Identification of Pneumonia from X-Ray Images using Machine Learning Techniques

Doshi Ruchit, Ghule Lalit, Shetty Akanksh, Woodford Jeffery

Abstract

Each year in the United States, more than 250,000 people go to the hospital because of pneumonia and around 50,000 people die from it [1]. Quick and accurate diagnosis of pneumonia is critical, as the symptoms often overlap with more mild conditions such as colds and the flu, meaning it can grow unsuspected until it is too late. According to the National Heart, Lung, and Blood Institute, the best test for diagnosing pneumonia is a chest X-Ray [2]. However, detecting pneumonia can be challenging and requires experienced radiologists, and radiologists are not consistent amongst each other when diagnosing pneumonia based upon X-Rays [3]. To help counteract this variability in performance of trained radiologists, a machine learning model can be created that can flag X-Rays for more in-depth evaluation for potentially having pneumonia. This project sought to create different models to assist radiologists when diagnosing pneumonia. The best model implemented in this project achieved an accuracy of 90% and an F1 Score of 0.93 while maintaining a Recall of 0.98.

1. Introduction

Pneumonia is a deadly disease that presents a high risk to many people around the world. The World Health Organization claims that nearly 4 million people die prematurely each year from pneumonia and other illnesses related to household air pollution each [4]. More than 150 million people get pneumonia each year and the rate of infection is especially prevalent in children under 5 [5]. The issue can be even more prevalent due to the lack of medical professionals in developing nations, such as the need for 2.3 million more doctors and nurses in Africa [6], [7]. By providing machine learning based tools for medical professionals to aid in the diagnosis of pneumonia, the issues related to the lack of medical professionals can be reduced and it can be easier to identify pneumonia while it is still treatable.

The implementation of clinical decision support algorithms for medical imaging faces many challenges with reliability and interpretability. Here, a diagnostic tool based on a deep learning framework for the screening of patients with pneumonia was developed. The framework utilized a Convolutional Neural Network which learned distinguishable features and subsequently fed them to a fully connected Neural Network. After applying this approach to a dataset of X-Ray images, the performance was comparable to that of human experts in classifying pneumonia and non-pneumonia. This tool may ultimately aid in expediting the diagnosis and referral of these treatable conditions, thereby facilitating earlier treatment, resulting in improved clinical outcomes. However, the model alone should not be used to make the diagnosis, and a trained professional should still look over the relevant information and data.

The traditional algorithmic approach to image analysis for classification previously relied on handcrafted object segmentation, followed by identification of each segmented object using statistical classifiers or shallow neural computational machine-learning classifiers designed specifically for each class of objects, and finally, classification of the image [8]. Creating and refining multiple classifiers required many skilled people, much time, and was computationally expensive [9], [10], [11]. The development of Convolutional Neural Network layers has allowed for significant gains in the ability to classify images and detect objects in a picture [12], [13]. There are multiple processing layers to which image analysis filters, or convolutions, are applied. The abstracted representation of images within each layer is constructed by systematically convolving multiple filters across the image, producing a feature map that is used as the input to the following layer. This architecture makes it possible to process images in the form of pixels as input and to give the desired classification as output. The image-to-classification approach in one classifier replaces the multiple steps of previous image analysis methods. In addition to CNN, shallow machine learning algorithms as well as ensemble techniques have been implemented in this paper. The shallow algorithms with raw pixels have the least accuracy. When the output after the convolution layers is fed to the shallow algorithms, they perform better. However, CNN has the best values for the metrics used to evaluate this model.

