A Nominal Deep Learning Approach to Classifying Disease in Chest X-rays

Zachary Seitz
CS505 – Introduction to Data Mining
Dr. Yuan
Southeast Missouri State University
12 December 2022

Abstract

Chest X-rays are a relatively cheap and reasonably effective tool for diagnosing various diseases of the heart and lungs, the combination of which represents the leading causes of death worldwide. Early detection of these types of diseases is crucial to their subsequent treatment and can dramatically improve patient outcomes. Recent advances in deep learning beg the question of how well a deep learning model is suited to the task of detecting diseases from medical images. This paper presents a method for exploring that problem space by way of training a deep neural network using various hyperparameters to detect three prevalent lung diseases from chest x-rays. The results of the experimental methodology indicate that it is certainly possible to find a set of hyperparameters that show great promise for the potentially lifesaving applications in the future.

Project repository: https://github.com/zachmichael14/data_mining_final

1. Introduction

1a. Background

The field of medical imaging has been around since 1895 when German scientist Wilhelm Röentgen discovered the X-ray [1]. While imaging technology has come a long way since the days of Röentgen, X-rays remain commonplace in medical diagnostics. Chest X-rays (CXRs) in particular are the most common radiologic procedure since they are capable of aiding medical professionals in the diagnosis of several types of diseases, including diseases of the lung and heart [2].

The problems of using CXRs in thoracic disease diagnosis are twofold. The first problem is that current thoracic disease detection methodologies (including CXRs technology) are insufficient at addressing the current need. Diseases of the lungs and heart remain the leading causes of death worldwide. Furthermore, diseases like malignant neoplasms (the second leading cause of death nationwide) may also occur in the lungs and heart, which further complicates treatment [3]. Despite this, CXRs are relatively cheap and quick diagnostic tools and thus have great potential to be useful in that space, since clearly there exists a persistent need for detection of these types of diseases [4]. The second problem with using CXRs is the shortage of medical professionals capable of effectively interpreting the results of a CXR [5]. In fact, many professionals in radiology are counting on the advent of effective AI-based tools to help manage their increased workloads [6].

1b. Motivation

Deep learning tools are uniquely positioned to solve both aforementioned issues. A well-trained deep learning algorithm could identify patterns of disease where a trained medical professional may not be able to, thus allowing for earlier detection. It's well known that early detection can dramatically improve patient outcomes [7], [8]. Furthermore, computer-assisted diagnostic (CADx) tools can decrease the workload of medical professionals, allowing them to spend more time on complex cases that require the domain knowledge of an expert [6]. The proposed methodology here serves to nominally explore the creation of model that addresses those problems by analyzing the effects of various hyperparameters on the efficacy of the model to detect disease from a given CXR.

This paper is organized into the following format. Section two presents a brief literature survey of the current landscape of medical image analysis using machine learning. Section three describes the methodology for developing a deep learning model to extract clinically relevant features from CXRs. Section four presents an analysis of the proposed methodology, including its limitations. Section five concludes this paper and suggests avenues for future research.

2. Related Work

There are several precedents for using deep learning models in medical diagnostics. For example, deep learning has shown great promise in assisting the field of medical image analysis in particular [9]. Nishtar et. al. showed the efficacy of using computer aided detection for tuberculosis (CAD4TB) on populations in Pakistan [10]. [10] states that this implementation lowers costs associated with diagnosing TB, which is especially important in developing areas, where there are more limited resources for detecting and treating such diseases.

However, using deep learning for medical image analysis comes with its own set of obstacles. For instance, medical image datasets tend to be much smaller. To combat this issue, [11] used transfer learning to fine-tune a model to detect pneumonia from CXRs to a clinically significant degree of accuracy. Still, other methods are available to increase the efficacy of deep learning models as applied to medical image analysis and diagnostics, some of which are outlined by [9]. Clearly, there is important and meaningful work being done in this space.

3. Materials and Methods

The performance restrictions of the experimental environment necessitated certain choices in the experimental design. In order to facilitate the acquisition of complete and meaningful results, both the scale and the scope of the experiment were necessarily small. The following section briefly outlines the experimental design and the reasoning upon which it is based.

