Deep Learning Approach for Thoracic Cavity Disease Feature Identification and Classification

Jeff Moore, MSc[1, 4], Matt Pallini, MSc[2, 4], Mitchell Thompson, MSc[3, 4]

[1]MBD Solutions, Los Angeles, CA
[2]Advocate Health Care, Chicago, IL
[3]Riverside Research, Dayton, OH
[4]Northwestern University, Master of Science Data Science, Chicago, IL

June 4, 2020

EXECUTIVE SUMMARY

In this work we present a deep learning approach for feature identification and classification of 14 different thoracic cavity diseases. The team instituted two data pipelines depending on local machine capacity and processing capability. Our approach consisted of building three competing models with the first being a custom built, deep-layered convolutional neural network (CNN). This model, hereafter referred to as the Baseline Model, uses code that has been improved upon from previous deep learning courses. The second model, an ensemble *Bucket-of-Models*, instantiates a loop over six state-of-the-art CNN architectures. The third model for comparison was an effort to recreate and surpass the recent success it has acquired in the field of medical imaging. The exploration of various models was driven by a need to determine which had the best performance as well as explore which would work in a resource constrained environment. Given the number of companies professing a model that can predict diseased or damaged lungs, we knew that our model must offer differentiation. To our surprise, very little limitations were found with model readiness, but significant gaps were seen in data availability, quality, and variety. In order to improve the existing model and how it performs, TensorBoard was instituted to visualize the metrics, architectures, and tensor distribution during training in each model. One significant challenge that the team did not fully mitigate was the large class imbalance within the original dataset. As a result, our effort of semantic segmentation was delayed by approximately two weeks along the critical path. More specifically, data curation and the construction of a data pipeline for all efforts involved in our approach proved both time-consuming and complicated. The goal with segmentation was to determine the shape of the lung disease in question for more rapid diagnoses as well as dimensionality reduction to further reduce loss and time in the training and testing phases. Upon discovering this shortfall, two things have become apparent. The data challenges to be discussed required the team to reprioritize the modeling goals as well as understand the need for further research into image processing techniques that would most likely increase the return on investment. This body of work resulted in nominal rates of accuracy and loss with a downsample version of the original dataset. More importantly however, were the practical application lessons and portfolio each author acquired as a result of the project.

Deep Learning Approach for Thoracic Cavity Disease Feature Identification and Classification

Problem Statement

The focus of this analysis is to understand how state of the art artificial intelligence tools can be deployed in health care organizations. Many industries are using advanced image detection models to serve the needs of their customers, but health care has been slow to adopt this same technology. A survey completed in December 2019 shows that barely one-third of hospitals or imaging centers in the United States use artificial intelligence (AI) in their organization (Kent, 2020). Two of the biggest reasons for the lack of adoption of AI are because those organizations lack strategic direction and technical expertise to implement these technologies. When AI is implemented, it leads to better patient outcomes and saves doctors and organizations time and money. Additionally, the COVID-19 pandemic has revealed the technological shortcomings that exist in the healthcare industry. When comparing organizations based on their analytical maturity, Davenport and Harris show that health care rates the lowest of any industry (Davenport & Harris, 2017). Also, expertise and resources are not evenly distributed across the population. This problem does not exist only in third world nations, but also developed countries like the U.S. and U.K. (Bassett, 2019). The increased use of remote doctor visits (Telehealth) that have come during the era of COVID-19 has shown that health care can quickly adopt new technologies. By accessing a set of 112,120 thoracic cavity x-rays made available by the National Institutes of Health (NIH), this study shows how medical imaging can adopt state of the art AI technologies. The analysis incorporates several AI and deep learning architectures, such as transfer learning, image classification, and image segmentation. Using these tools will have long term benefits for health care organizations and patients alike. AI models that can process images like this means that doctors will have more time to spend face to face with their patients instead of manually diagnosing x-rays.

Literature Review

Computer aided diagnosis in medical imaging was first introduced in 1966 by Dr. Gwilynn Lodwick (Lodwick, 1966). However, it wasn't until 2015 when Bar et al. illustrated sufficiency of deep learning in detecting abnormalities in the thoracic cavity utilizing transfer

learning (Bar, Diamant, Wolf, & Greenspan, 2015). Since that time, the advent of compute power available datasets, and implementation of deep learning on chest pathology has seen an exponential increase. A 2018 application of deep learning originally intended for aged-based degeneration and diabetic eye disease also illustrated generality when it classified childhood pneumonia with 96.8% accuracy (Kermany et al., 2018). In this case, the backbone of the model was a pre-trained Inception V3 architecture with frozen convolutional layers as the fixed feature extractors, the ADAM optimizer algorithm, and cross entropy as the loss function. Most likely to be the most comprehensive effort to date however is the work of Wang et al. in their comparison study of four DCNN architectures using the newly established NIH dataset (Wang et al., 2017). This study demonstrated moderate success across all four models with the ResNet50 architecture proving to rank the highest among seven of the eight lung abnormalities in the dataset. Additional work employs custom layers in comparison of complexities within the ResNet architectural family (Baltruschat, Nickisch, Grass, Knopp, & Saalbach, 2018), and an implementation of a deep Siamese based neural network (Anuja Kumar Acharya, 2020). Most interesting of late though is the work and datasets directly related to the ongoing COVID-19 pandemic. In their most recent work, Ilays, Rehman, and Nait-ali show high rates of success among the state-of-the-art deep learning architectures across different sizes of datasets (Ilyas, Rehman, & Nait-ali, 2020). The key to this success lies in appropriate use of transfer learning as initially demonstrated in 2015 and further validated by Greenspan et al. in 2016 (Greenspan, van Ginneken, & Summers, 2016)

