- Whitaker Behrens
- BrainStation Data Science

- Capstone Project Report

Pneumonia Image Classification with Neural Networks

Thank you for your interest in my BrainStation Capstone project, detecting Pneumonia infections using neural networks. This report is a summary of the question behind the project, brief background information on Pneumonia, how the data was sourced, the process of preparing and modeling the data, results from modeling and surprising or interesting insights discovered, and a reflection on how the original business problem was addressed as well as thoughts on how the project could be developed further in the future and expanded. For additional details on any of the steps outlined in this report, please reference the accompanying notebook with all code and specific steps taken.

The project started with a simple question: Can a computer be trained to accurately and quickly detect the presence of a Pneumonia infection in medical X-ray imaging. A system such as this could be used to process a large volume of medical images, acting as an initial screening at a much higher rate than would be possible for a human being. Potential cases could then be verified by a trained human professional, and further action taken if required. This type of system would add value to the medical field by helping to speed up the initial evaluation process and move the diagnosis process more quickly – meaning savings in time, money, and very much potentially human lives.

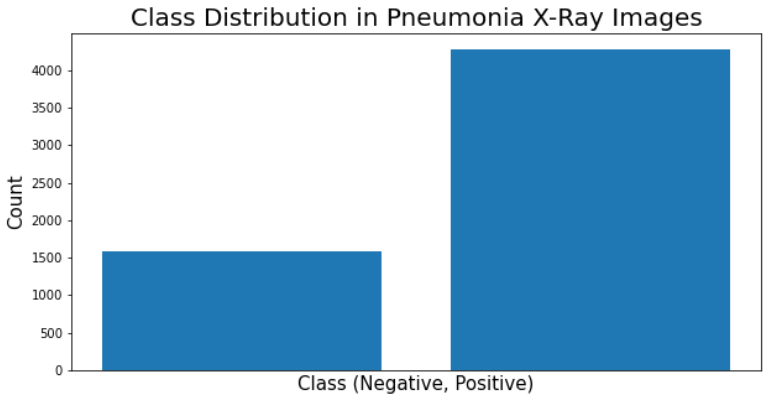
From left to right: Negative, Bacterial, and Viral diagnosed infection images. Note the relatively clear lung area in the Negative image, the overall more opaque appearance of the lungs in the Bacterial image, and the localized opacity in the Viral image – indicative of how each class presents visually in medical imaging on average.

Pneumonia is a type of infection that causes inflammation in the lungs. An infection can be caused by bacteria or viral means. Treatment is mostly similar, most often including a simple course of antibiotics for a bacterial infection or antivirals in the case of a viral infection. The most common practice for diagnosing a Pneumonia infection is medical imaging of the lung area which is then evaluated by a trained radiologist for the visual presentation of the infection. This makes computer image detection techniques a prime candidate for increasing the efficiency and decreasing the overall time the diagnosis stage requires. Computer vision systems are capable of processing and classifying images at a much higher rate than human

counterparts, and previous studies in this area indicate that they can be trained to achieve accuracy of detection at levels equal to if not exceeding trained human professionals. Multiple studies published by the Radiological Society of North America (RSNA) place the best-case radiological interpretation error rate by trained professionals at about 5%, with some studies estimating a 20% error rate or more – equating to one in five patients being incorrectly or entirely un-diagnosed. If a computer imaging system can be trained to maintain this level of accuracy while processing a much larger volume of images much more efficiently, more time can be saved for treatment of patients and result in more positive experiences and outcomes.

The data required to undertake this project was sourced from the public dataset aggregating website Kaggle. The dataset consists of slightly fewer than 5,900 X-ray images from pediatric patients between the ages of 1 and 5 years old. The images were collected during routine clinical care at the Guangzhou Women and Children's Medical Center in Guangzhou, China. Each X-ray was independently evaluated and diagnosed by two expert physicians and split into respective classes, and finally evaluated by an independent third party expert for clarity and quality. The images are compiled into two subdirectories, 'negative' images in one and 'positive' images in the other. The 'positive' images are a combination of bacterial and viral infections, each denoted by the name of the file itself.

