An Analysis of Deep Convolutional Neural Network Models to Classify Chest X-Rays: A Varied Approach to Catalogue Model Efficacy Factors

Maxwell Klaiman1,2

1. Northwestern University School of Professional Studies

Master’s of Science in Data Science Program

1. Church St, Evanston, IL 60201

2. Allina Health System

2925 Chicago Avenue, Minneapolis, MN 55407

Address to which correspondence should be addressed:

MaxwellKlaiman2023@u.northwestern.edu

**Abstract**

Chest X-ray Radiography (CXR) is one of the most prominent radiographic technique, with nearly 100 x-rays occurring annually per 1000 patients in the United States (Winder, 2021). While CXR is an inexpensive technique on its own, the inconclusive nature of x-rays often leads to more extensive imaging, such as computed tomography (CT) scans being used to confirm or make diagnoses. This dual imaging results in higher costs to patients and facilities, as well as occasionally unnecessary exposure to x-rays. To improve this problem, the NIH Clinical Center released de-identified images of over 30,000 patients, for analysis and improvement of diagnostic techniques using AI. A deep Convolutional Neural Network (DCNN) can be trained using the provided dataset, and compared to existing diagnostic models such as AlexNet. Additionally, investigation of different DCNN model features can answer relevant questions surrounding complexity, efficiency, and architecture. Specifically, modulation of model architecture to determine the impact of adding or removing convolutional layers, or fully-connected layers, is a key feature of the conducted variational analysis.

**Keywords:** Chest X-ray, Radiography, Artificial Intelligence, Deep Learning, Neural Networks, Convolutional Neural Networks, Diagnostics.

1. **Introduction**

Chest Radiography, more commonly known as Chest X-ray (CXR), is a key medical diagnosis tool, used in around one in every ten seen patients across western healthcare (Winder, 2021). As CXR is a relatively inexpensive treatment, it has long been an aspect of standard workup relating to pulmonary concerns (Spreets, 2006). While CXR is often favored due to its relative affordability, one primary issue is the lack of diagnostic clarity present in chest x-ray scans. This lack of diagnostic power has lead health systems to lean heavily on a more expensive alternative – Computed Tomography (CT) scans – to diagnose conditions, if not confirm CXR diagnoses. According to Winder, since 2010, the hospital in question ordered nearly double the amount of CT scans in 2019 (pre-COVID-19) as in 2010, while the rate of CXR has not changed. This leads to a combination of higher costs to run imaging departments, unnecessary exposure to X-rays, and ultimately, a higher expense to patients.

To address the issues around CXR accuracy, the National Institutes of Health (NIH) compiled and released scans from over 30,000 patients (over 100,000 scans in total), with 14 complex outcomes possible in these cases, called CXR8 (CXR8, 2017). Wang, et al. (2017) performed labelling and superficial classification of the dataset, using a combination of deep Convolutional Neural Networks (CNNs) and natural language processing techniques. The applicability of a DCNN-based technique is clear in this case, but only limited use of the CXR8 dataset has been done to date.

1. **Literature Review**

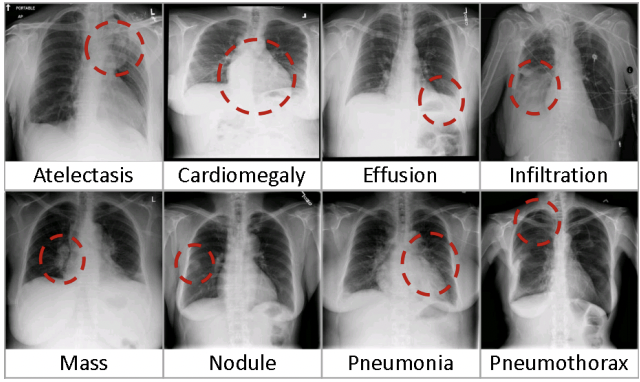
The use of CNNs in medical imaging is widespread, and the use of deep CNNs (DCNNs) to accomplish diagnostic tasks has been widely studied. One of the most widely-accessed papers, written by Tang, et al. (2020) in Nature, details the potential of DCNNs to automate chest radiography readings. Their research, which sought to assess the effectiveness of a variety of models in determining whether a CXR is normal or abnormal. Five models were trained and tested on a similar NIH dataset (NIH Chest X-ray 14), and each performed well, with AUC scores > 0.96.