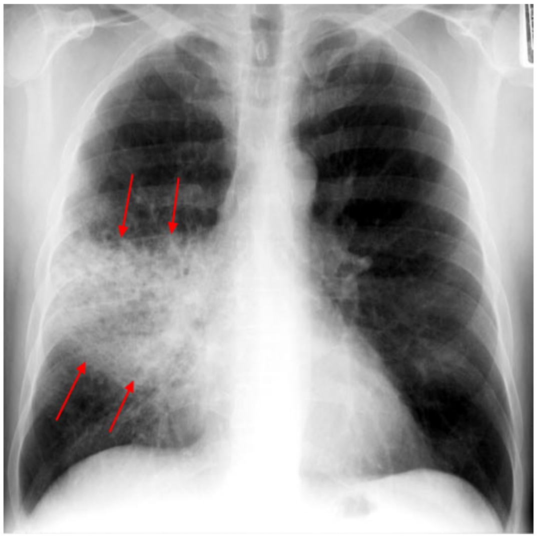


Figure 1. Pneumonia X-Ray image

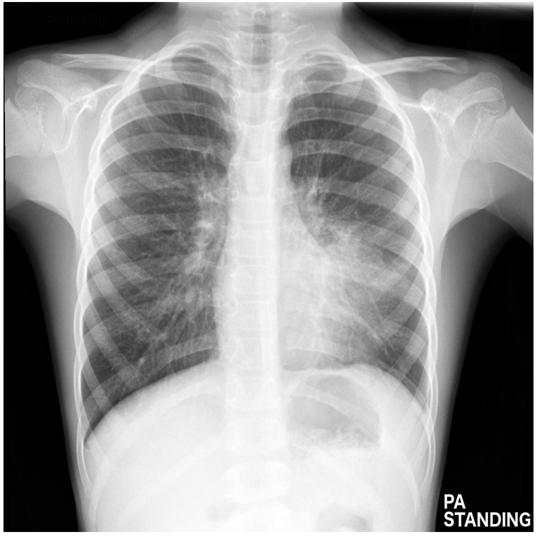


Figure 2. Normal X-Ray image

2. Related work

Researchers at Stanford University's Machine Learning group have used a similar X-Ray image dataset for classification. The dataset, released by the NIH, contains 112,120 frontal-view X-Ray images of 30,805 unique patients, annotated with up to 14 different thoracic pathology labels using NLP methods on radiology reports. Images that have pneumonia as one of the annotated pathologies were labeled as positive cases of pneumonia and all other images were labeled as negative for the pneumonia detection task.

ChexNet is a 121 layer convolutional neural network that inputs a chest X-Ray image and outputs the probability of pneumonia along with a heatmap localizing the areas of the image most indicative of pneumonia. For the pneumonia detection task, data set was randomly split into training (28744 patients, 98637 images), validation (1672 patients, 6351 images), and test (389 patients, 420 images). There is no patient overlap between the sets. Before inputting the images into the network, the images were downscaled to 224 by 224 and normalize based on the mean and standard deviation of images in the ImageNet training set. Also, training was augmented with random horizontal flipping.

The results obtained by CheXNet outperforms the four test radiaologists with respect to F1 score. Average F1 score of four radiologists was lesser than algorithm's F1 score of 0.43.

Automated diagnosis from chest radiographs has received increasing attention with algorithms for pulmonary tuberculosis classification [14] and lung nodule detection [15]. Islam et al. studied the performance of various convolutional architectures on different abnormalities using the publicly available OpenI dataset [16], [17]. Wang et al. released ChestX-ray8, an order of magnitude larger than previous datasets of its kind, and also benchmarked different convolutional neural network architectures pre-trained on ImageNet [18]. Recently Yao et al. exploited statistical dependencies between labels in order make more accurate predictions, outperforming Wang et al. on 13 of 14 classes [19].

All the above mentioned approaches mainly stick to the CNN approach and making it better for classification. This approach is different as shallow algorithms, ensemble techniques and other approaches have been tried in order to make predictions.

3. Data

A dataset from Kaggle was used for this project [20], [21]. It consists of three sets of images of chest X-Rays labeled as Test, Train and Validation. There are 3742 images in the training set, consisting the chest X-Rays of patients with and without Pneumonia. The Testing folder consists of 624 images with their labels which can be used to verify the accuracy of the models. In training images' set, there are 1371 normal images and 2371 images are of Pneumonia class. Out of 624 test images, 390 images are of Pneumonia class and the remaining of 234 are normal X-Ray images.