3a. Original Dataset

The experimental dataset used is a subset of the dataset used by [12]. The original dataset contains over 112,000 CXRs for more than 31,000 patients. Each image is 1024 x 1024 pixels and is labeled with one or more *findings*, where a finding is either *no finding* (representing a healthy patient) or some form of medical condition (pneumonia, hernia, atelectasis, etc.). The dataset also

contains the sex and age for each patient, among other information. There are 15 unique findings (including *no finding*) and 836 unique combinations of these findings throughout the dataset.

Because the amount of information contained in the original dataset is large and multi-dimensional, a smaller subset was created for the purposes of this experiment. In this way, the data could be simplified enough to avoid running out memory or otherwise halting the training process.

3b. Data cleaning and pre-processing

The original dataset was pared down to a subset of 18,431 images. Down sampling was applied to the "No Finding" and "Infiltration" records wherein only 9% and 50% were maintained, respectively. All images with a label of more than one finding were also removed. In a further attempt to decrease the dimensionality of the data, only those findings with more than 3,000 images were kept.

The final subset was made up of four roughly equal classes (findings). The remaining classes were *no finding* (30%), *infiltration* (26%), *effusion* (21%), and *atelectasis* (23%). From that subset, 1,500 images were randomly chosen and split into experimental sets with 900 images (60%) in the training set and 300 images (20%) in each of the testing and validation sets. In order to increase the quantity of images that could be used during training, the size of each image had to be smaller, so the images were resized to 256 x 256 pixels.

3c. Model architecture

The model's architecture was chosen with regard to the nature of data and the restrictions of the experimental design. For example, since the size of each image was so heavily reduced, the amount of detail that could be drawn from each image was also reduced. As such, the model had to be simple enough to be performant, but complex enough to attempt to draw meaningful conclusions.

Since it was anticipated that pixels in an image that are closer to one another are likely to be more meaningful to classification than pixels with more distance between them, the model's architecture contains two convolutional layers of 32 and 64 units followed by three fully connected layers containing 64, 32, and 4 units (Table 1).

Table 1. The experimental model architecture. All models were trained using this architecture with the notable exception of the 'no dropout' trial, in which only the dropout layers were removed.

Layer	Output Shape	Trainable Parameters
InputLayer	(256, 256, 3)	0
Conv2D	(254, 254, 32)	896
MaxPooling2D	(127, 127, 32)	0
Dropout	(127, 127, 32)	0
Conv2D	(125, 125, 64)	18,496
MaxPooling2D	(62, 62, 64)	0
Flatten	246,016	0
Dense	64	15,745,088
Dropout	64	0
Dense	32	2,080
Dense	4	132
	Total trainable parameters	15,766,692

3d. Model training

The model was trained using different values for the following hyperparameters: learning rate, batch size, and dropout rate. An early stop monitor was employed to halt training if validation loss failed to improve over 20 epochs. As such, many of the experimental models were trained for different lengths of time. The notable exception to the model architecture is the 'no dropout' trial, in which the dropout layers were removed.

4. Results and Discussion

4a. Learning rate

To determine a favorable learning rate (*lr*) for, one model was trained with a learning rate of $1x10^{-4}$ and another with a learning rate of $1x10^{-5}$. Since the validation loss of the former model decreased much more slowly than that of the latter model, using a learning rate of $1x10^{-5}$ yielded more favorable results (Figure 1).

