Using machine learning to diagnose chest X-rays and interpret patient symptoms and medical history

Rohan M. Bhansali

Loudoun Academy of Science

8 October 2020

Thoracic radiographs are the most commonly utilised medical diagnostic tool, with over two billion performed annually (Raoof, et al., 2012). However, there is a global shortage of radiologists to analyse these X-rays, as exemplified by the nearly two thirds of the world's population that lacks radiologists. This problem is further exacerbated in poor countries such as Rwanda, where eleven radiologists care for twelve million inhabitants, and Liberia, where, despite a population of four million, there are merely two practising radiologists (Johnson, et al., 2019).

The cardiopulmonary diseases that are typically detected through these radiographs tend to be among the most lethal; they include pneumonia, a contagion that hospitalises over a million Americans annually, of which approximately fifty thousand expire ("Pneumonia Can Be Prevented-Vaccines Can Help | CDC", n.d.); tuberculosis, which currently afflicts one fourth of the world's population and kills an annual average of 1.3 million people worldwide ("Data & Statistics | TB | CDC", n.d.); and lung carcinoma, the deadliest cancer for both men and women, with over one hundred and fifty thousand annual deaths attributed to the disease ("Lung Cancer Is the Biggest Cancer Killer in Both Men and Women" Infographic | CDC", n.d.).

The frequency of chest X-rays and the deadly nature of the diseases associated with them make their accurate diagnosis imperative. However, radiologists, though professionally trained, are subject to human limitations that include fatigue, inattentiveness and bias. Consequently, a model with automated diagnostic capabilities would have enormous consequences; for example, in areas with a deficiency of radiologists, the model could essentially replace the radiologists and provide a diagnosis of patients' X-rays with symptoms and history taken into consideration, just as a radiologist would. A model with this capability would also be extremely versatile as it could be implemented in areas where there are not insufficiencies of radiologists; in these areas, the model could act as a confirmation to the radiologists and could also expedite the diagnostic process and reduce the costs associated with it. However, to fully simulate the clinical process of diagnosis, a model must meet several criteria.

First, the model must differentiate between anteroposterior (AP) and posteroanterior (PA). Although AP and PA X-rays are both frontal radiographs, they are fundamentally different in terms of the method by which they are performed and the resultant radiograph. PA X-rays are taken from the back to the front

whereas AP X-rays are taken from the front to the back. AP X-rays are generally not preferred except in scenarios in which the patient is too weak or is unable to assume an erect position (Puddy & Hill, 2007). AP X-rays are much more difficult to read as radiologists must make several adjustments to account for the differing view. For example, AP X-rays tend to return the appearance of mild cardiomegaly (enlargement of the heart) because the X-rays diverge as they pass through the mediastinum, resulting in an overall magnification of the anterior structures of the thorax, among which is the heart.

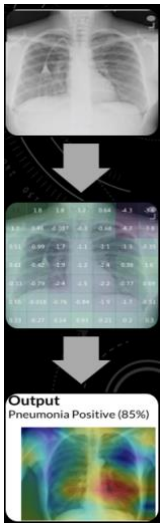
Second, the model must take into account the lateral X-rays as they are significant for at least 15% of diagnoses and often reveal information that a frontal X-ray does not (Ittyachen, Vijayan, & Isac, 2017). Lateral X-rays are especially useful in situations where the frontal X-ray is AP; due to their aforementioned difficulty in interpretation, the lateral view often provides clarification and further detail.

Third, the model would need to also account for patient information, including symptoms and medical history. The importance of these factors in the diagnosis can be highlighted by the example of a patient's chest X-ray showing signs of congested lung vasculature. The radiologist analysing the roentgenogram could potentially diagnose it as an acute illness like multifocal pneumonia; however, if the radiologist knew that the patient has a heart condition and is exhibiting shortness of breath, it would be much more likely that the patient is affected by pulmonary edema (Bhansali, personal communication, March 29, 2019). However, without the knowledge of the patient and the patient's symptoms and history, the radiologist would have provided a misdiagnosis which ultimately could have led to the expiration of the patient.

