# Pneumonia Detection with Weighted Voting Ensemble of CNN Models

Heewon Ko*
Dept. of Computer Science and Engineering
Ewha Womans University
Seoul, South Korea
e-mail: huiwonee96@gmail.com

Hyunsoo Ha*
Dept. of Software Engineering
Soongsil University
Seoul, South Korea
e-mail: dhy03196@naver.com

Hyuna Cho**
Dept. of Computer Science and Engineering
Kyung-Hee University
Seoul, South Korea
e-mail: Hannah3767@gmail.com

Kiwon Seo**
Dept. of Computer Science and Engineering
Sejong University
Seoul, South Korea
e-mail: rldnjs3258@naver.com

Jihye Lee**
Dept. of Information Security
Seoul Woman's University
Seoul, South Korea
e-mail: awlgpcpfl@naver.com

*Abstract*—**There is a clear correlation between the diagnosis of pneumonia and abnormalities on CXR. This association can be trained through performing object detection by a deep neural network. Therefore, applying deep learning to the detection of pneumonia can be a very effective diagnostic solution. This paper proposes a method to detect lung opacities, which can be identified as pneumonia, on chest radiographs (CXR) using an ensemble of deep convolutional neural networks. Furthermore, this paper applied an ensemble model using Mask R-CNN and RetinaNet to RSNA Pneumonia Detection Challenge on Kaggle and demonstrate that this method can effectively achieve high accuracy of prediction. Applying our voting ensemble methodologies outperformed all the individual models.**

***Keywords-CNN; object detection; pneumonia; ensemble; radiograph; CXR; mask R-CNN; RetinaNet; anomaly***

## I. INTRODUCTION

Since there is a considerable advance with image processing area such as classification and object detection, Convolutional Neural Network (CNN) is frequently used in automatic diagnosis with medical images. Moreover, an outstanding performance of CNN in a medical image such as X-rays, MRI and CT have been demonstrated in recent research [1], [2].

The most effective way to diagnose pneumonia is anomaly detection in Chest X-rays (CXR) images [3]. There are abnormal areas called 'Lung Opacity' in most of Pneumonia patient's CXR. Lung opacity refers to any area that preferentially attenuates the x-ray beam, therefore appears more opaque than the surrounding area. Usually, lungs are filled with full of air. However, when Pneumonia has occurred, the air inside the lungs is replaced by other materials such as fluids, bacteria, immune system cells. This phenomenon results in the area to be shown as opaque [4].

For the purpose of pneumonia detection, this paper constructs CNN based deep neural network model to detect lung opacities in CXR image with RSNA Pneumonia Detection Challenge dataset of Kaggle that consists of CXR images and train labels with detailed class information for about 25,000 patients. It is a subset of the National Institutes of Health (NIH)'s dataset. The primary endpoint of our project is achieving high accuracy on detecting the bounding box that has a possibility of pneumonia. The evaluation criterion of accuracy is based on mean average precision (mAP).

One of the effective ways to achieve the accuracy improvement of object detection in medical images is to utilize ensembles of multiple models [5], [6]. Ensemble learning [7] is a method for combining multiple prediction models and using their aggregated prediction results.

This paper proposes a weighted majority voting ensemble based on RetinaNet [8] and Mask R-CNN [9] with different weight for detecting pneumonia and describe our approach to yield high-quality predictability on CXR.

## II. RELATED WORK

### A. RetinaNet

One-stage detectors such as YOLO [10] and SSD [11] are fast and simple by using a fixed grid of boxes in substitute for region proposals but provide a relatively low accuracy of less than 10-40% over two-stage methods such as Mask R-CNN. To address this problem caused by the

978-1-7281-0831-5/19/$31.00 ©2019 IEEE

foreground-background class imbalance, the one-stage RetinaNet suggests focal loss.

$$CE(p_t) = -\log(p_t) \tag{1}$$

$$FL(p_t) = -(1-p_t)^{\gamma}\log(p_t)\ (\gamma > 0) \tag{2}$$

This is a new loss function modified on standard cross entropy criterion. This function drastically reduced the scaling factor of cross entropy loss to almost zero as well-classified class increases. Therefore, it is able to concentrate on mis- classified examples during training and prevent learning from being overwhelmed by most negative samples.