Discussion

Data: NIH Clinical Center Chest X-ray Dataset

The data source for this project will be a large collection of chest x-rays that are made available through the NIH (Health, 2017). The data includes over 100,000 images, which makes the size of the files a challenge to work with. Instead of using the files directly from NIH, they will be accessed via Kaggle. The images and the data are the same from both sources, but Kaggle offers multiple avenues that reduce and eliminate some of the system challenges that come with working with such a large data source. The images themselves are divided up into a series of compressed files that are 46 GB in size. By accessing these files via Kaggle, Google

Colaboratory (Colab) can be utilized to ease the resource strain of loading these files. Colab can use the Kaggle API to pull the files directly without the need to download them locally. For example, a sample .zip file is available through Kaggle that contains roughly 5,600 images and is 4.2 GB in size. By using Colab this entire file loads in well under one minute. Additionally, Colab offers free GPU and TPU processing support. These options will reduce the amount of time it takes to process the images and train the network.

In addition to using the NIH data to identify more common chest and lung conditions, this analysis can be expanded to identify COVID-19 as well. While there are not large, robust datasets of chest x-rays identifying COVID-19, there are smaller sets available that have shown early promise for image detection. This data is available through several authors on GitHub and Kaggle. While chest x-rays are not commonly used for COVID-19 patients at the moment, adding this data to the model is still very useful. This disease can be included with the larger list of conditions when a vaccine and treatment exist in the future. For the purposes of this research and modelling effort, the COVID-19 dataset provided by Chowdhury et al. comprises a blend of 190 COVID-19, 1345 viral pneumonia, and 1341 normal chest x-ray images (Chowdhury et al., 2020).

Exploratory Data Analysis. The NIH data provides not only the chest x-rays but also details about those images. These details include patient age, patient gender, and a label for each image that indicates what condition exists in the x-ray. In total, there are 112,120 unique images and 30,805 unique patients captured in this data. Within the patient population 56.8% of patients have only one x-ray. The remaining patients have multiple images assigned to them, with some having as many as 184 x-rays. When looking further at the patient information, 56.5% of the x-rays in the data belong to men. While the number of images is skewed towards men, the age distribution is roughly the same for both men and women. There were a very small number (16) of age outliers that were updated by imputing the population median for those values. As part of the EDA, age groupings were created to better classify each patient. These groupings were in 10-year increments from zero to 100. These groupings were then used to highlight the age distribution for the entire population as well as by each gender. One concern with this data is that the age distribution may not be the same for each gender, which could create problems. Older patients are more susceptible to a variety of conditions, especially infections like pneumonia,

which is part of this data set. However, this issue was not the case with this data set (Appendix A, Figure A1). The chest x-rays in this data are looking for 14 specific conditions. There is also an option for "No Finding," meaning there are 15 possible diagnoses to be identified in the images. However, the labels provided for each image had 836 unique values. This was a result of many images having multiple conditions in the same x-ray. In fact, there are 20,796 images that show multiple conditions (18.5% of the total). After identifying these images work was done on the data to show multiple condition labels for each image. Doing this keeps the list of conditions at 15, meaning that there will be 15 classes for the model to predict in the future. No Finding is the most common image result and Hernia is the least common result in this data set (Appendix A, Figure A2). Finally, it is important to identify any conditions that affect one gender more than another. This information, if it exists, can be useful in any future predictions. However, as with the age distribution, each condition affects each gender at roughly the same rate (Appendix A, Figure A3).

Significant Data Challenges. When performing any kind of data analysis, more data is always better. Having more data provides a larger sample size to draw conclusions from. Computer vision models are no different. In the case of computer vision for medical images, it is even more important. While many computer vision examples look to identify different types of objects (like dogs or cats), the differences between medical images are very subtle. For example, Figure 1 shows a side by side comparison of two thoracic cavity x-rays. The image on the left shows a person with cardiomegaly (enlarged heart), while the image on the right shows a clean scan (no conditions). It is difficult to determine the difference between the two. Given these very small differences, having as many examples of images as possible will be a benefit to the model performance. By having more images showing the various conditions, it will be easier for the model to identify the differences. However, labeling of such data is outside the realm of expertise of the authors and beyond the scope of this project.

