SMU Data Science Review

Volume 2 | Number 2 Article 15

2019

Automated Pleural Effusion Detection on Chest X-Rays

Nathan Wall
Southern Methodist University, nwall@mail.smu.edu

Muthu Palanisamy
Southern Methodist University, mpalanisamy@mail.smu.edu

John Santerre
jsanterre@mail.smu.edu

Automated Pleural Effusion Detection on Chest X-Rays

Muthu Palanisamy, Nathan Wall, and Dr. John Santerre

Master of Science in Data Science
Southern Methodist University
Dallas, Texas USA
mpalanisamy@smu.edu, nwall@smu.edu, jsanterre@mail.smu.edu

Abstract. In this paper we present a lightweight solution to help identify a pathological condition called Pleural Effusion using chest x-rays (CXR). Patients with Pleural Effusion have been found to have increased mortality rates, and if left undiagnosed effusion has been found to contribute to congestive heart failure, malignancy, pulmonary embolism, and tuberculosis [15] [13]. Using convolutional neural network architectures we developed a model to assist in the successful diagnosis of Pleural Effusion. The effectiveness of our model was evaluated against 200 studies manually labeled by consensus from 3 board certified radiologist. We demonstrate that our model is able to reproduce current baseline performance for this task with a model that is 10x smaller and 30x faster. This lighter architecture allows for more flexibility in deployment including the ability to deploy directly on an edge node. We present this model as a tool for the radiologists to diagnose the presence of Pleural Effusion from a diagnostic imaging study.

1 Introduction

Each year in the United States an estimated 1.5 million people develop Pleural Effusions [13]. Patients with Pleural Effusions have been found to have increased mortality rate and left undiagnosed contributes to congestive heart failure, malignancy, pulmonary embolism, and tuberculosis[15] [13]. Both frontal and lateral chest x-rays remain the primary means for initial diagnosis of this condition [10]. In this work we present a model to help assist in the accurate diagnosis of Pleural Effusion from diagnostic imaging studies.

Technological advancements in medical imaging devices allow radiologist to better diagnose a variety of diseases. With the increase in quality, volume and complexity of these images has also increased leading to larger workloads for radiologist [17]. This resulted in a shortage of qualified radiologists in both the US & UK and is directly impacting the quality of care [19] [17]. Based on the 2018 UK census of radiologist, only 20% of clinical directors feel their current staffing are able to deliver a safe & effective care. In addition, the UK is reporting that almost half of the imaging studies have not been reviewed by radiologists [17]. In the US these shortages are believed to adversely impact low-income and

rural communities in both quality of care and cost [19]. Forecasts indicate that these staffing shortages will continue to rise unless radiologists are able to be shifted around to under-served geographies [19] or governments begin offering significant funding to the training of new radiologists [17].

In addition to increases in staffing concerns, 2-20% of radiologist reports have been found with clinically significant or major errors [5]. Current research has even found a day-to-day error rate around 3-5% and much higher rates on targeted studies [1]. The most common type of errors uncovered are perceptual errors or failure to identify any abnormality in the images to begin with. These errors taken in combination the staffing shortages highlight two key issues with the current methods used to diagnose patient's x-rays. In this paper we present a mobile ready convolutional neural network trained using the CheXpert [8]. The input to the model is a chest X-ray and the output is a probability of the pathology, and an image highlighting the area of the X-ray most likely representing the pathology. These outputs are intended to help radiologists quickly assess the highest risk studies and closely review the region contributing to high risk studies.

Our paper is structured into several sections beginning with general information about Pleural Effusion. Specifically, on the importance of X-rays in the successful diagnosis and potential treatment. Information about the causes and contributing factors are outside the scope of this paper as patient history or medical records are not considered in our analysis.

The data section documents our data source used from training our model, CheXpert [8]. This dataset contains labeled 224,316 chest radiographs from Stanford hospital, as well as, 200 hand annotated studies based on the presence of 14 different pathologies. The annotation are determined by a consensus diagnosis from 3 board level radiologists. We also discuss several of the decisions made for our handling uncertain & missing labels and our data pre-processing steps.

The convolutions neural network (CNN) section details the high level topics specific to CNN and how they apply to our task of image classification. This review is intended for audiences unfamiliar to CNN architectures. However, further theoretical understanding may be necessary as we detail the specific hyperparameters used in our model.

