Identifying Pathologies from Chest X–ray Data

Alan Akil[1], Triet Duong[1], Sulaimon Oyeleye[1]

1: University of Houston

Abstract

Chest X–rays are widely used all over the world and have been proven to be incredibly useful for doctors to diagnose a large variety of pathologies. In 2017, NIH presented the largest chest X-ray database to date: "ChestX–ray8"[1]. Using this data set, we trained several convolutional neural networks to classify different thoracic pathologies from human chest X-ray images. Current state of the art results only classify 8 pathologies, however we achieved comparable results for classifying all 14 classes (13 pathologies and no pathology). This will allow for a faster and more accurate diagnoses of thoracic diseases, which will lead to a more efficient care of patients in hospitals around the world.

I. INTRODUCTION

A frequent and cost–effective medical imaging examination is the chest X–ray, since a clinical diagnosis via other methods such as CT imaging can be more challenging and costly. In fact, achieving clinically relevant computer-aided detection and diagnosis (CAD) 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.

Our desire is to train a neural network that can robustly detect and diagnose diseases. Ultimately, this artificial intelligence mechanism can lead to clinicians making better and quicker diagnostic decisions for patients by using a computer which has been taught to read and process extremely large amounts of scans, to confirm the results radiologists have found and potentially identify other findings that may have been overlooked.

In addition, this may also be able to: (1) help identify slow changes occurring over the course of multiple chest x-rays that might otherwise be overlooked; (2) benefit patients in developing countries that do not have access to radiologists to read their chest x-rays; and (3) create a virtual radiology resident that can later be taught to read more complex images like CT and MRI in the future.

Current state of the art results are in [2]. Since this is a multi–label problem, where one image can contain more than one disease, receiving operating characteristic (ROC) curves are better measures of performance than accuracy, and in this paper they report area under the curve (AUC) of the ROC in the range of 0.6–0.8. In [2], they achieved classified multi–label images well, but reduced the dataset considerably and only considered 8 pathologies.

Here, we achieved similar results but on all 13 pathologies and the healthy case (14 classes). We did this by training a deep convolutional neural network, where the convolutional layers were taken from pre–trained networks such as: ResNet, Xception, Inception, VGG16, etc. On top of each of these pre–trained network, a series of fully connected layers were used, and only trained the weights in the fully connected layers. After training the network AUC of the ROC for each disease was in the range between 0.5–0.8.

II. DATASET

This NIH Chest X-ray dataset [3] is comprised of $112,120$ X-ray images with resolution 1024×1024 and disease labels from $30,805$ unique patients. The 14 image labels correspond to the 13 diseases and the healthy case. Note that, as mentioned before, this is not a simple multi–class problem where one image has one label only. Here, each image may have more than one label. (Fig. 1 & 2). The labels were mined from the associated radiological reports using Natural Language Processing (NLP).

The 14 classes are common thoracic pathologies: Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pneumonia, Pleural_thickening, Cardiomegaly, Nodule, Mass, and Hernia. These text-mined disease labels are expected to have accuracy $> 90\%$. Meta data for all images includes: Image Index, Finding Labels, Follow-up , Patient ID, Patient Age, Patient Gender, View Position, Original Image Size and Original Image Pixel Spacing.

