Université Paris Saclay

Master 1 Computer Science, Artificial Intelligence

Project A Hygieia report

Group members: Phan Anh Vu, Alejandro de la Cruz López, Paavo Camps, Nour Jemli, Xinwen Xu

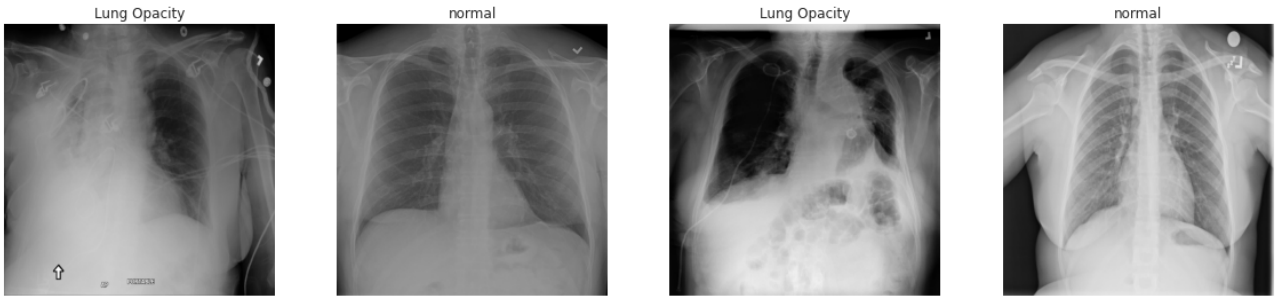

Figure 1: Sample of chest X-ray images

Challenge URL
Github URL
Video URL

Abstract

We propose a machine learning challenge of predicting pneumonia from chest X-ray images. From a collection of radiographs, participants need to classify the subjects as normal (healthy) or abnormal (lung opacity). We prepare 2 versions of input data. The 1st version contains only tabular features, which are activation extracted from the convolution part of a pretrained neural network. In the 2nd version, input features are raw image pixels. For this raw image pixel version, we also suggest a process to interpret the behavior of the model, as well as a method to systematically assess the trustworthiness of these interpretations. Our experiments with the dataset show quite good performance. However, by applying this explanation protocol, we find that these results may be less reliable than they appear. Despite the good performance, the model seems to rely too much on background and watermark to classify.

1 Background and Introduction

Among the medical imaging examinations, chest X-rays are the most common. Each year, there are 3.6 billion medical procedures involving ionizing radiation [2], of which over 2 billion are chest X-rays [6]. Pneumonia is one among the many diseases that chest X-ray can help to detect. Every year in the US, pneumonia causes hospitalization of more than 1 million adults, and kills around 50 000 [7]. Detecting pneumonia is a difficult task which requires expertise of radiologists. However, there is always a shortage of radiologists to interpret

