with the disease outcome, but were not novel or clinically relevant. There are two frontal positions that X-rays can be taken in, Chest Posterior Anterior (PA) where the patient stands up or Anterior Posterior (AP) where the patient lays down. The PA position is considered the "gold standard" measurement, which means that these positions have clinical significance since healthier patients will be encouraged to stand.[7] In our sample, the odds of being diagnosed with any disease is 1.58 times higher for those patients that received an AP X-ray (laying down) as opposed to those that received an PA X-ray (standing). The Chi-squared statistic comparing the contingency matrix of un/healthy patients with PA/AP positions was 316.8, which is statistically significant.

II. Pre-Processing

We followed several steps in order to get the data ready for training an image classifier. One of the issues we came across was the imbalance among the diseases within the NIH dataset.

Figure 6. Proportion of Category.

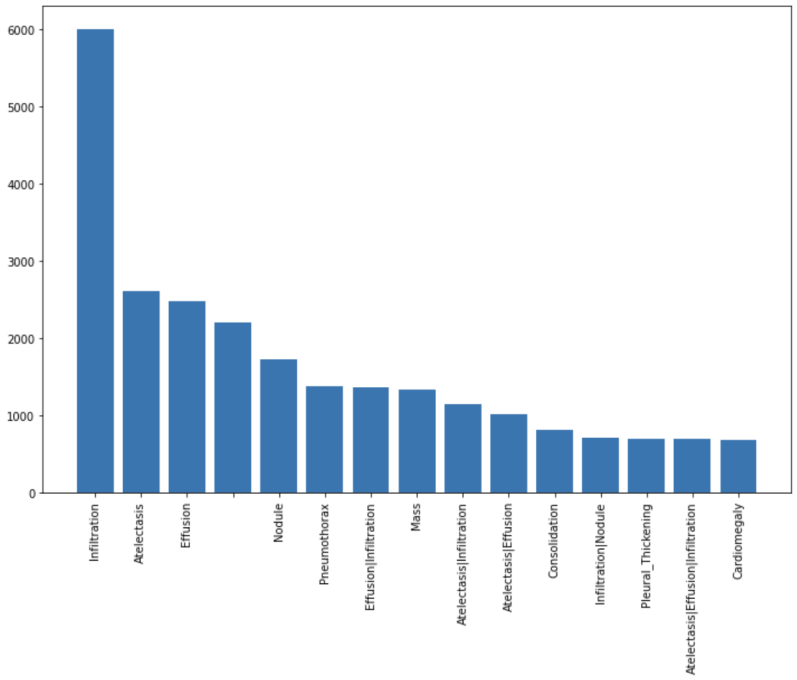


The figure above displays the counts of 15 most common labels we have in the raw dataset. Considering that the initial dataset consists of 112,120 records, more than half of these records belong to patients without any diseases that could be read from chest X-Rays by physicians. The

[7] International society of radiographers & radiological technologists, e-learning course titled "The difference between Chest Posterior Anterior (PA) and Anterior Posterior (AP) radiographs." https://www.elearning.isrrt.org/mod/book/tool/print/index.php?id=321

X-Rays that *were* associated with a disease also had imbalances. For instance, infiltration as a label accounted for about 100 times more observations compared to hernia. In order to account for these imbalances, first we drop the underrepresented diseases, i.e. labels that have less than 1,000 records associated with them. Then, we subsample the dataset based on calculated sample weights to include 40,000 images before undertaking the classification task. Figure 7 below shows the new distribution for the top 15 labels after these two steps:

Figure 7. New Distribution of Labels.



Later, we encode the diagnosis labels to fit into a multilabel classification format. Since one patient can be diagnosed with more than one disease, we create binary variables for each disease type we are left with after filtering out underrepresented diseases. This results in a label matrix of shape 40,000 * 13 where 40,000 is the number of records and 13 is the number of diseases we will try to predict. It should be noted once again that, unlike one-hot encoding, multiple elements in one row of this matrix can be 1 due to the nature of the dataset.

After sampling the dataset and formatting the labels, we preprocess the images to obtain an optimal balance between classifier performance and computational efficiency. First, we transform each image with a grayscale color mode as X-rays do not contain useful information that can be accessed through RGB color channels. Also, after experimenting with multiple image sizes, we settle with a 512*512 pixel resolution which gave us decent classification metrics for our classifier as well as enabling us to train the model in a reasonable timeframe.