Our model section discusses the pre-processing steps used to prepare our training images and review the architectures used in this analysis. We evaluate the performance of our model against 200 ground truth studies. We evaluate two separate architectures for this task, DenseNet121 & MobileNetV2. We have not seen other research evaluate the performance of any mobile architectures for this task. Our performance is compared against the published baseline performance reported with the original CheXpert dataset. The original authors reported an AUROC of .936 with a 95% confidence interval of (.904, .967). We find our model is able to produce similar results to those achieved by the original authors but at a much smaller size, with the MobilenetV2 network showing an AUROC of .90.

Finally, we will review some of the important ethical consideration with our model and others trained for similar tasks. We also will discuss the use of public healthcare datasets for AI tools like the model presented here.

2 Background

A Pleural Effusion is a build up of fluid in an area between the layers of tissue that line the lungs [21]. Pleural Effusion is classified into 2 groups transudative & exudative [16]. While both may require procedures to address the issue exudative Pleural Effusion is often related to a structural pleural involvement that impacts the chest mechanics of the patient. A diagram of this pathology is shown below in Figure 1.

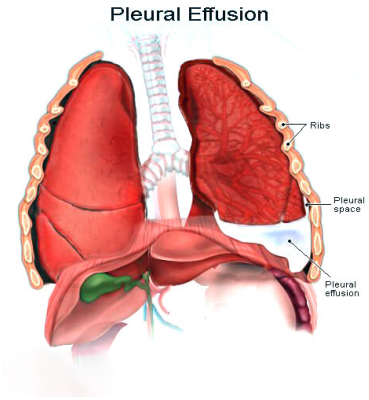

Fig. 1. Pleural Effusion: Fluid in the chest/lung

Routine bedside checks have approximately 8% chance of identifying this pathology, making x-rays the most common means of initial diagnosis of this condition [21]. Usually, radiologists are able to detect this condition in frontal x-rays from effusions with over 200ml of fluid, or 50ml in some lateral x-rays[10]. An example of frontal x-ray with Pleural Effusion is shown in Figure 2.

The mortality rates associated with Pleural Effusion seems to be dependent on a variety of factors however, the initial diagnosis of this condition is key for doctors to more efficiently assess the severity and best course [13].

The most common treatment for Pleural Effusion is thoracenteses which is a procedure to remove fluid from the space between the lining outside of the lungs and the wall of the chest. However, despite the 1.5 million patients diagnosed every year approximately 178,000 thoracenteses (12%) are performed and/or the underlying conditions are evaluated and treated for that condition. [13].

Fig. 2. Pleural Effusion: Frontal X-Ray

3 Data

Our model utilizes a collection radio-graph studies done at Stanford between October 2002 and July 2017, and has been made available by Stanford's Machine Learning group. CheXpert is the largest collection of labeled radiographs with over 220,000 different images from 64,000 different patients [8]. The closest comparable dataset of this size is the MIMIC-CXR dataset made available through MIT [9].

3.1 CheXpert

The patients data that we have for this study is made up of older men and women with the mean age of our patients at time of study at 60 years old and a median of 62. We have a somewhat equal distribution of men & women with 55% of our training sample made up of men. Additionally, about 47% of our patients in the training data have had more than one study in our data.

For each patient study we were provided the x-rays from the study in either the full high resolution or down sampled resolution of 330x330 pixels.

The majority of the images are frontal X-ray images (85%) however, some studies provide both a frontal and lateral x-ray. An example of the images from a patient study are shown in Figure 3.

In order to get a labeled data set of this size the CheXpert authors utilized an NLP algorithm to label the data. The labeler was trained using 1000 studies that were manually labeled by board certified radiologists to determine whether a specific pathology was mentioned in the report and if that pathology was present or not. The labeling process is described in three different steps, with the first step being mention extraction. This step utilizes a large lists of phrases curated by a group of radiologists that would signify the mention of one or more of the 14

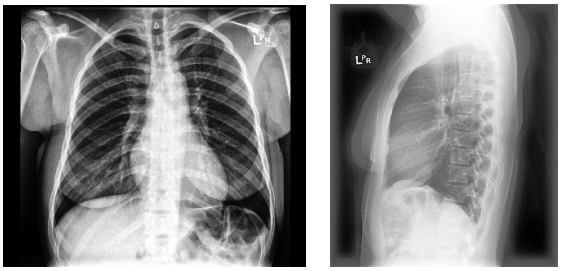

Fig. 3. Example Study Images

classes. Once extracted the mentions were classified into one of three potential categories; presence of the condition, condition not found, and uncertain. Lastly, the classified mentions are aggregated into the final labels. Any positive mention of the pathology is considered positive (1), if no positive mentions are found and and at least 1 uncertain mention it is uncertain (-1), if one or more negative mention is found it is negative (0), and if no mention is found the label is left blank. Table 1 shows the results of this labeling process on the complete dataset.

