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

Albina Cako
School of Continuing Studies
York University
Toronto, Canada

Joshua Dalphy
School of Continuing Studies
York University
Toronto, Canada

***Abstract*—Machine learning is a technique for recognizing patterns in data that is applied in various fields, including the medical industry. Research and implementations of machine learning models are rapidly increasing in the field of medicine due to their ability to automate tasks and rapidly derive insights from data, which can aid medical professionals. One growing application for machine learning models is X-ray image classification to aid in rendering medical diagnoses such as pneumonia or even COVID-19. This paper presents a structured experimental approach, based on the CRISP-DM framework, to develop a model which could be used to detect pneumonia in chest X-ray images, using convolutional neural networks. Using this process, a 2-layer, 7-layer and a 14-layer model were developed and optimized iteratively. The models' performance and behavior were assessed using three criteria: confusion matrix values, accuracy and learning curves. The analysis determined that the best performing model was the 14-layer, achieving an overall classification accuracy of 89%.**

Keywords—Convolutional neural networks, medical, image classification, machine learning, pneumonia

I. Introduction

A. Background

Pneumonia is a medical condition that involves inflammation of the alveoli in the lungs. It is usually caused by infectious agents such as bacteria, viruses or fungi. The most common symptoms of pneumonia are fever, headache, cough and shortness of breath. Effective treatments include antibiotics for treatment of bacterial infections or vaccinations for prevention. Pneumonia usually resolves with treatment in a couple of weeks, however, for some people it can take months to fully recover. Nevertheless, this is not always the case. 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 which can occur as a result of pneumonia include lung abscess, circulatory collapse, respiratory failure, heart failure or septicemia [2]. Pneumonia is one of the leading causes of hospitalizations and deaths in the world, even though there have been many medical advances to its treatment [1].

Pneumonia is diagnosed by observing the patient's symptoms and examining a chest X-ray. The X-ray is taken to determine how much of the lung tissue is inflamed, which aids doctors to define a course of action for treatment and asses whether hospitalization is needed. Currently, in 2021, the world is experiencing a pandemic brought about by the spread of the COVID-19 virus. Covid-19 causes pneumonia and is currently a major worldwide concern, as it has already caused over 2 million deaths in just over a year [3]. Hospitals are overloaded with patients and shortage of doctors and medical staff is a challenge which is affecting countries all over the world. Shortages of medical supplies and accommodations has created a demand for better detection systems to identify infections, which can aid hospitals with resource allocation.

B. Project Aim

The purpose of this paper is to present a model which could be used to classify chest X-ray images as either containing pneumonia or being pneumonia-free using a Convolutional Neural Network (CNN). CNN is a deep learning algorithm used mainly for image classification and was selected for this project due to its wide success in image classification problems [4]. The application and intended use of the model presented in this paper would be to aid doctors and medical professionals in detecting pneumonia, especially in cases where it might be difficult to assess X-ray results with the human eye. Additionally, in situations or locations where radiologists are not available or the hospital is overwhelmed with patients, the model could be used by medical doctors to rapidly assess whether a patient has pneumonia and decide on a treatment course. Having such a resource, especially during a pandemic such as COVID-19, can drastically improve resource allocation and treatment outcomes.

C. Related Work

A recent study used CNN to create a model for pneumonia X-ray image classification. The paper achieved a test accuracy of 98.43%. The researchers used transfer learning to fine tune the model in order to improve test and train accuracy. The study used the Guangzhou Women and Children's Medical Center pneumonia dataset [5].

D. Ethical Considerations

It is important to note that the model presented in this paper is for research purposes and to fulfill the requirements of the CSML1020 course. It should not be used as a replacement for medical advice.

II. Problem Statement and Dataset

A. Problem Statement

As previously stated, the objective of this paper was to build a model which could be used detect pneumonia in X-Ray images, using a CNN. The model's performance will be primarily assessed based on accuracy, confusion matrix and learning curves. Overall, successful detection of pneumonia can help doctors make better decisions in allocating their resources, especially during the COVID-19 pandemic.

B. Dataset

The dataset used for this project was obtained from Kaggle [6]. The dataset contains chest X-ray images which are classified as either NORMAL or PNEUMONIA. Each Image has a unique id to serve as an identifier. Images which are classified as NORMAL represent a healthy individual's chest X-ray, while PNEUMONIA indicates an X-ray of an individual who has been diagnosed with pneumonia. The training set contained 1341 NORMAL images and 3875 PNEUMONIA images, while the testing set contained 234 NORMAL images and 390 PNEUMONIA images. A validation set was also provided; however, it contained an insignificant number of images and was not used for this project.

