Detecting and Classifying Lung Diseases using X-ray images

Lu Lechuan, Ong Kuan Yang, Rhynade See Ey, Suyash Shekhar, Tan Jin Wei, Teo Ming Yi
lu.lechuan@u.nus.edu, kuanyang.ong@u.nus.edu, rhynade.see@u.nus.edu, suyashshekhar@u.nus.edu, jinweitan@u.nus.edu, mingyiteo@u.nus.edu

Abstract

In this project, we use Chest X-ray images to detect the presence of lung diseases and classify the images with diseases into 14 different types of lung diseases. Our goal is to build upon and improve the current best performing model - CheXNet. We did this by exploring different pre-processing techniques and experimenting with methods such as two-phase training. While we did not manage to surpass the performance of CheXNet in the end, we gained significant insight into the ways to tackle this problem and dataset.

Background/Introduction

Chest X-ray examinations are one of the most frequent and cost-effective form of medical imaging. However, the clinical diagnosis of chest X-rays can be challenging, and sometimes are believed to be harder than diagnosis via chest CT imaging. Furthermore, with the overwhelming number of medical diagnostic images coupled by the lack of medical professionals, patients are unable to receive immediate diagnosis and attention. Hence, through this project, we hope to improve disease detection through X-ray scans analysis so that medical diagnosis in the future can benefit in terms of both speed and accuracy.

The specific problem that we are studying is using chest X-ray images to detect the presence of lung diseases and classifying the images with diseases into 14 different types of lung disease.

Dataset

We used NIH's Chest X-ray dataset of 14 common Thorax diseases. ChestX-ray14 is the largest chest X-ray dataset currently available and it consists of 112,120 frontal X-ray images of 30,805 unique patients.

The dataset contains multi-label images - there are 20,796 images labelled with more than one disease.

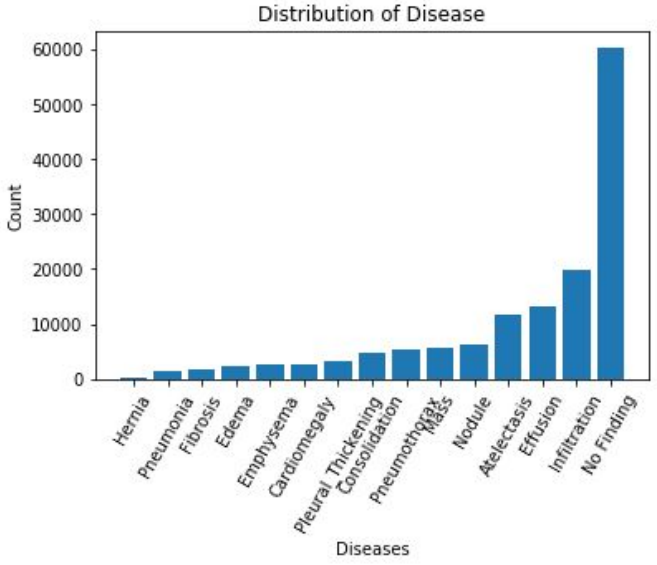


Fig 1: Distribution of Disease Labels

This dataset is severely imbalanced with the majority class - "No Finding" making up 53% of the dataset while the minority class - "Hernia" only makes up 0.2% of the dataset.

Related Work

Wang et al. [1] created the ChestX-ray8 dataset and initiated the use of the deep convolutional neural network (CNN) to classify these images. They used different pre-trained models including AlexNet, GoogLeNet, VGGNet-16 and ResNet-50.

Subsequently, Yao et al. [2] outperformed Wang et al. using a two-stage neural network that combines a CNN image encoder with a recurrent neural network (RNN) decoder. This allows for the exploitation of the dependencies between disease labels. They also trained the network from scratch to ensure that the best application-specific features were captured.