Data Quality. Data quality is an important part of any analysis. The NIH chest x-ray data is not of the highest quality images, making the analysis far more challenging. First, the quality of the images themselves vary, making it harder to read the scans. Looking at Image 1 again, the x-ray on the right (the clean scan) does not have the same resolution strength as the one next to it. These differences in quality present challenges in the model. Second, the data itself is very skewed. The NIH data deals with a very large class imbalance. There are 15 types of conditions that are picked up on these x-rays, but 53.8% of the images fall into the first class, no finding. The least populated class, hernia, represents only 0.098% of the images in the data (110 out of 112,120). Since the classes are so imbalanced, a model can predict "no finding" for every image and be correct more than half of the time. The poor image quality along with this class imbalance makes it very difficult for a model to learn the characteristics associated with each condition.

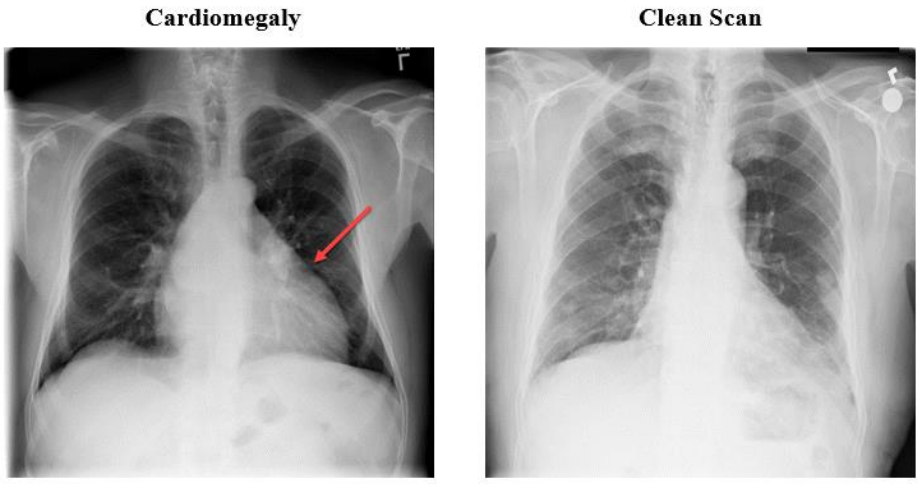

Figure 1. Side by side comparison of thoracic cavity x-ray scans

Hardware / Infrastructure. The final challenge when working with the NIH data is the amount of resources required to build a model for image classification. First, a large amount of storage is required for the images. The images for this data come in 12 compressed files that are a combined 42 GB in size. When the files are uncompressed, the size grows beyond what a normal personal computer or free cloud storage resources can handle. The use of an upgraded cloud storage account (through Google, Azure, or AWS) that offers at least several hundred GB of storage is a must for any computer vision project on this scale. Additionally, training a model for image classification requires a lot of compute resources. If the model uses image sizes of 224 x 224 pixels, that means that there are over 50,000 input variables for every image. Standard

CPU processors cannot handle this workload. High speed GPU processors are required for any image-based model. Google Colab provides free access to GPU, which was used for this model. However, the basic Colab service only provides 25GB of disk space. An upgraded platform such as Colab Pro would be even more beneficial for this project. Lastly, in an effort to model with more data, the use of the Kaggle service was brought to bear as the NIH dataset is also hosted on the site. Kaggle offers both the required storage space as well as Nvida Tesla P100 GPU resources. Unfortunately, access to the Kaggle GPU resource is not very straightforward, or guaranteed. All training sessions on Kaggle maximized available CPU and RAM resources while limiting any GPU activity to under 3% of capability. The Kaggle service also implements specific time limits per session. Initial training efforts utilizing the full dataset reached the time limit forcing subsampled data in subsequent sessions. In the end, modelling in the Kaggle environment was limited to roughly 40% of the dataset and predominantly CPU compute resources.

Research Design and Modeling Method(s)

Project design with regard to the development of a data pipeline and modeling encompassed a total of seven model comparisons utilizing transfer learning as well as semantic segmentation. First, several experiments incorporating various combinations of hyperparameters such as batch size, epochs, layers, dropout ratio, and optimizer selection were executed in the model comparisons. Second, the use of the UNet, originally developed by Olaf Ronneberger et al. for Bio Medical Image Segmentation, was used for the semantic segmentation effort (Ronneberger, Fischer, & Brox, 2015). Lastly, our efforts to model the entire multi-label, multi-class NIH dataset also incorporated modelling with the DenseNet-121 architecture due its inherent advantage of alleviating the vanishing-gradient problem. This section will discuss each modelling effort as well as the analysis and interpretation of the associated results.

