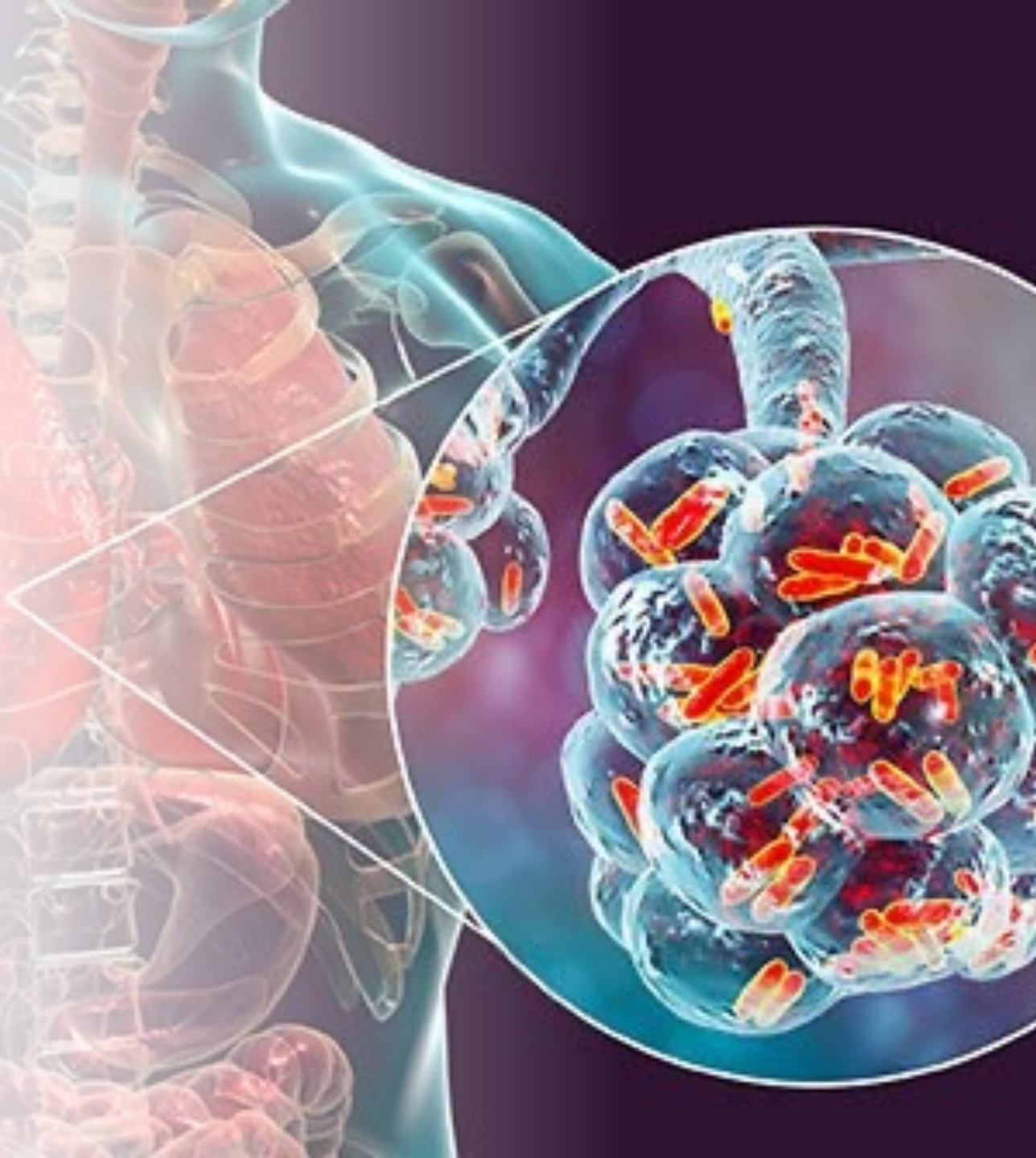
An Approach For Building an Image Classification Model for Pneumonia Detection Using Convolution Neural Networks

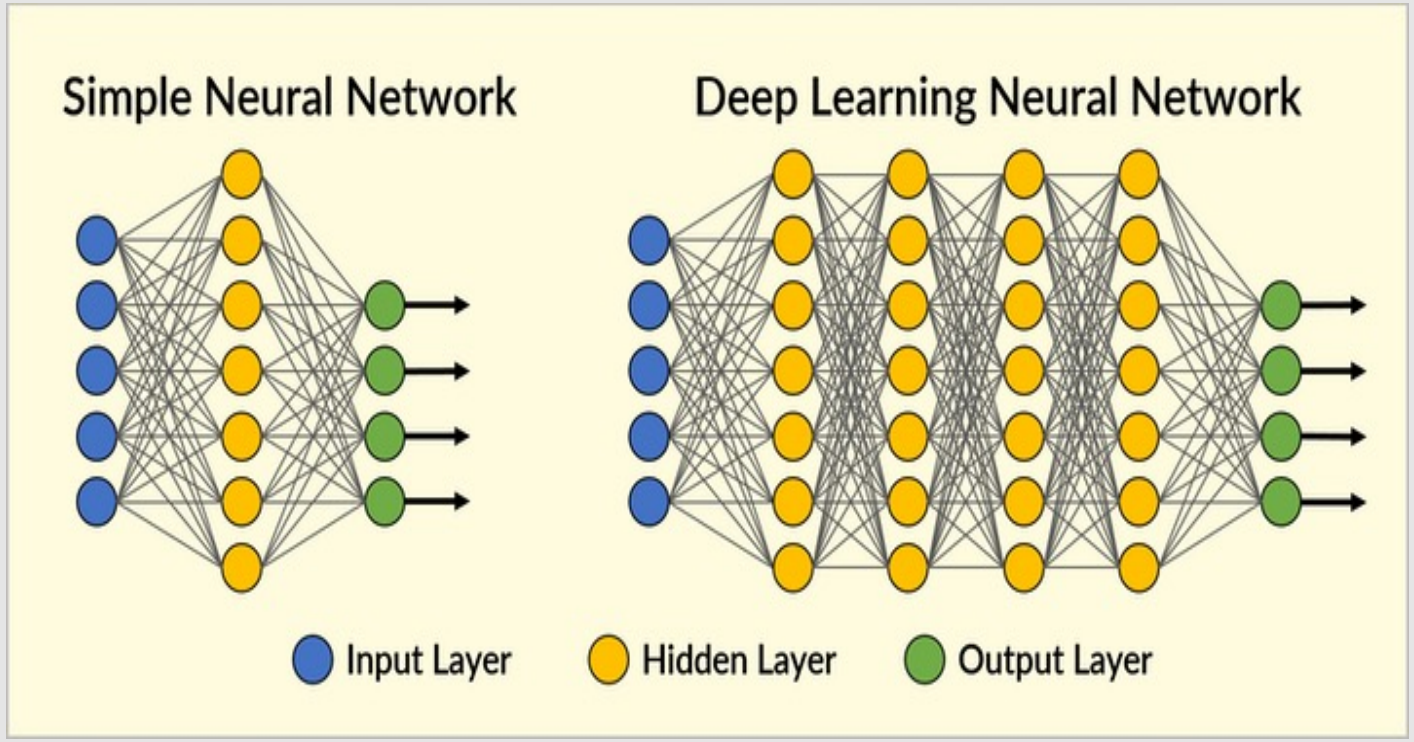
Albina Cako & Joshua Dalphy
CSML1020 Final Project
York University