As the RetinaNet network adopts Feature Pyramid Network (FPN) backbone [12] on top of the ResNet architecture [6], it generates multi-scale feature map layers with high resolution and rich semantic information. Additionally, the model achieves dense coverage of boxes by using anchors of 3 scales and 3 aspect ratios at each pyramid level in FPN implying the higher potential that two-stage systems may not provide. Based on these factors, RetinaNet outperforms the accuracy of two-stage methods such as Faster R-CNN as a one-stage detector.

### B. *Mask R-CNN*

Mask R-CNN [9], which is an extension of the previous algorithm Faster R-CNN [13], simultaneously detects objects and generates segmentation masks. Mask R-CNN is similar to Faster R-CNN from a structural point of view in that both of them consist of two stages for detection. In the first stage, both Mask R-CNN and Faster R-CNN propose candidate object bounding boxes called region proposals using Region Proposal Network (RPN). Subsequently, in parallel to performing bounding box classification and regression, Mask R-CNN also applies pixel-level semantic segmentation on each of candidate boxes while Faster R-CNN only predicts the class and box offset and does not carry out image segmentation.

Since RoIPool [13] used in Faster R-CNN yielded low-performing of pixel-to-pixel alignment in the process of extracting feature map, Mask R-CNN proposes RoIAlign to fix the misalignment and to preserve exact spatial locations instead of RoIPool. RoIAlign applies bilinear interpolation to calculate the pixel value on the feature map while passing through CNN, which in turn reduces the decimal error of RoIPool. It improved mask accuracy by a relative 10% to 50% and achieved top results in all tracks of the COCO suite of challenges, including instance segmentation and bounding box object detection.

### C. *One-Stage Detection and Two-Stage Detection*

One-stage detectors such as YOLO and SSD treat object detection as a simple regression problem. They do not use a sliding window or region proposal techniques. This makes the one-stage detector more simple and faster than two-stage Region-based CNN (R-CNN) [14]. YOLO is comprised of a single neural network. The single convolutional network can predict the bounding boxes and the corresponding classes on a full image [10]. SSD encapsulates all computations in a single network by eliminating proposal generation and feature resampling stage [11]. To sum up, one-stage detectors have lower accuracy, but faster than two-stage detectors.

Two-stage detectors such as R-CNN have complex network architecture because they generate bounding box proposals and subsequent pixels. However, this process of offering candidates improved the detection accuracy. Then the network extracts feature from each candidate box and performs classification and bounding box regression in the second stage. Additionally, in the case of Mask R-CNN, image segmentation is performed by creating a binary mask of the object in this step.

## III. Methodology And Approach

### A. *Ensemble of Models*

To acquire improved results from classification problems, ensemble learning is primarily used by combining diverse models. In ensemble learning, if all the individual detectors produce the same output, there is a problem that they cannot complement one another by correcting the different possible mistakes of each model. Although an ensemble of classifiers is not always superior to the top-level individual classifier, it certainly closes an overall gap between misclassified examples against the ground truth elements. Hence, it is reasonable to secure the classifier diversity for the sake of reducing the total error. The diversity can be achieved by means of using a strategic combination of independent or negatively correlated classifiers. On that account, the models with the highest mAP score in different types of detector were selected for ensemble learning.

TABLE I. Comparison of Object Detect Model

| Model Name | COCO mAP | Inference Time | Type of Detector |
|---|---|---|---|
| SSD | 31.2 | 125ms | One Stage Detector |
| YOLO v3 | 33 | 51ms | One Stage Detector |
| Faster RCNN | 34.9 | Very Slow | Two Stage Detector |
| Mask RCNN | 35.7 | Very Slow | Two Stage Detector |
| RetinaNet | 39.1 | 198ms | One Stage Detector |

In this work, the ensemble model used both a one-stage detector and two-stage detector. One-stage detectors such as RetinaNet produce a fixed number of predictions called anchor boxes on a grid to cover possible spaces of positive examples. Unlike RetinaNet, Mask R-CNN can classify bounding boxes in any range of scales and aspect ratios at any positions with pixel-level segmentation [15]. Thus, Mask R-CNN can easily perform image localization and detect more fine-tuned proposals of opacity that may not be considered by RetinaNet due to the shape of objects.