Pathology	Positive (1)	Uncertain (-1)	Negative (0)
No Finding	22381	0	201033
Enlarged Cardiomediastinum	10798	12403	200213
Cardiomegaly	27000	8087	188327
Lung Opacity	105581	5598	112235
Lung Lesion	9186	1488	212740
Edema	52246	12984	158184
Consolidation	14783	27742	180889
Pneumonia	6039	18770	198605
Atelectasis	33376	33739	156299
Pneumothorax	19448	3145	200821
Pleural Effusion	86187	11628	125599
Pleural Other	3523	2653	217238
Fracture	9040	642	213732
Support Devices	116001	1079	106334

Table 1. Results of Pathology Labeling

In addition to the large training set CheXpert also contains 200 ground-truth examples to assess our model performance. This set is made up of 200 labeled studies from 200 different patients. Each study was labeled based on a consensus of three different radiologists. These labels are only positive or negative as the three radiologists were able to reach a definitive consensus on all 200 studies.

One consideration important with the use of this data is the potential demographic biases built into the studies captured in this data. While ideally, any

training would be representative of the larger population we know that this set is specific to Stanford Hospital patients. While much of the demographic information is masked in this study we are able to review the age and sex of the patients in figure 4 below.

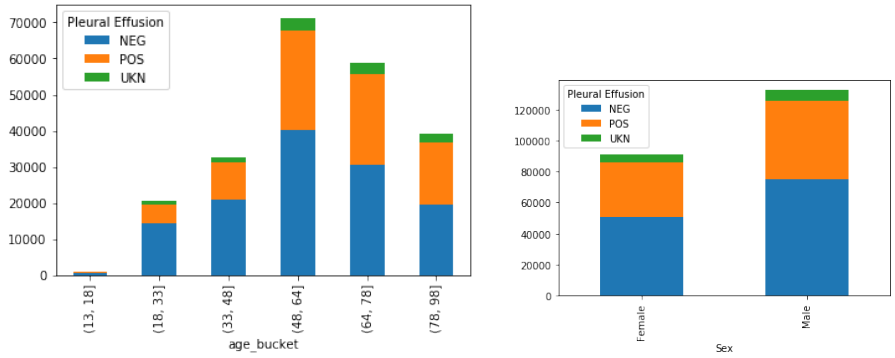

Fig. 4. Age & Gender of Patients in Training Data

This does give us some indication that our training data may not accurately represent the general population, as we see that our data appears skewed to older males. This is important to note when reviewing our results and considering the ethical implications of using a model trained with minimal observations on the younger age groups.

3.2 Data Preparation

In our model we have chosen to work with the down sampled images (11GB) and not the high resolution images (400GB). We did not test any potential improvements in model performance by working with the original images. The baseline we are comparing against was trained using the down sampled 320x320 pixel images.

In our training set we have a third "uncertain" class in our labels that represents 5% of our training data. There is no uncertain class in the validation set. We have considered various methods for handling this, however our models were all trained with these observations excluded from our training set. This is based on the findings from the original authors where a more thorough examination of how to handle these uncertain labels [8]. Additionally, with our labels we just have extracted the Pleural Effusion labels and will train our models against this as a binary target rather than predict across all 14 pathologies.

For some of the studies we were provided both the frontal and lateral x-rays. For our model we have chosen to train the same model for both of these image types as only 15% of our images are from the lateral view. Although lateral views have been seen to allow earlier detection of Pleural Effusion [10]. This may be

an important distinction and future work could train different models based on the type of x-ray.