Background

- Pneumonia involves inflammation of the alveoli in the lungs
- The most common symptoms: fever, headache, cough and shortness of breath
- Treatments include antibiotics (bacterial infections) or vaccinations for prevention
- For younger children, the elderly or immunocompromised patients with a pre-existing medical condition such as diabetes, kidney disease or cardiovascular disease, pneumonia can be life threatening
- Complications include lung abscess, circulatory collapse, respiratory failure, heart failure or septicemia
- Pneumonia is one of the leading causes of hospitalizations and deaths in the world
- Pneumonia is diagnosed by observing the patient's symptoms and examining a chest X-ray



Project Aim

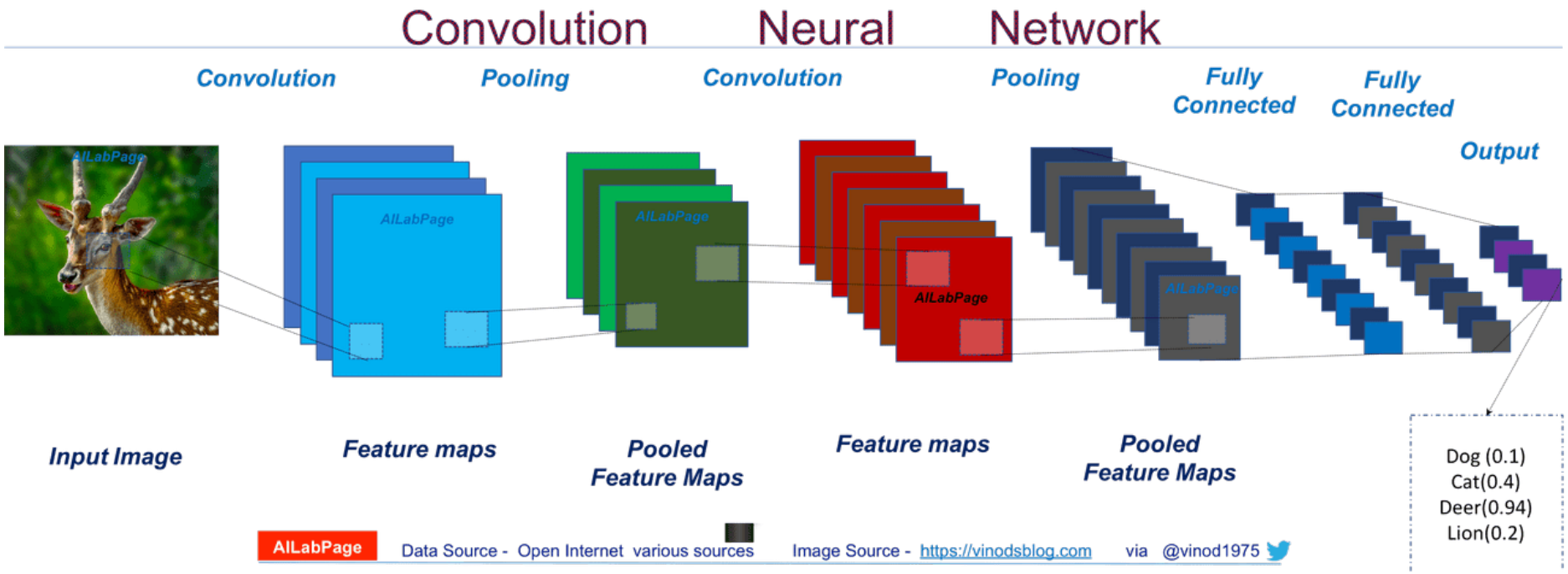


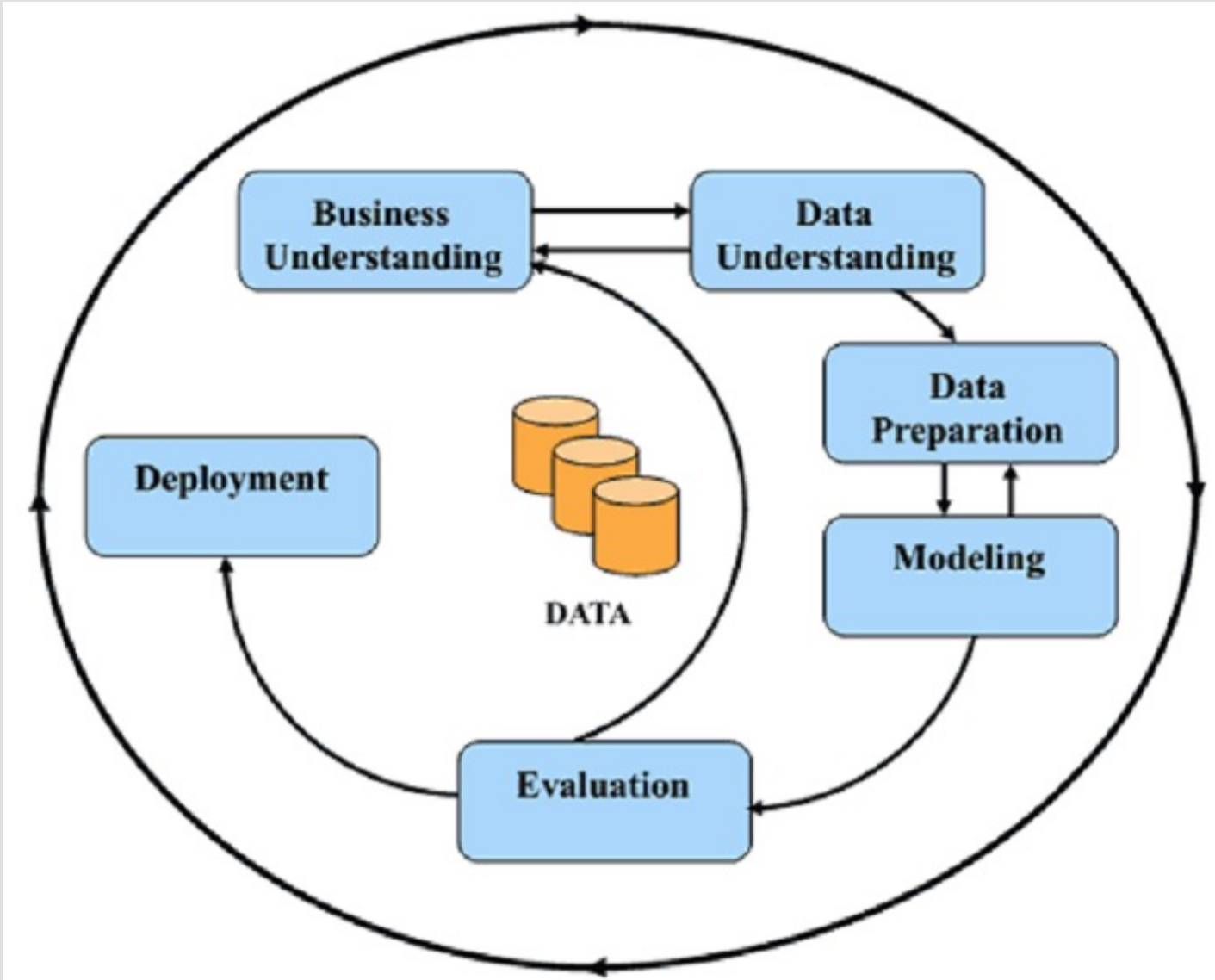
https://www.securityinfowatch.com/video-surveillance/video-analytics/article/21069937/deep-learning-to-the-rescue

