Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 192 (2021) 658–665

Procedia Computer Science

www.elsevier.com/locate/procedia

25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Chest X-Ray Image Preprocessing for Disease Classification

Guy Caseneuve, Iren Valova*, Nathan LeBlanc, Melanie Thibodeau

Computer and Information Science Department, University of Massachusetts Dartmouth, Dartmouth, MA 02747, USA

Abstract

Research has proven that images with blur, low contrast or other deficiencies are detrimental to accuracy of classification models. Working with a dataset of blurry, hard-to-see images creates an uncertainty in whether the machine learning model is able to accurately identify objects. These hard to identify images require the use of metadata on the image to be able to truly understand and classify the image accuracy within a model. To overcome the necessity of metadata for classification of hard to identify images, we propose a method of identifying the unfit images for purposes of cleaning the dataset prior to neural network training. In our work, we use Sobel and Scharr operators-based edge detectors to produce an image with detected boundaries among the elements of the original X-Ray. The resulting images are used to train a shallow CNN to classify the image as clear or lacking in quality. Our proposed method identifies the clear, quality images with 95% accuracy and can be utilized in future disease classification models. We define the proposed approach as image preprocessing aiming to weed out infit, blurry, low/high intensity X-Rays and create a dataset suitable for disease identification.

Keywords: edge detection, Sobel/Scharr operators, shallow CNN, image preprocessing, chest X-Ray film

1. Introduction

The power of convolutional neural networks (CNN) in classifying images in various settings is indisputable. There are many CNN variants designed to improve the performance based on the peculiarities of a specific dataset. In addition, CNNs are designed to automate the image preprocessing work of preparing the inputs for classification.

In our presented work we are focusing on a chest X-Ray dataset collected by the authors of [1]. While machine learning methods have evolved significantly in recent years, openly available medical image databases are still scant.

* Corresponding author. Tel.: +1-508-999-8502; fax: +1-508-999-9144.
E-mail address: ivalova@umassd.edu

10.1016/j.procs.2021.08.068

The ChestX-ray8 [1] dataset is the largest one with annotated and labeled images. Unlike many other subjects, pathology labels cannot be obtained through crowdsourcing. The dataset has been utilized in several studies so far, staring with its creators who offer a proof of concept for classifying nine common thoracic diseases via image processing combined with textual metadata analysis. The authors share their approach to image preprocessing as confining to resizing the images down to the more manageable 1024x1024 as well as developing bounding boxes for the specific manifestation of the disease on the particular image. While bounding boxes are the norm in many CNN applications for the purposes of localization before identification [2], in cases of poor image quality, the localization task may prove insurmountable and lead to weak classification performance.

The chest X-Ray dataset is used in another study [3] for the diagnosis of pneumonia with 85% accuracy. The images are further downscaled and additional metadata is attached for the pneumonia cases. The emphasis is on the disease identification glossing over details of the image preparations. As with all very large datasets collected from variety of sources to induct various classes, the quality of images is not guaranteed. By elaborating on the process used for eliminating hard-to-identify data, a novelty in understanding of the data is created that has not yet been explored in literature. This leads to a repeatable study that readers can perform on the preprocessing step of the reported research.

Another study [4] focuses on improving the CNN classifier performance by incrementing the size of the dataset. The final selected images are reported to be preprocessed for contrast enhancement, histogram equalization, mean-standard deviation normalization and downscaling. While the reported results are impressive at 96%, the amount of image rejection due to various consideration is also prohibitive for smaller teams of researchers.

Very recent work with chest X-Ray images reports on excellent classification results, with the dataset preparation stage featuring a slew of image enhancing techniques coupled with handcrafted feature extraction. The study employs principal component analysis, morphological segmentation, shape and texture feature extraction, feature fusion and selection to name a few [5].