The current state of the art neural network is CheXNet. CheXNet is a 121-layer Dense Convolutional Network (DenseNet) developed by Rajpurkar et al. [3] from the Stanford ML group. Unlike Yao et al., the network is pre-trained on ImageNet.

The last fully connected layer produces a 14-dimensional output, after which an elementwise sigmoid nonlinearity is applied to get the final output, which is the set of predicted probabilities of the presence of each of the 14 disease classes.

Methods

Preprocessing

Extracting features from images is crucial to perform well in disease classification. Several image enhancement techniques were used in an attempt for feature extraction.

Histogram Equalisation and Contrast Stretch: Both methods aim for contrast adjustment. As seen in Fig 2 below, image pixel intensities are more spread out after applying histogram equalisation or contract stretch. Images are expected to look clearer after the enhancement.

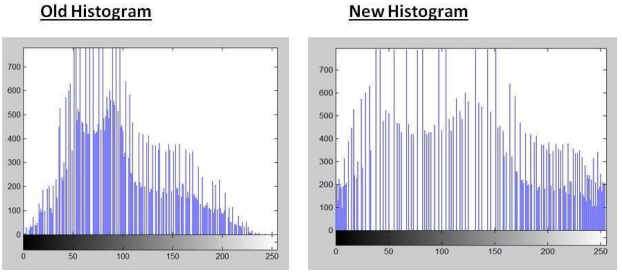


Fig 2: Distribution of pixel intensity before and after histogram equalization

Edge Detection: an image processing technique for finding the boundaries of objects within images. It works by detecting discontinuities in brightness.

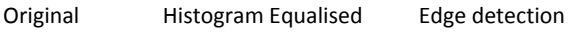


Fig 3: Preprocessed Images

Model Architecture
With the imbalanced dataset problem where 53% of the images are labelled as *No Finding*, and the other 47% a combination of all 14 diseases, we decided to use a 2-stage classifier. In the first stage, our model classifies the images into *Finding* vs *No Finding*. Images classified as *Finding* will then be fed into a second model where it is further classified as one or more of the 14 different diseases.

First Stage: Disease Vs Non-Disease (Binary Classifier)
In this first stage, we perform a binary classification that aims to classify the images into either of the 2 classes, *Finding* or *No Finding*.

Before feeding our data into a model, we first converted the target column into the 2 classes. Images originally labelled as No Finding will remain while images with a disease (both single and multi-label) will be labelled as Finding. Our label column is $y_i \in [0, 1]$ where y = 0 if No Finding, 1 otherwise.

Baseline Model
As our baseline, we used Naive Bayes to classify the images into the 2 classes. Naive Bayes works well as a baseline since it is relatively easy to implement.

Instead of using the original pixel values as features for the model, we used two better features, namely Haralick Texture and Histograms, and used the output as features for the model. Haralick Texture was chosen since it is a well-known mathematical method to detect lung abnormalities. It shows how often each gray level occurs at a pixel located at a fixed geometric position relative to each other pixel, as a function of the gray level (Srinivasan and Shobha 2008). On the other hand, histograms are used to show the gray level intensities of the X-Ray images.

We first resized each image to $500 \text{ x } 500$ and then applied 2 feature extraction functions to the images. The output - a one-dimensional vector of size 269, is fed into the Naive Bayes model.

Deep Learning Model
We used the original CheXNet architecture and added two more layers - one fully connected layer that matches the number of labels and another sigmoid layer. For the optimizer, we used SGD with initial learning rate set to 0.02, momentum 0.9 and weight decay $1e^{-4}$. While training, we reduced the learning rate by a factor of 2 when there is no improvement in the validation loss over a number of epochs.

The training loss function used is Binary Cross Entropy: $l(x, y) = -[y\ log(x) + (1-y)\ log(1-x)]$
where y = actual label, x = predicted probability.