- Use Convolutional Neural Network (CNN) to design a model that will classify X-ray images as “pneumonia” or “pneumonia-free”
- The model would aid doctors and medical professional in detecting pneumonia
  - Hard to detect with human eye
  - Lack of radiologists
  - COVID-19 pandemic
  - Proper allocation or resources and course of treatment

What is CNN

- Convolutional Neural Network is a deep learning algorithm used mainly for Computer Vision (image processing)





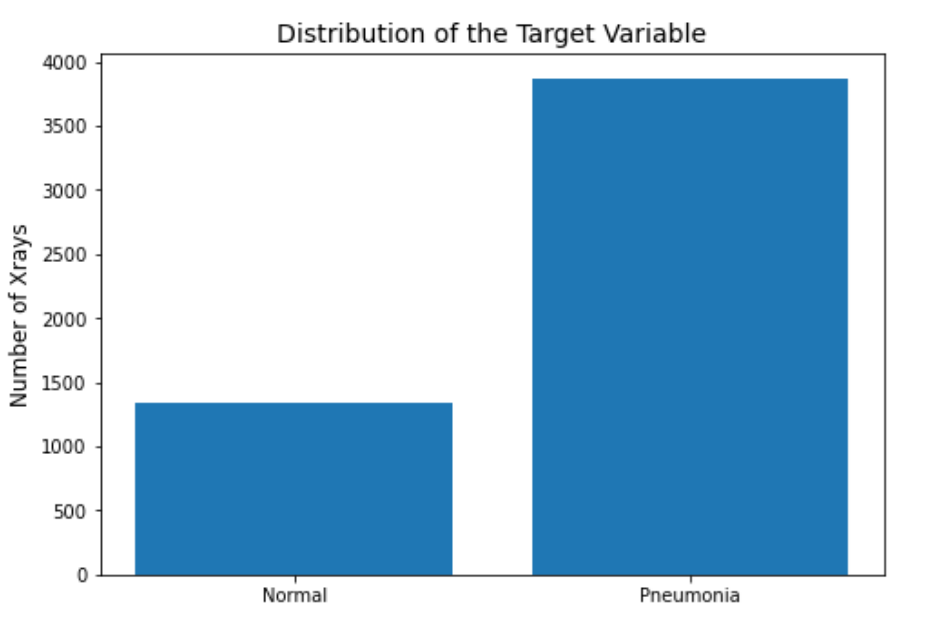
Methodology: CRISP-DM

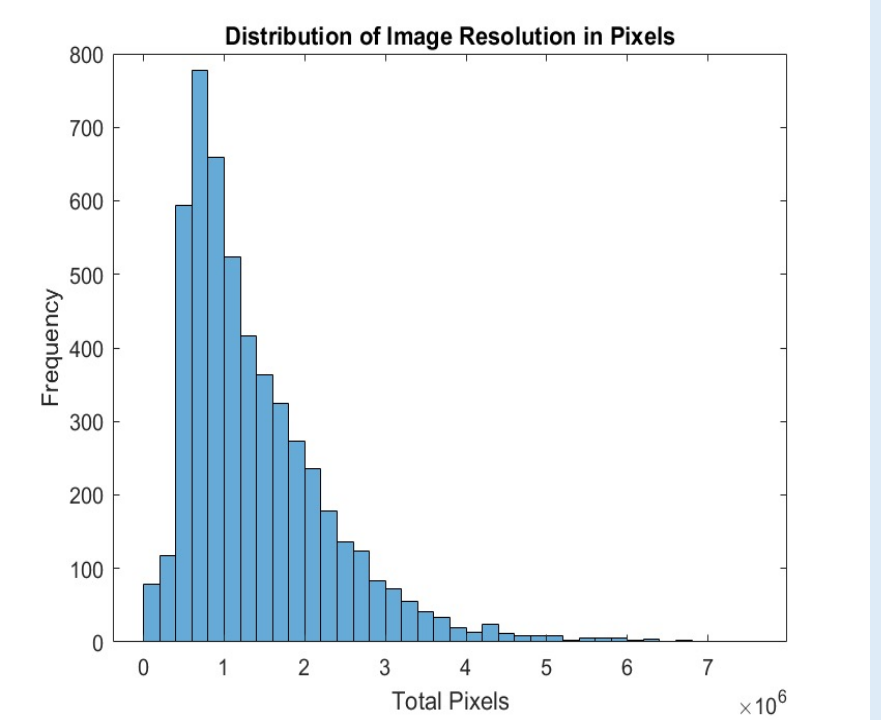
Data Exploration

Label	Training Set	Testing Set	Validation Set
Pneumonia	1341	234	8
Normal	3875	390	8
Total	5216	624	16

- Data exploration is used to further our understanding of the data and identify any underlying patterns, characteristics or features.
- The data was obtained from Kaggle and contained 3 datasets of labelled chest X-ray images.
- The data provided used a 90/10 split ratio between training and testing datasets.
- It is important to note that the validation set were not used in the analysis.