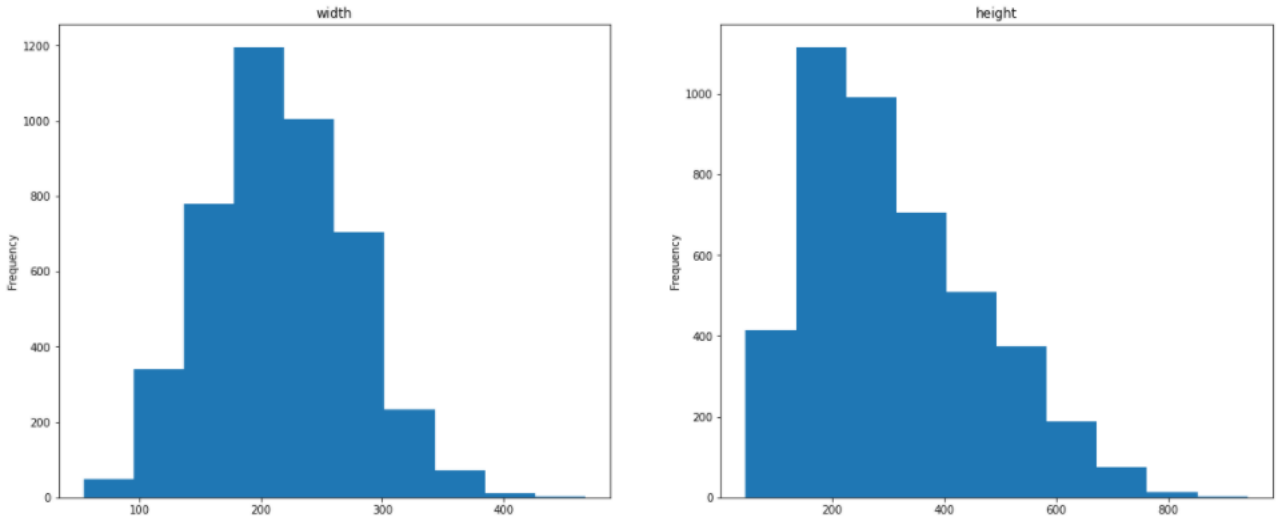

Figure 2: Distribution of bounding box width and height

the X-ray results [6]. Moreover, fatigue due to heavy workload may also deteriorate the diagnostic accuracy of radiologists [8]. Automatic interpretation of chest X-ray would bring numerous substantial benefits, but it is also a challenging task. Lung opacity may be vague and very similar to minor benign anomalies. Variation in radiation settings may also make an image look more hazy in general.

A survey [3] published in 2020 shows the steady rise of Convolution Neural Network as feature extractor for lung disease detection. More specifically, in the CheXNet paper [7], the authors reported results for the Chest X-ray 14 dataset [12]. The selected model is DenseNet 121, and the images are resized to 224x224 px. The AUROC metric ranges between 0.7 and 0.8 for the 3 classes associated with pneumonia (Pneumonia, Consolidation, Infiltration).

Another application of convolutional neural network as feature extractor is implemented by M. Toğaçar et al [11]. AlexNet, VGG-16 and VGG-19 were utilized to extract features. The number of deep features was reduced from 1000 to 100 by using the minimum redundancy maximum relevance algorithm. Then, they feed their features to some classical machine learning models, such as decision tree, k-nearest neighbors, linear discriminant analysis, linear regression, and support vector machine. All models displayed promising results, showing that the deep features are robust and consistent input for pneumonia detection. Minimum redundancy maximum relevance method was beneficial tool to reduce the dimension of the feature set.

The challenge was inspired by 2018 RSNA Pneumonia Detection Challenge.[10] Excluding the No lung opacity / Not normal category, we propose a binary classification task of normal vs pneumonia for our participants. We don't provide the original images. Instead, We apply 100 x 100 px crop on original images which is helpful to reduce bias. And we also prepare a tabular version data which is features extracted from the last convolution layer of a pre-trained MobilenetV2 PyTorch model on Imagenet.

Our participants will work on both of these data to train their own models. But results on the tabular data might be severely biased because these features are extracted from the original images. Our participants are encouraged to find another way to solve this

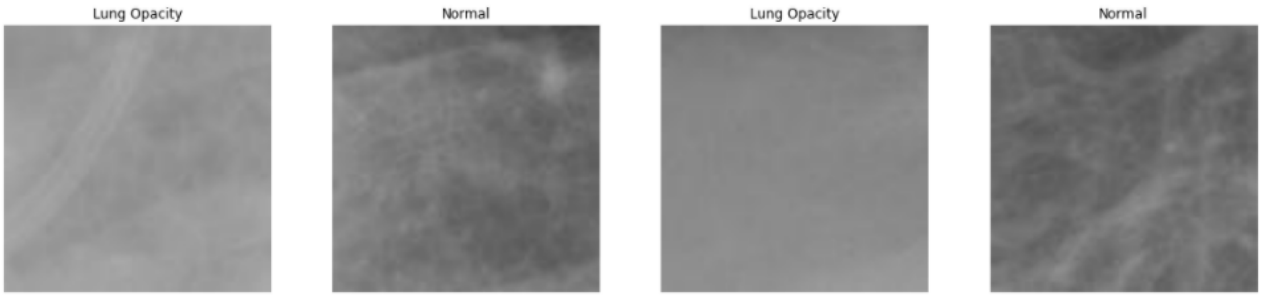

Figure 3: Examples of cropped images

problem. And they are also encouraged to consider other insightful metrics, such as recall, precision or fbeta2, to evaluate their models instead of only accuracy.