Finally, there are a large proportion of our patients with multiple studies in our data. Due to the method of labeling this data we must consider how the reports have been written up for patients with multiple studies. For example a patient seen at 83 may have a report identifying and diagnosing the existence of a given condition. If they are seen again a few months later and the same condition is present we can not be certain how that is represented in the reporting and if the labeler will be able accurately extract and classify that report. We have chosen to trust the labels in our training set and use all patient studies with labels positive or negative for Pleural Effusion.

4 Convolutional Neural Networks

In order to understand the model presented in this paper some understanding of convolutional neural networks is expected. We will review some high level concepts that are important to understand why these algorithms are ideal for our problem. To explain these concepts we will assume we have a gray scale image similar to the type of image in our data. That image is converted to $i \times j$ matrix with each value representing it's gray scale for the corresponding ij^{th} entry in the matrix.

4.1 Convolutional Layer

Convolutional neural networks function very similarly to neural networks, the primary difference being the convolutions employed in the network layers for extracting and mapping features to be used in the training of different classifiers or regressors [14].

Convolutions move across the image matrix extracting smaller matrices and mapping it to a single value using the same weight matrix, referred to as a kernel or filter. The resulting output is commonly referred to as a feature or activation map of the prior input. A visual explanation of this process is shown in Figure 5 [6]. Through the training process those kernel weights are adjusted through forward and back-propagation to minimize a loss function using different optimizers, learning rates, and other hyper parameters. A detailed understanding of all these parameters is outside the scope of this tutorial although we will discuss the parameters used in our model in a later section.

The convolution layer is controlled primarily by 4 hyper-parameters; spatial extent, filters, padding, and stride. The spatial extent is the size of each sub-matrix that is extracted from the originally image. Filters control the depth of the resulting output and can be used to extract features of the spatial extent. Padding is another parameter that helps control the output size. This allows the resulting output feature map to be the same width and height as the input feature map by adding additional rows and columns to the input. Strides are often another parameter that can be uses as a means of down-sampling. This

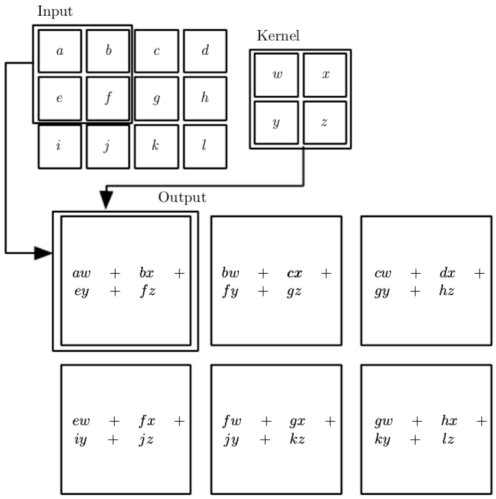

Fig. 5. Convolution Operation from [6]

controls the step size of spatial extent as it moves through original matrix. The default is normally 1, meaning it moves through one pixel at a time. An important consideration is that these parameters cannot be set arbitrarily as the value of one often relies on the values set in the others. Figure 6 shows how the dimensions of an original image are mapped to the final output map.

The corresponding calculations for determining the output dimensions of a layer can be found in Equation 1. For these equations n_{ch} is the number of channels, p is the padding, s is the stride and the filter is described by $f \times f \times n_{ch}$.

$$(n + 2p - f)/(s + 1) \times (n + 2p - f)/(s + 1) \times n_{c_h}^{(1)}$$

4.2 CNN for Computer Vision Problems

One of the major benefits of CNNs for image problems is their ability to learn invariant features [11]. Thought of simply this means that the features learned in one portion of the image or in a spatial extent can be identified in a different location in another image. This allows the CNN to learn fewer features improving the overall performance, for example a network that learns a pattern for a cat in the bottom left of an image, would be able to identify a cat in the top right of another image.

The second largest benefit is their ability to learn the hierarchies of images when run through several convolutional layers [2]. As discussed above each convolutional layer outputs an activation map from a provided input. One way to think that is each value in the output map represents the activation of certain type of feature. So for example the activation map has learned how to identify