From their work, Bressem, et al., (2020) did a more direct comparison of deep learning architectures. Inception v4, DenseNet, and AlexNet were evaluated, and, perhaps counterintuitively, shallow networks (i.e. AlexNet, ResNet-34, or VGG-16) outperformed deeper CNNs in this study. Their final conclusion posits that more complex models are not required for state-of-the-art results. This result is somewhat in conflict with both previous, highly-esteemed results, and what may seem logical – that a less nuanced modeling method can produce similar or even superior results.

Key in making the assertion that more complex models are able to produce results of the highest order, Krizhevsky, Sutskever, and Hinton (2012) produced a markedly high-accuracy model in the ImageNet Classification Challenge, which comprised 60 million parameters and 650,000 neurons. Logically, the finding that such a complex CNN was able to outperform simpler varieties is reasonable. In synthesizing the two conflicting conclusions, models of advanced complexity may not be required, but likely play a valuable role in making marginal improvements in the space. Ian Goodfellow, Yoshua Bengio, and Aaron Courville, in their book *Deep Learning* (2016) note that a more complex CNN will likely have an analogous drain on resources (often a greater runtime), and that simpler, more efficient, alternatives with comparable performance are preferable to highly complex models.

1. **Data**

Data for model construction is freely-available through the NIH Clinical Center (NIHCC) Box. The primary data is 108948 images, with accompanying metadata. The images are pre-labeled, with accuracy >90% (Per the NIHCC). While this accuracy does pose a possible challenge in the rare cases where a label may be inaccurate, model performance over more than 100,000 images should be unchanged. A set of example images provided by the NIHCC, detailing 8 of the 14 possible anomalies, is below:



The metadata provided alongside the images contains some basic demographic information. That includes: Image ID, Findings, Patient ID, Follow-up Number (for repeat imaging), Age, Gender, Image orientation, and information about the image size.

Of crucial importance is the preprocessing of the dataset images. To begin, the provided images must first be processed, and an edge must be added to each image. As each provided image may not be of the exact same size, an added challenge will be normalizing all images. Additionally, since images may not be viewed from an identical angle (options are AP and PA), further processing is required to ensure that like comparisons are being drawn between images. Finally, images must be paired to their metadata to ensure proper modeling occurs.

1. **Methods**

To develop an effective CNN, considerations must be made regarding each layer of set-up. Almezhghwi, et al. (2021) investigates the framework for popular AlexNet and VGGNet-16 methods in interpreting CXR images. In both cases, feature extraction occurs over a series of convolutional layers (5 for AlexNet, 12 for VGGNet-16). From there, extracted features pass into two fully connected layers. From there, the authors passed the output to an SVM classifier, which output the result in question. Interestingly, the authors posit that an SVM would outperform a Soft-max function for output, so investigation into this claim would be valuable as well.

In approaching a modeling method for these CXRs, it would be prudent to consider the highly prevalent AlexNet and VGGNet-16 features. From there, comparisons can be made to determine the most effective set-up. From an initial testing perspective, a CNN will be setup in Keras which passes data through 6 convolutional layers for feature extraction. From there, two fully connected layers will be developed to connect the extracted features to outcomes. To consider the validity of the claim made by Almezhgwhi, et al., both a Soft-max and SVM classifier will be used, and the outcomes will be compared.

Additionally, in attenuating the developed CNN, a couple areas will be investigated further. First, an analogous model with additional convolutional layers (9 or 12) will be developed. Additionally, a model with 3 or 4 layers will be developed to investigate the tension between efficiency and effectiveness present in simpler CNNs. Finally, a proper investigation into the value of fully connected layers, and the impact of adding or removing from the initial two would be worthwhile.