Data Exploration (cont.)



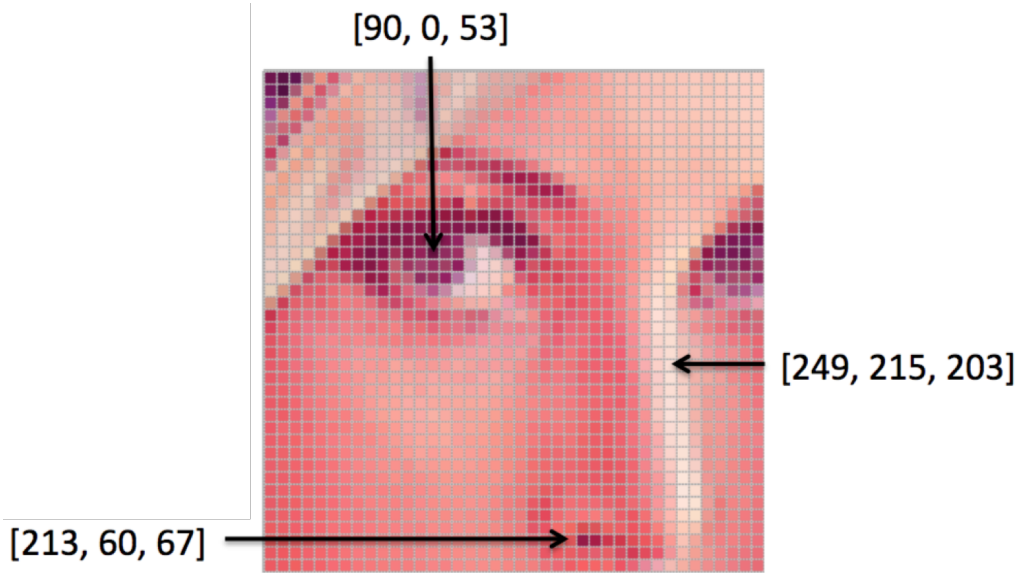


- The dataset is imbalanced
  - 76% of X-rays contain pneumonia
  - 24% of X-rays contain normal results

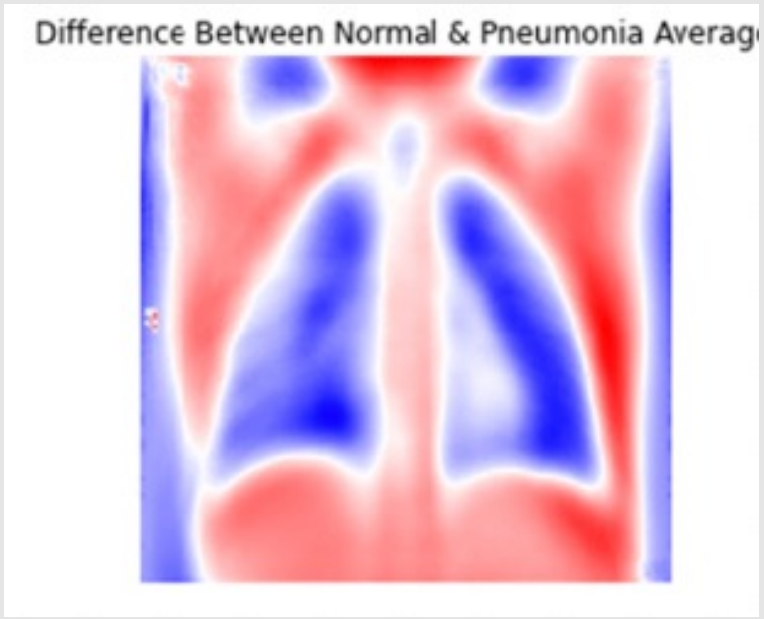
- The dataset contained images with different resolutions and sizes
  - The minimum resolution for the pneumonia class was 384x127 pixels
  - The minimum resolution for the normal class was 912x672 pixels

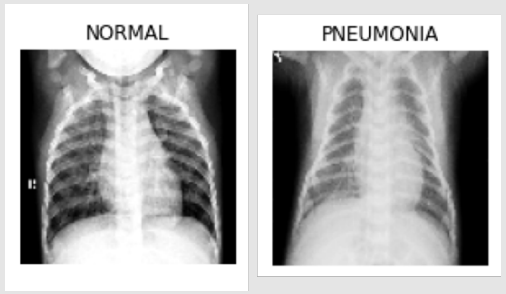
Data Preparation

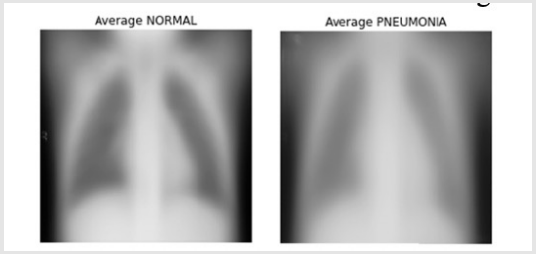
- This process varies depending on the selected modelling approach. For our study CNNs were used and thus required a 2-step preparation
    1. Resizing/Reshaping the data
    2. Normalizing the data
- Originally, the images were of inconsistent sizes and resolutions. The images were all resized to 128x128 pixels, which is a common image size
- In order to use CNNs, the images needed to be reshaped from a 2D matrix to a 1D array where each element contains a pixel value (0-255)
    - Pixel value = 0 for black
    - Pixel value = 255 for white



Data Preparation (cont.)