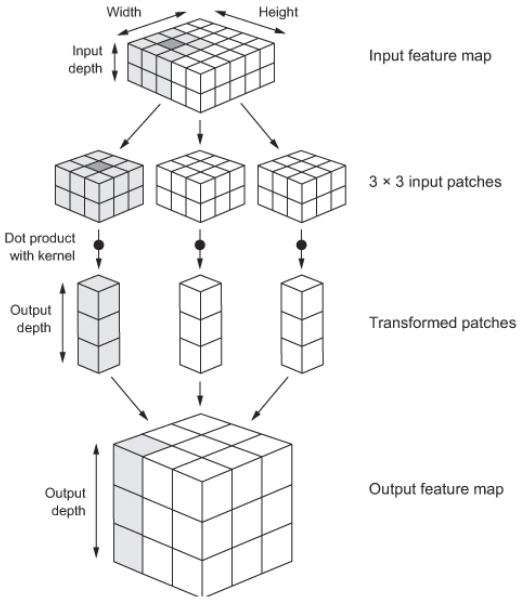

Fig. 6. Convolutional Layer Input to Feature Map from [4]

edges of an object. The values from that activation fed into another convolutional layer that layer could learn that these edges together form an eye, ear, or nose. Then fed through another layer it could begin to learn a face. So when thinking of the an entire convolutional network it is common that some of the early layers identify things like edges and shapes, with the deeper layers beginning to learn more complex features like faces. To highlight this we output what the different layers our network was learning through out a small network in Figure 7.

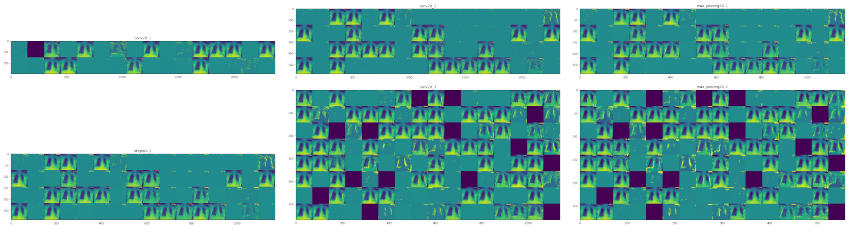

Fig. 7. Convolutional Layers

5 Prior Work

Prior to the release of the CheXpert there has been a large number of studies done using convolutional neural networks to classify some of the common pathologies in chest X-rays [22, 18]. With findings showing models performing well at predicting Pleural Effusion using the ChestX-ray14 data set. Specifically, the model CheXNet which showed a 0.86 AUC against it's validation set.[18]. This was considered state of the art for this dataset.

However, in 2019 CheXpert was released by the Stanford ML group that was seen to be a much larger data set with higher accuracy labels than what was previously available from the ChestX-ray14 set [8]. With this dataset the authors also created baseline models for 5 different pathologies for researchers to compare against. The baseline model was trained using the same training and validation data used in this paper. Although, the final model was evaluated against a separate 500 images that were hand labeled by a consensus from 5 board certified radiologists. The results of their original baseline models for Pleural Effusion were very strong, reporting their strongest ROC value of 0.936 with a 95% CI between (.904 , .967). We will compare our results to this baseline moving forward.

Additionally, we could not find any work utilizing mobile architectures for this data as most work focuses solely on performance. We introduce a mobile architecture to this problem to evaluate potential trade-offs between performance and model size. Our goal is to determine the feasibility of a potential edge deployment of these models using MobileNetV2.

6 Model

In this section we present our methodology for training and evaluating our models for this task. The goal of each model is to predict the probability of the Pleural Effusion from a patient diagnostic imaging study. Both models were trained using a stack of Keras on top of Tensorflow, trained on an AWS GPU instance.

6.1 Model Training

For our model we have chosen to train models using two different architectures DenseNet121 & MobileNetV2 architecture. During our research we found that it DenseNet121 outperformed other common CNN architectures such as ResNet for this task [8] [18]. As no mobile architectures were compared in the original work we have chosen to compare the currently utilized dense architecture to a lighter mobile architecture for this task.

Google's MobilNetV2 builds off a concept known as Depthwise Separable Convolutions which essentials divides the convolutional layer that allows for less costly feature creation from the convolution with a very minimal performance trade-off. Additionally it utilizes uses a more memory efficient layer called a

residual bottleneck layer capable of minimizing the memory usage only the input & output operations. This model was used with a width-multiplier equal to 1 and depth multiplier equal to 1 which are the default parameters used by the MobileNet team [20].

The DenseNet121 architecture provides dense connections between the convolutional layers in the network which is shown to reduce compute resources and alleviate vanishing gradient issues common in deep networks [7]. Max pooling was chosen for the pooling layers, differing slightly from the original published paper, but not uncommon. All other parameters match the original architecture.