1. **Results**

With a dataset exceeding 100,000 images, the computational burden on a single laptop is both extraordinary and limiting. For that reason, all models were evaluated on an identical slice of the entire NIHCC CXR dataset. A total of four Convolutional Neural Networks were developed: one AlexNet CNN, which serves as a control, and three experimental CNNs derived from the AlexNet architecture. Each experimental CNN had its own modified architecture, to further investigate the impact of additional fully-connected or convolutional layers within the CNN architecture. Experimental CNN 1 had the same setup as the AlexNet control, except with only one fully-connected layer (versus 3). Experiments two and three both had three fully-connected layers, but with 10 and 3 convolutional layers, respectively. A table summarizing some key results is below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model Name | Number of Convolutional Layers | Number of Fully-Connected Layers | Training Epochs | Optimal Validation Accuracy | Test Accuracy |
| AlexNet CNN | 6 | 3 | 50 | 0.2157 | 0.1429 |
| Experimental Model 1 | 6 | 1 | 50 | 0.2353 | 0.1400 |
| Experimental Model 2 | 10 | 3 | 50 | 0.2157 | 0.0800 |
| Experimental Model 3 | 3 | 3 | 50 | 0.2157 | 0.1114 |

Some additional information, such as number of trainable parameters and overall training time, is included in Table 2 below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model Name | Total Parameters | Trainable Parameters | Total Training Time | Final Model Loss | Final Model Accuracy |
| AlexNet CNN | 58,327,818 | 58,325,066 | 1,134s | 4333908480 | 0.1429 |
| Experimental Model 1 | 4,008,938 | 4,005,994 | 872s | 27.2410 | 0.1400 |
| Experimental Model 2 | 53,301,546 | 53,292,842 | 3,454s | 1177238144 | 0.0800 |
| Experimental Model 3 | 84,152,554 | 84,151,658 | 949s | 6380248064 | 0.1114 |

1. **Analysis and Interpretation**

In deeper investigation of each of the models, and comparing them with the AlexNet CNN control, a couple of interesting trends arise. While the correlation between model complexity and model training time may be obvious, each of the models confirmed that a less complex model (one with fewer layers, not necessarily with fewer trainable or total parameters) is able to be trained in a shorter timeframe. With a complete dataset spanning over 110,000 images, the value of a simpler model that does not compromise on accuracy is substantial.

However, serious analysis reveals additional, valuable information about the structure and functionality of CNNs, analyzing smaller-than-optimal samples of highly complex data. First, it is worth noting that training models using a categorical cross-entropy loss function results in very high loss values, which only increases across each epoch. Additionally, with a training dataset of only 400 images, experimental CNNs were able to accomplish somewhat respectable classification accuracies, surpassing 21% in each case (11/51). Presumably, over the course of the entire CXR dataset, these accuracies could grow substantially.

Questions around the impact of additional or fewer CNN layers, or decreasing fully-connected layers yielded interesting results. In general, simpler models had slightly diminished performance when compared to their more complex counterparts. When looking at experimental model one, final accuracy was about 2.1% lower than the control group, despite having a slightly larger optimal validation accuracy (12/51 vs. 11/51). This additional validation accuracy, however, was only a transient showing, and it is safe to say that the simpler experimental model one was outperformed by the AlexNet Control.

Comparing experiments two and three to the AlexNet control yields somewhat surprising results. In both cases, optimal validation accuracy was not exceeded, and final accuracy was not matched. This proposes a bit of a quandary: both a simpler and more complex model does not outperform the AlexNet control. However, some of the models are able to produce substantially lower losses, which may be solely a function of the continuous cross-entropy loss function used.

1. **Conclusions**

There are several takeaways that can be concluded from this high-level analysis. First, there is the potential for further substantiation of the work by conducting a more advanced (and computationally challenging) analysis of the CXR-8 Dataset. From a superficial perspective, this experiment serves to confirm the effectiveness of AlexNet CNNs for classification tasks of this nature. However, it also raises a question around the areas in which improvements can be made on this highly prevalent DCNN methodology. Perhaps, a more complex model which takes in all 110,000 images would outperform the elegant but fairly simple AlexNet model. However, that modeling would require run-times in the order of days to weeks, and is undoubtedly a limiting factor.