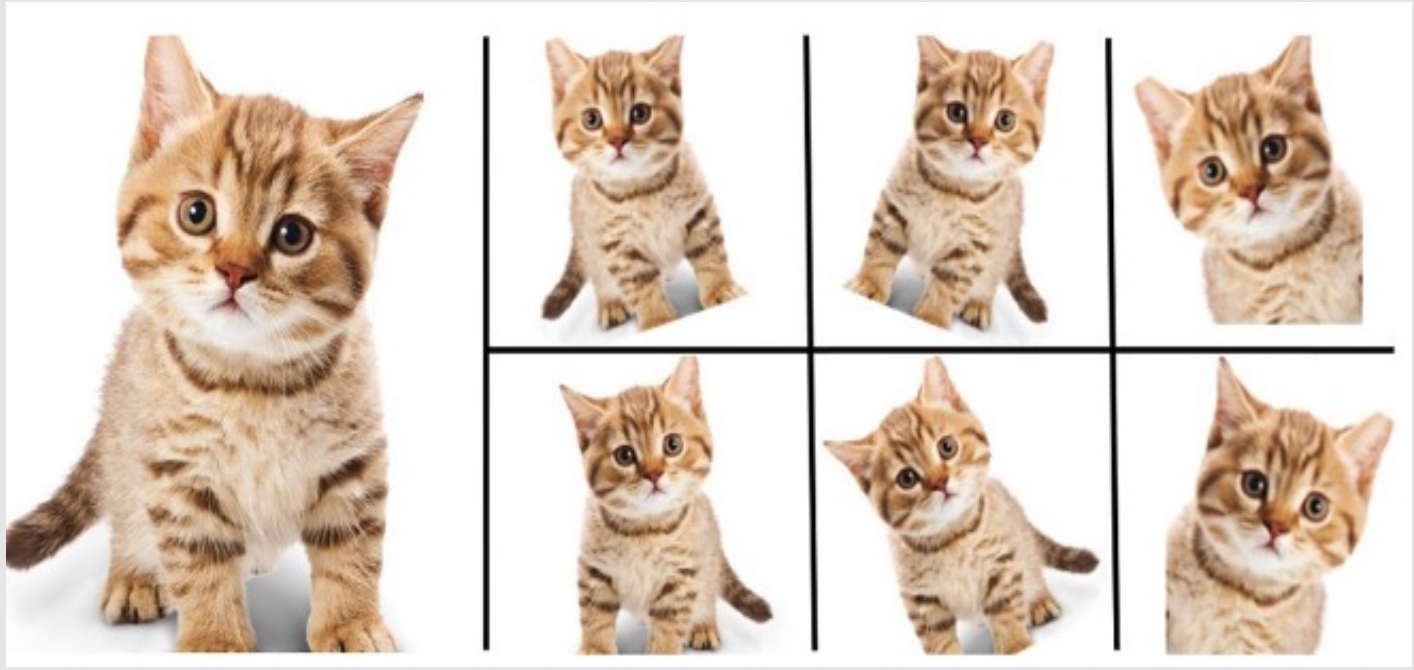






- Using the 1D array representation of the images, additional insights can be extracted from the dataset
- Min-Max normalization was applied to the dataset
- The maximum and minimum pixel values were represented by 1 and 0, respectively
- Normalization is a technique which aims make every observation in the dataset have the same scale, making each feature equally important

Data Preparation (cont.)

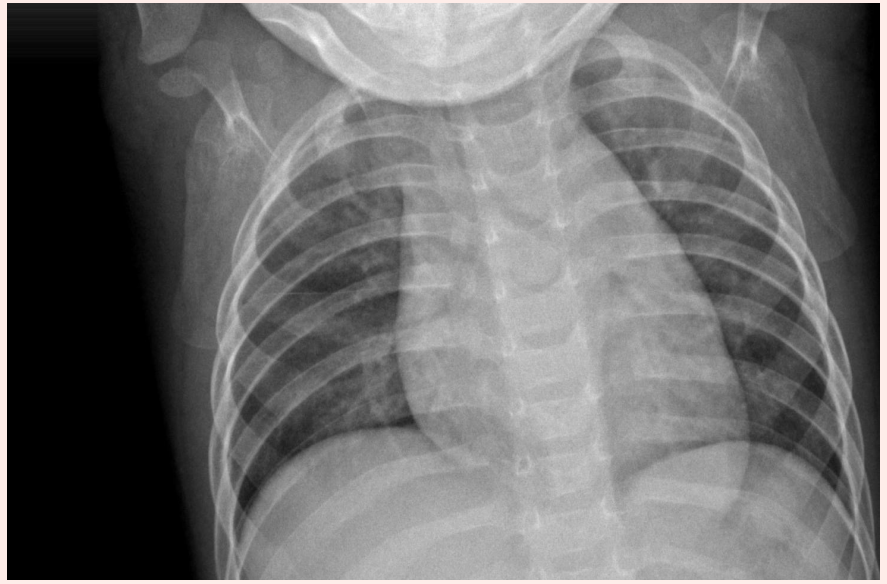


- The training set contained imbalanced data which can negatively impact the results.
- Ideally new data would be gathered; however, this was not possible. Instead, image augmentation was used.
- Image augmentation is a technique which artificially expands the size of the training dataset by creating variations of existing images.
- Increasing number of observations in the dataset:
  - results in better fitted model
  - improves the ability of fitted models to generalize what they have learned and apply to new images

Data Preparation (cont.)

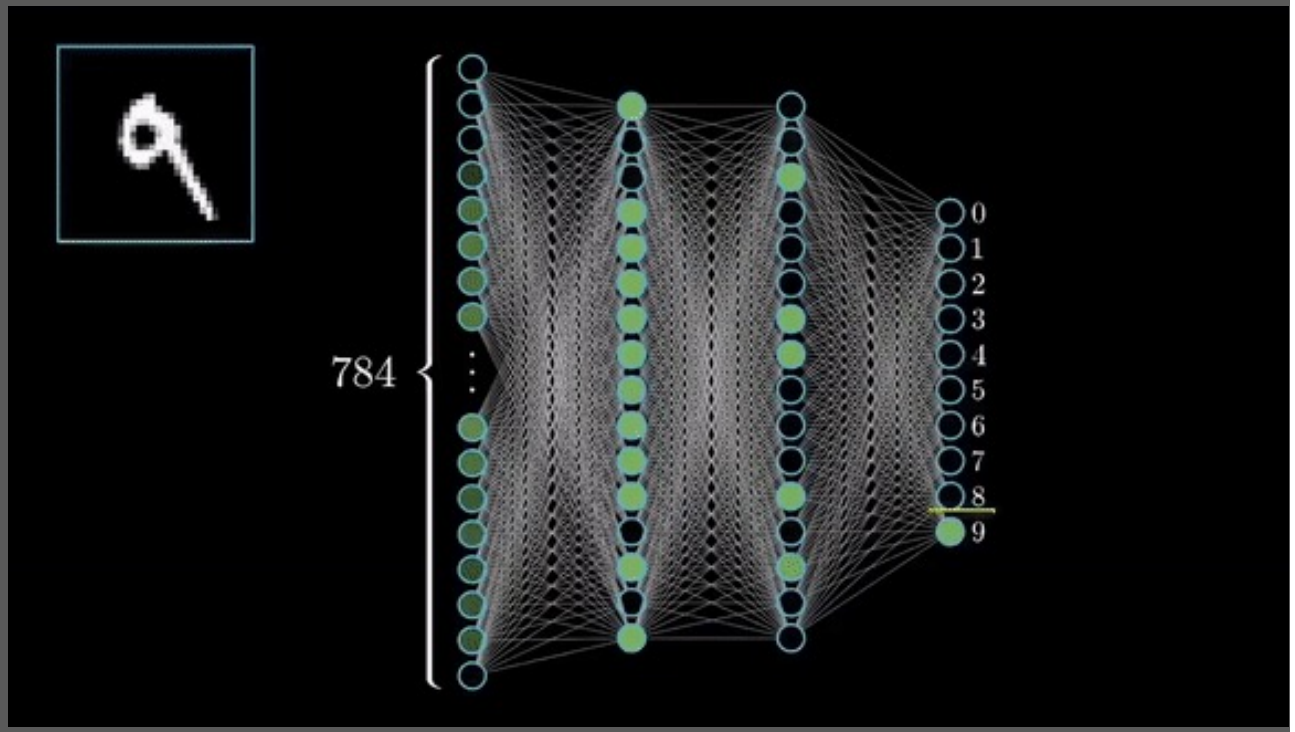
- Image augmentation was implemented using the ImageDataGenerator library in python