To summarize our contributions, we prepared a cropped image version which may reduce the bias in data. Moreover, we also proposed a protocol to assess systematically the attention heatmaps generated by class activation map methods. Thus, participants are incited to inspect the reliability of their model, from a qualitative point of view. Variation in background, watermarks and other artifacts are common in X-ray images. We also hope that participants will find new ideas to deal with the bias issue.

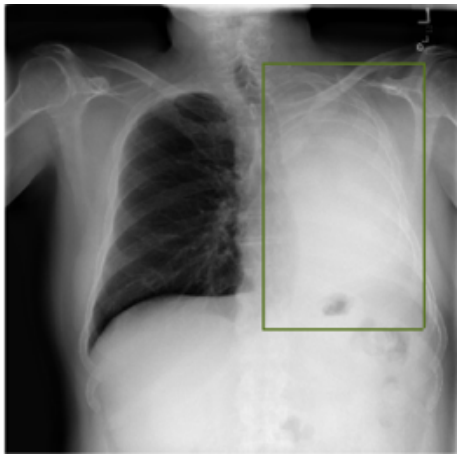

Figure 4: Example of lung opacity with bounding box

2 Material & Method

2.1 Data

We start by selecting the application domain and finding datasets. In the fields of health, biology and medicine, medical imaging is a major terrain for machine learning algorithms. Radiography stands out as a popular and effective method for diagnosis. Our first selection was the MURA [8] and LERA [4] datasets. These 2 collections contain X-ray images of upper and lower body extremities. The original task is to classify a radiograph as either normal or abnormal. We intended to combine this binary classification problem with the task of identifying the body part in the image. After studying the 2 datasets, we realized that the usage license is quite restrictive. We fear that the restrictions may hinder the preparation of our own challenge later, so we decided to look for other alternatives.

We quickly gravitate toward the most common medical imaging examination: chest X-ray. We begin to inspect the Pediatric Pneumonia dataset [5]. The associated task is to classify the images as healthy or pneumonia. We used this dataset to implement our first convolution model, and to estimate the difficulty of the task. The total number of images is around 5000. We worry about having too few images for a stable estimation

of performance. Therefore, we began looking for a larger chest x-ray dataset.

We finally settle on the RSNA Pneumonia Detection Challenge 2018 [10]. This dataset contains publicly available chest X-ray from the NIH (National Institutes of Health). Board-certified radiologists located the opacity region with bounding boxes. There are frontal view chest X-ray of 26 684 unique patients in the original collection. Each image belongs to one of 3 categories: Normal, Lung opacity, No lung opacity / Not normal. The Normal class contains healthy lungs. Lung opacity contains potential pneumonia patients (signs of Infiltration or Consolidation). No Lung opacity / Not normal corresponds to radiographs without opacity related to pneumonia. This class still contains other types of opacity (e.g. nodule, mass) and other kinds of anomalies. To prepare a preprocessed data version, we extract the activation from the last convolution layer of a convolution neural network.

Based on the distribution of bounding box width and height (Figure 2), we decide to apply 100 x 100 px crop on images. For pneumonia cases, we crop to bounding box. For healthy cases, we randomly sample a bounding box from the lung opacity class to crop (Figure 4 and 3).

We hope to reduce bias by excluding the surrounding background and including only the lung region. We also prepare a tabular version of the data by extracting activations from the last convolution layer of a convolution neural network. For this purpose, we use a MobilenetV2 PyTorch model pretrained on Imagenet.

The final data set we provide to our participants contains three sets: 10000 samples in train set, 2000 samples in validation and test set (table 1).