Several datasets have been released to further the development of machine learning in thoracic radiograph diagnosis. One of the earliest and most significant datasets was the ChestX-ray14 dataset, a large set of thoracic radiographs released by the National Institute of Health. The dataset, which was, at the time of its initial publication, the largest dataset of chest X-rays, contained 112,210 X-ray images in DICOM format from 30,805 patients (Wang, et al., 2017). The dataset was especially notable for the role it played in the development CheXNeXt, a deep learning algorithm designed by researchers belonging to Stanford University's Machine Learning Group. CheXNeXt is a 121-layer convolutional neural network that was trained and validated using the aforementioned ChestX-ray14 dataset (Rajpurkar, et al., 2018). The model

took a frontal X-ray as an input and outputted a vector of disease probabilities and a heat map of where the findings of the radiograph were localised.

Figure 1: A visual representation of CheXNeXt, with the inputted X-ray and outputted heat map and disease probability



The model was tested with a set of 420 chest X-rays, which it diagnosed in 90 seconds; conversely, the four board-certified radiologists against whom CheXNeXt was being compared required approximately four hours. Although CheXNeXt was remarkable in terms of the accuracy it achieved, it was limited in that it did not differentiate between AP and PA X-rays, account for lateral chest X-rays or consider patient information. However, despite its shortcomings, CheXNeXt was revolutionary as it was the first model to conclusively provide evidence for the potential of such a model matching, and occasionally surpassing) the accuracy of radiologists.

Table 1: A comparison of the AUC (area under the receiver operating characteristic curve) achieved by radiologists and CheXNeXt for each pathology

Pathology	Radiologists	CheXNeXt	Advantage
Atelectasis	0.808	0.862	CheXNeXt
Cardiomegaly	0.888	0.831	Radiologists
Consolidation	0.841	0.893	No significant difference
Edema	0.910	0.924	No significant difference
Effusion	0.900	0.901	No significant difference
Emphysema	0.911	0.704	Radiologists
Fibrosis	0.897	0.806	No significant difference
Hernia	0.985	0.851	Radiologists
Infiltration	0.734	0.886	No significant difference
Mass	0.886	0.909	No significant difference
Nodule	0.899	0.894	No significant difference
Pleural thickening	0.779	0.798	No significant difference
Pneumonia	0.823	0.851	No significant difference
Pneumothorax	0.940	0.944	No significant difference

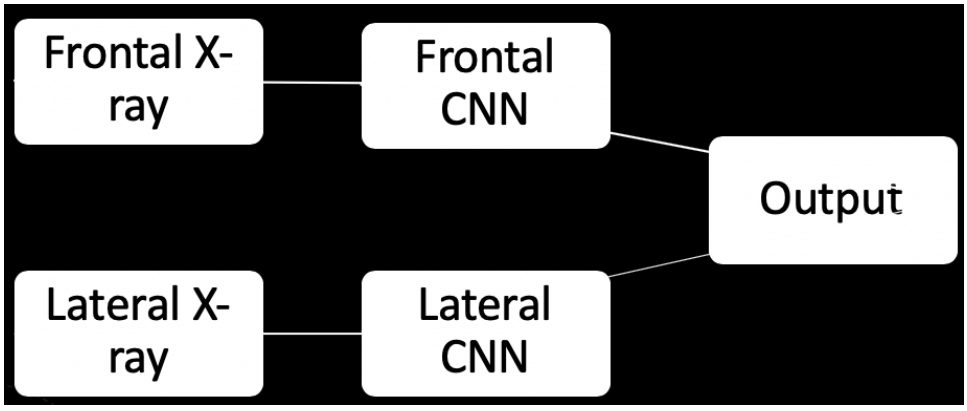
Another prominent dataset is MIMIC-CXR, the largest collection of thoracic radiographs released to date (Johnson, et al., 2019). The dataset contains 371,920 thoracic radiographs with positive and negative labels for the following diseases/findings: no finding[1], enlarged cardiomediastinum, cardiomegaly, airspace opacity, lung lesion, pulmonary edema, consolidation, pneumonia, atelectasis, pneumothorax, pleural effusion, pleural other, fracture and support devices. The DICOM-formatted images were obtained from