For both of our models we tested using both 224x224x3 and 320x320x3 as the resolutions for our input dimensions. For this work, we did not explore how the original hi-res x-rays performed for this task. Using the compressed images would significantly reduce the memory requirements at time of inference but understanding the performance trade-offs could be valuable future work. For our dense architecture we found that the 223x224x3 performed slightly better, while the mobile performed slightly better with 320x320x3, although neither differences were significant. The only pre-processing that was performed on our data was a scaling of of the pixel values between 0,1. No other image pre-processing or enhancements were used in this model. Based on our experimentation, we found that our performance was only minimally impacted when including pre-processing steps commonly used for smaller training sets, although an we did not perform an exhaustive search of pre-procesing steps.

The output of the final layers from both of the models were fed into a global average pooling layer then into a dense layer of size 1 with a sigmoid activation. This is a fairly standard architecture decision for this task, although no other final layers were tested. This final layer's output provides us with the predictions for our binary class.

When training we used the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and a learning rate initialized at 1×10^{-2}. Adam is a stochastic gradient-based optimization algorithm that improves the efficiency of our learning process when adjusting the weights of our kernels throughout the network. Adam is ideal for high-dimensional learning problems like ours and has been shown to require very little memory [12]. Adam is shown to perform most closely to RMSProp it was observed to have a lower training cost on a variety of tasks[12].

Each model was trained for 6 total epochs. For the DenseNet121 model we trained the first 3 epochs with a batch size of 32. The next 3 epochs trained with a batch size of 16, at a learning rate of 1×10^{-3} which we reduced by a factor of one from the initial learning rate as the loss appeared to plateau after the first three epochs. For our MobileNetV2 model we trained all six epochs with a fixed batch size of 16. We used the same learning rate initially as the DensetNet, 1×10^{-2} reducing by a factor of 1 if the loss did not improve over one epoch, which occurred after the 2nd epoch. We checked in with out model performance against our training set at the end of every epoch for the loss, accuracy, & AUROC and stored the weights if the loss performed better than

the prior epoch. By the final epoch our training loss had not reduced since the last epoch.

6.2 Model Evaluation

Once we trained our two models we evaluated their performance against the 200 ground truth studies contributed by the CheXpert authors. Using the binary predictions we see our results for all 200 studies relative to the actual label in Table 2.

DenseNet121

		Positive	Negative	Total
Actual	Positive	44	20	64
	Negative	6	130	136
	Total	50	150	200

MobileNetV2

		Positive	Negative	Total
Actual	Positive	49	15	64
	Negative	16	120	136
	Total	65	135	200

Table 2. Confusion Matrices for the Two Models

Based on the comparisons of the two tables we see that while our mobile model has a lower overall accuracy it has fewer false positives than the dense model. The error in DenseNet121 seems to driven primarily by the model's false negative rate which is concerning as a false negative indicates the model failed to identify a harmful condition from the image. MobileNetV2 shows a lower false negative rate, but a much higher rate of false positives. Each model appears to have their own merits when evaluating across a fixed decision point of 0.5 however that is an unlikely scenario for this task. To better understand our model fit we evaluate the precision-recall curves & ROC curves in Figure 8.

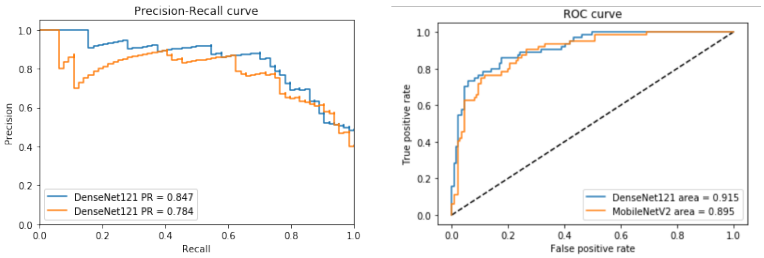

Fig. 8. PR & ROC Curves

The precision recall curve on the left of Figure 8 shows that we are able to sacrifice some of our recall score in favor of a higher precision. Similarly, our ROC curve from these prediction, on the right of Figure 8, better illustrates our trade off between true positive rate & false positive rates across the various decision points. As you can see that the two different models both perform very similarly with even several decision points along the graph that would yield the same results. An even further indication that the two models have similar performance overall, but the differences in the curves have learned different structures in the data they are utilizing to make predictions. We also compare these results to current baseline results for this task.