III. Methods and Models

In this section, the approach used to explore and prepare the dataset will be discussed. Then, the methodology employed to build and optimize the model will be presented. This project implemented a structured experimental approach based on the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework. CRISP-DM is a standard process model used to help plan, organize and implement data mining projects [7]. The iterative process consists of six main steps, shown in Figure 1. Each step will be addressed in this paper.

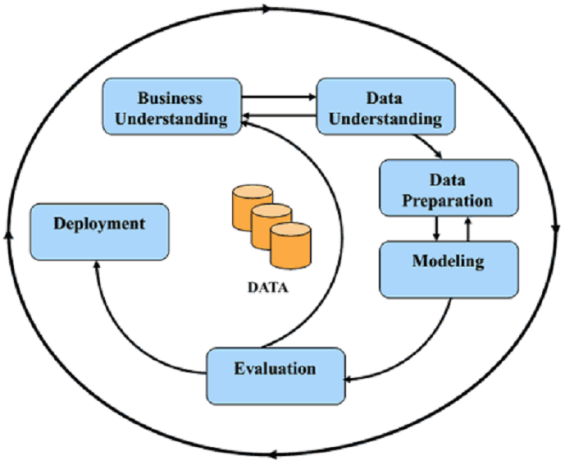

Fig. 1. The CRISP-DM Framework [7]

A. Data Exploration

The business understanding has already been addressed in the Section I. Data exploration was undertaken to aid in understanding the X-ray image dataset and to identify any potential patterns, characteristics or relevant features which could be used to better understand the problem statement, as well as aid during the modelling activities. The approach taken to explore the dataset was based on the methodology presented in [8]. Figure 2 shows two sample X-rays, one for a patient with pneumonia and another for a patient without. All images were grey scale, as shown.

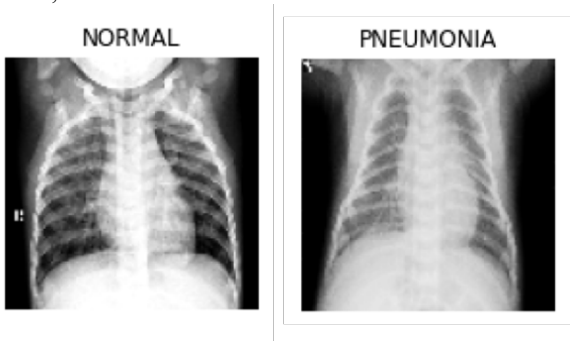

Fig. 2. Sample chest X-Ray images for a patient without pneumonia (left) and with pneumonia (right).

As previously stated, the dataset used for this project was sourced from Kaggle [6] and contained 5,856 labeled images, subdivided into training, testing and validation datasets. The images were classified as either NORMAL or PNEUMONIA. For data exploration and modelling activities, these labels were changed to binary values, where 0 represented a normal X-ray and 1 indicated an X-ray which was positive for pneumonia. A detailed breakdown of the datasets is provided in Table I.

TABLE I. SUMMARY OF THE CHEST X-RAY IMAGEDATASET

Label	Training Set	Test Set	Validation Set
Normal	1,341	234	8
Pneumonia	3,875	390	8
Total	5216	624	16
% of Total Data	89.1%	10.7%	0.2%

From Table 1, the split ratio between the training and testing set is approximately 90/10 and observing the breakdown of the training set, it can be seen that the data is imbalanced with approximately 74% of the images belonging to the pneumonia class and 26% corresponding to the normal class. Prior to modelling, the data imbalance will need to be addressed. The former will be discussed further on in the report during data preparation. During data exploration, it was observed that the images in each dataset were of varying sizes. Image size is a characteristic used to describe the height and width of an image, in pixels. Additionally, the image size can also be described by the total amount of pixels in an image [9]. Figure 3 illustrates the distribution of the total amount of pixels per image in the training dataset.