Table 1 shows comparison of various object detection models. RetinaNet is one stage detector that obtain highest mAP score, and Mask R-CNN is the most accurate two stage detector [8] [9]. Therefore, RetinaNet and Mask R-CNN were selected for ensemble learning.

TABLE II. HYPERPARAMETERS FOR RETINANET

| Model | Backbone | Epochs | Steps per epoch | Learning Rate | Mini-batch size | Optimizer |
|---|---|---|---|---|---|---|
| RN 178 | ResNet-101 | 25 | 2500 | 0.00001 | 8 | Adam |
| RN 184 | ResNet-50 | 25 | 2500 | 0.00001 | 8 | Adam |
| RN 201 | ResNet-101 | 25 | 2500 | 0.00001 | 8 | Adam |

TABLE III. HYPERPARAMETERS FOR MASK R-CNN

| Model | Backbone | Epochs | Steps per epoch | Learning Rate | Learning Momentum | Mini-batch size | Optimizer |
|---|---|---|---|---|---|---|---|
| Mask 150 | ResNet-50 | 15 | 1000 | 0.0001 | 0.9 | 8 | SGD |
| Mask 162 | ResNet-50 | 25 | 500 | 0.001 | 0.9 | 8 | SGD |

Majority weighted voting was used in order to ensemble those models applied to this experiment. The paper applied intersection over union (IoU) between bounding boxes which each of model predicted and set the IoU threshold as 0.3. That is, if the IoU value of prediction bounding boxes exceeds the threshold value, the average coordinates of two bounding boxes (x, y-coordinates, width, and height) are considered to be in the prediction candidate group. Then the voting threshold was set which is to determine whether the predicted bounding box is included in the final prediction. If a number of votes for a particular prediction bounding box by each model exceed the voting threshold, the box would be considered as the final prediction. The voting threshold depends on the number of models contained in the ensemble learning.

In addition, it is important to select the suitable weights for the classifiers to be used for voting because ensemble can be optimized by giving appropriate weights. At first, the highest weight was given to one of the RetinaNet model which scored the highest mAP among the models handled on this experiment. Next, we slightly reduced the weight of the classifier and instead increased the weight of other models especially Mask R-CNN. This was done in order to exercise relatively more voting authority than other models. Whenever changing weights, it is necessary to ensure that the effect of an ensemble is maximized while balancing the weights between Mask R-CNN and RetinaNet.

### B. *Implementation Environment*

Both training and testing were performed on Google Cloud Platform Machine Learning Engine. It has 39GB memory, 6 vCPUs, and NVIDIA Tesla V100 on the Google Cloud virtual machine instance. TensorFlow machine learning framework was used with hyperparameters in Table II and Table III to develop all the processes.

## IV. EXPERIMENT

### A. *Data*

The dataset for this experiment is a subset of the NIH dataset, which is available from the Kaggle. The total number of CXR images is 25,684 and the total number of lung opacity bounding boxes is 28,989. The bounding boxes within an image are classified into three classes based on the presence or absence of Pneumonia. That is, each of them will be labeled as 'Lung Opacity' if bounding box contains lung opacity which is considered as pneumonia, otherwise no pneumonia. In the case of absence in pneumonia, each box is labeled as 'Normal' if bounding box does not contain lung opacity and 'No Opacity/Not Normal' if bounding box does not contain lung opacity but involves any abnormalities except lung opacity. The last class, 'No Opacity/Not Normal' is added with the intention of improving the accuracy of learning. Figure 1 shows the number of each class.