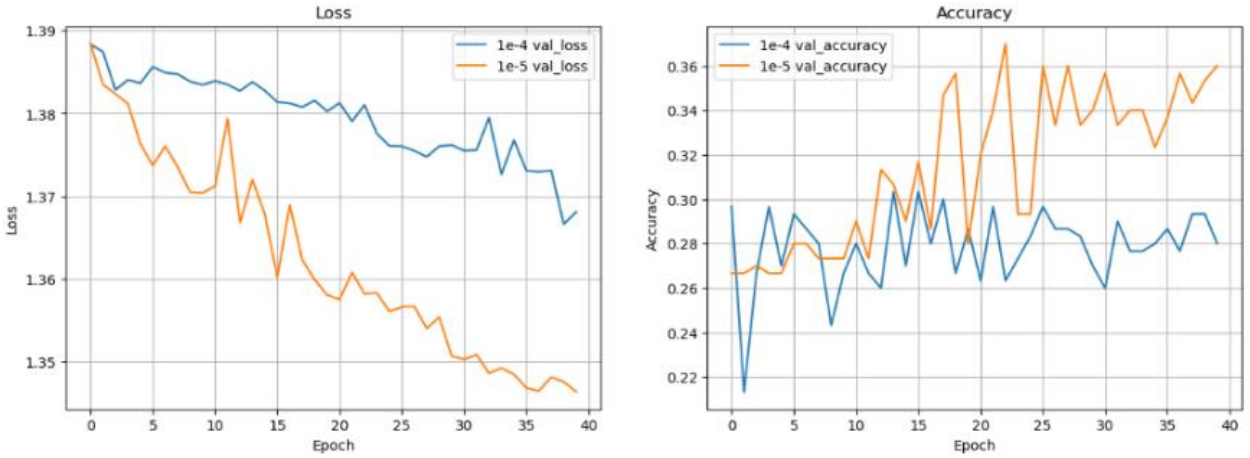

Figure 1: Loss and accuracy plotted for each epoch for the validation set using the experimental model architecture. One model was trained with a learning rate of $1x10^{-4}$ (blue) and the other with a learning rate of $1x10^{-5}$ (orange).

4b. Batch size

Two separate batch sizes were tested, 64 and 32 (learning rate = $1x10^{-5}$); using a batch size of 64 appeared to decrease validation loss more than a batch size of 32, indicating that a batch size of 64 is more favorable in this instance (Figure 2).

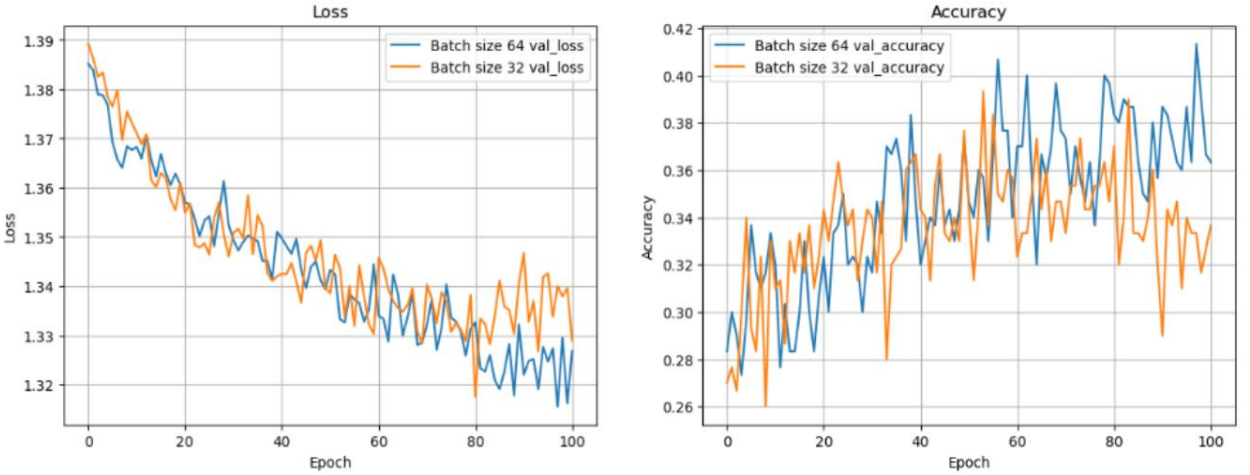

Figure 2: Loss and accuracy plotted for each epoch for the validation set using the experimental model architecture. One model was trained with a batch size of 32 (orange) and the other with a batch size of 64 (blue).