It is frequently said in the data mining world that data preparation is 70% of the job. Indeed, as we tackle the chest X-Ray dataset, many images of poor quality are encountered – blurring, lack of contrast, shearing and other defects. Relying on such images for the successful classification of disease without the use of textual related metadata is a insurmountable task. This led us to seek additional image preparation, rather selection, of “qualified” images to eventually address the final classification task. This work focuses on the image selection method developed in our quest.

2. Data set

The original data set was built to identify common thoracic pathology disease [1] and contains over 108,948 frontal-view X-ray images from 32,717 patients. Each X-ray image is assigned to one or multiple labels revealing some connection between pathologies. Eight of the diseases are featured in Fig.1 [1]. It is evidenced that these exemplars cleanly display both lungs without blur, shearing or contrast deficiencies. Fig. 2 offers some of the images in the dataset and the reasoning behind the approach we propose in this paper.

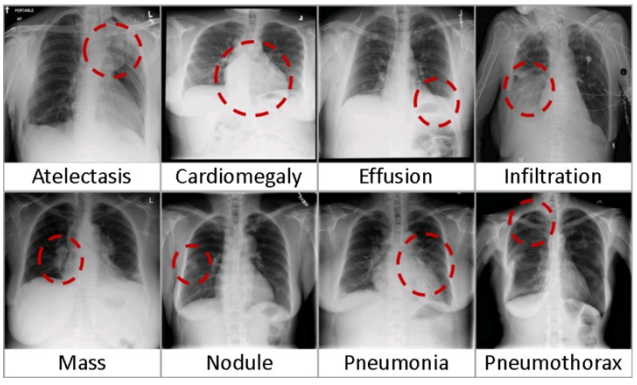
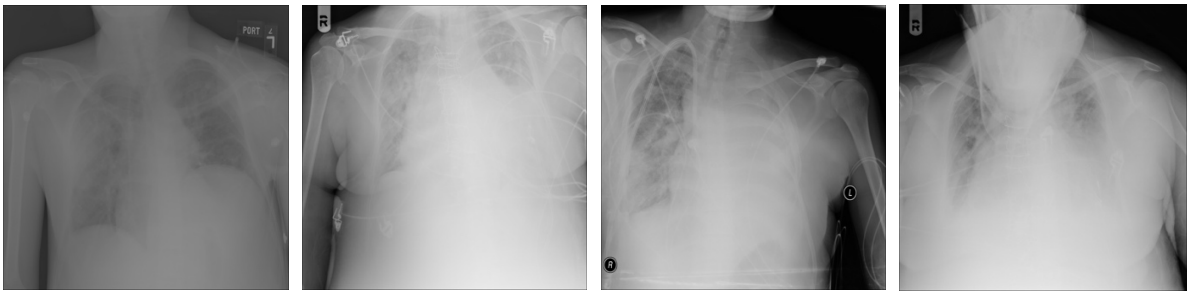
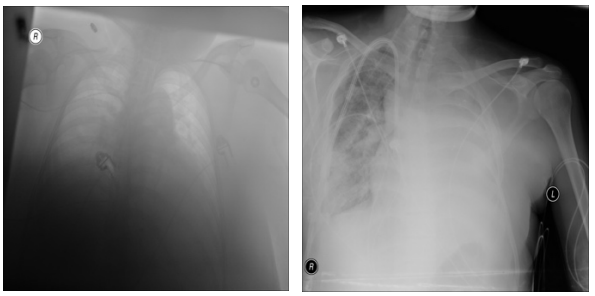


Figure 1. [1] Frontal chest X-Ray film with identified diseases



(a) (b) (c) (d)

Figure 2. Films identified as lacking in various aspects and labeled as: a) pleural thickening pneumonia; b) cardiomegaly edema effusion infiltration; c) no disease findings; d) no disease findings



(a) (b)

Figure 3. Examples of images with: a) low brightness; b) noisiness and blurriness

The images in Fig.2 showcase some of the issues that could eliminate the image from being successfully utilized in a CNN-based image classification task. The many artifacts and noise are factors in the inability of machine learning methods to explore fully the information contained in the pixels. Fig.3 further illustrates the need for a more detailed approach in identifying clear images where encoded information is available to the automated expertise of CNNs. Conversely, Fig.4 offers an example of clear image, the likes of the showcased images in [1].

