# Detecting and Understanding Pneumonia with Deep Learning

**Nikolaos Altiparmakis**
Stanford University
nikosa@stanford.edu

## Abstract

Artificial intelligence is perceived as a possible solution to problems requiring highly trained human expertise, such as in healthcare. The idea is to automate tedious tasks related to detection of health issues, as well as pave the way to understanding new complex issues to which humanity has not yet given an answer. We make use of a ResNet model in order to automate the detection of pneumonia through Chest X-ray images. At first, we define, formulate and understand the problem. Second, we make use of limited data in order to train the model and generalise the results. Finally we try to show what the model has learned, by using the integrated gradients method. We envisage our work to change the mindset of healthcare professionals across the globe, who might start to perceive deep learning as a step towards solving previously stunningly tough healthcare issues.

## 1 Introduction

Pneumonia is one of the most dangerous and prevalent diseases worldwide. Such an important and highly prevalent disease is putting increased load on hospitals and medical staff worldwide. Therefore, there is a growing need for technology and automation to assist in the diagnosis and, later on, in the treatment of such conditions.

In such conditions, diagnosis is usually done through the use of chest X-rays, which is then evaluated by a relevant doctor and results are then provided. The increased use of deep learning in this effort has shown promising results. Therefore, harnessing the power of neural networks to automate the task of classification using X-ray images yields both an interesting project, as well as a piece of important work towards improving healthcare in the future.

Through the use of deep learning in this field, we hope to create a useful tool for experienced doctors and also help practicing doctors understand how to correctly interpret a case of pneumonia, which will in turn provide them with more experience.

## 2 Related Work

There have been several attempts to automate the detection of pneumonia using artificial intelligence.

In [1], the development of a convolutional network achieved accuracy equal to $76.8\%$, thus outperforming previous research, as well as statistically significantly improved performance compared to radiologist performance. $100,000$ frontal-view X-ray images were provided, that could potentially have $1$ of $14$ related diseases. That dataset was labelled by $4$ academic radiologists and their performance was then compared to that of the convolutional network.

In [2], a convolutional network achieved validation accuracy $93\%$. In this paper, it is agreed that given related datasets are very small, therefore other methods must also be explored in order to avoid

overfitting. In this case, a special network is designed from scratch and data augmentation techniques are explored.

# 3 Automated Pneumonia Detection

## 3.1 Problem Definition

The problem of pneumonia detection should generally be formulated as follows. The input $X$ should be a frontal-view chest X-ray image of the patient, while $\hat{y} \in \{0, 1\}$ should be a binary label that indicates whether the patient has pneumonia or not. The main purpose of automated pneumonia detection is the use in real-life scenarios. For that reason, we will formulate the problem as having two requirements:

- **Satisficing**: Recall should be equal to 1.
- **Optimizing**: F1-score should be as high as possible.

It is our suggestion that applying deep learning in healthcare should work under the above doctrine, because the medical staff usually function under the same principle. When a doctor is not sure about a patient's condition (whether they are suffering from a specific illness or not) they will normally tend to edge on the worst-case outcome.

## 3.2 Baseline

Our baseline will be a slightly modified kernel taken from Kaggle [3]. The algorithm uses dropout and early stopping as regularization techniques.

For the seed we used in the implememtation, the algorithm achieved $96.70\%$ accuracy in the training set and $62.5$ accuracy in the dev set. Finally, the accuracy on the test set was $71.6\%$.

## 3.3 Architecture

We will use the ResNet50 model for this problem and we will modify the top layers of the network: a global average pooling layer will be used and, finally, a single logistic node with a sigmoid activation will provide the prediction.

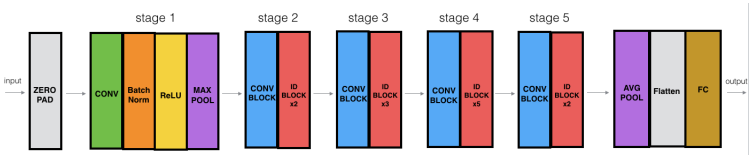


Figure 1: The structure of the neural network used for pneumonia detection.

We train the model using mini-batches of size $64$. The model is trained end-to-end using Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. the learning rate is equal to $\lambda = 0.003$ with a decay equal to $\alpha = 3e - 6$. The repository can be found below. [1]

# 4 Dataset

The dataset has been taken from Kaggle [2] and contains $5,856$ high quality chest X-ray images. In order to get a glimpse of what a case of Pneumonia would look like, we will provide samples from the same dataset.

As might be visible in Fig. 2, a normal X-ray image would normally be clearer and the general area of the chest would normally be more visible in such pictures than pictures of cases of pneumonia.