Second Stage: Identifying Disease Labels

Multi-label Multi-class Setup
In this multi-label multi-class setting, our team has defined a 14-dimensional label vector $Y = [y_1, \ldots, y_i, \ldots, y_n]\ where\ y_i \in [0, 1]$ and n = 14 for each image. y_i is a binary variable that indicates the presence of the disease class i, 0 - absent, 1 - present.

Model Overview
DenseNet is a 121 layers convolutional neural network. DenseNet ensures a maximization of information flow between layers and it connects all layers (with matching feature map sizes) directly with each other. To fit our use case, we added two more layers: a fully connected layer with an output size of 14 and a sigmoid layer to model the binary cross entropy loss.

The weights of the layers are initialized with pre-trained ImageNet weights. The model is trained with stochastic gradient descent optimizer with an initial learning rate of $1e^{-2}$ and weight decay of $1e^{-4}$. In order to avoid stagnating at a local minimum, a new optimizer is generated with a decay rate of 0.1 in the learning rate when it detects that there is no improvement in the validation loss over 3 consecutive epochs.

Two-Phase CNN
The disease only data is still imbalanced. Infiltration - the major disease class has 19,894 images while there are only 227 images for the Hernia class. The imbalanced ratio is approximately 100:1. In order to resolve the issue, our team proposed a two-phase convolutional neural network training architecture.

In the first phase, minority classes such as Hernia are oversampled along with data augmentation techniques like horizontal flipping while the majority classes are undersampled to create a balanced dataset. We proceeded to train a 121-layers DenseNet with the balanced dataset. This is to ensure that the neural networks to have sufficient information to learn all the disease classes.

In the second phase, we freeze the weights of the neural networks except for the last two layers - fully connected layer and sigmoid layer. The freezing of weights helps to ensure that the information for which the neural network has learnt from the balanced dataset is retained. The network is then trained on the original dataset with the actual distribution of classes. In this phase of training, only the weights of the fully connected layer and sigmoid are modified. This ensures that the model learns the frequencies of each disease class in the dataset.

Simulated Annealing
In order to avoid being trapped at a local minimum cost, Simulated Annealing is added in the process of training the model. Previously when the model generated a new epoch with a higher validation loss as compared to previous epochs, it will be discarded immediately. Simulated Annealing allows a certain probability to accept a worse epoch (epoch with higher validation loss than before) so that there exists a chance to escape from the local minimum cost. The probability of acceptance decreases exponentially, meaning epochs can gradually converge to the global minimum cost.

Evaluation

Metrics
Similar to many of the existing literature, we used the Area Under the Receiver Operating Characteristic (AUROC) curve as our evaluation metric. We plotted the Receiver Operating Characteristic (ROC) curve with the True Positive Rate (TPR) as the y-axis and False Positive Rate (FPR) as the x-axis. The points are generated from setting the threshold for the probabilistic output of our model at various different values and plotting the TPR and FPR for each value. We then evaluated the Area Under the Curve (AUC) of the ROC curve. An AUROC value of 0.5 corresponds to random guessing while a value of 1 corresponds to a perfect classifier.

In this project, we are more concerned with False Negatives than False Positives because incorrectly diagnosing a patient with lung disease as healthy has much more severe consequences than misdiagnosing a healthy patient as having lung disease. As such, we would ideally strive for a threshold that maximizes TPR as much as possible while maintaining a reasonable FPR.

Results

First Stage (Binary Classifier):