Figure 4. Clear frontal chest X-Ray with labelling for cardiomegaly emphysema

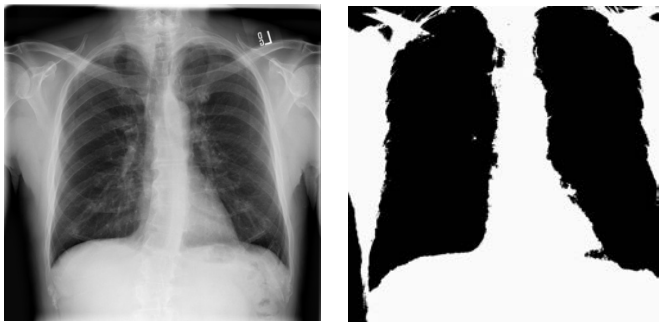
Based on reviewing 500+ images, we identified four major causes of identifying the image as lacking: noise, blur, crop, brightness. Additional image processing methods were applied before handing the images to a CNN to classify them as clear or lacking. The dataset used for this work is derived from chest X-Ray dataset designed in [1] and includes 3,402 edge-detected images.

3. Methodology

One of the overarching goals is to identify images with two clearly visible lung lobes. As such, the first thought goes to thresholding. In addition, we tested Otsu's method and selected Scharr's Operator for edge detection. The so processed images are then submitted to a shallow CNN for final classification into clear and lacking.

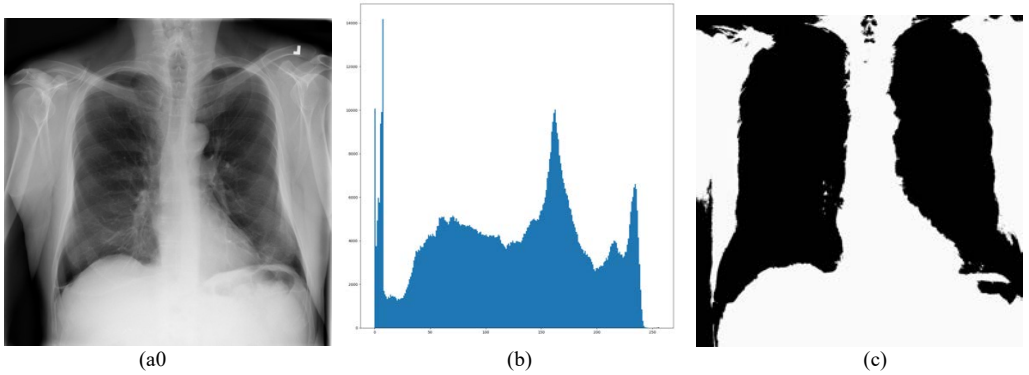
3.1. Threshold approach

Problems abound as ribs, collar bones, and disease artifacts can be easily lost under such approach as shown in Fig.5.



(a) (b)

Figure 5. Original and thresholded images



(a0 (b) (c)

Figure 6. Otsu threshold experiment and results: a) original image; b) histogram; c) threshold result

Clearly a similar approach, but one that affords for emphasis of important features, needs to be selected. The next candidate for image processing is Otsu's threshold [6] as the actual threshold value is adaptive. The critical issue here being the adaptiveness of the threshold value. Otsu's method searches for the threshold intensity to maximize the between-class variance. While we detect the lung shapes as shown in Fig.6 with few more details, it is still not reaching the result needed to train a CNN to differentiate into clear and lacking images.

3.2. Edge detection

Edge detection is another tried and true method in image processing. We applied the Sobel/Scharr operator. The foundation of the Sobel operator is a gradient operation – the goal of it is the same as ours – to extract the boundaries within the image. Generally, a 3x3 kernel is used, so that horizontally the image is convolved via G_x and vertically via G_y (Eq.1).

$$G_x = \begin{matrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{matrix} * Image, G_y = \begin{matrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{matrix} * Image \quad (1)$$

Figure 7. Sobel/Scharr operator edge detection

Sobel is essentially investigating the presence of boundaries between the upper/lower part of the image or left/right in the horizontal and vertical gradient calculations [7]. In other words, Sobel operator will show whether changes in the image are abrupt or smooth.