[1] The "no finding" label indicates the absence of any of the diseases associated with the other labels

227,943 radiologic imaging studies conducted at the Beth Israel Deaconess Medical Centre and were de-identified using an algorithm that removed dates and potential patient identifiers. Subsequently, the images were labelled with information from their corresponding radiology report using the CheXpert labeller developed by researchers at Stanford University. The dataset has been published for researcher use and is intended to be fully disseminated in the near future. As a prerequisite to gaining access to the data, completing a course in human ethics is mandatory; additionally, the researchers must agree to citing the data in any publication that makes use of it.

A recently published model developed by researchers at Philips Research Institute was able to build upon the research done by Rajpurkar, et al.; in their research, three separate networks were trained for PA, AP and lateral X-ray images (Rubin, et al., 2018). These networks were then paired together for two different models. One model was composed of a PA and lateral network, while the other model was composed of an AP and lateral network. The model operated by accepting two different inputs: a frontal X-ray and a lateral X-ray. Subsequently, the images were independently analysed through separate networks with the final output being a fusion of the outputs of the individual networks. Each network was designed based on the DenseNet-121 architecture. A sigmoid operation was applied to each of the fourteen outputs and the networks were trained using a binary cross-entropy loss function.

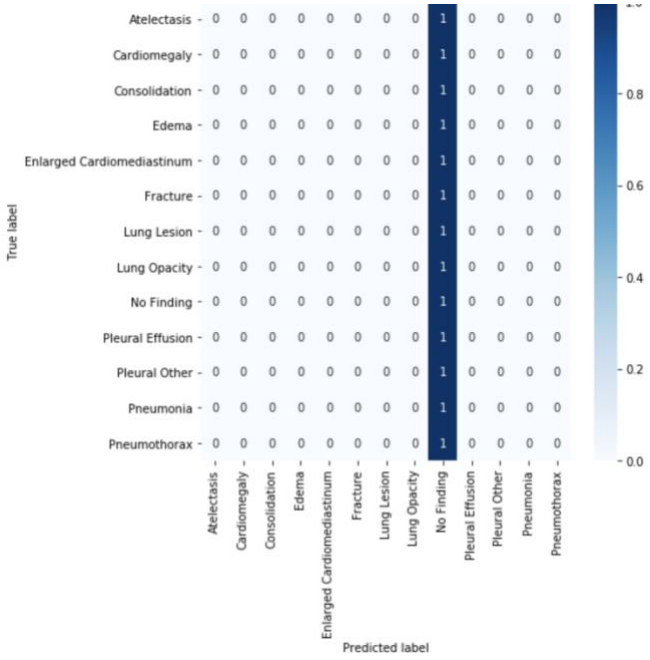
Figure 2: A visual representation of the dual convolutional neural networks



While this model was successful in implementing lateral X-rays, something that prior research had been unable to accomplish, it still did not account for patient information, which would be the focus of future research in this field (Rubin, et al., 2018).

A common issue with multi-class classification models comes when extracting a confusion matrix from the model. This is an issue I ran into myself, as my model consistently predicted that all of the images it encountered belonged in a single class. A confusion matrix showing this tendency is shown in *Figure 3* below.

Figure 3: Single-class prediction confusion matrix



Thus, to work around this problem I decided to embark on this project from a drastically different approach. In order to simplify the task at hand, I decided to reduce the number of classes from fourteen to two; seeing as the most relevant disease classification at this time is COVID-19, I decided to approach this project from a COVID-19 diagnosis perspective.