Baseline Model. The baseline model is an image classification model that was built from scratch to analyze the NIH data. This model was used to identify a simple performance baseline before exploring more advanced architectures, such as those that utilize transfer learning. As mentioned previously, the data presented two unique challenges to building a brand-new model for this data. The heavy class imbalance of the images makes training a model

for each class challenging given the available data. The second challenge is that 18.5% of the x-rays show multiple conditions in them. Accounting for each of these images makes the overall analysis a multi-label multi-class classification problem. To simplify the model so that it will produce results, the first step has been to create a binary classification model. This model treats all images that have a condition as a single class. This has removed some of the initial struggles that came with training a model for an imbalanced multi-label multi-class solution. The classes are now close to being of equal size, and there is only one label output for each image.

The model itself uses a convolutional neural network (CNN) architecture to achieve its results. It uses three separate dropout layers to reduce overfitting. It also utilizes several layers of Conv2D and MaxPooling2D to pick up important features in each image to differentiate the conditions from clean scans (Figure 2). The initial baseline model uses a sample of 5,600 out of the 112,000 total images, making it easier to update and train the model. The model uses a train/test split of 80% for training and 20% for testing. Additionally, the model trained through 15 epochs, each with a batch size of 44. This produced a training accuracy of 63.5% and a validation accuracy of 60.5%. Also, using this model to make predictions on the test data produced a ROC score of 0.578.