[1] https://github.com/nikaltipar/CS230-Pneumonia-Detection

[2] https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

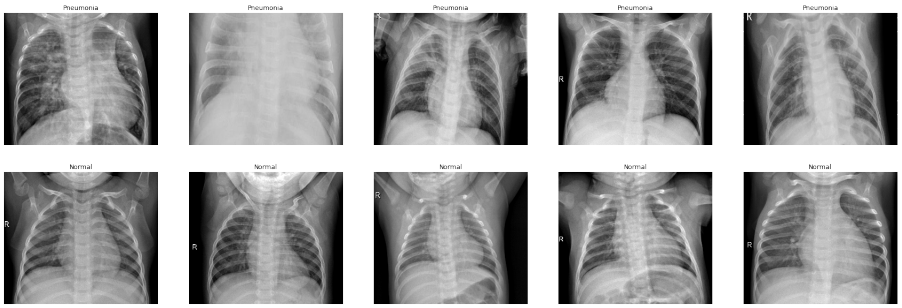


Figure 2: Sample X-Ray images.

The dataset is already split in training, dev and test set. The training set contains $5216$ pictures, the dev set $16$ and the test set $624$. We have combined all pictures into one dataset and then applied a $85\%/7.5\%/7.5\%$ split. We have tried to keep as much data in the training set, since the dataset is very small.

### 4.1 Augmentation

One important aspect of AI in healthcare is privacy. Due to that factor, we should not expect to ever (in the short time, at least) have enough data in order to be able to train a proper neural network. For that reason, we should make sure to augment the current dataset as much as possible. The dataset will be augmented as follows:

- **Horizontal mirroring**: Although the chest is not symmetric (due to how the organs are distributed in the body, we will examine this strategy and evaluate its impact in the results.
- **Brightness Scaling**: Each picture has its brightness value randomly scaled with a value $\alpha \in [0.9, 1.2]$
- **Random rotation of small angle**: Each X-ray is taken with a person having a particular pose (we can think of this pose as the angle of the spine with the horizontal axis). We will randomly rotate a portion of the dataset in order to account of cases a person has a different pose. The maximum angle an image can be rotated by is $15$ degrees.

### 4.2 Avoiding overfitting

We have tried several ways to avoid overfitting in the dataset. First of all, the use of ResNet50 allows us to achieve remarkable results with such a small dataset. Second, we have also tried L2-regularization and Dropout as a means to maintaining the same accuracy in the train and the dev/test datasets.

There are several other regularization techniques which we will describe below:

- **Class Scaling**: Since the dataset is skewed, the binary cross-entropy loss terms are scaled according the number of sample in each class.
- **Neural style transfer**: Some layers of the neural network are initialized using weights from imagenet. This is configurable; ideally, we would not want to preload all layers with weights from imagenet, since imagenet deals with a different problem.
- **Image data generator**: Instead of creating a static augmented dataset, we create new samples at runtime. During each iteration, each image of the dataset is replaced with randomly generated augmented counterpart. This means that each iteration deals with a different dataset, thus having a regularizing effect.

## 5 Results

In terms of regularisation, our final model did not make use of L2-regularisation, Dropout and Neural style transfer. The first two reduced the accuracy, while neural style transfer (which was usually used

only in the bottom layers) elongated the training process, which, in a sense, meant it had no effect, since the provided weights were completely changed during the training process. The image data generator was used in the training set, which was also shuffled on each epoch. We trained the model for 75 epochs. The resolution of the images was $(224, 224, 3)$.

### 5.1 Metrics

Our model achieved $98.36$ accuracy on the training set, $96.35$ on the validation and $96.61$ on the test set.

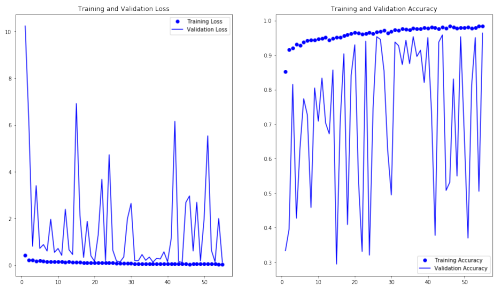


Figure 3: Loss and Accuracy on the training and validation sets.

The training process was generally very noisy (Fig. 3). This was due to the use of mini-batches and the size of the validation set, but means the results were hard to reproduce. Therefore, we had to run the training process several times in order to get accurate results.