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has been identified as the virus causing coronavirus disease (COVID-19) (Lai, C. C., Shih, T. P., Ko, W. C., Tang, H. J., & Hsueh, P. R., 2020). It

is thought to have originated in the city of Wuhan, China. The disease has caused the World Health Organization to declare a global public health emergency and officially classify the disease as the cause of a pandemic. Similar to previous coronaviruses, SARS-CoV-2 is originally found in bats, and is generally consistent across different regions. COVID-19 can cause mild to severe illness that if becomes more severe, can cause pneumonia, organ failure, and death. Symptoms of the disease include fever, dry cough, shortness of breath, and fatigue. The disease spreads through respiratory particles which are produced when an infected person either sneezes or coughs. These particles must be taken into the body through the nose, mouth, or eyes. COVID-19 has spread globally to affect at least 2 million people in 184 countries at the time of publication. The worldwide extent of COVID-19 has resulted in over 130,000 deaths. Those at risk belong primarily to older demographics with historically compromised immune systems or pre-existing medical conditions such as asthma, diabetes, and heart disease, as well as those with compromised lungs.

Although many countries began preparations for the spread of the disease relatively early, the effects of the pandemic are still widespread and proliferating. In many parts of the world, a massive shortage of Reverse Transcription Polymerase Chain Reaction (RT-PCR) tests is contributing to an inability to properly combat the pandemic. One way in which this problem could be alleviated is through the use of X-ray scans for diagnosis. In X-rays cans, COVID-19 manifests itself as ground-glass opacities (GGOs) and consolidation with or without vascular enlargement, interlobular septal thickening, and air bronchogram sign. X-ray scans are much faster and more accessible than RT-PCR tests, allowing for easier and more quickly acting diagnostics.

More importantly, using X-ray scans has been shown to be an effective and accurate method for diagnosing COVID-19. RT-PCR tests, in addition to being scarce, take a long time to return results, and even then are prone to frequent false negatives; this results in many people with COVID-19 going undiagnosed. Unfortunately, as with many other medical classification problems, there is a lack of publicly available COVID-19 X-ray scans. A lack of training data can significantly restrict the performance of deep neural models and lead to overfitting. One potential solution that has been explored

to this problem is the use of generic data augmentation techniques. Generic data augmentation involves manipulating the original images in the dataset through various methods of cropping, rotating, and zooming, in order to artificially grow the dataset while preserving the distinguishing features that are present in the images. Generic data augmentation has been shown to be especially useful in fine-grained datasets, or datasets that have low sample sizes and high degrees of similarities between images. For example, extensive use of rotating training images to increase CNN task performance for galaxy morphology classification using a fine-grained data-set [Dieleman *et al.*, 2015]. The primary concern in the application of data augmentation is that of over-fitting occurring. Thus, it is important that studies which employ data augmentation analyze differences in training and validation metrics over the runtime in order to rule out overfitting. In our study, we utilized multiple rotations in order to artificially grow our fine-grained dataset due to their success in previous classifier models (Dieleman et al., 2015) (Taylor and Nitchschke, 2017).

In mathematical terms, the Laplace filter is a filter that is defined by the divergence of a scalar field's gradient. In image processing, it is used for enhancing an image's edges to help in its detection. Because derivative filters, among them the Laplace filter, are sensitive to noisy images, they are often performed in association with a smoothing filter to remove noise. The filter operates by calculating a sum of differences across multiple neighboring pixels to replace the magnitude of each individual pixel and is defined as $Laplace(f) = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}$. This is how it effectively locates edges and simultaneously removes noise from images. In the presence of a bright spot located in a dark region of the image, the Laplace filter will return an even brighter spot to highlight the disparity. The Laplace filter will be utilized in Image Preprocessing since it has shown the ability to improve the recall rate of CNNs when tested on X-ray images (Chen, 2019).

A model with the capability to instantaneously and accurately diagnose X-rays would have immense benefits in health-care environments as they would drastically reduce the number of diagnostic-related deaths, lower health-care costs and decrease the amount of time needed by radiologists to analyse the

radiographs. Additionally, areas with minimal medical staff would greatly benefit from a model that effectively replaces radiologists.