One of the major problems with this data set is the image size and image content. Every image is of different size. Also, the actual area captured by X-Ray is different. Some of the images have captured the complete ribcage and some of them have captured only a part of it. To overcome this, the images were resized. Initially, images were resized to 224 by 224 size. Also, 227 by 227 image size was also tried for AlexNet. In the end, a 150 by 150 image size was used as it provided better performance and helped in efficient computation. The images were converted to greyscale and then they were normalized before feeding them to any of the algorithms. Code was written to identify corrupt images amongst the entire dataset. These images were removed from the dataset before training.

4. Methods

4.1. Shallow Algorithm

Different Shallow Machine Learning algorithms were used to classify the images between with and without Pneumonia. The algorithms selected were Random Forest, Support Vector Machine, Logistic Regression, and XG-Boost. Initially, the normalized pixel values of the images were used as the input to these algorithms. Each algorithm was applied to the data set and the hyper parameters of the algorithms were tweaked to get better results. Tuning of the hyper-parameters was done as follows:

- Random Forest: The number of estimators were tweaked, keeping all the other parameters as default. Number of estimators was tweaked from 10 to 30. The accuracy of this model was fluctuating between upper 70's and the low 80's. The best accuracy of the random forest model was found to be 81.66% with the value of n_estimator= 20.

- Support Vector Machine: The algorithm was tested using different kernels such as Linear, Poly and Radial Basis Function (RBF). The accuracy of this model on the given data set was varying in the 70's. The best accuracy of this model was obtained using the Linear kernel, which was equal to 75.48%.

- Logistic Regression: Multiple different solvers were tried for logistic regression, and L1 and L2 regularization were also tested. The max number of iterations was reduced from 100 to 5 to prevent overfitting. The accuracy of this model was between 75% and 83%. Using the saga solver and L2 regularization, the best accuracy was 82.85%.

- XG_Boost: Two parameters were tweaked in the The XG_Boost model, namely: the learning rate and the number of estimators. The value of the learning rate was iterated from 0.001 to 0.1 and the value of the N_estimators was tweaked from 100 to 500. The results were found to be in the mid 70's and the best accuracy was found to be equal to 77.40%.

Along with raw pixel values as input to the shallow algorithms, the Convolutional Neural Network's output was also used as the input. For this, the flattened output of the final convolution layer was fed as the input. It makes more sense to feed these values as compared to just feeding raw values as they are learned features. The accuracy improved compared to the raw pixel accuracy. A point to be noted is that even though there was an increase in accuracy, the increment is not significant enough. For most of the shallow algorithms, the increment over base accuracy was approximately 2% to 5%. Apart from convolution layer's output as input to the shallow algorithm, one more way is to feed the activation of fully connected neural net's hidden layer as input (in this case, 1024 neurons). The results obtained by this method were not satisfactory and accuracy was even lesser than the convolution layer's output. Detailed results such as accuracy, recall and F1 score are showed in Figures 7, 8, and 9.

Algorithm	% Accuracy	F1 Score
Random Forest	81.66	0.87
Logistic Regression	75.8	0.85
Support Vector Machine	75.84	0.83
XG-Boost	77.4	0.85

4.2. Deep Algorithm

A Convolutional Neural Network was implemented using Keras with TensorFlow backend [22]. Many different architectures were implemented before arriving at the one shown in Figure 3 and Figure 4. Each image was first converted to greyscale and then normalized. The image then went through multiple convolution layers. Each convolution layer had a relu activation applied to it and a padding layer was used to keep spatial sizes constant after convolution. All the convolution layers had a kernel size of 3x3, and all the maxpooling layers had a size of 2x2. A total of eight convolution layers were used before the flattened output was fed into the first layer of fully connected network. The first two convolution layers had a filter size of 16. The first two convolution layers were followed by a maxpooling layer. A dropout of 0.3 was applied after these initial operations. The next set of two convolution layers had 32 filters. This was followed by a maxpooling layer and then batch normalization was performed. The third set of two convolution layers had 64 filters. This was followed by a maxpooling layer and then batch normalization was performed again. The final set of two convolution layers had 128 filters. This was again followed by the final maxpooling layer and then batch normalization.