Technologies Used

For the classification task, we make use of the pretrained MobileNetV1 architecture which can be used for a variety of computer vision applications such as detection, embeddings and segmentation. MobileNet architecture has a relatively lower number of parameters to train as it is optimized to run on mobile devices with limited computing power. However, on many tasks it can outperform heavier and more complex architectures and, overall, provides a fair tradeoff between agility and predictive power[8]. The architecture of the MobileNetV1 model can be seen in the image below (Figure 8).

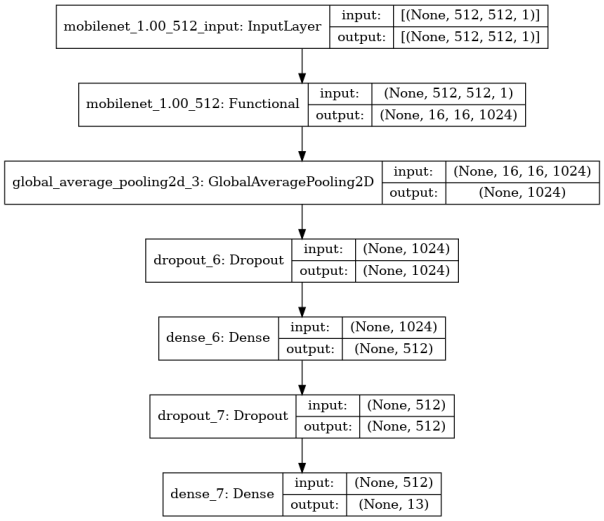
Figure 8. Architecture of MobileNetV1.

Type / Stride	Filter Shape	Input Size
Conv / s2	$3 \times 3 \times 3 \times 32$	$224 \times 224 \times 3$
Conv dw / s1	$3 \times 3 \times 32$ dw	$112 \times 112 \times 32$
Conv / s1	$1 \times 1 \times 32 \times 64$	$112 \times 112 \times 32$
Conv dw / s2	$3 \times 3 \times 64$ dw	$112 \times 112 \times 64$
Conv / s1	$1 \times 1 \times 64 \times 128$	$56 \times 56 \times 64$
Conv dw / s1	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 128$	$56 \times 56 \times 128$
Conv dw / s2	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 256$	$28 \times 28 \times 128$
Conv dw / s1	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 256$	$28 \times 28 \times 256$
Conv dw / s2	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 512$	$14 \times 14 \times 256$
$5\times$ Conv dw / s1	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
$5\times$ Conv / s1	$1 \times 1 \times 512 \times 512$	$14 \times 14 \times 512$
Conv dw / s2	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
Conv / s1	$1 \times 1 \times 512 \times 1024$	$7 \times 7 \times 512$
Conv dw / s2	$3 \times 3 \times 1024$ dw	$7 \times 7 \times 1024$
Conv / s1	$1 \times 1 \times 1024 \times 1024$	$7 \times 7 \times 1024$
Avg Pool / s1	Pool 7×7	$7 \times 7 \times 1024$
FC / s1	1024×1000	$1 \times 1 \times 1024$
Softmax / s1	Classifier	$1 \times 1 \times 1000$

Using Keras deep learning API with TensorFlow backend, we add additional pooling, dropout, and fully connected layers to the MobileNet architecture to have the ability to fine-tune and train it with our set of X-Ray images. Our final model, as shown below in Figure 9, contains a total of 3,737,869 trainable parameters. Since this is a multilabel image classification problem we use the binary cross entropy loss function for each label and Adam algorithm to optimize our model.

[8] Howard, Andrew G., Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." *arXiv preprint arXiv:1704.04861* (2017).

Figure 9. Final Model.



