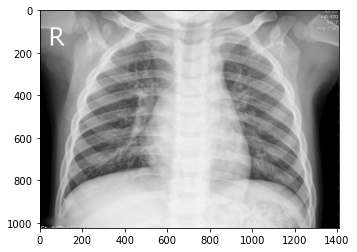
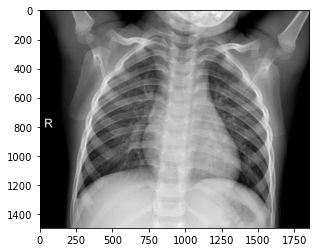
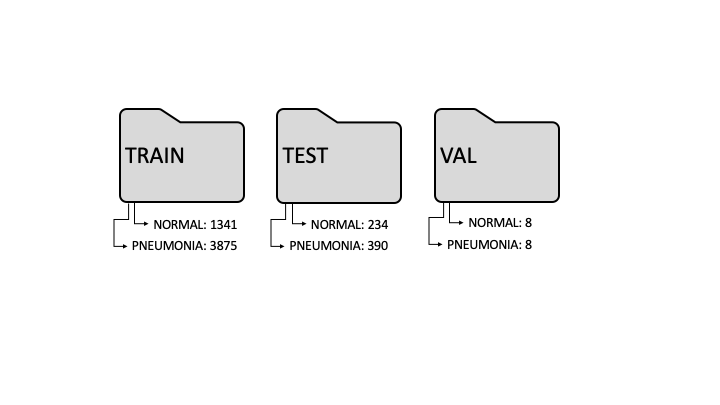
**Objective**

Create four machine learning models to decipher between x-rays of normal lungs vs. x-rays of lungs with pneumonia and see which model performs best relative to one another. The four types of classifier models are: 1) **Convolutional Neural Network (CNN)**, 2) **Gradient Boosting Classifier**, 3) **Random Forest Classifier**, 4) **k-Nearest Neighbours Classifier (KNN)**. The Kaggle *Chest X-Ray Images (Pneumonia)* dataset can be found [here](https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia).

  
*X-ray of* ***normal*** *lungs (left) and X-ray of lungs with* ***pneumonia*** *(right)*

**Data Preparation**

When extracted from the zip file, the images are already separated into *train*, *test* and *val* folders. Within each folder is a pair of sub-folders: one for NORMAL images and one for PNEUMONIA images. The image below shows the folder structure along with the number of x-ray images.

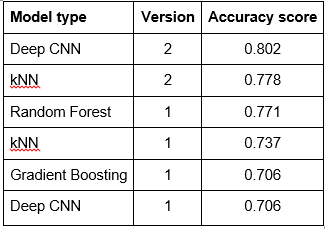
The OpenCV command *cvtColor* was applied to convert RGB images into a one-dimensional grayscale image. This conversion reduced computational complexity when the neural network was being trained.

**Model Design**

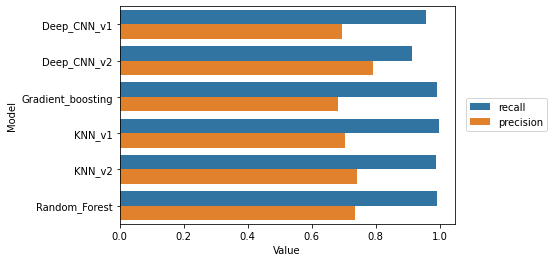
The models and their descriptions that were used to train the images can be seen in the table below:

|  |  |
| --- | --- |
| **Model type** | **Description** |
| Convolutional Neural Network (CNN) | A neural network that uses convolutional layers, where neurons are not connected to every data point in the input, but only to inputs in their “receptive fields” (Geron, 2017, p. 474). |
| Random Forest Classifier | “Is an ensemble of Decision Trees, generally trained via the bagging method (or sometimes pasting)” (Geron, 2017, p. 255). |
| Gradient Boosting Classifier | “In contrast to the random forest approach, gradient boosting works by building trees in a serial manner, where each tree tries to correct the mistakes of the previous one” (Müller, A. C., & Guido, S., 2017, p. 88). The default gradient boosting has no randomization and uses pre-pruning to work through a series of weak learners. The learning rate of a gradient boosting classifier determines the degree of correction from one weak learner to the next (Müller, A. C., & Guido, S., 2017, p. 89). |
| k-Nearest Neighbours Classifier (KNN) | Considers the *k* closest points “to the point we want to make a prediction for” (Müller, A. C., & Guido, S., 2017, p. 21). |

**Model Evaluation**

The table below lists the accuracy scores and version for each model.

Deep CNN and kNN were modified, resulting in a second version. The modification of the Deep CNN model yielded the best result, followed by kNN, as measured by accuracy.

The chart below compares the **recall** and **precision** scores of each model. All of the models do better on recall than they do on precision. Overall, this is reassuring as a false negative, in this project, is someone who has pneumonia in reality, but the model predicts that they do not. The Deep CNN v2 has the lowest recall of the models tested. This model, despite its overall accuracy, produced 34 false negatives. The cost of this increase in false negatives would ultimately be defined clinically. However, in context of this project, a recall of 0.91 is small relative to the other models.



**Conclusion**

Pneumonia detection through x-ray images could advance methods in using machine learning for medical diagnosis. By using the chest x-ray image dataset, the team has demonstrated that our classification models can, at a minimum, do better than a random classification of 0 or 1. Moreover, certain models are able to do better than others. The second version of the CNN is able to reach an accuracy greater than 80% and its accuracy score is roughly 7% better than the six-model average accuracy of 75%.

In order to achieve greater accuracy, a machine with greater graphic memory, or with a cloud service that can handle the VGG16 model is necessary, which is what inspired our CNN. Pre-trained weights could also be used to extract more detailed information from the x-rays.