After these operations were performed, the output of the final convolution layer is flattened and passed as the input to the fully connected network. This network had four hidden layers with 1024, 512, 128 and 64 neurons respectively. Relu activation was used throughout this network. A dropout of 0.3 was applied on the second, third, and fourth hidden layers to aid in avoiding overfitting. The final output of the last hidden layer was then connected to the final output layer of the network. The final layer of the fully connected network had only one output to which a sigmoid nonlinearity was applied. The network was trained using the Adam optimizer. A Mini batch of size 32 was used along with Binary Cross entropy loss in this model. Figure

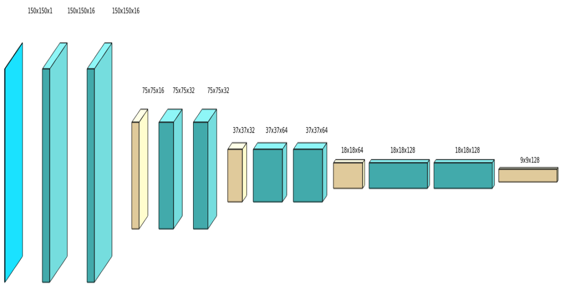


Figure 3. Convolutional Network Architecture

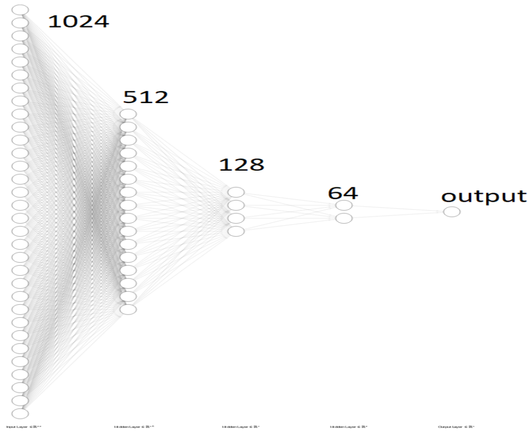
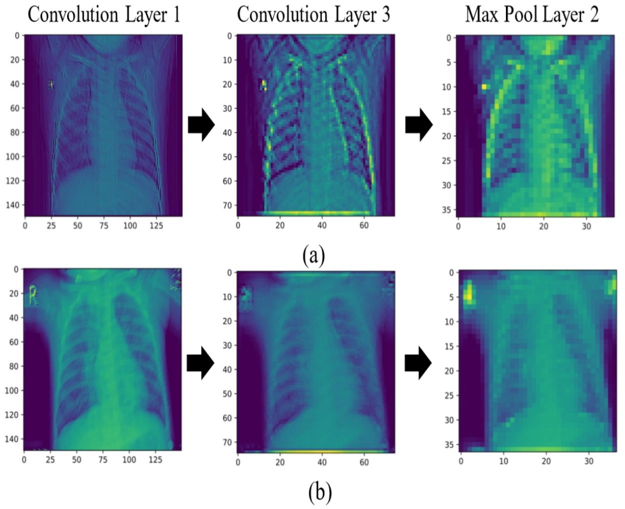


Figure 4. Fully Connected Network

5 shows the transformation of an image as it goes through the above described deep learning network.

Many different parameters were tweaked before arriving at the network described above. Different numbers of filters, types of activation functions, number and order of convolutional layers, as well as the fully connected layers were changed. Adaptive Learning Rate Method (adadelta) optimizer was also tried while training the network. In addition to this, the last layer was changed to 2 neurons with a softmax activation. All the above combinations resulted in lower values for metrics used to evaluate the model. Apart from this, transfer learning technique was also implemented. AlexNet was also implemented on this dataset [12]. For this, the images were reshaped to 227 by 227 as AlexNet uses this image size. The accuracy obtained by the ALexNet structure and fully connected neural network discussed earlier was 85%. Hence, eventually, AlexNet was changed to the above discussed Convolution Network.