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_1 (Conv2D)            (None, 126, 126, 32)      896
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 124, 124, 32)      9248
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 62, 62, 32)        0
_________________________________________________________________
dropout_1 (Dropout)          (None, 62, 62, 32)        0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 60, 60, 64)        18496
_________________________________________________________________
conv2d_4 (Conv2D)            (None, 58, 58, 64)        36928
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 29, 29, 64)        0
_________________________________________________________________
dropout_2 (Dropout)          (None, 29, 29, 64)        0
_________________________________________________________________
conv2d_5 (Conv2D)            (None, 27, 27, 128)       73856
_________________________________________________________________
conv2d_6 (Conv2D)            (None, 25, 25, 128)       147584
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 12, 12, 128)       0
_________________________________________________________________
dropout_3 (Dropout)          (None, 12, 12, 128)       0
_________________________________________________________________
flatten_1 (Flatten)          (None, 18432)             0
_________________________________________________________________
dense_1 (Dense)              (None, 128)               2359424
_________________________________________________________________
dense_2 (Dense)              (None, 2)                 258
=================================================================
Total params: 2,646,690
Trainable params: 2,646,690
Non-trainable params: 0
_________________________________________________________________
```

Figure 2. Base model CNN architecture

Early positive signs for this model are that the training and validation accuracy metrics are close together, suggesting that the model is not overfitting. However, the accuracy and ROC scores are not very high. While tuning hyperparameters and utilizing callbacks (such early stopping or a scheduled learning rate) would produce better results, these improvements would

be marginal at best. These results showed that in order to see significant improvements, as well as accurately address the class and label challenges, a transfer learning architecture was needed.

Ensemble Model. The secondary model is a *Bucket of Models* method of ensemble learning. By utilizing a model selection algorithm, we are directly applying the concept of transfer learning and attempting to discover the best network architecture for the problem. The aforementioned challenges and early mitigation efforts remain the same, however the *Bucket of Models* provided a rapid prototype capability to the team. The rapid prototyping illustrated early signs of improved performance with the VGG16 architecture achieving higher rates of accuracy (Table 1). We allowed certain hyperparameters to either remain as default or modified with the exception of the following. These remained static during all training and testing runs.

Ensemble Model Constants:

- *Epochs 50*
- *Binary Classifier*
- *HeadModel with BaseModel Architecture*
- *Weights - Imagenet*
- *Image Count: 450*

Model	Precision (COVID Label)	Precision (non-COVID Label)	Accuracy	Sensitivity	Specificity
VGG16	.94	.98	.97	.97	.97
VGG19	.96	.92	.93	.84	.98
MobileNet	1	.82	.86	.59	1
InceptionV3	1	.65	.66	.03	1
ResNet50	.36	0	.36	1	0
Xception	1	.84	.88	.66	1

Table 1. Ensemble of model comparisons

Forging ahead with a VGG16 based transfer learning model, the use of Tensorboard was relied upon to visualize and hyperparameter and optimizer changes. Integrated with every installation of TensorFlow, Tensorboard is the platform's visualization toolkit which affords the real time tracking of important ML metrics like loss and accuracy. The toolkit also allows the developer to visualize the model in graph form and view histograms of the weights and biases as they change over time. Our experimentation with the toolkit during the project was an effort to visualize the