4c. Dropout

The employment of dropout with a drop probability of 25% appeared to slightly decrease the chance of overfitting compared to no dropout, the validation loss of which seemed to increase from epochs 60-80, potentially indicating overfitting (Figure 3).

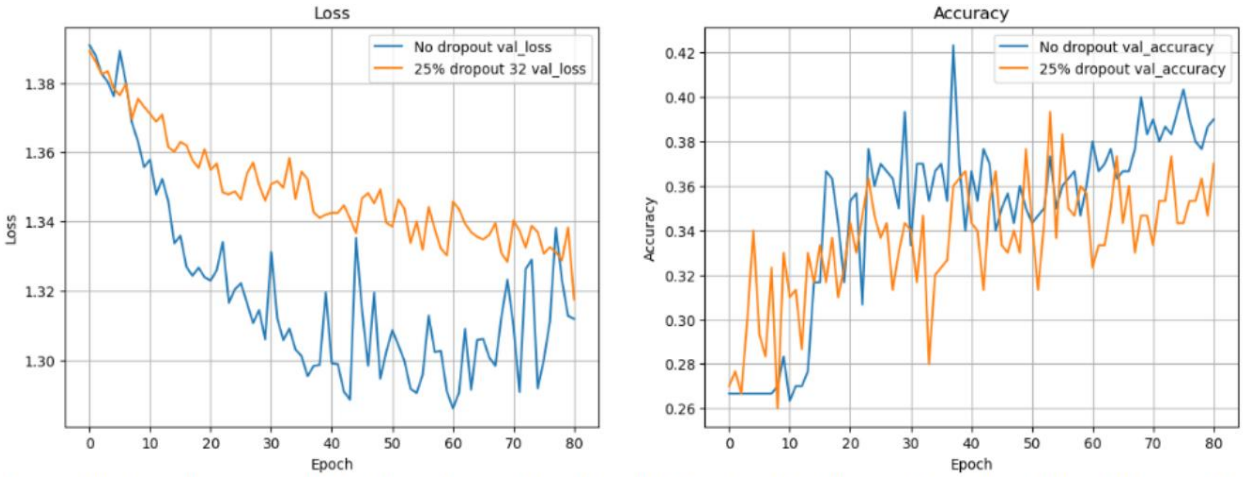

Figure 3: Loss and accuracy plotted for each epoch for the validation set using the experimental model architecture. One model was trained with a dropout rate of 25% (orange) for both the convolution portion and the fully-connected portion. The other model was trained without dropout (blue).

Ultimately, the best possible model used a learning rate $1x10^{-5}$, a batch size of 64, and a dropout rate of 25%, attaining a minimum validation loss 1.32 of and a maximum validation accuracy of 0.41.

4a. Limitations

Though the model yielded promising results, it is certainly not ready for deployment in a clinical setting, as it currently lacks both the specificity and the sensitivity required for such a task. The model is only capable of detecting three types of findings, and without further training, does so less accurately than a coin flip. It also pays no attention to the demographic data of the patient, some of which is likely to be pertinent to the diagnosis. Additionally, the performance of the model on images of other sizes and formats remains unknown.

There are also important clinical details neglected by this methodology. For example, the finding *infiltration* can have many different causes and thus require different treatments [13]. It's possible that, for a given patient, an earlier visit was labeled *infiltration* and then diagnosis was

clarified to a specific cause during a follow-up visit. These details, necessarily neglected by the proposed methodology, likely provide for important clues for feature determination.

5. Conclusions

Despite the limitations of experimental model, it does appear possible to train a model capable of finding features that are diagnostically significant. It further demonstrates the promise and importance of this area of research, as creating more effective tools to detect diseases at an early stage could save many lives. Having said that, the experimental methods certainly need improvements if they are to generalize well in a clinical setting. Some suggestions for future research are outlined below.

5a. Suggestions

The training and use of a multi label classifier would likely yield more clinically significant results. Many of the data points included in the original dataset are labeled with multiple findings, which not only increases the size of useful training data, but it also creates a model better capable of generalizing on previously unseen data. Such an approach may also incorporate patient demographic data as input, which could help classification performance.