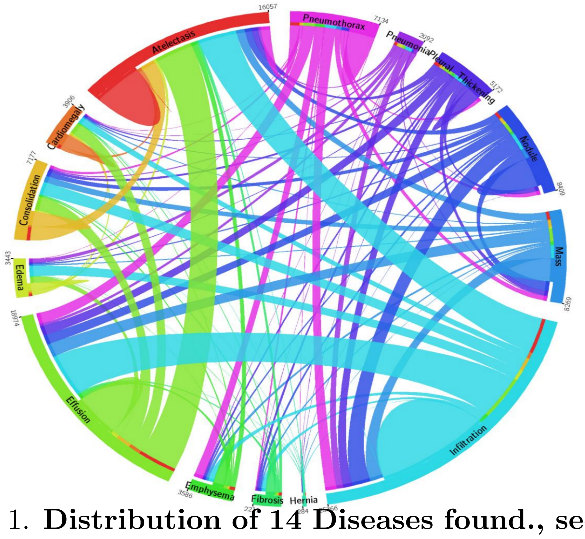

FIG. 1. **Distribution of 14 Diseases found., see [2]**

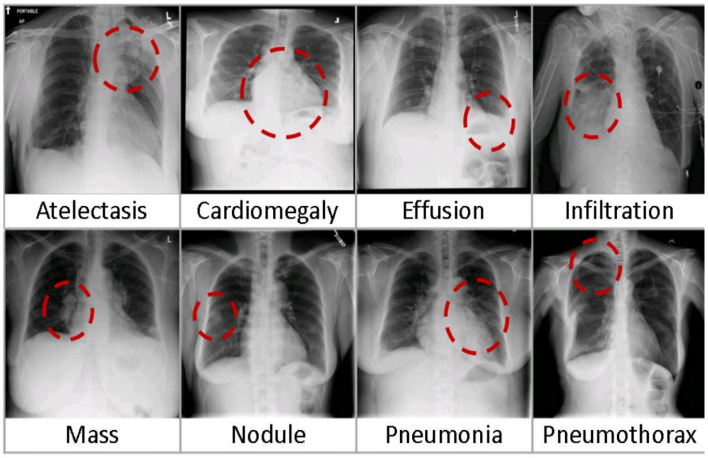

FIG. 2. **Eight common thoracic diseases found in the chest x-rays, see [2].**

III. METHODS

We sampled 200 images for each class with only one disease, including the 'No findings' (healthy) class, from the full dataset. We also sampled 300 images where the patient had more than one disease (multiple labels). We also resized the images to 79x79. At this resolution, the images were still clear and the diseases were identifiable. There are a number of reasons why we sampled only 200 images per class: (1) To keep the classes balanced – the lowest occurring classes had about 200 images each; (2) To train the models in a decent computational time without the access to high performing GPU's.

We trained a number of pre–trained architectures: ResNet, VGG16, Inception, and Xception. On top of each of these, two fully connected hidden layers were stacked containing 256 and 128 neurons each. At each layer, the inputs were normalized (Batch Normalization), and 30% of the neurons were switched off to prevent overfitting (Dropout). And the last layer was a softmax that gave probabilities for each class.

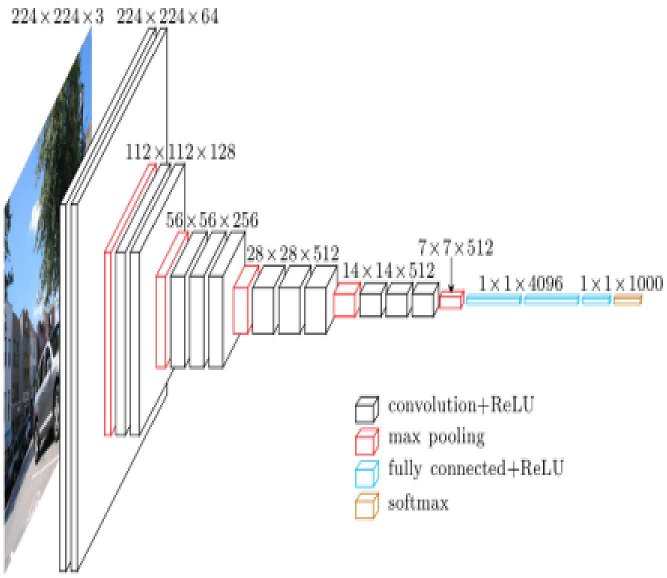

FIG. 3. **VGG16, see** [4] **.**

First, we trained the network to only classify images into single label (excluding the 300 samples from classes that have more than one disease). We trained the networks with several different hyperparameters (number and sizes of the hidden layers, learning rate, activation function etc) and several network architectures (VGG16, LeNet5, Xception, ResNet50 etc) in order to achieve the best results. Next, we trained on images with multi–labels as well: in this case, the last layer was a fully connected layer of 14 nodes (each representing a class) and with "sigmoid" as the activation function, which is commonly used for multi–label classification problems. The sigmoid outputs a number between 0 and 1, we set a threshold at 0.5, and each image was classified into every class with output higher than 0.5.

IV. PREVIOUS RESULTS

The original paper [2] also classified images into multi people diseases, but only within 8 classes. They trained their network using four commonly used pre–trained models: ResNet50, GoogleNet, VGGNet and AlexNet. ResNet50 was found to be best these 8 classes according to the paper, and performed best on the classes: Cardiomegaly (AUC=0.8141) and Pneumothorax (AUC=0.7891).

V. RESULTS

We performed two approaches, multi–label approach which was similar to the original work and a multi–class (or single label) approach. In both approaches, we focused on all 14 pathology classes instead of 8 classes.

We first trained the networks to identify one pathology only, in other words to classify each image to one label. After testing on different architectures, we obtained the best results using VGG16. The train accuracy reached 100% and the test accuracy only 23% (Fig. 4), which is a clear sign of overfitting. These results were actually improved by the use of two techniques: BatchNormalization and Dropout. The loss on the training set converged to nearly zero, and we computed the confusion matrix in order to understand how the network was behaving (Fig. 5). We see that the the largest values lie along the diagonal, which suggests good classification performance, but since the dataset was small, the accuracy ends up being small.