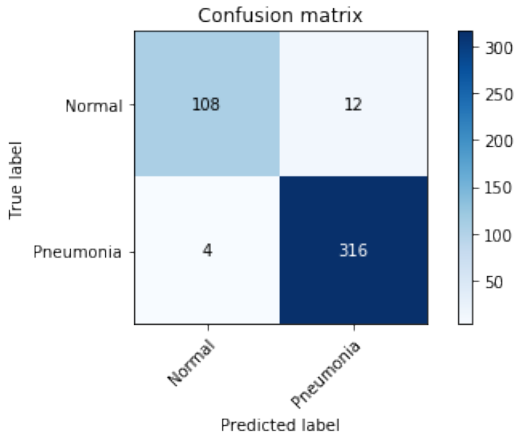


Figure 4: Confusion matrix for the test set.

The confusion matrix of the test set (Fig. 4) highlighted the results were positive. with $recall = 0.9875$, we got very close to our requirements. Also, with $F1 = 0.9417$, we understand our classifier is not trivial. We generally got much better results when using data augmentation. It helped raise the accuracy of the dev and test sets, therefore was a successful decision.

### 5.2 Visualisation

We have used the integrated gradient method [4] in order to visualise what the neural network learnt for each picture in the dataset. We preferred this to class activation maps, since, in this case, the several small parts of the patient's X-ray are evaluated (which means the features are not highly localised). When visualising the results, we observed the algorithm was not as successful as we thought it was.

In Fig. 5, the neural networks is highly sensitive in the lower left and upper right areas of the picture. This generally seems to be correct. First of all, this image highlights the patient had pneumonia, since

it was quite foggy. The network, in this case, is neither a very sensitive to the patient's organs, nor to unrelated parts of the image.

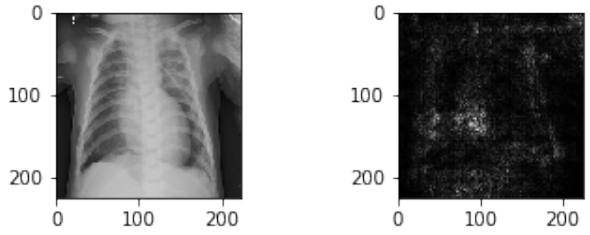


Figure 5: Prediction based on relevant parts of the image.

Fig. 6 highlights a case of overfitting. The network ignored the whole image and focused on an unrelated landmark on the upper-left part of the image.

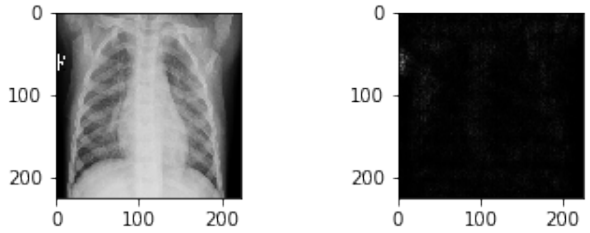


Figure 6: Prediction based on irrelevant parts of the image.

## 6 Conclusion and Reflection

The results generally highlighted the dataset was too small for a reliable tool. We used a complicated neural network, achieved remarkable results without overfitting the training set, but we weren't able to create a tool that can be trusted, which was the primary purpose of this endeavor. However, we firmly believe the choices we made would have a more positive effect, if applied to a larger dataset.

There were also mistakes in our general workflow. It is not enough to judge the efficiency of a network through metrics. Those can often be misleading, but visualisation methods can always provide more useful insights. Therefore, when conducting research, random images of the dataset should be displayed along with the neural network visualisation (class activation maps or saliency models, for instance), in order to help the researcher understand the efficiency of the model on each step.

We would also recommend changing the way we state the problem. Using just the image would not be enough in order to provide reliable results. Experienced doctors could provide their prediction, alongside most important parts of the image. The training process should then result in the network matching those parts, instead of just making the same predictions.

Finally, with our effort, we hope to change the mindset of people towards artificial intelligence. In this case, we used a classification model, but we were also able to visualise the results and thus expand our knowledge about pneumonia, just by looking at what the model believe the important parts of images to be. This means that we can use neural networks for healthcare issues, the cause of which we do not know.

## 7 Contributions

The whole assignment was completed by the only member of the team (nikosa). We got help from a trainee doctor in order to understand the problem (as "we" I mean myself only).

## References


[1] 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.

[2] Okeke Stephen, Mangal Sain, Uchenna Joseph Maduh, and Do-Un Jeong. An efficient deep learning approach to pneumonia classification in healthcare. *Journal of healthcare engineering*, 2019, 2019.

[3] Aakashnain. Beating everything with depthwise convolution, Jun 2018.

[4] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3319–3328. JMLR. org, 2017.