Perhaps of additional value is the investigation of less complex modeling techniques, reflected by the first experimental model analyzed. This analysis, which was able to produce (1) lower loss, (2) shorter runtime, and (3) higher optimal accuracy, showed conclusively that the removal of two fully-connected layers did not negatively impact the model’s performance. While complex modeling architecture provides a basis for the highest-performing models both logically and historically, the value of a competitive or slightly worse model that requires substantially fewer resources cannot be overstated. In future analyses, the importance and relative value of developing low-cost models (that is to say, models requiring less computational resources) must be considered.

1. **Directions for Future Work**

Looking towards future work, there are a couple developments that could both further inform the optimal use of CNNs to categorize CXRs, and directions from which deeper analysis can be conducted. As a first step, gathering more substantial computational resources, which would allow for the full analysis of the NIHCC CXR-8 Dataset would be the most logical progression from current methods. While the inferences drawn from conducted experiments hold water in this case, it is not necessarily guaranteed that the same generalities will hold true as more complex models are developed and trained. Additionally, a full-scale implementation of the most effective models on an outside dataset would be a valuable and informative experiment, which could possibly validate the generalizability of CXR interpretation models to outside sources (i.e. specific hospitals). Finally, more substantial hyperparameter modification experiments could be developed, with an eye towards optimizing the performance of existing models in a more schematized manner. Undoubtedly, a comprehensive analysis of hyperparameter options and their corresponding outcomes would lend additional strength to existing experimental models.

1. **References**

Almezhghwi, Khaled, Sertan Serte, and Fadi Al-Turjman. 2021. “Convolutional Neural Networks for the Classification of Chest X-Rays in the IoT Era.” *Multimedia Tools and Applications* 80 (19): 29051–65. https://doi.org/10.1007/s11042-021-10907-y.

Bressem, Keno K., Lisa C. Adams, Christoph Erxleben, Bernd Hamm, Stefan M. Niehues, and Janis L. Vahldiek. 2020. “Comparing Different Deep Learning Architectures for Classification of Chest Radiographs.” *Scientific Reports* 10 (1). https://doi.org/10.1038/s41598-020-70479-z.

“CXR8 Data Set” 2017. NIH Clinical Center. Nihcc.app.box.com. https://nihcc.app.box.com/v/ChestXray-NIHCC/folder/36938765345.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. Cambridge, Massachusetts: The Mit Press. https://www.deeplearningbook.org/.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. “ImageNet Classification with Deep Convolutional Neural Networks.” *Communications of the ACM* 60 (6): 84–90. https://doi.org/10.1145/3065386.

“NIH Clinical Center Provides One of the Largest Publicly Available Chest X-Ray Datasets to Scientific Community.” 2017. National Institutes of Health (NIH). September 27, 2017. https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community.

Speets, Anouk. M., Yolanda van der Graaf, Arno W. Hoes, Sandra Kalmijn, Alfred P.E. Sachs, Matthieu J.C.M Rutten, Jan Willem C. Gratama, Alexander D. Montauban van Swijndregt, and Willem P. Th. M. Mali. (2006). “Chest radiography in general practice: indications, diagnostic yield and consequences for patient management.” *The British journal of general practice: the journal of the Royal College of General Practitioners* 56 (529): 574–578. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1874520/

Tang, Yu-Xing, You-Bao Tang, Yifan Peng, Ke Yan, Mohammadhadi Bagheri, Bernadette A. Redd, Catherine J. Brandon, et al. 2020. “Automated Abnormality Classification of Chest Radiographs Using Deep Convolutional Neural Networks.” *Npj Digital Medicine* 3 (1). https://doi.org/10.1038/s41746-020-0273-z.

Wang, Xiaosong, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 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)*, July. https://doi.org/10.1109/cvpr.2017.369.

Winder, Mateusz, Aleksander Jerzy Owczarek, Jerzy Chudek, Joanna Pilch-Kowalczyk, and Jan Baron. 2021. “Are We Overdoing It? Changes in Diagnostic Imaging Workload during the Years 2010–2020 Including the Impact of the SARS-CoV-2 Pandemic.” *Healthcare* 9 (11): 1557. https://doi.org/10.3390/healthcare9111557.

‌