As I was working with purely image data without accompanying text data, cleaning and processing of the images was relatively limited. After importing the images were normalized in overall size and pixel intensity was scaled for ease of training. Initial data exploration did reveal a notable class imbalance, roughly a 3:1 ratio between the 'positive' and 'negative' classes.



While significant enough to be of importance, I made the decision this imbalance did not warrant immediate action to remedy before beginning the initial modeling phase. My first model was instantiated as a sequential Convolutional Neural Network (CNN) with what could be considered relatively 'standard' CNN architecture (see accompanying notebook for specific details). After training the model for ten epochs using a 50-25-25 train-validation-test split of the entire dataset (this became standard for all model iterations during the project), initial results were very promising with a training accuracy of 96.3%, validation accuracy of 95.1%, and test set accuracy of 94.3%. The high training accuracy is indicative of a model that was not under-fit, and the small two percentage point difference with the test set accuracy indicates the model isn't over-fit to the training data. In addition, the initial model displayed high precision and recall with 96.7% and 95.3% respectively. Of particular note is the recall figure, or the total

number of Positive images that were correctly classified by the model. Special attention was given to this figure as the practical purpose of this system indicated it should be – in the medical field, a False Positive that is then reviewed and corrected by a trained professional is much less potentially costly than a False Negative, which could result in a patient in actual need of treatment remaining undiagnosed to potentially catastrophic consequences. These initial accuracy and recall metrics were used as baseline performance that I attempted to improve on by iterating different variations of the model from that point.

Although several additional models were trained with variations attempting to compensate for the identified class imbalance (see accompanying notebook for specific details on each), overall accuracy and recall figures never improved over the initial modeling baseline metrics. Processing techniques applied to the dataset included weighting the classes for training, the Synthetic Minority Over-sampling Technique (SMOTE), and dividing the single 'positive' class into the component 'bacteria' and 'virus' classes, turning the original binary classification problem to a multiclass question. One very surprising result from this multiclass modeling approach was the determination that the model could relatively easily distinguish between the 'negative' class and either of the two 'positive' classes – the computer had difficulty, however, distinguishing between the 'bacteria' and 'virus' classes present in the dataset. This finding was further supported by training a model on only the two 'positive' class datasets, which resulted in the worst performing model by far. Given the original question posed by the project of simply identifying the presence of Pneumonia, and the relative similarity in treatment between the two types of Pneumonia infections, I decided to forego attempting to further increase the ability to accurately discern between the two types of infections and return to the original binary classification model.

Additional evaluation of the best-performing model was attempted by visualizing the misclassified images, however it was at this time my personal lack of domain expertise became rather apparent. Although the model itself is capable of identifying the presence of Pneumonia with nearly 95% accuracy, I was unable to visually identify any patterns or trends in the misclassified images due to my lack of experience reading medical images (please see accompanying notebook for example images and further explanation). With slightly more time, I would have included further model evaluation using the SHAP Values library, a tool which helps to open up the 'black box' of neural network models and indicates which pixel values in a dataset are most influential in the final model – this could help to better identify trends in the misclassified images despite my current lack of domain expertise. Another possibility would be to begin with a dataset that includes not only images but additional (anonymized) patient information, including stated symptoms, environmental variables, other test results, etc.

In summary, I set out to attempt to train a computer vision model to accurately and efficiently identify the presence of Pneumonia infections in X-ray images while attempting to maintain human-level error rates. I was able to accomplish this goal, resulting in a model equal to the quoted best-case error rate of trained human professionals, while being able to increase the classification rate to nearly 500 images per second. A system such as this could reasonably be deployed in the medical field to significantly increase the efficiency and even accuracy of diagnoses, leading to time, money, and potentially human lives saved. For further details and descriptions of each step please reference the included coding notebook, and feel free to contact me with any questions regarding the project.