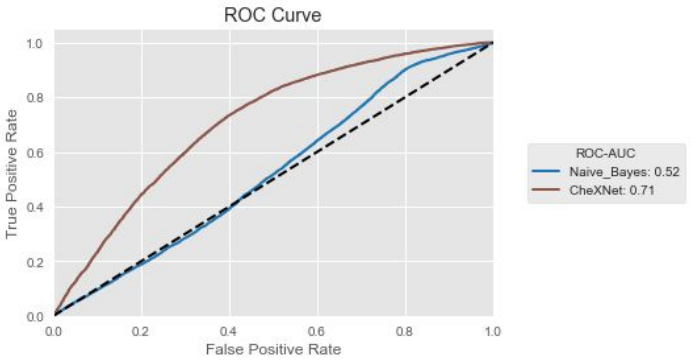


Fig 4: ROC curve of first stage classifier

From the AUC of 0.52 for Naive Bayes, we conclude that a Naive Bayes model does not adequately capture the complexity of the dataset. Despite performing feature extraction, it seemed like the features chosen are not sufficient for this binary classification problem. On the other hand, using CheXNet, AUC obtained increased to 0.71. This is not surprising since CNNs are known to perform much better in image classification.

Second Stage

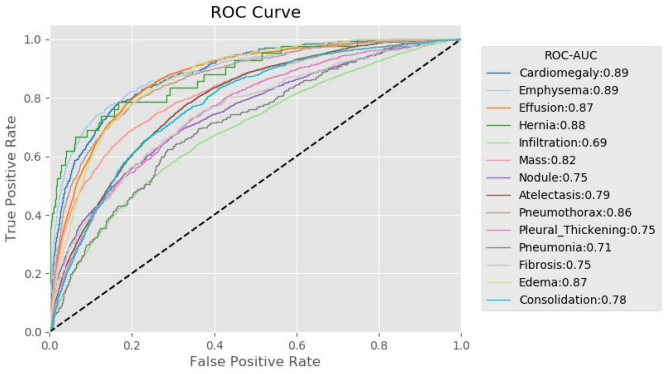


Fig 5: ROC curve of second stage classifier

Pathology	Wang et al. (2017)	CheXNet (2018)	Ours
Atelectasis	0.716	0.809	**0.789**
Cardiomegaly	0.807	0.925	**0.886**
Consolidation	0.708	0.790	**0.778**
Edema	0.835	0.888	0.866
Effusion	0.784	0.864	**0.872**
Emphysema	0.815	0.937	**0.890**
Fibrosis	0.769	0.805	0.750
Hernia	0.767	0.916	**0.881**
Infiltration	0.609	0.735	**0.686**
Mass	0.706	0.868	**0.818**
Nodule	0.671	0.780	**0.747**
Pleural Thickening	0.708	0.806	0.755
Pneumonia	0.633	0.768	**0.710**
Pneumothorax	0.806	0.889	**0.864**

Table 1: Comparison of model performance with the previous state of the art. (Bold font indicates an increase by more than 0.05 as compared to Wang et al.)

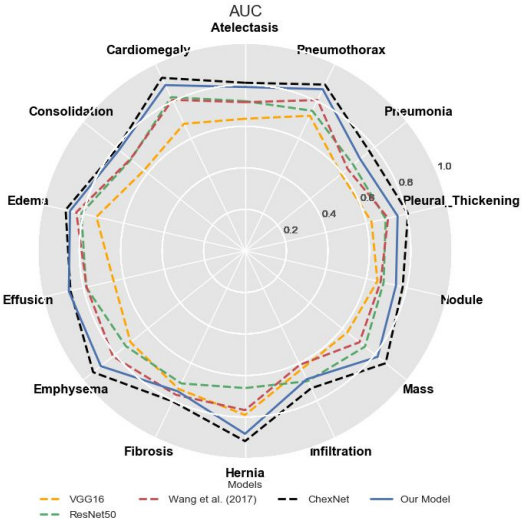


Fig 6: Comparison of model performance

From Table 1, the average AUC obtained is approximately 0.807. The performance of our model is better than the previous state of the art (Wang et al. 2017). From the AUCROC results, we can see that the our model's performance is better at detecting at least 11 out of the 14 disease classes. Atelectasis, Cardiomegaly, Consolidation, Effusion, Emphysema, Hernia, Infiltration, Nodule, Pneumonia and Pneumothorax have an AUCROC of more than 0.05 as compared to Wang's model.