aforementioned metrics and tensors as a result of a change in optimizers. Runs of 20200516* use `adam` as the optimizer while named runs of 20200517* use `rmsprop` (Figure 3). The x-axis shows time steps and the y-axis shows the predicted accuracy over time. This is quite similar to plotting out `history = model.fit()` except that Tensorboard offers a great level of interactivity. The model overfits as expected but it is interesting how much earlier the `rmsprop` optimizer begins to overfit compared to `adam`.

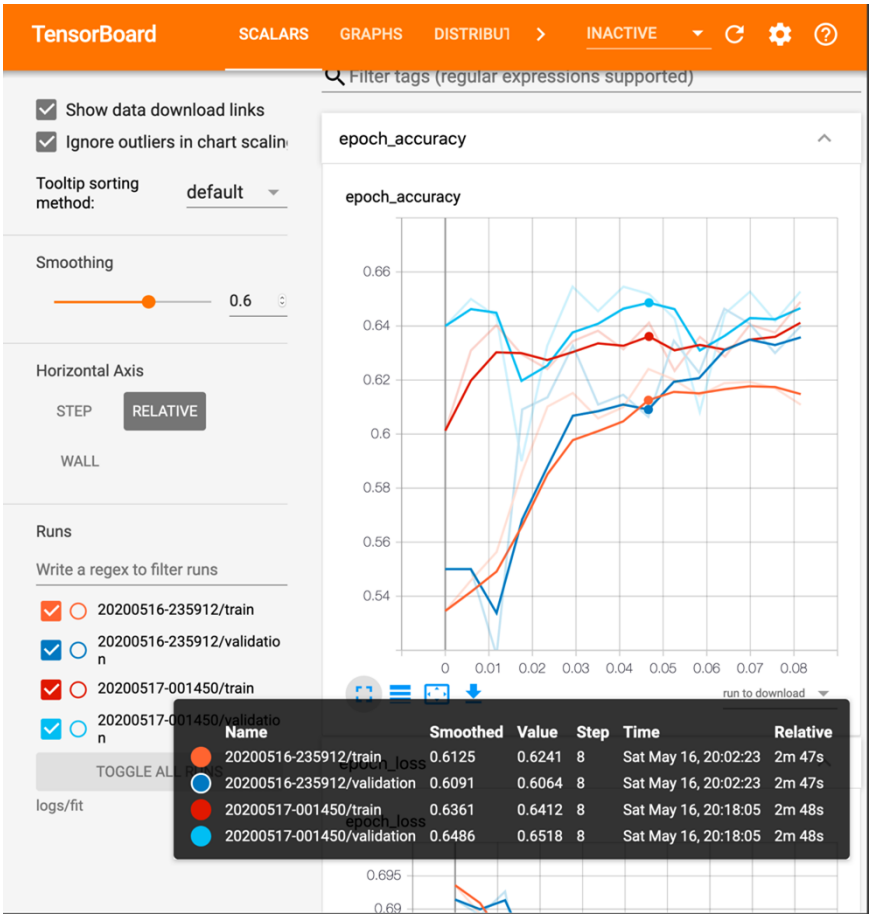

Figure 3. Comparison of optimizer in VGG16 transfer learning

Semantic Segmentation. While the team was working to find bigger and better datasets, we reviewed our initial list of objectives. One of the initial objectives was to build a model that could perform semantic segmentation. The UNet architecture was selected as it was built for this purpose. After several attempts the team had some success in getting the model up and running. However, model performance was uneven at best as initial predictions fell short of expectations. Analysis suggests that part of the issue is fitting our dataset into this code set. Although several

modifications were made to the UNet model, more would be needed to improve predictive power. In addition, image quality might also be a concern as the source code was built for color images vs. our 128 x 128 grayscale photos (Figure 4).

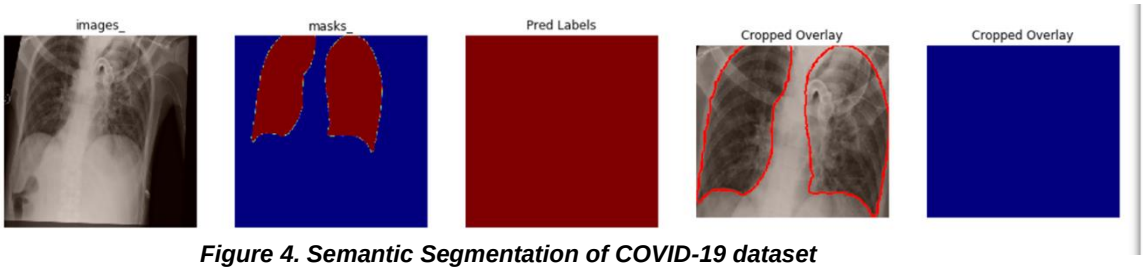

Figure 4. Semantic Segmentation of COVID-19 dataset

DenseNet-121 Model. Having a relative amount of success with both the custom baseline model and the transfer learning ensemble of architectures provided space to experiment further with the entire NIH dataset. To accomplish this, we turned to an architecture that has been successfully utilized quite frequently in medical imaging (Xu, Lin, Tao, & Wang, 2018) (Ozturk et al., 2020). In addition to the successful uses in the medical field, the DenseNet architecture enables a multi-path flow for input feature maps and backpropagation between each layer, thereby addressing the problem of a vanishing gradient. In other words, the output feature maps of one layer and incoming feature maps are not summed as in more traditional architectures like ResNets. Instead, the feature maps are concatenated which provides access to the loss function gradient to each layer. The architecture is also known to reduce the dimensionality of required parameters which has a direct relation to the computational cost. Such a reduction in cost was considered a positive trait and mitigation effort on the resource constraint problem mentioned previously. For the experimentation against the NIH full dataset, a pre-trained DenseNet-121 architecture was selected as yet another mitigation tactic against limited compute resources due its efficiency in floating point operations per second during testing against benchmark datasets (Huang, Liu, van der Maaten, & Weinberger, 2016).Experimentation with DenseNet-121 on the full NIH dataset occurred within the Kaggle platform to take advantage of both the storage and compute resources. The pre-trained DenseNet-121 model was used as a feature extractor, which is one form of transfer learning. In place of the last fully connected layer of the original DenseNet-121 architecture, we placed two Dense layers with the final layer serving as the classifier for 13 classes of lung disease (Figure 5). Due to the

aforementioned challenges with the Kaggle platform, final training of the pre-trained DenseNet-121 model used 40,000 out of the 112,000 total images. A data pipeline utilizing data generators was used to achieve a train/test split percentage of 80/20. Training ran for 10 epochs with callbacks to break out of training if validation loss did not improve after three epochs. Unfortunately, this session resulted in a model that underfit the data. Fine tuning of the model followed to include modifications to the optimizer and adding L2 Regularization. A second round of training broke out of the cycle after only five epochs, again underfitting to the data. At this point, additional tuning of the model was put on hold due to time and resource constraints. Next steps for further are discussed in the next section.