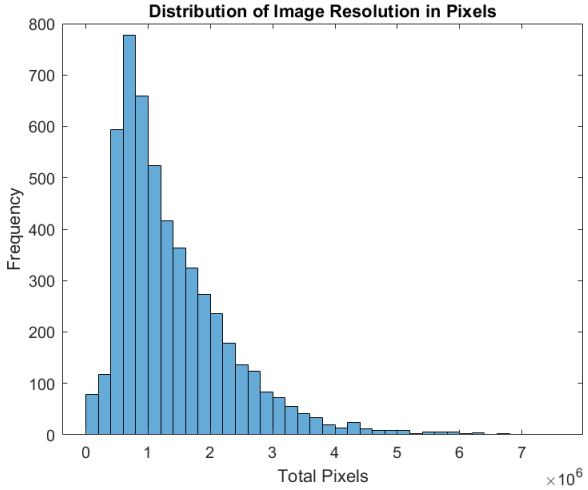

Fig. 3. Distribution of image resolution represented in total pixels per image

Using the data presented in Figure 3 and a custom python function, it was determined the minimum image size for the normal and pneumonia classes were 912x672 and 384x127 pixels, respectively. Initially, the intent was to resize all images in the datasets to the smallest size identified (384x127). However, following an initial test, it was observed that the majority of the resized images were left quite distorted and blurry, which made distinguishing whether the image belonged to the normal or pneumonia class difficult. Instead, the images were resized to 128x128 pixels, which is a common image size. It is important to note that by reducing an image's size, there is the possibility that there might be some loss of detail, which could impact the results. Additionally, by reducing the size of the image, there are less features present while training a model, which could also affect the results.

B. Data Preparation

The next step in the CRISP-DM framework is data preparation, which involves cleaning and formatting the data. This step varies based on the selected modelling approach. For this paper, three different CNN models will be presented and compared. The CNN models required two preparation steps: resizing/reshaping the images and normalization of the data. For the reshaping process, the images are converted from 128x128 pixels to a one-dimensional array representation. The newly created array has shape of 1x16,384, where each element contains a pixel value ranging from 0-255, with 0 representing the color black and 255 representing white, for a greyscale image. Using the pixel values, additional insights can be obtained from the dataset such as the average image. The average image for both the normal and pneumonia classes as well as the difference between them is shown in Figure 4.

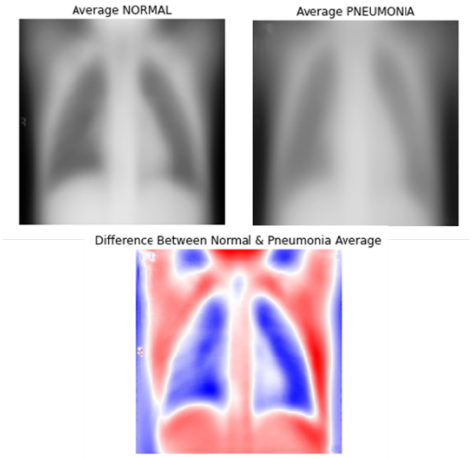

Fig. 4. Average Chest X-Ray Image

Normalization is a data preparation technique which aims to make every observation in the dataset have the same scale, making each feature equally important [10]. Min-max normalization was applied to the X-ray image dataset, resulting in the maximum and minimum pixel values being represented by 1 and 0, respectively. To achieve this, each image pixel was divided by 255. As previously mentioned, the training dataset is imbalanced and had to be addressed prior to modelling. Ideally, new data would have been collected to address this issue. However, this was not possible given the short time span of the project and the available resources. Instead, image data augmentation was used. The former is a technique used to artificially expand the size of the training dataset by creating variations of existing images [11]. By increasing the number of observations in the dataset, this can result in better fitted models and improve the ability of the fitted models to generalize what they have learned and apply it to new images [11]. The Keras python library's ImageDataGenerator class was used to implement image augmentation and the parameters that were used are summarized in Table II.

TABLE II. DATAGEN PARAMETER LIST

Parameter	Value
Rotation range	30
Zoom range	0.2
Width shift range	0.1
Height shift	0.1
Horizontal flip	True

C. Modelling

The next step in the CRISP-DM framework is modelling and a structured experimental approach was used to develop the selected model architecture. Initially, existing literature and models were reviewed to identify commonly used structures and methodologies for CNN model development. The surveyed literature showed that when developing CNN models, it is important to start with a shallow neural network (NN) and gradually add layers as needed. This approach is suggested to prevent overfitting, as using an unnecessarily complicated model structure can lead to the model performing well on the training set, but poorly on the testing set. The first model, shown in Figure 5, used was a 2-layer neural network.

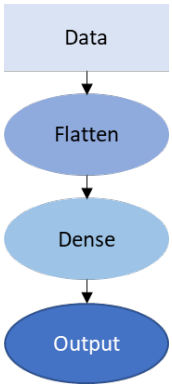

Fig. 5. 2 -layer CNN model architecture used

In [12], the article describes an approach to creating a CNN model used for X-Ray image classification and was used as the basis for the final model presented in this paper. The 14-layer CNN model's structure is shown in Figure 6.