(a)

(b)

Figure 5. (a) Normal X-Ray image after Convolution and (b) Pneumonia X-Ray image after Convolution

4.3. Ensemble methods

In machine learning, ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. Evaluating the prediction of an ensemble typically requires more computation than evaluating the prediction of a single model, so ensembles may be thought of as a way to compensate for poor learning algorithms by performing a lot of extra computation. In this study, the ensemble technique used was based on normal majority voting. A total of five models were combined. Combining the models actually decreased the overall accuracy when compared to the Convolutional Neural Network alone. As other algorithms do not perform as good as the Convolutional Neural Network, they actually mis-predict in the ensemble technique. Hence, eventually, only the Convolutional Neural Network's output was considered as the final prediction.

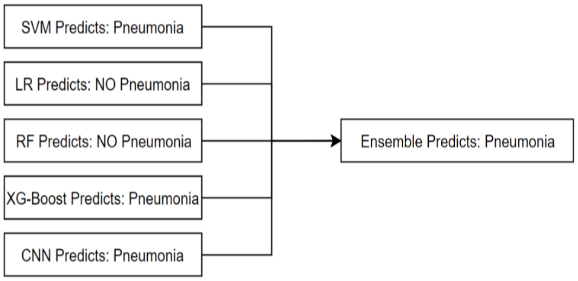

SVM Predicts: Pneumonia
LR Predicts: NO Pneumonia
RF Predicts: NO Pneumonia
XG-Boost Predicts: Pneumonia
CNN Predicts: Pneumonia
Ensemble Predicts: Pneumonia

Figure 6. Ensemble Method

5. Results

For evaluating the performance of the models, a different method is needed than just comparing accuracy alone. For medical applications, recall is very important. This is because it is very important that the model does not mispredict X-Ray images of patients that do have pneumonia as not having pneumonia. This would mean that the model would tag an X-Ray image as being safe when the patient needs further medical attention. Such a scenario should be avoided at all costs. Thus, the recall of the models is more important than the accuracy because it is associated with how well the model can catch cases of pneumonia. But recall should be balanced with precision so that the model is useful and not predicting that every X-Ray indicates pneumonia. Therefore, to compare models, the F1 Score is also used.

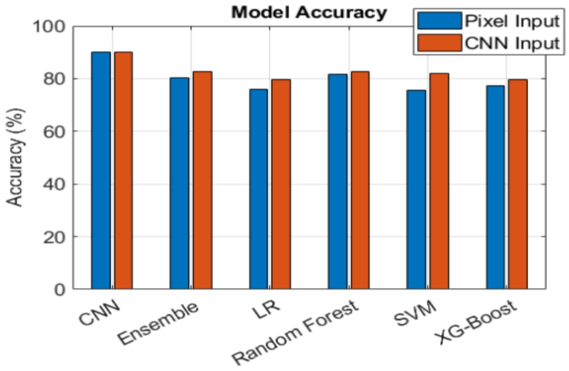


Figure 7. Final Accuracy

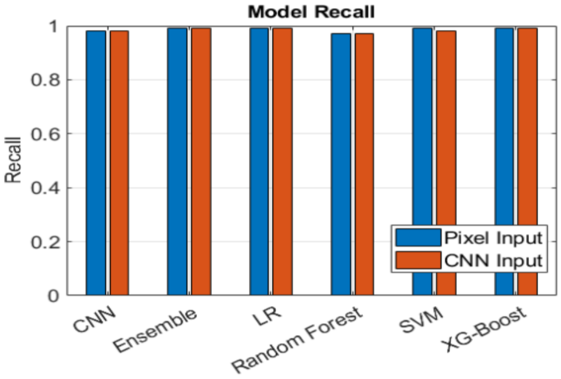


Figure 8. Recall

Changing the input of the shallow algorithms to the Convolutional Neural Network output increased the individual accuracies and F1 scores without significantly changing the Recall, but the overall changes were not statistically significant. For the ensemble, the accuracy increased by 2% and the F1 Score increased by 0.02 while maintaining the Recall at 0.99. Nonetheless, the input of the models should be the output of the convolutional layers. However, the recommended method of use would be to drop the shallow models and ensemble and exclusively use the Convolutional Neural Network. If only the Convolutional Neural Network is used, the model will have an accuracy of 90% and a F1 Score of 0.93 while maintaining a Recall of 0.98.