The results of their original baseline models for Pleural Effusion were very strong, reporting their strongest ROC value of .936 with a 95% CI between (.904, .967) [8]. For comparison purposes we constructed 95% bootstrap confidence intervals for our AUC by taking 10,000, 50 observation samples with replacement from our test set. We calculate our DenseNet121 AUC at 0.92 (95% CI of 0.819, 0.984) and our MobileNet AUC at 0.90 (95% CI of 0.789, 0.975). While the results are slightly lower we do find that we are able to reproduce current baseline performance with both a single dense network as well as the mobile network.

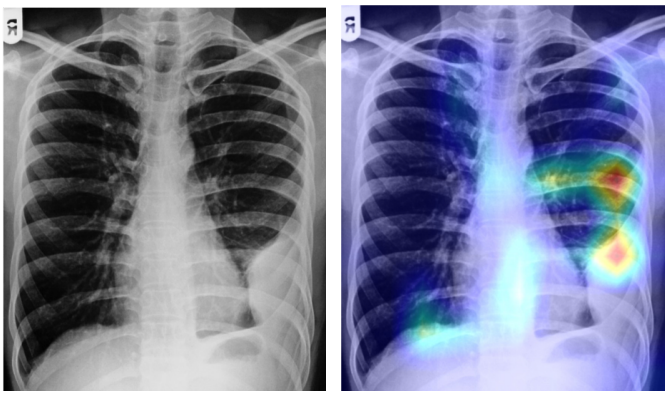

Fig. 9. Original X-Ray Image and Heat maps of class activation

In addition to the prediction, our model can be used to output have created class activation heat maps. Figure 9 shows the original x-ray and super imposing the class activation heat map on the original picture. This visualization points the most activated portion of the image for this specific class. Meaning the area contributing to the positive prediction the most. This helps radiologist see why the model thinks the x-ray has positive pathological condition of Pleural Effusion and where it is located for them to make there own diagnosis. Proper evaluation of the model requires the guidance of a radiologist to validate the areas located by the mappings are in fact Pleural Effusion, however after a review of several

mappings we were able to determine several patterns in the images resulting in heavy activation of our model.

While model performance is a key metric for these models it is important to understand the trade-offs between performance relative to the number of parameters and floating point operations (FLOPs). This provides more information about the hardware requirements necessary to deploy this model and thus guides decisions on how these predictions can be implemented into operations. We show the the estimated size and model speeds in Table 3.

Model	AUROC	Model Parameters	GFLOPs
Baseline	0.93	21,115,515*	34.76*
DenseNet121	0.92	7,038,505	5.69
MobileNetV2	0.90	2,259,265	1.23

Table 3. Model Performance, Size & Speed (Baseline values estimated based on available information)

As you can see that our models seem to degrade slightly in terms of performance as the number of parameters & floating point operations decrease. Although the overall model performance remains comparable, we are able to reduce the number of parameters by a factor of 10 and the amount of FLOPs by a factor of 30.

7 Ethical Considerations

One of the primary ethical considerations of working with personal medical information is with respect to the sensitivity of this information. All identifying information was masked in our training data with the exception of age, gender, and the diagnosis. Additionally access to this data was granted to us through the Stanford University School of Medicine and all use of this data is in line with our signed research use agreement. There is no indication that written consent was provided from the patient for use of this data so we take reasonable means to handle this data responsibly.

While this model is intended for research use only we also recognize implicit bias from the demographics of the population used in training. If our sample is not from a representative sample of the population the results and potential diagnosis would likely be impacted. Size, shape and lung capacity varies based on race [3] as much as 17-20% difference between groups. Demographics, variations in equipment, different imaging characteristics influences the outcome. These factors increase the risk of misdiagnosis and potential harm to patients. For our particular training data we recognize that our population may be underrepresented of younger population and slightly over-represented by males. No other demographic information was provided with this data which limits our

ability to assess the models robustness across various demographic characteristics.