Parameter	Value
Rotation Range	30
Zoom Range	0.2
Width Shift Range	0.1
Height Shift	0.1
Horizontal Flip	True



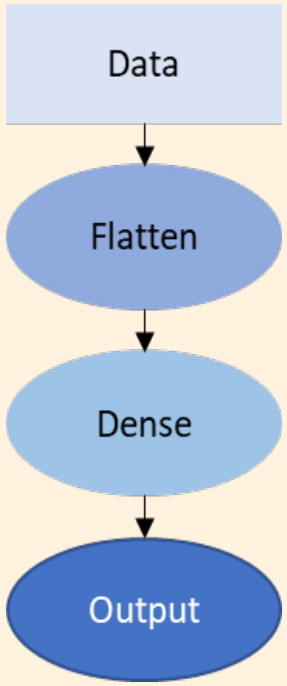
Modelling

- Modelling was based on a structured experimental approach.
- There were 2 different approaches taken for the modelling activities:
    1. Existing literature was surveyed to identify any potential CNN structures, which could be leveraged and adapted for the current application.
    2. A model was built from the ground up, starting with a simple NN and gradually adding one layer at a time and observing its impact on the results.
- Once baseline models were developed, the 2 best models were identified and further improved through hyperparameter tuning.

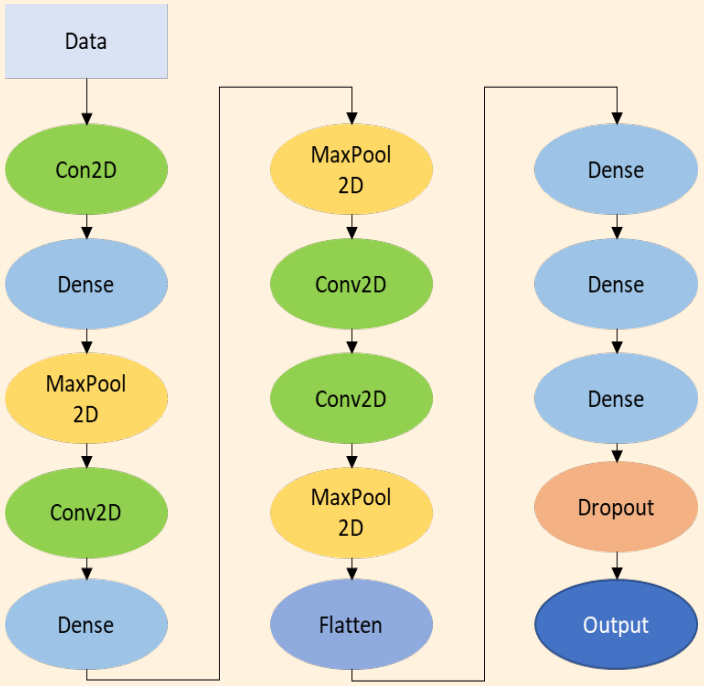


https://gfycat.com/discover/deep-learning-gifs

Modelling (cont.)



2-layer model



14-layer model

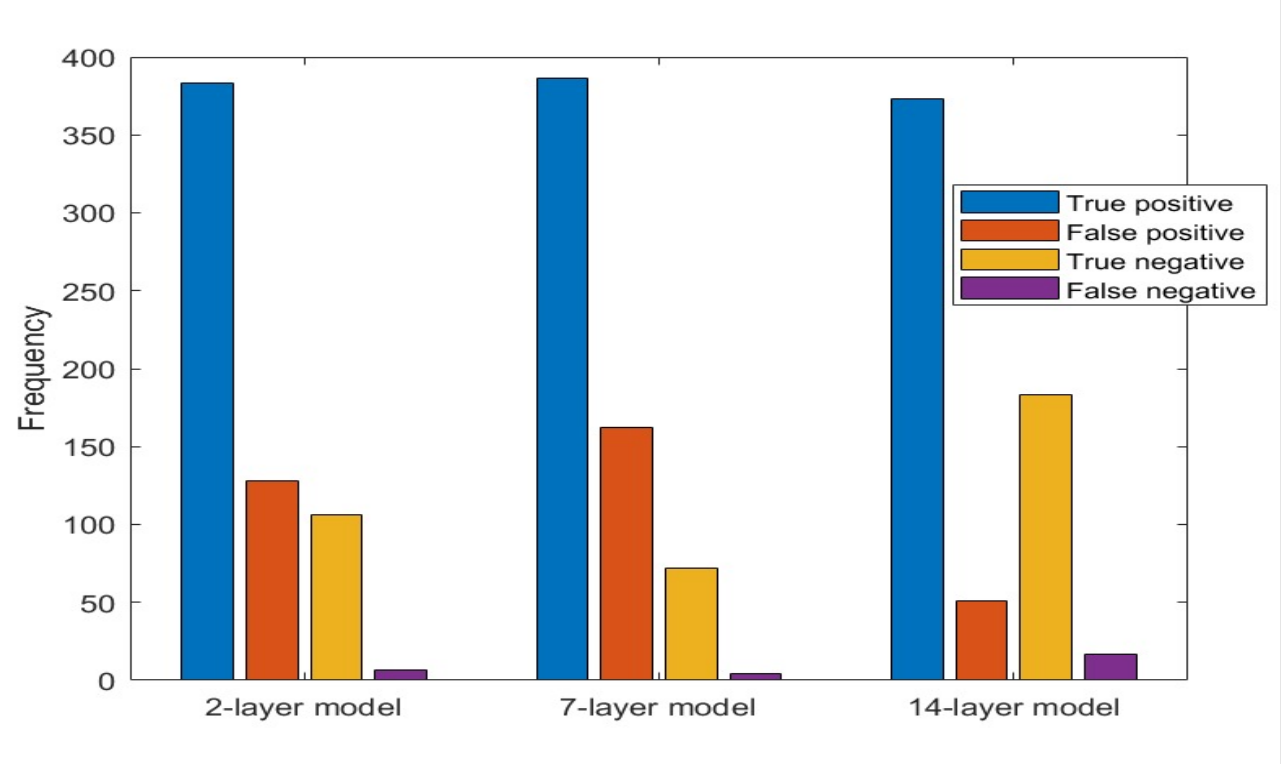
- Two top performer model architectures
- The 14-layer model has a regularization Dropout layer