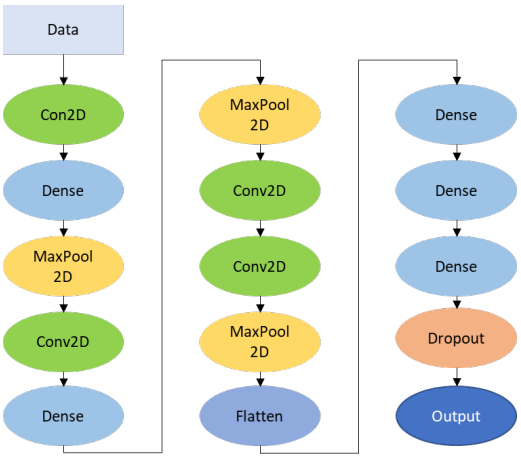

Fig. 6. 14-layer CNN model architecture used

It is important to note that this model contains a Dropout layer. Dropout is a regularization technique that randomly shuts down some nodes on the network. This allows to spread out the weights of the features and helps prevent overfitting.

The process used to develop these models was iterative and involved starting with a simple structure (Figure 5) and gradually adding one layer at a time and observing its effect on the model's accuracy and loss function. Once complete, the two best models were identified and further fined-tuned by hyperparameter tuning to determine the best parameter combinations. Additionally, to highlight the fact that greater model complexity does not always result in better performance, an additional model which contains seven layers is presented.

IV. RESULTS AND DISCUSSION

A. Results

During modelling, several model structures were developed and trained iteratively. Initially, the models were evaluated solely on accuracy. After initial training, three models were chosen for further investigation: the 2-layer, 7-layer and 14-layer models. To evaluate model performance, the number of images that were correctly identified were analyzed using a confusion matrix, whose results are illustrated in Figure 7, shown below.

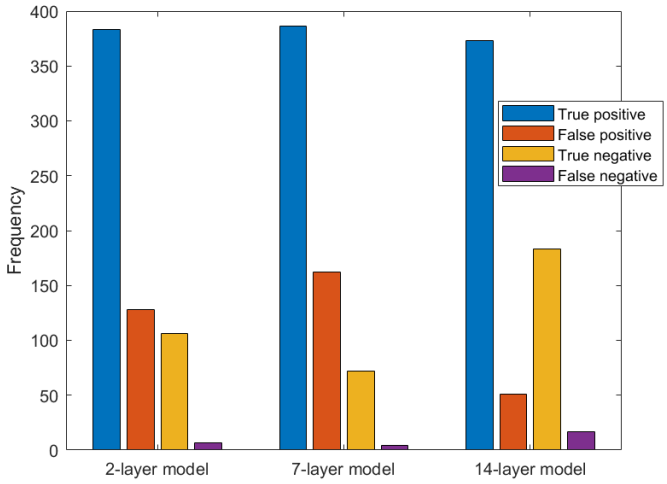

Fig. 7. Confusion Matrix Comparison

Observing Figure 7, all three models performed similarly when classifying images which contained pneumonia (true positives), with the 2-layer and 7-layer models slightly outperforming the 14-layer model. When classifying images which did not contain pneumonia (true negatives), the 14-layer model significantly outperformed both the 2-layer and 7-layer models. Given that the intended application of these models is in the medical industry, an important element to consider is the number of false negatives produced because in the real world, failing to detect or misdiagnosing an illness can have severe consequences, including the patient's health worsening or in certain cases death. All three models produced a small number

of false negative values, with less than 3% of the predicted results categorized as false negatives.

The second metric that was used to evaluate the model performance was the accuracy, which quantifies how well the models were able to predict the results from the testing set. The three models' accuracy are summarized in Table III.

TABLE III. MODEL ACCURACY

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

The results summarized in Table III show that the best performing model was the 14-layer architecture with an accuracy of 89%, followed by the 2-layer at 78% and the lowest accuracy was obtained by the 7-layer model, which was 73%.

The last evaluation metrics used to assess the performance of the models were learning curves, which are used to demonstrate a model's performance over time. Learning curves are a useful diagnostic tool used to assess whether a model is biased (underfitting), has high variance (overfitting) or is a good fit for the dataset (low bias and low variance). In this paper, two learning curves were created, the first used accuracy as a metric, and the second used loss. Both plots are shown in Figures 8 and 9.