These results are intended as a tool for practicing radiologist and by no means intended as a replacement. Marking an x-ray with Pleural Effusion positive whereas it is not could lead to inaccurate diagnosis or if it occurs frequently would reduce confidence in the model. Inversely, the model's failure to correctly identify Pleural Effusion could result in variety of outcomes that could effect the ability to deliver the correct treatment. Reducing false negatives should be primary concern and the trade off between false negatives and false positives can be determined through how we select decision points in our ROC & PR curve. Legal liability issues would be raised if this model were to be used without the supervision of radiologist as any misinterpretation leads to patient harm. Because of these concerns this tool is relies on human participation in the decision loops. One additional benefit of incorporating humans into this loop is the ability to obtain feedback on predictions. Capturing the radiologist diagnosis relative to the prediction provides another mechanism to improve model performance.

8 Conclusion and Future Work

We found that both models were able achieve comparable performance to the current baseline models for this data using two different single network models (DenseNet121, MobleNetV2). The original authors were able to achieve state of the art model for the task of identifying Pleural Effusion using an ensemble of models. The models introduced are able to produce results with an AUC of .92 using Densenet121 and .90 using MobileNetV2. By reproducing the baseline performance we showed that lightweight architectures are able to produce comparable results with up to 10x fewer parameters and 30x fewer floating point operations. This positive contribution would allow for a lot more flexibility in terms of deployment strategies. For example, these models could be integrated into portable x-ray machines as portable radiographs are one of the main tools to monitor patients in intensive care unit (ICU) and in-patient facilities.

Future work could include the exploration of various lightweight architectures to understand the trade-offs between model performance and model size & speed. These trade-offs could help guide deployment decisions such as the directly deploying to edge devices. These architectures also should be tested against other pathology beyond effusion to determine if a similar trade-off between performance and size exists.

Additionally, our model only performance is evaluated on a small subset from the same population sample as our training images. We recognize the concern with these types of data sets and the potential for over fitting [1]. As future work we propose to collect additional test or validation images to asses the robustness of our model to help address some of these concerns. Specifically, a collection from a source outside of the Stanford hospital.

[1] `https://lukeoakdenrayner.wordpress.com/2018/01/24/chexnet-an-in-depth-review/`

References

1. Adrian P Brady. Error and discrepancy in radiology: inevitable or avoidable? *Insights into imaging*, 8(1):171–182, 2017.
2. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
3. PM Donnelly, TS Yang, JK Peat, and AJ Woolcock. What factors explain racial differences in lung volumes? *European respiratory journal*, 4(7):829–838, 1991.
4. Chollet Francois. Deep learning with python, 2017.
5. P Goddard, A Leslie, A Jones, C Wakeley, and J Kabala. Error in radiology. *The British journal of radiology*, 74(886):949–951, 2001.
6. Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. Deep learning, volume, 2016.
7. Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
8. Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.
9. Alistair EW Johnson, Tom J Pollard, Seth Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr: A large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.
10. Vinaya S Karkhanis and Jyotsna M Joshi. Pleural effusion: diagnosis, treatment, and management. *Open access emergency medicine: OAEM*, 4:31, 2012.
11. Koray Kavukcuoglu, Rob Fergus, Yann LeCun, et al. Learning invariant features through topographic filter maps. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1605–1612. IEEE, 2009.
12. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
13. Anna S Kookoolis, Jonathan T Puchalski, Terrence E Murphy, Katy LB Araujo, and Margaret A Pisani. Mortality of hospitalized patients with pleural effusions. *Journal of pulmonary & respiratory medicine*, 4(3):184, 2014.
14. Yann LeCun, Koray Kavukcuoglu, and Clément Farabet. Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pages 253–256. IEEE, 2010.
15. Richard W Light. The undiagnosed pleural effusion. *Clinics in chest medicine*, 27(2):309–319, 2006.
16. Richard W Light, M Isabelle Macgregor, Peter C Luchsinger, and Wilmot C Ball. Pleural effusions: the diagnostic separation of transudates and exudates. *Annals of internal medicine*, 77(4):507–513, 1972.
17. Royal College of Radiologists. Clinical radiology uk workforce census 2018 report, 2018.
18. Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

19. Andrew B Rosenkrantz, Danny R Hughes, and Richard Duszak Jr. The us radiologist workforce: an analysis of temporal and geographic variation by using large national datasets. *Radiology*, 279(1):175–184, 2015.
20. Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.
21. Chuin Siau. *Pleural Disease and Pneumothorax*, pages 1752–1756. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.
22. 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. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.