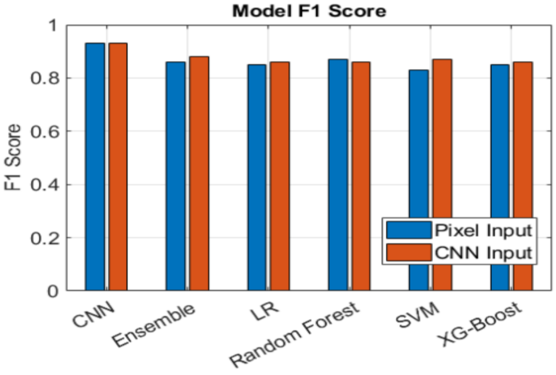


Figure 9. F1 Score

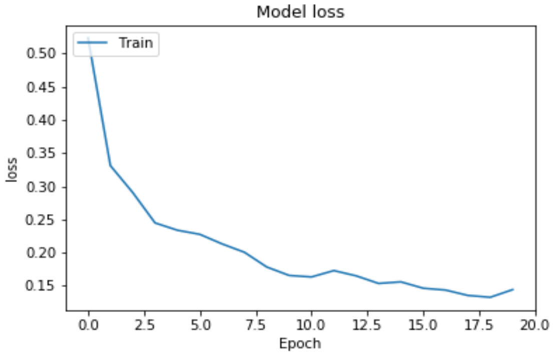


Figure 10. Overall training loss

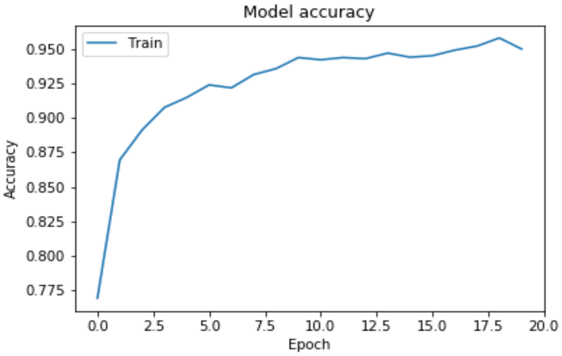


Figure 11. Training Accuracy

6. Conclusion

6.1. Summary

The Convolutional Neural Network performing the best is not surprising as Convolutional Neural Networks are preferred when handling large inputs and images because of the convolution layers that aid in extracting the important features from the images. With an accuracy of 82.69 % a F1 Score of 0.88 and a Recall of 0.99 the combined model did not perform as well, but is still useful. To improve the combined model, different methods of extracting features from images should be implemented, as was found by changing the input to the output of the convolution layers rather than raw pixel values. Overall, both methods could be used by radiologists to assist in the diagnosis process, but care should still be taken when using these models and a radiologist should not rely on the model to make a diagnosis, but rather use it as a tool help identify X-Rays that may indicate the presence of pneumonia.

6.2. Future Work

Future work in this project would be to add a softmax layer after the last hidden layer in order to perform multi-class classification. Similar networks can be used to flag X-Ray images for many different health issues, not just pneumonia. Also, a weighted average of each algorithm can be incorporated in majority voting while computing the final prediction. This will be useful as more importance would be given to algorithms that perform better.

7. References

[1] "Pneumonia Can Be Prevented-Vaccines Can Help," Centers for Disease Control and Prevention, 21- Nov-2019. [Online]. Available: https://www.cdc.gov/pneumonia/prevention.html?CDC$_{A}A_{r}efVal = https://www.cdc.gov/features/pneumonia/index.html.$

[2] "Pneumonia," National Heart Lung and Blood Institute. [Online]. Available: https://www.nhlbi.nih.gov/health-topics/pneumonia.