Set	Num samples	Num features	Num classes
Train	10000	100 x 100	2
Validation	2000	100 x 100	2
Test	2000	100 x 100	2

Table 1: Cropped image data

Each unique patient has only one image which is different from the original data, and there is no overlap between the 3 subsets. The ratio of normal to lung opacity in each subset is around 3:2.

2.2 Classification

Starting from the RSNA Pneumonia dataset, we propose a binary classification task of normal vs pneumonia. We select the normal images from the original collection as the negative class, and the lung opacity examples as the positive class. We decide to exclude the No lung opacity / Not normal category, since this class has a lot of variations. Another reason is to limit the size of the input data.

We use MobileNetV2 and MnasNet for our experiments. Both of them are lightweight CNN models. Classification performance varies with different choice of image size and model architecture.

2.3 Evaluation

Due to time constraint and in order to simplify deployment, we use accuracy as the basic metric. Nevertheless, we encourage participants to consider other insightful measures such as recall, precision or fbeta2. Besides, accuracy alone is not enough as assessment. John Zech [13] demonstrated the remarkable ability of deep convolutional neural networks to detect confounding factors and take advantage of them. Hence, we suggest calculating an explanation score as a measure of trustworthiness. We generate a class activation heatmap for a set of explanation images,

Figure 5: Left: target bounding box in red. Middle: heatmap. Right: rectangle holding at least 95% of activation in orange. IOU = 0.65

using the GradCAM (Gradient-weighted Class Activation Map) method [9].

Inspired by section 8.1 of the GradCAM paper, we leverage the provided bounding box in the original data to assess the quality of the GradCAM heatmap. For evaluation of heatmap, we select the examples having both positive predicted and positive target class (Prediction = Lung Opacity & Target = Lung Opacity). There is around 400 examples in this explanation collection.. We compute the proportion of activation inside the bounding box:

$$\texttt{Ratio of activation inside box} = \frac{\texttt{Activation inside box}}{\texttt{Total activation}}$$

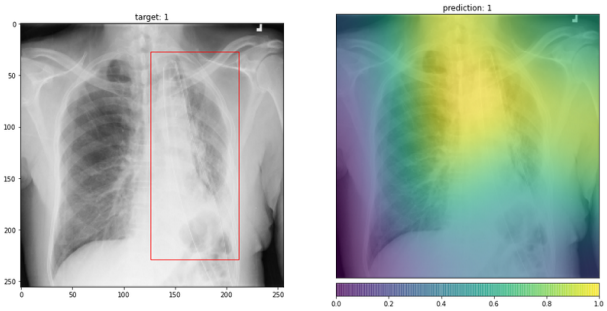

Figure 6: Ratio of activation inside box = 0.33

A trustworthy heatmap should have high activation inside the box, and weak or almost 0 activation elsewhere. Hence, this ratio will be close to 1. To summarize this measure, we simply average over all heatmaps. Figure 6 shows the best ratio of activation inside box from the explanation set. The bounding box holds around one third of total activation. We can notice that this measure fails to reflect faithfully the quality of the attention heatmap.

For this reason, we attempted another protocol introduced by the ProtoPNet authors [1]. We start by calculating the smallest rectangle containing at least 95% of activation from the attention weight matrix. Next, we evaluate this attention rectangle with the Intersection Over Union (IOU) metric. Figure 5 shows the best IOU score of the explanation collection. This IOU protocol appears to be more trustworthy than the activation inside box protocol. Nevertheless, the IOU protocol still has some flaws. Specifically, this algorithm fails when the weight matrix contains multiple non-contiguous regions of strong activation. Instead of identifying these separate areas, this algorithm can only detect a large enclosing rectangle. Figure 7 illustrates this case.