In the multi-label approach, we also used the pre–trained VGG16 network and classified each X–ray image into multiple labels if more than one disease is present. Here, we reached results comparable to the original paper [2]. Our best classified classes were Hernia(AUC=0.83), Consolidation(AUC=0.60), Edema(AUC=0.85), Emphysema(AUC=0.66), Pleural Thickening(AUC=0.69) and Fibrosis(AUC=0.66).

Whereas, in the original paper, the best classified classes were: Cardiomegaly (AUC=0.8141), Pneumothorax (AUC=0.7891), Effusion (AUC=0.7362), and Atelectasis (AUC=0.7069). Even though our network did not perform as well when classifying Cardiomegaly(AUC=0.53) and Pneumothorax(AUC=0.65), we still were able to classify the above mentioned classes decently.

VI. DISCUSSION

Throughout this work, we run into a number of issues. The first issue was the imbalance of classes. We fixed this by sampling a reduced number of images from the full dataset keeping the classes uniformly distributed. We found a large number of Kaggle results claiming to have high accuracy on single–label classification ($\sim$ 85%), but when we analyzed each of those results, we found that those network were just classifying everything into the largest class (which was the healthy case).

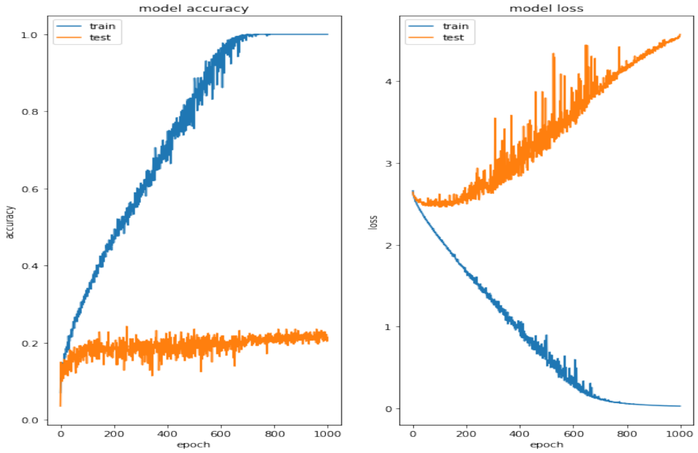

FIG. 4. **Accuracy curve.**

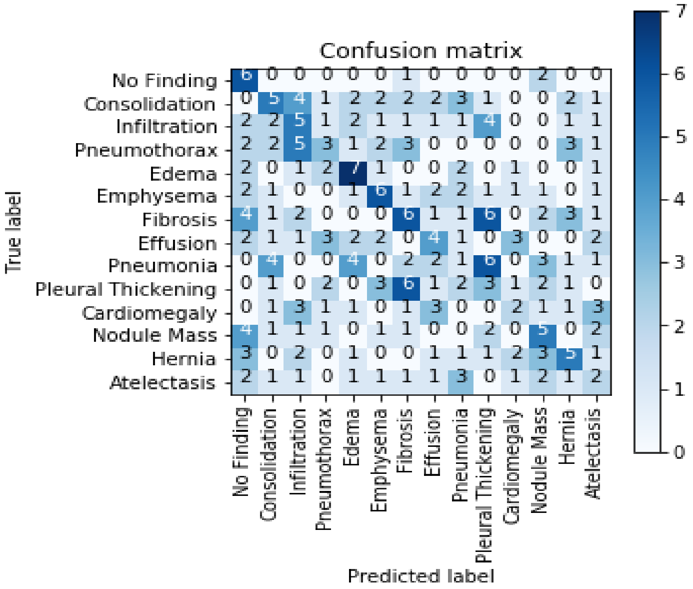

FIG. 5. **Confusion matrix.**

The second issue that presented was now the small number of images to train the network. We believe that such small number of examples may have been the cause of why the networks were overfitting so easily (See confusion matrix in Fig. 5).

The third issue, and one of the main issues, was the fact that the labels were obtained through NLP with accuracy above 90%. This causes the obvious issue of possibly confusing the network during training and testing. And it might be one of the reasons why test accuracy was low on the single–label case.

VII. CONCLUSION

In conclusion, we introduced two models that can predict pathologies from images of chest X–ray. Even though do not perform as well we intended, still perform close to state of the art. In the literature,