[3] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, R. L. Ball, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," 2017. arXiv preprint arXiv:1711.05225.

[4] "Household air pollution and health," World Health Organization. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/household-air-pollution-and-health.

[5] Rudan I, Tomaskovic L, Boschi-Pinto C, and Campbell H, WHO Child Health Epidemiology Reference Group. Bull World Health Organ. 82(12):895-903, 2004.

[6] Narasimhan V., Brown H., Pablos-Mendez A., et al. Responding to the global human resources crisis. The Lancet. 2004;363(9419):1469–1472. doi: 10.1016/s0140-6736(04)16108-4.

[7] Naicker S., Plange-Rhule J., Tutt R. C., Eastwood J. B. Shortage of healthcare workers in developing countries. Africa, Ethnicity Disease. 2009;19:p. 60.

[8] Goldbaum, Michael Moezzi, S. Taylor, A. Boyd, J. Hunter, E. Jain, Ramesh. (1996). Automated diagnosis and image understanding with object extraction, object classification, and inferencing in retinal images. 695 - 698 vol.3. 10.1109/ICIP.1996.560760.

[9] S. Chaudhuri, S. Chatterjee, N. Katz, M. Nelson, and M. Goldbaum, "Detection of blood vessels in retinal images using two-dimensional matched filters," IEEE Transactions on Medical Imaging, vol. 8, no. 3, pp. 263–269, 1989.

[10] Hoover, Adam Goldbaum, Michael. (2003). Locating the optic nerve in a retinal image using the fuzzy convergence of the blood vessels. IEEE transactions on medical imaging. 22. 951-8. 10.1109/TMI.2003.815900.

[11] A. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," IEEE Transactions on Medical Imaging, vol. 19, no. 3, pp. 203–210, 2000.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Communications of the ACM, vol. 60, no. 6, pp. 84–90, 2017.

[13] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," Computer Vision – ECCV 2014 Lecture Notes in Computer Science, pp. 818–833, 2014.

[14] P. Lakhani and B. Sundaram, "Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks," Radiology, vol. 284, no. 2, pp. 574–582, 2017.

[15] P. Huang, S. Park, R. Yan, J. Lee, L. C. Chu, C. T. Lin, A. Hussien, J. Rathmell, B. Thomas, C. Chen, R. Hales, D. S. Ettinger, M. Brock, P. Hu, E. K. Fishman, E. Gabrielson, and S. Lam, "Added Value of Computer-aided CT Image Features for Early Lung Cancer Diagnosis with Small Pulmonary Nodules: A Matched Case-Control Study," Radiology, vol. 286, no. 1, pp. 286–295, 2018.

[16] M. T. Islam, M. A. Aowal, A. T. Minhaz, and K. Ashraf, "Abnormality Detection and Localization in Chest X-Rays using Deep Convolutional Neural Networks," 2017. arXiv preprint arXiv:1705.09850.

[17] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. Mcdonald, "Preparing a collection of radiology examinations for distribution and retrieval," Journal of the American Medical Informatics Association, vol. 23, no. 2, pp. 304–310, Jan. 2015.

[18] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[19] Yao, L., Poblenz, E., Dagunts, D., Covington, B., Bernard, D., Lyman, K. "Learning to diagnose from scratch by exploiting dependencies among labels," 2017. arXiv preprint arXiv:1710.10501.

[20] P. Mooney, "Chest X-Ray Images (Pneumonia)," Kaggle, 24-Mar-2018. [Online]. Available:

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

[21] Kermany, Daniel; Zhang, Kang; Goldbaum, Michael (2018), “Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification”, Mendeley Data, v2 http://dx.doi.org/10.17632/rscbjbr9sj.2

[22] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S. ”TensorFlow: Large-scale machine learning on heterogeneous systems,” 2015. Software available from tensorflow. org.