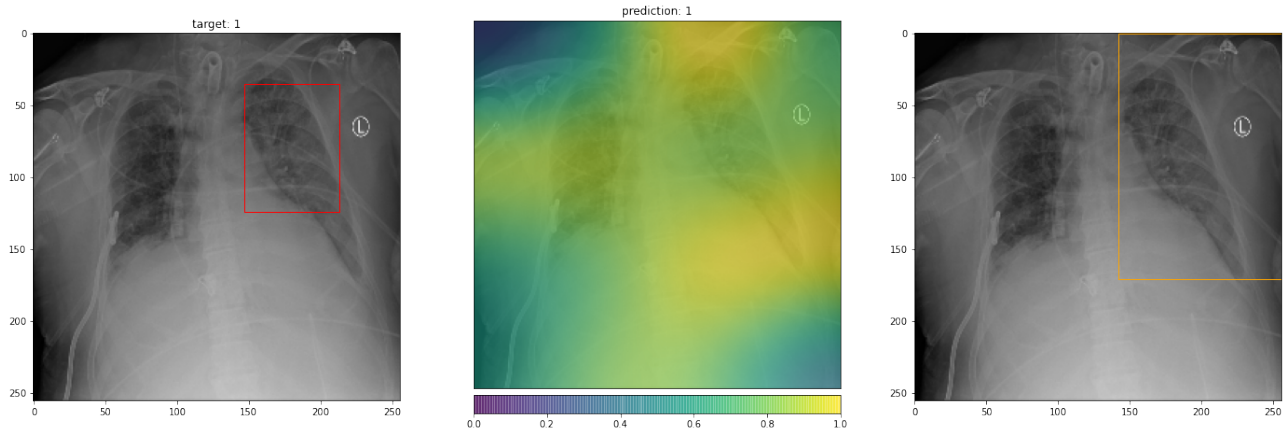

Figure 7: Left: target bounding box in red. Middle: heatmap. Right: rectangle holding at least 95% of activation in orange. IOU = 0.31

3 Results

For the experiments, we use Pytorch models pretrained on ImageNet with the convolution part fixed. Then, we add a linear layer after the convolution part and train for 5 epochs with around 13 000 images. We use the SGD optimizer with cross entropy loss, learning rate = 0.001, and momentum = 0.9. We save the parameters having the highest validation score. During preprocessing, images are resized with `torchvision.transforms.Resize`.

The scores are computed for a validation set of around 2000 images. Unless stated otherwise, the results presented here are for MobilenetV2 with image size 128. Using accuracy = 0.8 and with test set size = 2000, we estimate participant score to be 0.8 ± 0.0089. We calculate this value by considering the accuracy score as a Bernoulli random variable p. Hence, the variance is p(1-p)/N, for N = size of dataset.

Table 2 shows the results with different image size and model combinations.

Model	Image size	Accuracy
MobilenetV2	128	0.87
MobilenetV2	192	0.89
MobilenetV2	256	0.90
Mnasnet	128	0.64

Table 2: Baseline results

Increasing image size brings some improvement in performance. As the variation of performance between the image resolutions is quite small, we stop experimenting at 256 x 256 px.

Figure 8 and Figure 9 shows the confusion matrices.