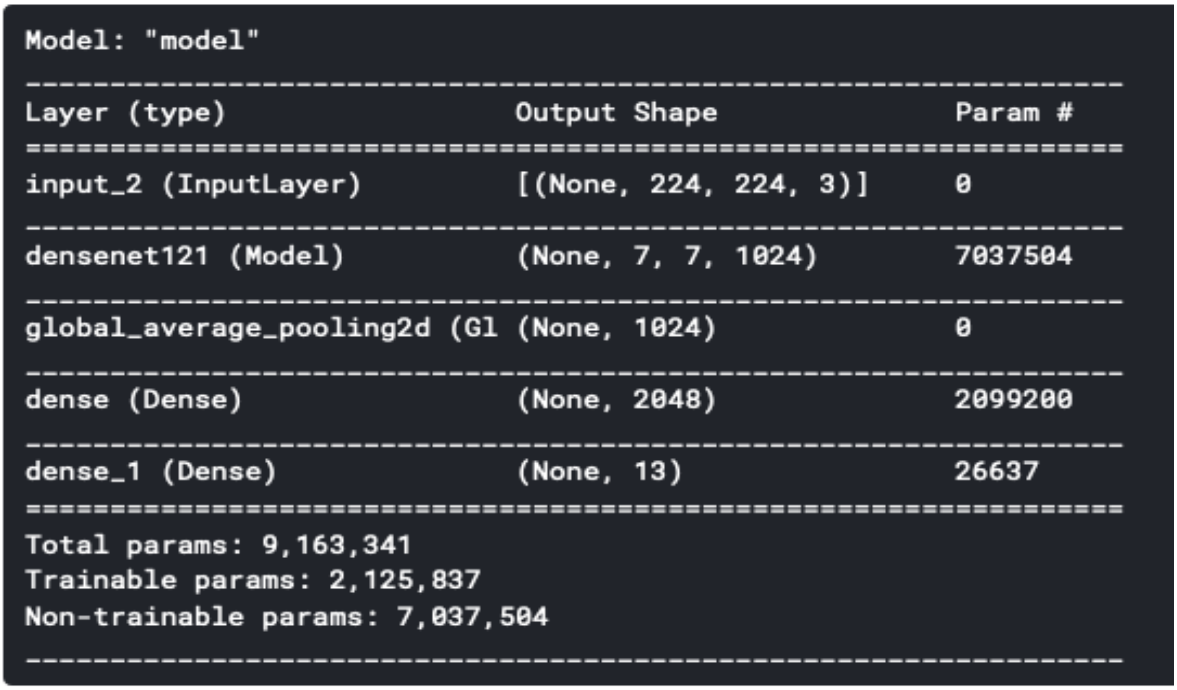

Figure 5. Pre-trained DenseNet-121 architecture

Recommendations and Future Work

Several recommendations and initiatives for future work exist, beginning with the acquisition of appropriate storage and compute resources to take advantage of the full dataset. Closely following this top priority is the acquisition of more reliable data having higher quality pixel intensity distributions and balance of targeted labels. The NIH dataset in particular involved a labelling method consisting of natural language processing which has both an

inherent error rate and an irreducible error. Utilizing larger and more robust datasets made available by Stanford (224,000 images) and MIT (371,000 images) would help overcome many of the data challenges with the NIH data. However, access to these datasets required upfront permissions and training on proper use. Additional plans for future work also include further fine-tuning of pre-trained layers within the ensemble model as well as the DenseNet-121. Lastly, the creation of additional masks for use in semantic segmentation would likely be of significant utility to reduce the dimensionality when placed upstream of disease classification. However, any effort to create additional mask labels would absolutely require radiological subject matter expertise. With regard to professional development and advancement of career, the closing paragraphs serve to illustrate how this deep learning approach to thoracic cavity disease has emboldened each author.

Jeff Moore started his machine learning journey, not knowing if the skills necessary to gain employment could be learned. Having achieved some level of professional success, he craved freedom and independence that typically comes from launching a business. Jeff knew he didn't have all of the skills and that time was not on his side. Specializing in AI (more specifically, computer vision) is providing the skills and confidence to compete in the job market and to change his side hustle into a software and services startup. At this point, Jeff has the 'ring' around his neck and is looking for 'friends' who will join him on the hero's journey to create AI solutions that open access to healthcare for more people around the world.

Matt Pallini aims to use this project as a starting point for his organization to adopt AI technologies. Working for a large healthcare provider, he sees the untapped potential that exists in health care for AI and computer vision. The experience from this project, both the successes and failures, are a great learning opportunity to continue pursuing this work. This analysis will increase his skills as a thought leader and help bring strategic vision and technical expertise on AI to the organization. By identifying and utilizing publicly available data and research, the project has shown what can be achieved in the healthcare field when applying artificial intelligence. As technology becomes a bigger and bigger part of patient care, utilizing artificial intelligence for medical images will allow the organization to be at the technological forefront of the industry.

In the midst of a career change, Mitchell Thompson will capitalize on the confidence and knowledge gained during this project. As a team and independently, we tackled what field