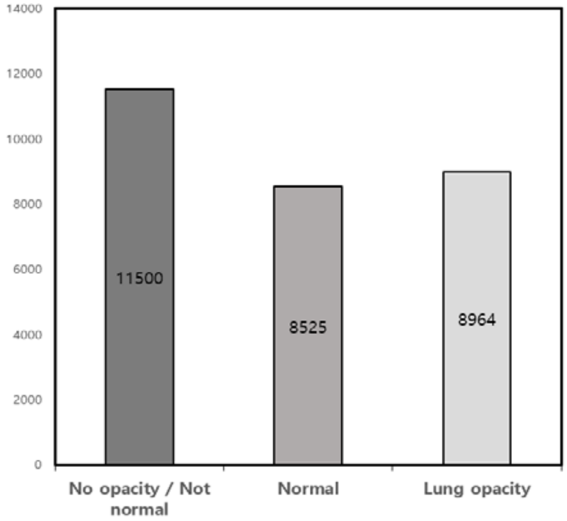

Figure 1. Number of each class.

For the purpose of improving the performance of learning, data augmentation was done on each instance within a training dataset [16]. Data augmentation is a method of applying distortion to original dataset, therefore creating a bunch of altered copies. The distortions applied on each instance of our dataset include scaling, rotating, sharpening, shearing and gaussian blur.

### B. *Metrics*

Our dataset is derived from the Kaggle challenge. The competition suggests its own rule of evaluating metric for each submission. This paper follows the rule of finding the mAP at each IoU thresholds range from 0.4 to 0.75 with a step size of 0.05 (0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75).

*1) IoU:*

$$IoU(A, B) = \frac{A \cap B}{A \cup B} \quad (3)$$

The way to know the accuracy of a predicted bounding box is IoU. The IoU between the predicted box and the ground truth box is calculated as (3). The overlap sizes of the Predicted boxes and the ground truth boxes are divided into the total area of the two objects.

True Positive means that the predicted object exceeds the threshold value and is matched to the ground truth. False Positive indicates that the predicted object is not associated with a ground truth object. False Negative indicates that there is no predicted object associated with the ground truth object.

If the IoU is above the specified threshold value, it is a True Positive, otherwise a False Positive [17]. At each threshold value t, C is measured based on all objects that are predicted when compared to True Positive (TP), False Negatives (FN), and False Positives (FP) [18]. It is calculated as (4) and it is called Precision value.

*2) AP:*

$$C(t) = \frac{TP(t)}{TP(t) + FP(t) + FN(t)} \tag{4}$$

While the IoU threshold of this range is calculated on one image, Average Precision (AP) is measured. The threshold value is increased from 0.4 to 0.75 by 0.05 to find 'hits' and 'misses'. AP for a single image is calculated as the average of Precision Values at each IoU thresholds. T is the set of specific thresholds [18].

$$AP = \frac{1}{|T|} \sum_{t}^{T} C(t) \tag{5}$$

*3) mAP:*

The evaluation score given to us in this competition is the mean taken over the individual average precisions of each image in the test dataset. Thus, all of the models are evaluated on the mAP at different IoU thresholds.

### C. Experimental Results

The table IV shows the mAP of each Mask R-CNN and RetinaNet models that applied to the ensemble. Two Mask R- CNN models and three RetinaNet models were used respectively for an ensemble.

TABLE IV. MAP PER THRESHOLD

| Model | mAP |
|---|---|
| Mask 150 | 0.15035 |
| Mask 162 | 0.16211 |
| RN 178 | 0.17813 |
| RN 184 | 0.18467 |
| RN 201 | 0.20147 |

TABLE V. MASK R-CNN ENSEMBLE RESULTS

| Mask 150 | Mask 162 | mAP |
|---|---|---|
| 1 | 1 | 0.16444 |
| 1 | 2 | 0.16661 |

*1) Mask R-CNN Ensemble:*

Table V shows the result of ensemble using two Mask-R-CNN models. The result of mAP using the equal weight for two Mask R-CNN models was 0.16444. Then doubled the weight to the classifier which gained the best mAP among Mask R-CNNs. When ensembled with Mask R-CNN only, the mAP obtained by majority weighted voting method is 0.16661. It performed better than the majority voting without weight.