Results

- The models were evaluated based on 3 criteria:
  1. Accuracy
     - To evaluate the models' overall performance
  2. Confusion Matrix
     - To evaluate the models' ability to properly predict images containing pneumonia and pneumonia-free
     - To evaluate the false negatives
  3. Learning Curves
     - To evaluate and assess each model's behavior and determine whether there was overfitting, underfitting or a good fit for the dataset

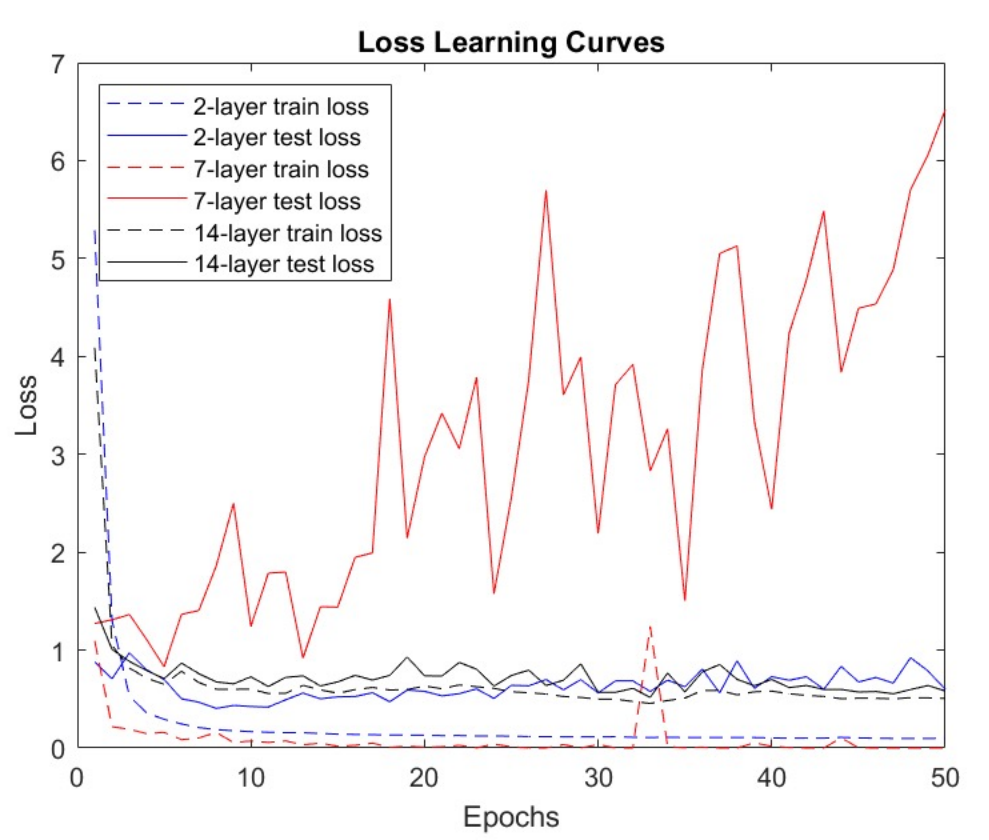
https://mytitleguy.com/2012/08/good-vs-bad-real-estate-content/

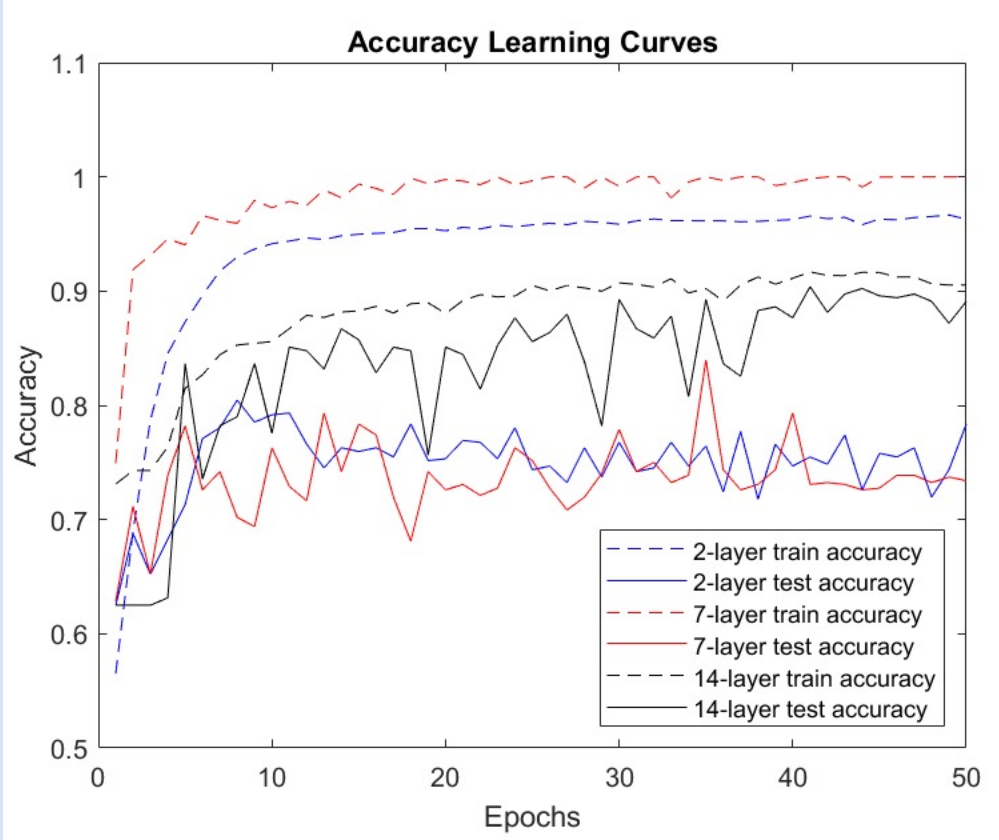
Results (cont.)