experts hypothesized as the next frontier in AI/ML: transfer learning. Within that context we successfully compared 6 different deep learning architectures against a subset of a very challenging dataset. We then exercised image segmentation on a COVID restricted dataset. In the successes and failures within this project, Mitchell has acquired an elevated mental high ground. Moreover, the specific shortcomings of the DenseNet-121 application has proved to be highly educational in that it drove further investigations into reasons behind the failure. This directly addresses personal and professional motivations for success and learning. The implementation of all three modeling efforts required a heavy amount of code conversion and troubleshooting. By way of this conversion, the physicality of professional skillsets as well as patience, confidence, integrity, and devotion advanced a great deal. The above-mentioned attributes, coupled with domain expertise remotely sensed, earth observation data, furnish a solid foundation to the career transition and will most definitely be pragmatic to a data science role.

Finally, we'd like to thank Dr. Alianna J. Maren for her steadfast dedication, devotion, and encouragement to the members of this project. The combination of philosophical, theoretical, and practical application behind this deep learning approach to thoracic cavity disease proved to be most advantageous.

References

Anuja Kumar Acharya, R. S. (2020). A Deep Learning Based Approach towards the Automatic Diagnosis of Pneumonia from Chest Radio-Graphs. *Biomedical & Pharmacology Journal, 13*(1), 449-455. doi:10.13005/bpj/1905

Baltruschat, I. M., Nickisch, H., Grass, M., Knopp, T., & Saalbach, A. (2018). Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification.

Bar, Y., Diamant, I., Wolf, L., & Greenspan, H. (2015). Deep learning with non-medical training used for chest pathology identification. In (Vol. 9414, pp. 94140V-94140V-94147).

Bassett, M. (2019). Radiologist Shortage in the U.K. Continues to Deepen. (May 24). Retrieved from https://www.rsna.org/en/news/2019/May/uk-radiology-shortage

Chowdhury, M. E. H., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M. A., Mahbub, Z. B., . . . Al-Emadi, N. (2020). Can AI help in screening viral and COVID-19 pneumonia? *arXiv preprint arXiv:2003.13145*.

Davenport, T., H., & Harris, J., G. (2007). *Competing on Analytics: The New Science of Winning* Harvard Business School Press.

Greenspan, H., van Ginneken, B., & Summers, R. M. (2016). Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique. *IEEE Transactions on Medical Imaging, 35*(5), 1153-1159. doi:10.1109/TMI.2016.2553401

Health, N. I. o. (2017). NIH Clinical Center provides one of the largest publicly available chest x-ray datasets to scientific community [Press release]. Retrieved from https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2016). Densely Connected Convolutional Networks.

Ilyas, M., Rehman, H., & Nait-ali, A. (2020). Detection of Covid-19 From Chest X-ray Images Using Artificial Intelligence: An Early Review.

Kent, J. (2020). One-Third of Orgs Use Artificial Intelligence in Medical Imaging [Press release]. Retrieved from https://healthitanalytics.com/news/one-third-of-orgs-use-artificial-intelligence-in-medical-imaging

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., . . . Shi, A. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell, 172*(5), 1122-1131.e1129. doi:10.1016/j.cell.2018.02.010

Lodwick, G. S. (1966). Computer-aided diagnosis in radiology. A research plan. *Investigative radiology, 1*(1), 72. doi:10.1097/00004424-196601000-00032

Ozturk, T., Talo, M., Yildirim, E. A., Baloglu, U. B., Yildirim, O., & Rajendra Acharya, U. (2020). Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Computers in Biology and Medicine,* 103792. doi:10.1016/j.compbiomed.2020.103792

Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-net: Convolutional networks for biomedical image segmentation*.

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-*, 3462-3471. doi:10.1109/CVPR.2017.369

Xu, X., Lin, J., Tao, Y., & Wang, X. (2018). An Improved DenseNet Method Based on Transfer Learning for Fundus Medical Images. In (pp. 137-140).

Appendix A

Figure A1

Groupings by Gender in the NIH dataset

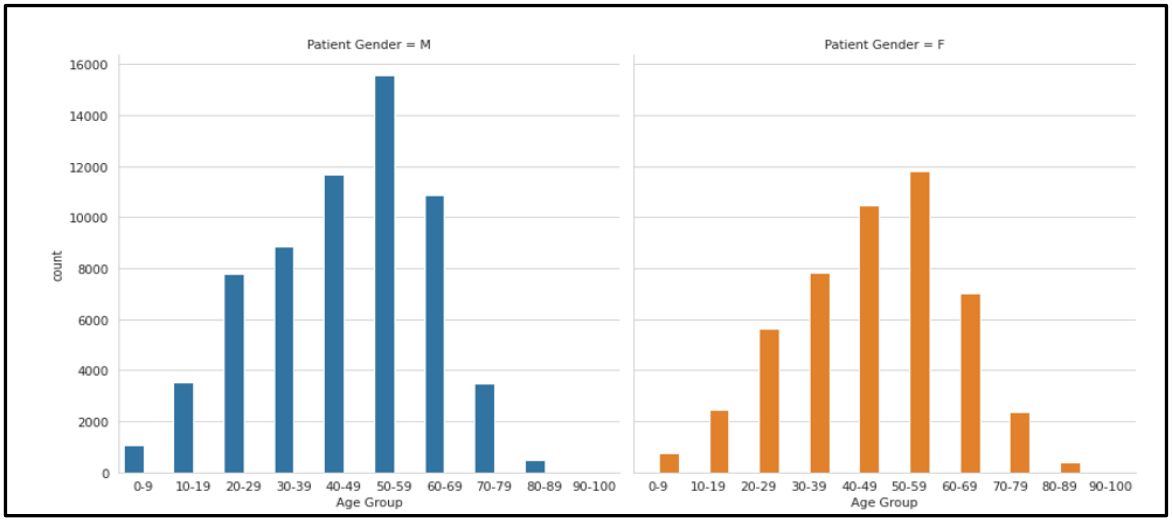

Figure A2

NIH Dataset: Group Counts by Thoracic Disease

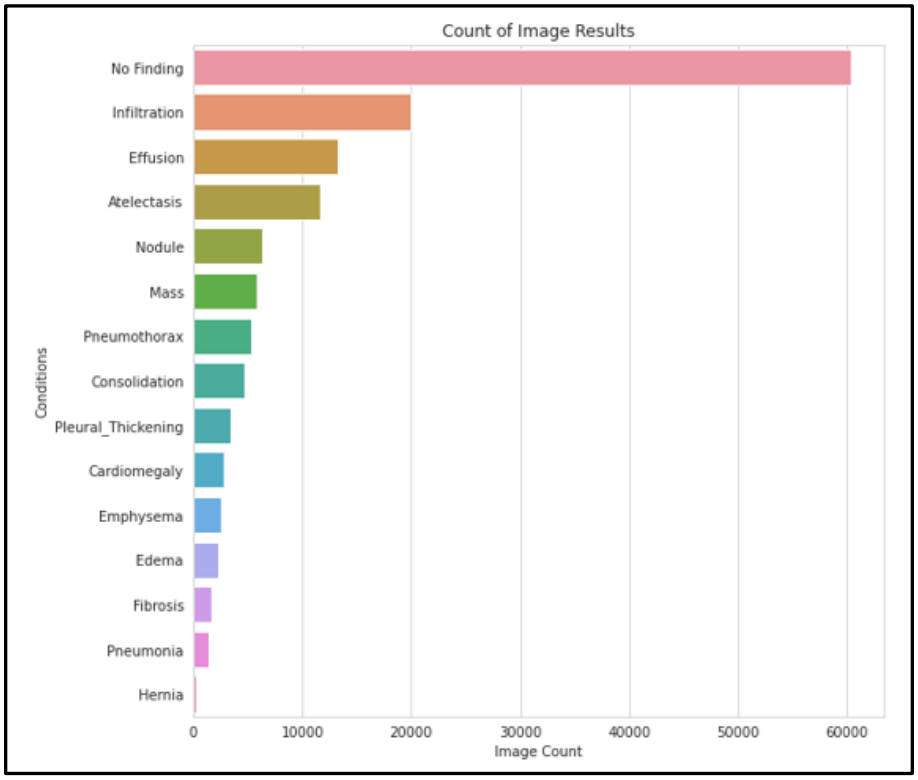

Figure A3

NIH Dataset: Thoracic Disease Group Counts by Gender