As an additional optimization and introduction of increased sensitivity, the Scharr operator represents a modification where the mean squared angular error is measured. This allows the kernels to be of derivative nature, better approximating Gaussian filters [8]. The result of this edge detection step is shown in Fig.7. The bones are now more visible and the boundaries will supply the needed visual information for the CNN stage.

3.3. CNN

The utilized CNN has a shallow architecture with four convolutional layers, with ReLu activation functions. Input of 500 x 500 x 1 with an initial 5 x 5 convolution layer with a stride of 1 is applied. After each convolution, 2 x 2 maxpooling layer is reducing the dimension of the future maps (architecture shown in Fig.8). To prevent overfitting, about 20% of neurons are dropped during training. Stochastic gradient optimizer is proven to be the best performing. We did not apply batch normalization in the convolutional layers, but to the training set. After the last convolution, the outputs (186,624 nodes) are passed to fully connected layers, which have two hidden layers (120 nodes, 84 nodes) and produce one output. Batch cross entropy loss function with sigmoid layer is followed by batch cross entropy loss function to help with the numerically unstable problems. The output is then applied to a softmax function within the loss function.

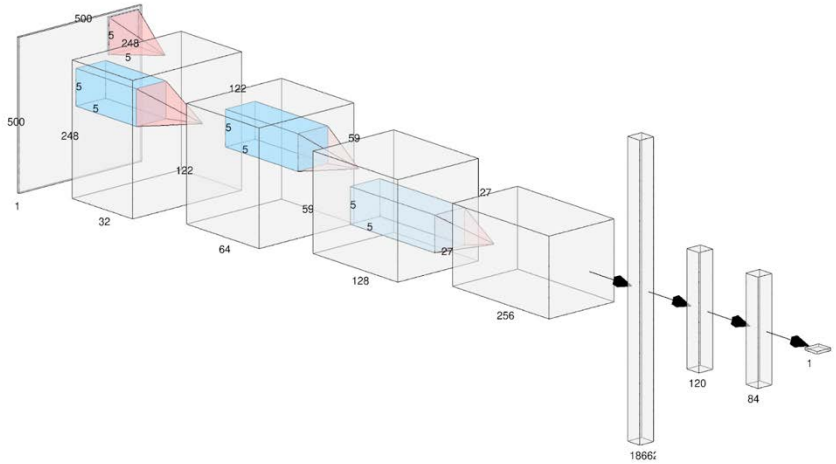


Figure 8. CNN architecture

The dataset used for the work is derived from chest X-Ray dataset designed in [1] and includes 3402 edge-detected images: 2,381 training, 681 validation, and 340 testing images. During the dataset creation, the mean and standard deviation of the image are calculated for normalization purposes. Finally, the images are resized to 500*500. The images in the dataset are prelabeled as clear (fit film) and lacking (unfit film) based on the examples of images in other research works.

The training sessions are divided into runs of 20 epochs each. The training accuracy and loss graphs (Fig.9) represent the following runs: orange line (run 1 with learning rate of 0.01 and batch size of 32); red line (run 2 with

learning rate of 0.001 and batch size of 32); blue line (run 1 with learning rate of 0.01 and batch size of 64); ocean blue line (run 1 with learning rate of 0.001 and batch size of 64).

The training/validation performance data for learning rate of 0.01 and batch size of 32 is further analyzed through box plots as shown in Fig.10. The training and validation accuracy contains right-skewed data. The best training and validation accuracy produced by the models with learning rate of 0.01 stand at 99% and 94% respectively. The conclusion is that the CNN architecture is more likely to produce better models when using low learning rates with small batch sizes. For the models trained on over 3k images, we saw no signs of overfitting. Finally, Fig.11 provides a snippet of the validation accuracy.