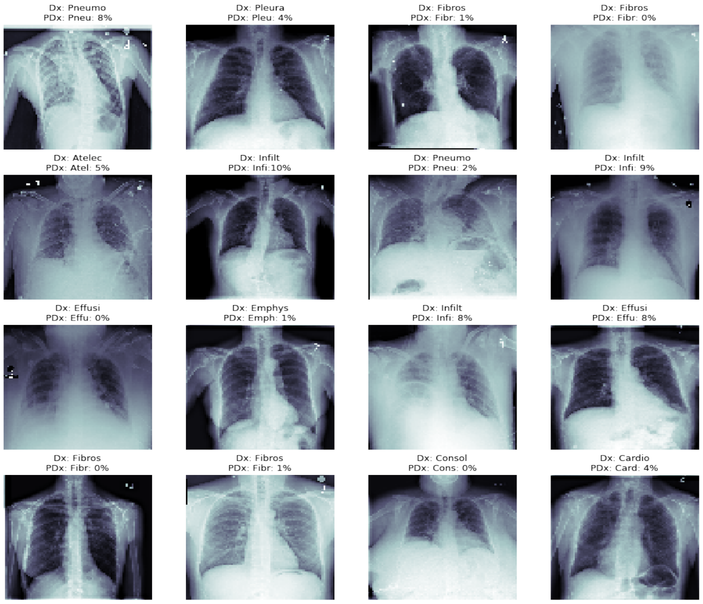

FIG. 6. **Multi-Label Chest x-rays results.**

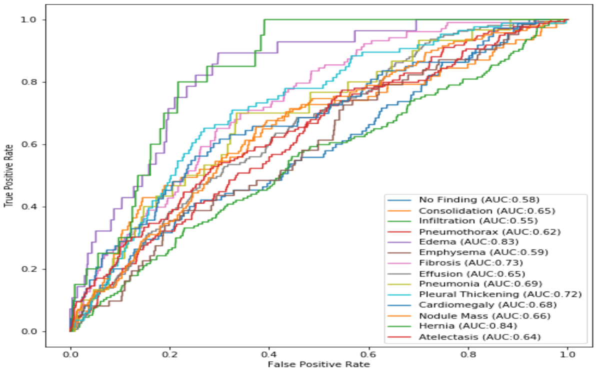

FIG. 7. **Multi-label classification performance using the VGG16 model on the training set.**

we did not find anyone approaching this problem as single–label and obtaining reasonable results on the 14 classes (Most models were classifying all images into the largest class).

On the other hand, in the case of images with multiple labels, we performed comparably to state of the art, even though we classified into 14 classes instead of 8 classes.

We believe that our developed models are obviously not ready for diagnosing patients, but they take us a step forward in that direction. Ultimately, these types of models will help improve and speed up diagnoses in hospitals all over the world. And in the end, they will help build models for identifying pathologies in more complex imaging techniques such as CT scans.

Note: Training of all the models in this paper were trained in a laptop Acer Aspire E 15 with Intel Core i5–6200U. This laptop contained a GPU NVIDIA FORCE 940MX that was utilized for training the neural networks.

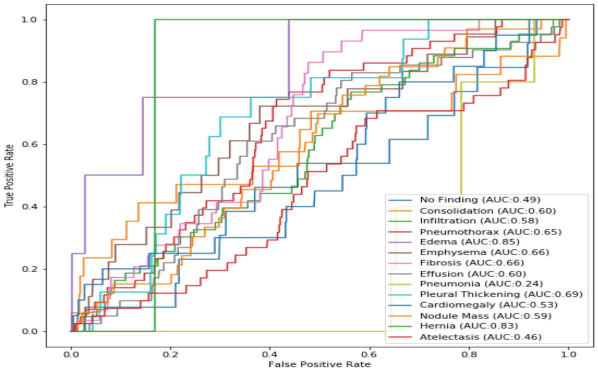

FIG. 8. **Multi-label classification performance using the VGG16 model on the test set.**

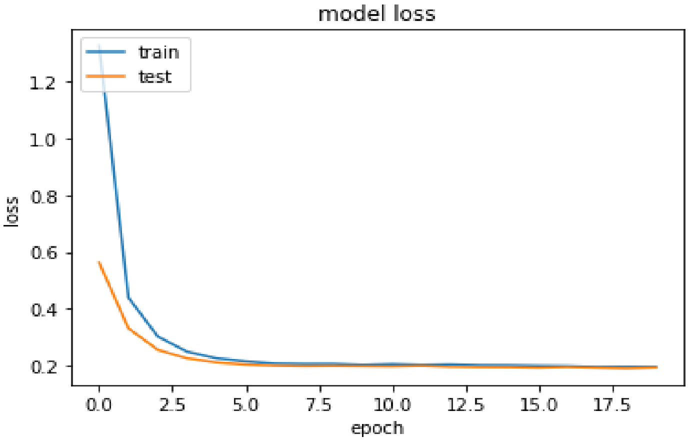

FIG. 9. **Loss function.**

[1] "NIH Chest X-Ray data," https://nihcc.app.box.com/v/ChestXray-NIHCC/folder/36938765345 ().

[2] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017) pp. 3462–3471.

[3] "Random Sample of NIH Chest X-ray Dataset," https://www.kaggle.com/nih-chest-xrays/sample ().

[4] "VGG 16," https://towardsdatascience.com/step-by-step-vgg16-implementation-in-keras-for-be