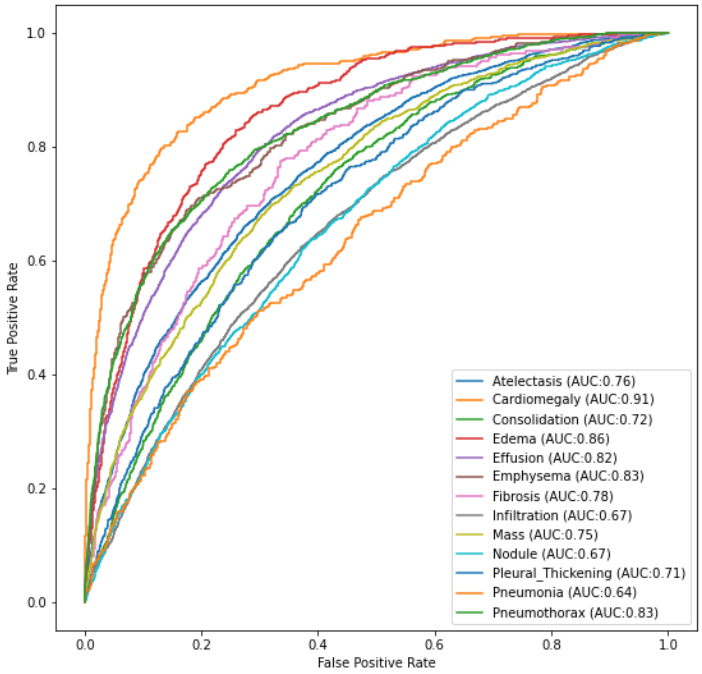
Innovations

As mentioned in the background section of this paper, there is a fairly robust literature surrounding the use of deep learning models in classifying lung pathology using this data. Our contributions to lung pathology classification models are in pairing the largest X-ray imaging dataset with the MobileNet architecture to maximize AUC. Compared to the literature using the Openi dataset prior to 2017 (roughly 7,000 images), our model performs as well or better than those we have seen in maximizing the false positive/false negative tradeoff in classification. It is clear that having a larger dataset really boosts the predictive power of our model.

Results

Initial models we tried that made use of ResNet architecture as well as the ones that we built from scratch did not perform well and the AUROC scores were indicating that the models were not doing much better than baseline. Only with the MobileNet architecture we started obtaining AUC scores well above 0.5. The figure 10 below summarizes our results for each disease type found in the final dataset.

Figure 10. AUC by Pathology Type.



As seen in the ROC curves above, model performance varies based on the disease. Since our model made 13 probability predictions between 0 and 1, looking at the ROC curve, we can tell the disease types where the model performed well and where it performed only slightly better than random guess.

The model performed the best when making predictions on cases of cardiomegaly which is commonly referred to as the condition of "enlarged heart." Cardiomegaly is a relatively common incidental finding on chest X-rays; if left untreated, it can result in significant complications. Using our model for diagnosing cardiomegaly could be beneficial, as this pathology can be underreported, or overlooked, especially in busy or under-staffed settings.

Our model also performed very well in diagnosing edema. The diagnosis of pulmonary edema is usually confirmed via X-ray, which shows increased fluid in the alveolar walls. Kerley lines, increased vascular filling, pleural effusions, upper lobe diversion (increased blood flow to the higher parts of the lung) may be indicative of cardiogenic pulmonary edema. It makes sense that our model would perform best when classifying pathologies that result in the most stark visual differences.

Our model performed worst in classifying Pneumonia, which we believe could have been the result of following: pneumonia was often paired with another pathology in our dataset, making it more difficult to classify pneumonia as our model was identifying the pathology that resulted in a more pronounced change in the X-ray image.

Lessons Learned

We were able to learn and perform image classification on 112,120 X-ray images and achieve fairly high AUC with variations by disease type. However, there were some meaningful limitations and opportunities for future work. The dataset is imbalanced by disease label, which made it difficult to train appropriate models. We tried to account for this by sampling labels in different proportions. Future approaches could try weighting the categories to achieve better accuracy or modifying the loss function to make the model learn proportionally from each class. Additionally, disease labels often co-occur in a single X-ray image (as demonstrated in Figure 4). This further complicates classification efforts as multiple categories are required to describe one image. We were also interested in building a classification model that considers both features of the X-ray images themselves and of the patients simultaneously. This dataset includes some important patient characteristics and information about the X-ray structure (positioning) that are associated with disease status, but were not accounted for in our classification.

We were able to construct and apply a novel convolutional neural network (MobileNet) to the disease classification of chest X-ray images and achieve comparable results to prior studies. Although significant work and improvement would be needed before these technologies could be used in a clinical setting, this study demonstrates the potential applications machine learning techniques can have for important medical and public health problems.

The notebooks used to conduct this analysis can be found as additional PDFs.

The data exploration is named “Exploration_Notebook.pdf” and the Kaggle notebook can be accessed here: https://www.kaggle.com/jadebenson/cnn-xray-exploration

The classification is named “notebook.pdf” and the Kaggle notebook is: https://www.kaggle.com/egemenpamuku/cnn-xray-multilabel-classification-v2?scriptVersionId=81294128