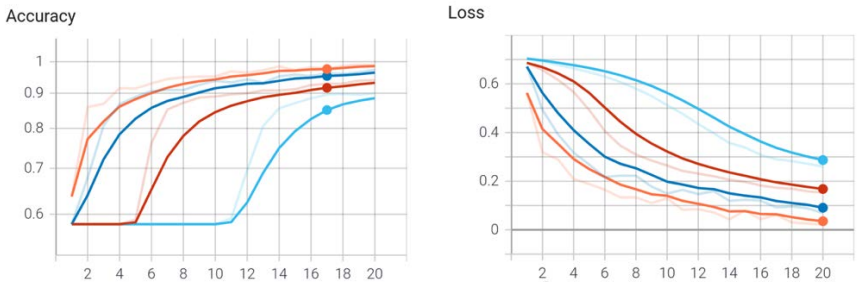


Figure 9. Training accuracy and loss functions for four runs with varied learning rate and batch size

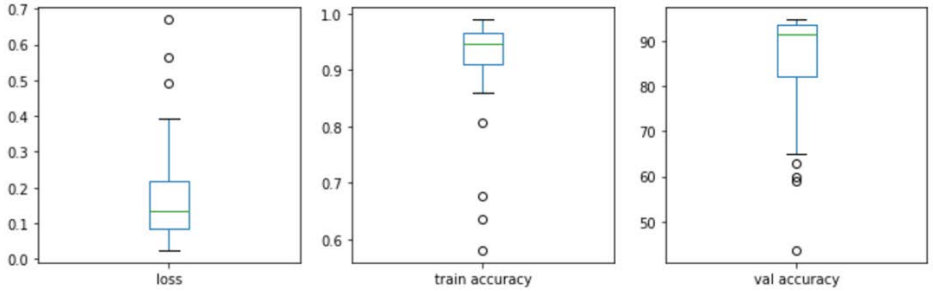


Figure 10. Statistical representation of the training/validation performance

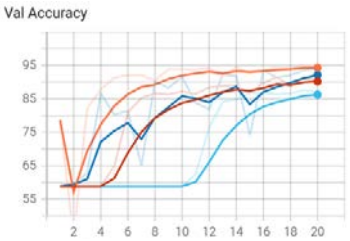


Figure 10. Validation accuracy

4. Results

All models described in the training/validation process have been tested on 340 images (10% of the dataset developed for this work). Overall accuracy as an average among all models is 95%. The model which performed best in testing learned with a rate of 0.01 and batch size of 32. The confusion matrix documenting a top performance is shown in Fig.12. Clearly, images that are suitable for classification are labeled as unfit.

In addition, we tested our model by randomly selecting 50 additional images outside of the utilized 3400 (chest X-Ray in [1]), applied edge detection as with our dataset and tested them with the CNN. A sample of the original X-Ray and the corresponding edge-detection CNN input are shown in Fig.13. This additional testing achieves 90% accuracy, as five of the 50 images are incorrectly classified by the CNN. Two misclassifications are shown in Fig.13g, h.

	Predicted label – Fit image	Predicted label – Unfit image
True label – Fit image	97.16%	2.84%
True label – Unfit image	0.0%	100%