TABLE VI. RETINANET ENSEMBLE RESULTS

| RN 178 | RN 184 | RN 201 | mAP |
|---|---|---|---|
| 1 | 1 | 1 | 0.19952 |
| 1 | 1 | 2 | 0.19940 |
| 2 | 2 | 3 | 0.19983 |
| 1 | 2 | 3 | 0.19984 |

TABLE VII. MASK R-CNN AND RETINANET ENSEMBLE RESULTS

| Mask 150 | Mask 162 | RN 178 | RN 184 | RN 201 | mAP |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 0.21378 |
| 2 | 2 | 2 | 2 | 3 | 0.21592 |
| 2 | 3 | 2 | 3 | 4 | 0.21626 |
| 2 | 3 | 3 | 3 | 4 | 0.21699 |
| 2 | 3 | 2 | 2 | 3 | 0.21746 |

*2) RetinaNet Ensemble:*

Table VI shows the weight given to each classifier and the mAP of an ensemble using three RetinaNet models. The result of mAP using equal weight for each models was 0.19952. Then doubled the weight of the model with the highest mAP value. The mAP obtained from this case is 0.19940. Then fine- tuning was done with the ratio of the weights given to each model in more detail by 2:2:3, 1:2:3 and obtained the mAP of 0.19983, 0.19984. These results demonstrate that combining multiple models should not always be better performed than the best individual classifier in the ensemble.

*3) Mask R-CNN and RetinaNet Ensemble:*

Our final attempt is to combine two of the Mask R-CNN models and three of the RetinaNet models handled on this experiment.

Table VII shows the mAPs of each experiment conducted by giving different weight ratios to multiple classifiers. The mAP is 0.21378 when only the majority voting method is applied. According to the two previous ensemble experiments, it is indicated that applying relatively more weights to the model of the highest mAP helps to improve the mAP. Thus, the highest weight was given to the model of RN 201. As a result of the experiment, the value of mAP increased to 0.21592. Then we adjusted the ratio of weights given to each of the models in more detail and gained the best mAP of 0.21746. The process of combining the results of these different CNN architectures in this way allows each model to capture complementary information and, consequently, to increase the completeness of the overall prediction. Arguably this integrated set of prediction improves mAP and reduces the total error rate compared to applying each separate model.