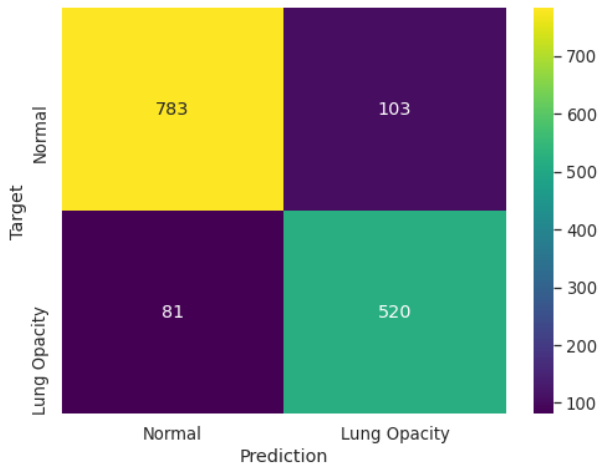

Figure 8: Confusion matrix for 128x128 images, 1400 examples

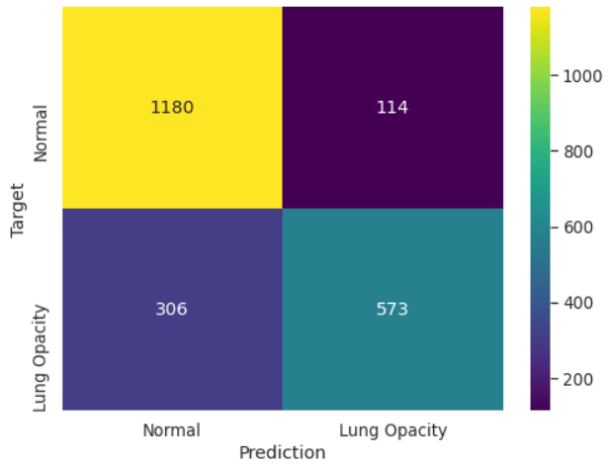

Figure 9: Confusion matrix of cropped images

We also compute other metrics for MobilenetV2 with image size 128.(Table 3 and Table 4)

Accuracy	Recall	Precision	fbeta2
0.87	0.87	0.83	0.86

Table 3: Results for 128x128 images

Accuracy	Recall	Precision	fbeta2
0.81	0.65	0.83	0.68

Table 4: Results for 100x100 cropped images

Figure 10: Cross entropy loss per epoch for 128x128 images

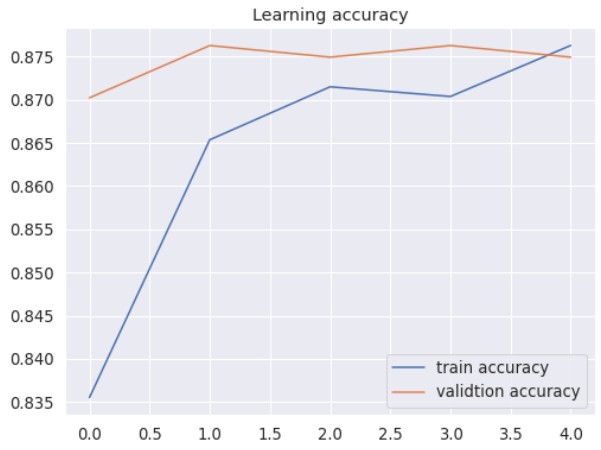

Figure 11: Accuracy per epoch for 128x128 images

Figure 10 and figure 11 show the binary cross entropy loss and classification accuracy. The horizontal axis shows the epoch, and the vertical axis shows cross entropy loss or accuracy.

Figure 12 and 13 shows the distribution of the activation inside box and IOU score for around 400 Lung Opacity images with annotated bounding box.