Figure 12. Confusion matrix for the testing of the proposed method

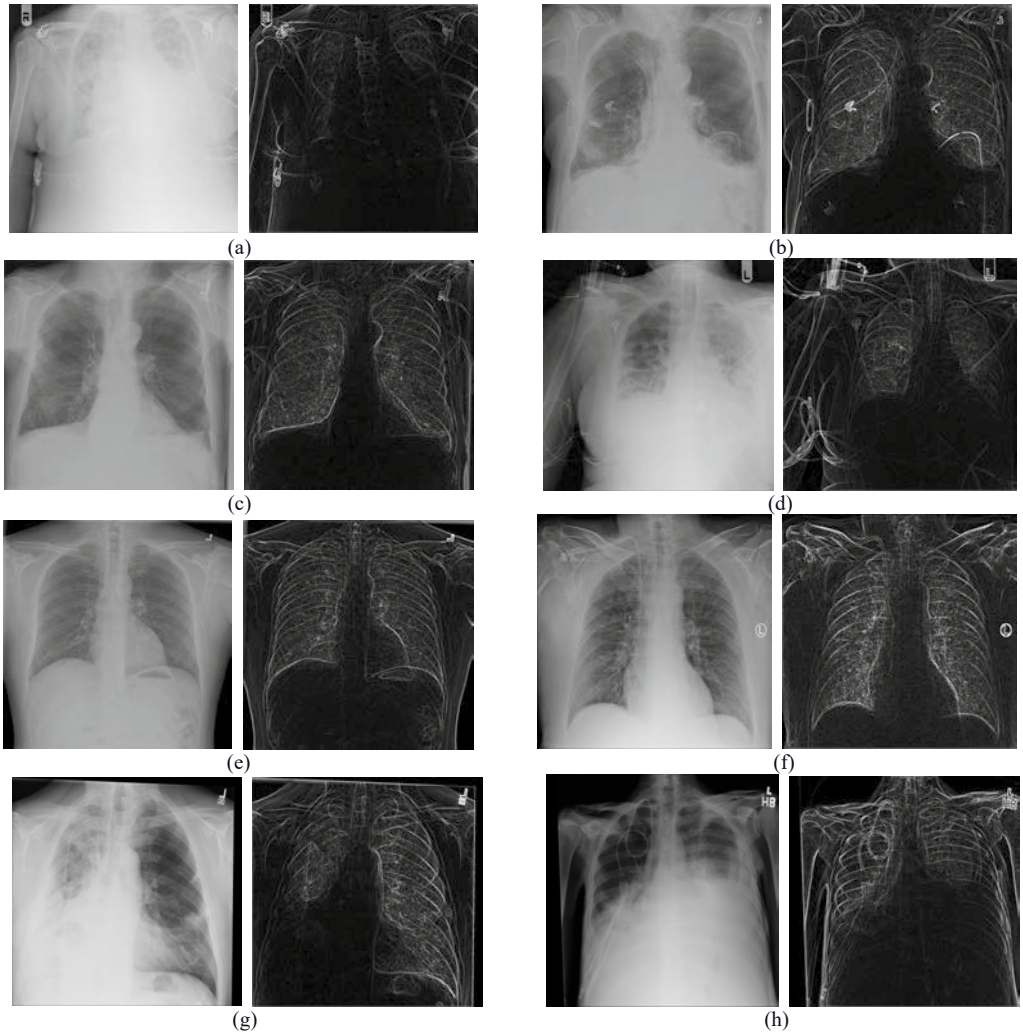


Figure 13. Additional test images from chest X-Ray dataset [1]: a) original image and edged image, correct classification as lacking; b) correct classification as clear; c) correct classification as clear; d) correct classification as lacking; e), f) correct classification as clear; g), h) the heavy shadows obscuring one lobe and the some of the artifacts assumes the image lacks the properties to be correctly classified, our approach determined the image is clear for disease classification

5. Discussion and Conclusions

This paper presents work on the significance of image processing and preparation for the purposes of identifying diseases in frontal chest X-Ray films. While existing research works document success in the classification process, their authors have utilized images along with textual information on disease labeling and additional image information. Such data are processed by a variant of CNN along with a textual classifier for the final disease decisions. This approach is understandable as images that are blurry or of low contrast or otherwise sheared can hardly provide the needed knowledge for an effective performance of a disease classifier. To overcome the need of textual information