The significance of the follow-up data for each patient is also an intriguing area to explore, and it is especially important when aiming for early detection. For example, if a new finding is discovered during a follow-up visit, there may have been indicators of the finding in the visits leading up to the diagnosis; that is precisely the type of task deep learning classifier are great at solving.

It’s worth noting that there are plenty of other aspects of this problem that lend themselves well to exploration and improvement. Though the level of confidence in this particular model does not meet the strict scrutiny required for deployment in a clinical setting, the methods outlined previously, with improvements, could nevertheless lead to the development of tools that dramatically improve outcomes for patients.

References

[1] W.G. Bradley, “History of medical imaging,” *Proceedings of the American Philosophical Society*, vol. 152, no. 3, pp. 349-361, Sept. 2008.

[2] M. Bretz. “Chest X-rays and children.” Cincinnati Children’s. https://www.cincinnatichildrens.org/health/c/chest-x-ray (accessed Dec. 10, 2022).

[3] “Underlying Cause of Death, 1999-2020 Results,” Centers for Disease Control and Prevention, National Center for Health Statistics, Atlanta, GA, USA, https://wonder.cdc.gov/controller/saved/D76/D266F024 (accessed Dec. 10, 2022).

[4] K. Ziegler, J.M. Feeney, C. Desai, D. Shapiro, W.T. Marshall, and M.Twohig, “Retrospective review of the use and costs of routine chest x rays in a trauma setting,” *Journal of Trauma Management & Outcomes,* vol. 7, no. 1, pp. 1-6, 2013, doi: 10.1186/1752-2897-7-2.

[5] W.A.A. Omar, R.M. Al-Shahrani, M.A. Almushafi, and H.M. Boraie, “Factors affecting patients’ waiting time at the Radiology Department,” *Middle East Journal of Family Medicine,* vol. 20, no. 11, pp. 62-68, Nov. 2022, doi: 10.5742/MEWFM.2022.95251376.

[6] “The global future of imaging,” British Institute of Radiology, London, UK https://bir.org.uk/media/408496/the_global_future_of_imaging_a4_24pp-hr.compressed.pdf (accessed Dec. 10, 2022).

[7] N. Hardikar, “The importance of early detection of ovarian cancer: Epidemiology and risk factors,” *Basic & Clinical Cancer Research*, vol. 13 no. 3, pp. 1-9, 2021.

[8] J. Yuen, V. Varadarajan, M. Stavrakas, S. Muquit, and H. Khalil, “A case of invasive sinonasal carcinosarcoma: The importance of early detection,” *Case Reports in Otolaryngology*, vol. 2018, 2018, doi: 10.1155/2018/2745973.

[9] D. Shen, G. Wu, and H. Suk, “Deep learning in medical image analysis,” A*nnual Review of Biomedical Engineering,* vol. 8, pp. 221-248, June 2017, doi: 10.1146/annurev-bioeng-071516-044442.

[10] T. Nishtar, S. Burki, F.S. Ahmad, and T. Ahmad, “Diagnostic accuracy of computer aided reading of chest x-ray in screening for pulmonary tuberculosis in comparison with Gene-Xpert,” *Pakistan Journal of Medical Sciences*, vol. 38, no. 1, pp. 62-68, Jan. 2022, doi: https://doi.org/10.12669/pjms.38.1.4531.

[11] M.F. Hashmi, S. Katiyar, A.G. Keskar, N.D. Bokde, and Z.W. Geem, “Efficient pneumonia detection in chest xray images using deep transfer learning,” *Diagnostics*, vol. 10, no. 417, June 2020, doi: 10.3390/diagnostics10060417.

[12] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R.M. Summers, “ChexX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases,” *IEEE CVPR 2017,* 2017, pp. 2097-2106, doi: 10.1109/CVPR.2017.369.

[13] J. Carver. “Lung infiltration: What are the causes.” Red Deer First Aid. https://firstaidreddeer.ca/lung-infiltration-what-are-the-causes/ (accessed Dec. 12, 2022).