References

A. Krizhevsky, I., S. Nitish, H., W. Karl, M., Shao, L., J. Yosinski, J., D. Erhan, Y., . . . DJ. Drown, T. (1970, January 01). A survey on Image Data Augmentation for Deep Learning. Retrieved July 29, 2020, from https://link.springer.com/article/10.1186/s40537-019-0197-0

Chen, X. (n.d.). Image enhancement effect on the performance of convolutional neural networks. Retrieved from http://www.diva-portal.org/smash/get/diva2:1341096/FULLTEXT02.pdf

Data & Statistics | TB | CDC. (n.d.). Retrieved from https://www.cdc.gov/tb/statistics/default.htm

Hall, L., Goldgof, D., Paul, R., & Goldgof, G. M. (2020). Finding COVID-19 from Chest X-rays using Deep Learning on a Small Dataset. doi: 10.36227/techrxiv.12083964

Holshue, M. L., Doremalen, N. van, Vaduganathan, M., Fineberg, H. V., & Epidemic Intelligence Service. (2020, March 17). First Case of 2019 Novel Coronavirus in the United States: NEJM. Retrieved from https://www.nejm.org/doi/full/10.1056/NEJMoa2001191

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L., & Aerts, H. (2018, August). Artificial intelligence in radiology. Retrieved July 29, 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6268174/

Hurt, B., Kligerman, S., & Hsiao, A. (2020). Deep Learning Localization of Pneumonia. Journal of Thoracic Imaging , 1. doi: 10.1097/rti.0000000000000512

Ittyachen, A. M., Vijayan, A., & Isac, M. (2017). The forgotten view: Chest X-ray - Lateral view. Respiratory Medicine Case Reports, 22, 257-259. doi:10.1016/j.rmcr.2017.09.009

Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C., . . . Horng, S. (2019). MIMIC-CXR: A Large Publicly Available Database of Labeled Chest Radiographs. Retrieved from https://arxiv.org/abs/1901.07042.

Lai, C. C., Shih, T. P., Ko, W. C., Tang, H. J., & Hsueh, P. R. (2020). Severe acute respiratory syn drome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges. International journal of antimicrobial agents, 55(3), 105924. https://doi.org/10.1016/j.ijantimicag.2020.105924

Litjens, G., Kooi, T., Bejnordi, B., Setio, A., Ciompi, F., Ghafoorian, M., . . . Sánchez, C. (2017, July 26). A survey on deep learning in medical image analysis. Retrieved July 29, 2020, from https://www.sciencedirect.com/science/article/abs/pii/S1361841517301135

"Lung Cancer Is the Biggest Cancer Killer in Both Men and Women" Infographic | CDC. (n.d.). Retrieved from https://www.cdc.gov/cancer/lung/basic_info/mortality-infographic.htm

Pneumonia Can Be Prevented-Vaccines Can Help | CDC. (n.d.). Retrieved from https://www.cdc.gov/pneumonia/prevention.html

Puddy, E., & Hill, C. (2007). Interpretation of the chest radiograph. Continuing Education in Anaesthesia Critical Care & Pain, 7(3), 71-75. doi:10.1093/bjaceaccp/mkm014

Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., . . . Lungren, M. P. (2018). Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLOS Medicine, 15(11). doi:10.1371/journal.pmed.1002686

Raoof, S., Feigin, D., Sung, A., Raoof, S., Irugulpati, L., & Rosenow, E. C. (2012). Interpretation of Plain Chest Roentgenogram. Chest, 141(2), 545-558. doi:10.1378/chest.10-1302

Rubin, J., Sanghavi, D., Zhao, C., Lee, K., Qadir, A., & Xu-Wilson, M. (2018). Large Scale Automated Reading of Frontal and Lateral Chest X-Rays using Dual Convolutional Neural Networks. Retrieved from https://arxiv.org/abs/1804.07839.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (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 (CVPR). doi:10.1109/cvpr.2017.369

Zhao, J., Zhang, Y., He, X., & Xie, P. (n.d.). COVID-CT-Dataset: A CT Scan Dataset

about COVID-19. Retrieved from https://arxiv.org/pdf/2003.13865.pdf