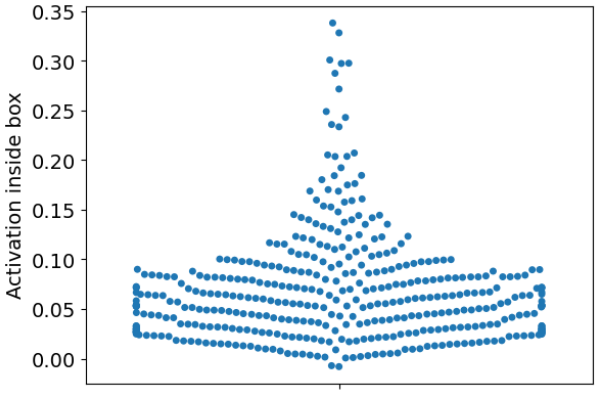

Figure 12: Distribution of activation inside box score

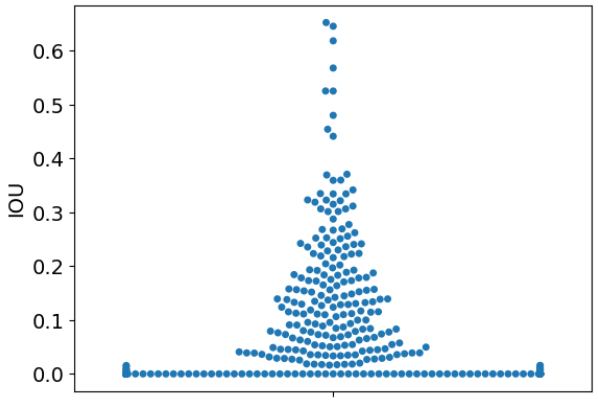

Figure 13: Distribution of IOU score

4 Conclusion

Organizing a machine learning challenge is perhaps even more difficult than solving one. As with many data science task, collecting data is the 1st critical issue. Medical datasets often require experts' knowledge for annotation. But even specialists may also have contradicting opinions about the same case. Once we find a suitable dataset, we need to decide how to preprocess it.

Each application domain and each specific task may require a different evaluation protocol. In medical settings, we would prefer to sound a few more false alarms than to overlook an infected patient. Some metrics allow us to attribute more importance to false negative, or wrongly predicted as healthy. Meanwhile, they may be harder to inpterpret for the participants.

Our explanation protocol aims to inspect the attention mechanism of the trained model. From our experiments, we observe a divergence between the annotations from expert and the focus of the model. Although our measure is far from satisfactory, it provides a rough estimate for the trustworthiness of the trained model.

References

[1] Chaofan Chen et al. *This Looks Like That: Deep Learning for Interpretable Image Recognition*. 2019. arXiv: 1806.10574 [cs.LG].

[2] Rachel Draelos. *Radiology: Normal Chest X-Rays*. URL: https://glassboxmedicine.com/2019/02/10/radiology-normal-chest-x-rays.

[3] Stefanus Tao Hwa Kieu et al. "A Survey of Deep Learning for Lung Disease Detection on Medical Images: State-of-the-Art, Taxonomy, Issues and Future Directions". In: *Journal of Imaging* 6.12 (2020). ISSN: 2313-433X. DOI: 10.3390/jimaging6120131. URL: https://www.mdpi.com/2313-433X/6/12/131.

[4] *LERA- Lower Extremity RAdiographs*. URL: https://aimi.stanford.edu/lera-lower-extremity-radiographs-2.

[5] Kermany Daniel; Zhang Kang; Goldbaum Michael. *Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification*. 2018. URL: http://dx.doi.org/10.17632/rscbjbr9sj.2.

[6] Nick A. Phillips et al. *CheXphoto: 10,000+ Photos and Transformations of Chest X-rays for Benchmarking Deep Learning Robustness*. 2020. arXiv: 2007.06199 [eess.IV].

[7] Pranav Rajpurkar et al. *CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning*. 2017. arXiv: 1711.05225 [cs.CV].

[8] Pranav Rajpurkar et al. *MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs*. 2018. arXiv: 1712.06957 [physics.med-ph].

[9] Ramprasaath R. Selvaraju et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *International Journal of Computer Vision* 128.2 (Oct. 2019), pp. 336–359. ISSN: 1573-1405. DOI: 10.1007/s11263-019-01228-7. URL: http://dx.doi.org/10.1007/s11263-019-01228-7.

[10] George Shih et al. *Augmenting the National Institutes of Health Chest Radiograph Dataset with Expert Annotations of Possible Pneumonia*. URL: https://www.rsna.org/education/ai-resources-and-training/ai-image-challenge/RSNA-Pneumonia-Detection-Challenge-2018.

[11] M. Toğaçar et al. "A Deep Feature Learning Model for Pneumonia Detection Applying a Combination of mRMR Feature Selection and Machine Learning Models". In: 41 (4 2020), pp. 212–222. ISSN: 1959-0318. URL: https://doi.org/10.1016/j.irbm.2019.10.006.

[12] Xiaosong Wang et al. "ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (July 2017). DOI: 10.1109/cvpr.2017.369. URL: http://dx.doi.org/10.1109/CVPR.2017.369.

[13] John Zech. *What are radiological deep learning models actually learning?* URL: https://jrzech.medium.com/what-are-radiological-deep-learning-models-actually-learning-f97a546c5b98.