However, our model performs slightly worse than the current state of the art (CheXNet. 2018), despite using a similar underlying architecture - 121-layer DenseNet.

Discussion

Preprocessing

The set of preprocessed images does not yield a better result than the original one. There are several possible reasons for being so.

As shown in the diagram below, the original cumulated histogram is similar to a straight line, this means the original images are near "histogram equalised" and their contrasts are sufficient. Therefore histogram equalisation and contrast stretch techniques add little values to them.

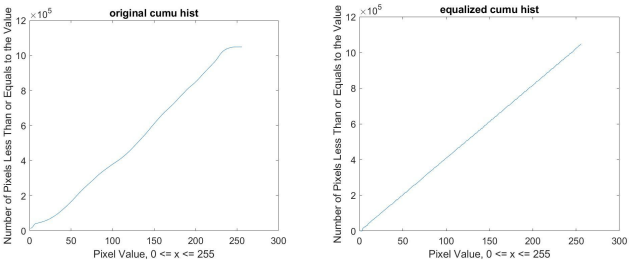


Fig 7: Cumulative histogram before and after histogram equalization

The edge detected images set does not work well as diseases features might be lost. Since edge detection is to find the boundaries of objects within images, clear boundaries like bones are captured well using edge detection while unclear regions including lesion regions are neglected.

Model: First Stage Classifier

While Naive Bayes models are efficient in terms of training time and easy to build with no complicated iterative parameter, there are two main weaknesses of using it in this problem of image classification.

Firstly, it is not possible to feed an image directly into a Naive Bayes model. Rather, the input has to be a vector so the 2D array of pixel values has to be flattened into a vector. This causes a great deal of information to be lost, specifically the spatial relationship within the image. However, spatial relationship is most likely needed to perform well for this binary classification problem since the position of lesion regions is crucial in identifying whether a lung disease is present and the type of disease present.

Secondly, Naive Bayes operates on the assumption that features are independent. Again, this might not be true in our case where the pixel values are likely to be dependent on its neighbouring values.

CNNs address both of these problems. During the convolution step, it preserves the spatial relationship between pixels by learning image features using small squares of input data (also known as a filter). This filter is applied to the image in a sequential order, so preserving the spatial relationship.

In addition, CNNs are known to do well in image classification because it can learn abstract concepts present in images. By applying different filters, the network can learn a lot of low-level features such as straight, diagonal lines, edges. These features are further stacked together to make intricate patterns (e.g shape of lungs) which are useful in detecting X-ray images that consist of lesion regions. Traditional classification methods such as Naive Bayes are unable to do so.

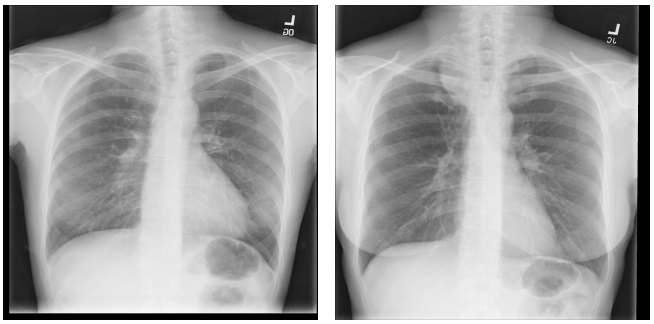
Model: Second Stage Classifier

The fundamental difference between Wang's attempt and our model is the underlying neural network architecture. Wang used ResNet50 while our model uses a 121-layer DenseNet. DenseNet is an extension of the ResNet architecture. DenseNet creates a direct route for information to flow backwards through the network by creating direct connections between layers of the same height and width. In that sense, the model can benefit when a low-level layer that recognizes more granular features such as a presence of an edge is combined with a high-level layer that recognizes more general features such as the presence of nodules.