Model	Accuracy
2-Layer	78%
7-Layer	73%
14-Layer	89%

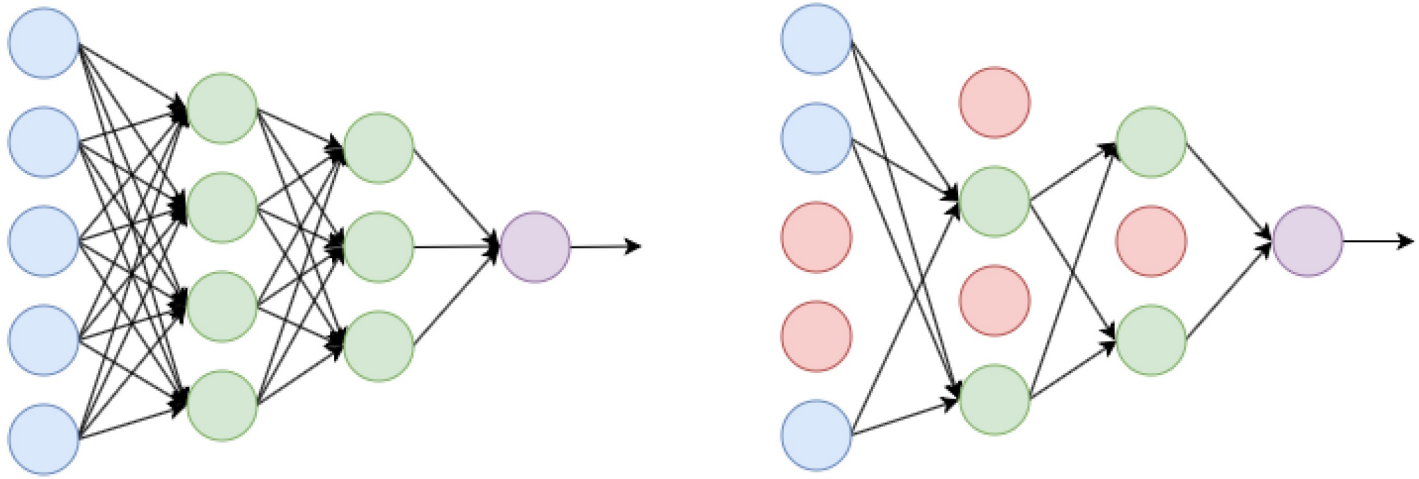
Results (cont.)



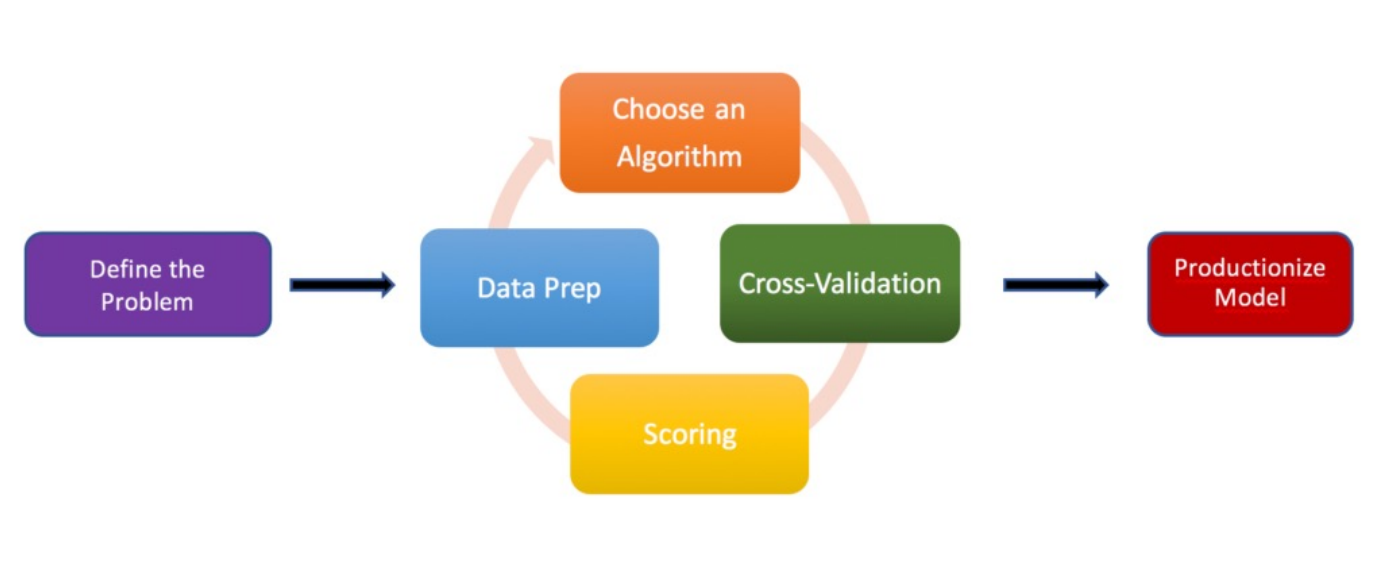


Discussion

- Based on the results examined in the previous section, it is evident that simply adding layers to a CNN and increasing model complexity, does not necessarily result in improved performance
  - Comparing the results obtained for the 2-layer and 7-layer underscores this point
- It also reinforces the concept that when building a CNN model, it's important to begin with a shallow NN and gradually add layers as need
- It is important to use a regularization method such as Dropout to reduce overfitting in a deep learning model
- It shows the importance of examining multiple metrics when evaluating model performance. Looking at multiple metrics allows for a more complete understanding of model performance and behavior
  - A model might have a high accuracy, but it could be overfitting the data, which would result in a poor overall model



Conclusions



https://randalscottking.com/machine-learning-overview/

- The 14-layer CNN model was determined to have the best performance and behavior when classifying chest X-ray images
- Machine learning lifecycle is not a linear process and even though the 14-layer model performed best, further improvement could be achieved through further hyperparameter tuning or obtaining more data
- The final step for this project would be to deploy this model into a user-friendly application

Thank you!