for classification of hard to identify images, we propose a solution of automatically identify images as clear/high-quality for further classification. To perform this identification, we propose Sobel/Scharr operator edge detection followed by a shallow CNN to classify the dataset images as clear or as lacking and provide a solid foundation for classification. Sobel edge detection has been shown to be a state-of-the-art feature extraction algorithm for defining complex features [9]. The edge detection operators provide the internal boundaries of the image objects and allow the CNN classifier to determine whether the image is clear enough to hold the needed information for disease classification. Using a method of Sobel edge detection in identifying the number of features that exist in data shows promises in creating comprehensive methods for future models that rely on the quality of features that exist within in the observed data for a neural network. The shallow CNN is able to classify the images with 95% average accuracy. While comparative analysis is key in validating results, the current literature does not make it clear what preprocessing was performed to ensure the use of high-quality images in their final model. With our approach, the low-quality images will be disregarded in order to further improve the accuracy of the applied disease classifier. The novelty of the presented model is in providing future work in the area with automated image validation for classification of disease.

The research in [10] is a very comprehensive review of existing methods. It is evident that the process of manipulating images to improve the results of disease classification is a crucial step. Also of import is the purpose of the image preprocessing. Majority of the works reviewed in [10] refer to either extracting relevant features from the image or augmenting the image information with additional characteristics as harvested by image segmentation. The reviewed works report comparable accuracy to the method proposed here, but many of them depend closely on the dataset, quality of imaging, etc. The approach we propose here is focused on the automated removal of low-quality images – we are not augmenting or focusing on regions of interest in the image pertaining to a particular disease. Rather, we aim to remove the blurry, high/low intensity images that cannot produce any of the features or augmenting information. Such dataset cleaning would lead to improved usability of the images for further processing and classification.

Acknowledgements

This work could not have been done without the help of Intramural Research Programs of the NIH Clinical Centers and National Library of Medicine. Also, thank you to Stanford for including the dataset so others could use it for their research as well!

References

[1] Wang, Xiaosong, and Yifan Peng, et al. (2017) “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, 10.1109/CVPR.2017.369.

[2] Ker, Justin, and Lipo Wang, et al. (2018) “Deep Learning Applications in Medical Image Analysis”, *IEEE Access*, vol.6, pp 9375-9389, 10.1109/ACCESS.2017.2788044

[3] Rajpurkar, Pranav, and Jeremy Irvin, et al. (2017) “CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning”, https://arxiv.org/pdf/1711.05225.pdf

[4] Dunnmon, Jared, and Yi Darvin, et al.(2018) “Assessment of Convolutional Neural Networks for Automated Classification of Chest Radiographs”, Radiology, Vol 290 No 2, https://pubs.rsna.org/doi/full/10.1148/radiol.2018181422

[5] Bhandary, Abhir, and Prabhu Ananth, et al. (2020) “Deep-learning framework to detect lung abnormality – A study with chest X-Ray and lung CT scan images”, Pattern Recognition Letters, Volume 129, pp 271-278

[6] Zhang, Jun and Hu, Jinglu (2008). "Image segmentation based on 2D Otsu method with histogram analysis". 2008 International Conference on Computer Science and Software Engineering. 6: 105–108. doi:10.1109/CSSE.2008.206.

[7] H. Farid and E. P. Simoncelli. (1997) “Optimally Rotation-Equivariant Directional Derivative Kernels”, Int'l Conf Computer Analysis of Images and Patterns, pp. 207—214.

[8] Scharr, Hanno. (2007) “Optimal Filters for Extended Optical Flow” In: Jähne, B., Mester, R., Barth, E., Scharr, H. (eds.) IWCM 2004. LNCS, vol. 3417, pp. 14–29. Springer, Heidelberg.

[9] Singh, Anamika, and Manminder Singh, and Birmohan Singh. (2016) "Face detection and eyes extraction using sobel edge detection and morphological operations" in Conference on Advances in Signal Processing pp. 295-300, doi: 10.1109/CASP.2016.7746183.

[10] Qin, Chunli, and Demin Yao, et al. (2018) “Computer-aided detection in chest radiography based on artificial intelligence: a survey”, BioMed Eng OnLine 17, 113, https://doi.org/10.1186/s12938-018-0544-y.