Our model performs slightly worse than the current state of the art (CheXNet. 2018). We hypothesized that the performance was affected due to the imbalanced dataset but the two-phase training did not seem to improve the classification of minority disease classes.

Our team concluded that perhaps the skewness in the dataset within the disease classes did not severely impair the performance of the model. We compared our predicted labels with the true labels and we found that there might be instances of ambiguity in the dataset that is affecting the model performance.

Apart from the architecture, a model's performance is also limited by the quality of the dataset that it is trained on. The two images in Fig 8 look similar and they also share a similar histogram distribution of pixel intensity. However, their true labels differ. The image on the left is labelled as no finding while the image on the right is diagnosed with mass and pleural thickening. In this case, it would be difficult for the neural network to distinguish between these two images given that their similarity in pixel intensities.



Label: No Finding Label: Mass|Pleural Thickening

Fig 8: Comparison of similar Chest X rays with different labels

Simulated Annealing

Our model with Simulated Annealing did not perform better than CheXNet. 2018. Simulated Annealing works best for scenarios where the model validation cost gets stuck at a local minimum and is unable to escape it because of a lack of flexibility. However, our model without simulated annealing also uses an optimiser that is able to escape local minima with the weight decay and learning rate decay when no change in validation loss is detected. This is why adding Simulated Annealing did not give us any significant improvements. However, given that the problem of local minima is common when training machine learning algorithms, Simulated Annealing can be considered when designing our own optimiser.

Transfer Learning

	Pretrained (ImageNet) Weights	Tuned Weights with Chest X-Rays
Mean AUC	0.807	0.699

Table 2: Comparison of model performance with pre-trained weights

Transfer learning allows us to apply a model trained on a task with more data to a separate task with fewer data. As indicated in Table 2, the performance of Transfer Learning is better than tuning weights from ground up in such computer-aided diagnosis problems. As our project uses Pytorch's 121-layer DenseNet as the underlying neural network architecture - which has millions of trainable parameters, a large number of images are required to fine tune these parameters.

In image classification, the similarities in the functions of different layers facilitates transfer learning. Knowledge such as edge detection in the earlier layers, shape recognition in the middle layers and task specific recognition in the final layers are transferable to other image classification tasks. Hence, despite the differences between ImageNet dataset and the chest X-ray images, the pre-trained weights obtained from ImageNet data may still be relevant in the medical image dataset.

Conclusion

The last two decades have seen a staggering increase in lung diseases which has led to a shortage of medical professionals needed for imaging and diagnosis[4]. We hence set off with a goal to make build upon and improve over CheXNet to improve disease detection using chest X-ray images.

In the end, although we did not manage to surpass CheXNet, we believe that we have made a significant contribution towards this problem by augmenting old attempts with new techniques and algorithms.

Our first major contribution is the attempt to overcome data imbalance using our two-stage and two-phase approach. Another contribution would be our attempts to improve feature extraction through image enhancements and preprocessing using methods such as Histogram Equalization. Finally, we also tried algorithms such as Simulated Annealing for better model optimisation.

With the advancements in medical imaging and computer vision techniques, we will potentially have a cleaner and better dataset to train these algorithms. With more labelled chest X-ray images made available, the problem of a skewed dataset can possibly be solved such that training can be better optimised. All in all, we hope that our attempts, both successful and unsuccessful, will provide useful lessons for others who choose to tackle this problem. Together, we can play a part in improving computer aided diagnosis for a better world.

References

[1] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8:Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: CVPR (2017)

[2] Yao, L., Poblenz, E., Dagunts, D., Covington, B., Bernard, D., Lyman, K.: Learning to diagnose from scratch by exploiting dependencies among labels. In: CoRR (2017)

[3] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M.P., Ng, A.Y.: Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. In: CoRR (2017)

[4] The burden of lung disease: The European Lung Whitebook. https://www.erswhitebook.org/chapters/the-burden-of-lung-disease/