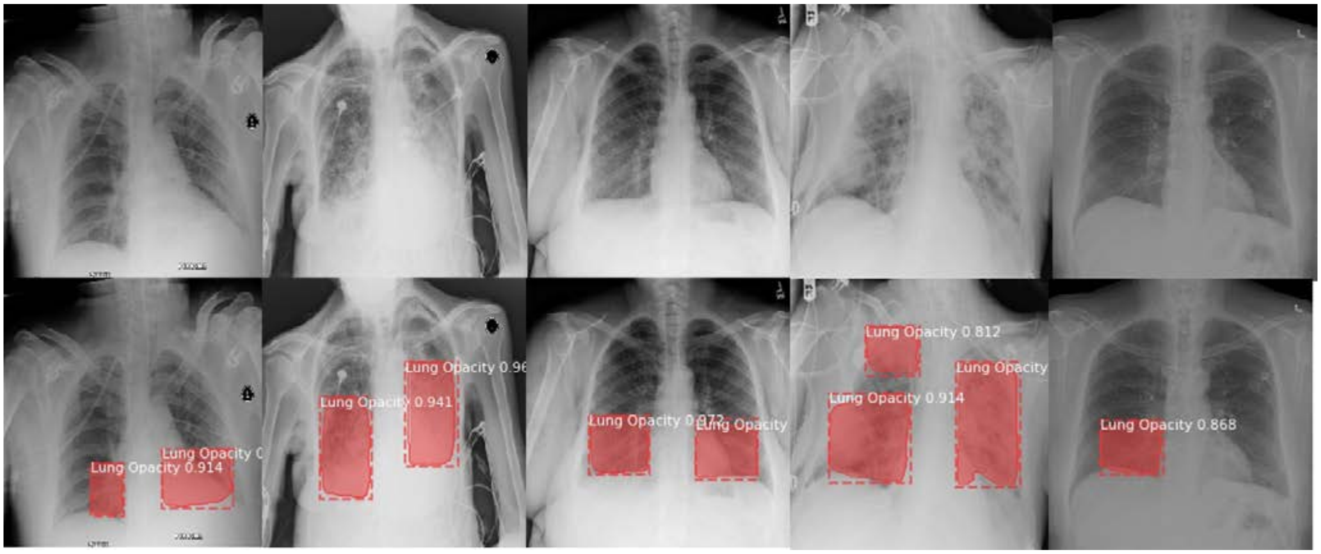

Figure 2. Result of prediction.

## V. Conclusion

This paper proposes multiple ensemble approaches based on Mask R-CNN and RetinaNet to detect lung opacity in CXR. Figure 2 shows our experiment result, which is the coordinate values and confidence of lung opacity on predicted bounding boxes for each patient. We demonstrated the efficiency of our approach by combining multiple classifiers with majority weighted voting ensemble method.

Various attempts to ensemble Mask R-CNN and RetinaNet models were done respectively, gaining slightly higher mAP compared to adopting only a top-performing classifier involved in the ensemble. Then combining several Mask R-CNN and RetinaNet models was done for the ensemble. The suitable ratio of weights given to each classifier played an important role in our competition score. By fine-tuning the ratio of the weights given to each classifier, the mAP was increased to 0.21746 with a late submission in Kaggle RSNA Pneumonia Detection Challenge, approximately could be ranked as a 21st place out of 1499 in the competition private leaderboard.

## References

[1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sa´nchez, “A survey on deep learning in medical image analysis,” *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[2] D. Ravı, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, “Deep learning for health informatics,” *IEEE journal of biomedical and health informatics*, vol. 21, no. 1, pp. 4–21, 2017.

[3] W. H. Organization *et al.*, “Standardization of interpretation of chest radiographs for the diagnosis of pneumonia in children,” 2001.

[4] L. R. Goodman, Felson’s principles of chest roentgenology: a programmed text. Elsevier Health Sciences, 2014.

[5] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[7] R. Polikar, “Ensemble learning,” *Scholarpedia*, vol. 4, no. 1, p. 2776, 2009, revision #186077.

[8] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dolla´r, “Focal loss for dense object detection,” *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[9] K. He, G. Gkioxari, P. Dolla´r, and R. Girshick, “Mask r-cnn,” in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2980–2988.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779– 788.

[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[12] T.-Y. Lin, P. Dolla´r, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, “Feature pyramid networks for object detection.” in *CVPR*, vol. 1, no. 2, 2017, p. 4.

[13] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *Advances in neural information processing systems*, 2015, pp. 91–99.

[14] P. Soviany and R. T. Ionescu, “Optimizing the trade-off between single- stage and two-stage object detectors using image difficulty prediction,” *arXiv preprint arXiv:1803.08707*, 2018.

[15] H. Jung, B. Kim, I. Lee, M. Yoo, J. Lee, S. Ham, O. Woo, and J. Kang, “Detection of masses in mammograms using a one-stage object detector based on a deep convolutional neural network,” *PloS one*, vol. 13, no. 9, p. e0203355, 2018.

[16] A. Mikołajczyk and M. Grochowski, “Data augmentation for improving deep learning in image classification problem,” in *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, May 2018, pp. 117–122.

[17] C. Rene´ and V. Hager, “Temporal convolutional networks for action segmentation and detection,” in *IEEE International Conference on Computer Vision (ICCV)*, vol. 1, no. 2, 2017, p. 3.

[18] T. D. Team, “Pneumonia detection in chest radiographs,” *arXiv preprint arXiv:1811.08939*, 2018.