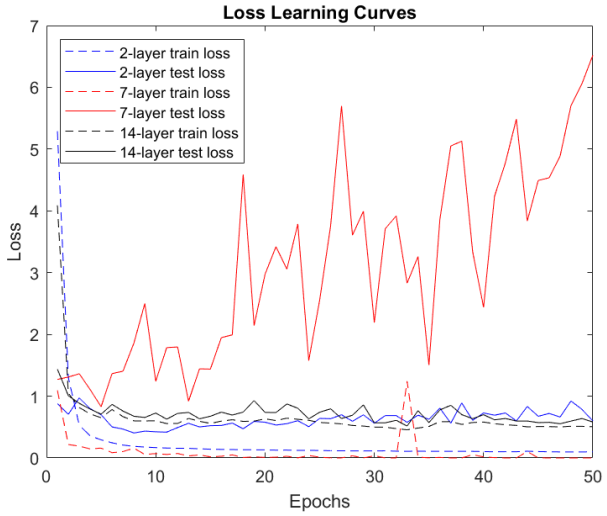

Fig. 8. Loss Learning Curves

Observing the learning curve for the loss of the 7-layer model, the training loss steadily decreases as the number of epochs increases, while the validation loss does initially decrease, it quickly begins to increase and diverges from the training loss curve. This indicates that the 7-layer model is overfitting and will not perform well on unseen data. A good fitting model is characterized by seeing the training and validation loss decrease to a point of stability with minimal difference between the two [13]. The learning curves for both the 2-layer and 14-layer models indicate that the model is a good fit, with the 14-layer being the best fit to the dataset. Lastly, the learning curve for the accuracy, which indicates how the model improves during training, was plotted. An ideal curve will show a model's increase over time (in this case, with each epoch), until it reaches a maximum accuracy. The learning curves for model accuracy are shown below.

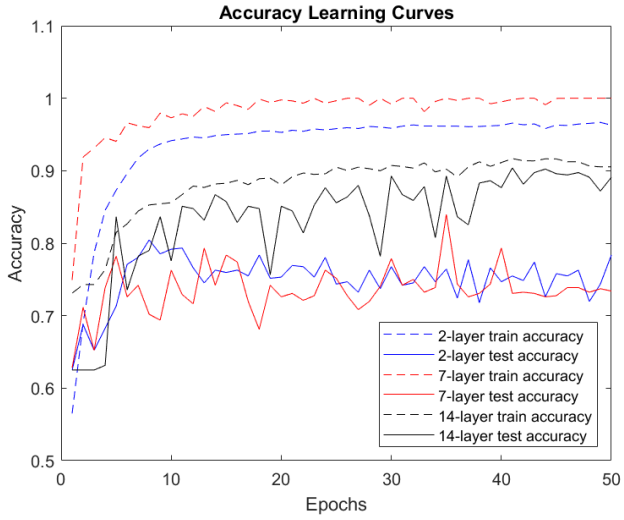

Fig. 9. Accuracy Learning Curves

Observing Figure 9, the 7-layer model ultimately achieved an accuracy of approximately 100% on the training set. Additionally, the model did not perform as well on the testing set, obtaining an accuracy of approximately 73%. Observing a high accuracy on the training set and low accuracy on the testing set, indicates that the 7-layer model is overfitting and can be considered as a poor performing model. The 2-layer model achieved an accuracy of approximately 95% on the training set while obtaining an accuracy of approximately 78% on the testing set. Like the 7-layer model, this behavior could be an indication of overfitting. However, the 2-layer model performed slightly better then the 7-layer model. Lastly, observing the learning curve for the 14-layer model, the accuracy observed on the training set plateaued at approximately 90%. The accuracy obtained for the testing set was determined to be approximately 89% and upon further inspection, the learning curves for both the training and testing sets have a similar profile, which indicates that the 14-layer model is a good fit to the dataset. In addition, it is important to note that the added Dropout layer could be one of the factors that has allowed for the model to not overfit the data.

B. Discussion

This paper presents a comparison of three different CNN models developed to classify X-ray images as either containing pneumonia or being pneumonia-free. 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. Additionally, the work presented in this paper shows the importance of examining multiple metrics when evaluating model performance. Looking at metrics such as accuracy, confusion matrix values and learning curves allows for a more complete understanding of model performance and behavior. Solely relying on a single metric is often not enough to understand the full picture as a model might have a high accuracy, but it could be overfitting the data, which would result in a poor overall model.

Based on the chosen evaluation metrics, the 14-layer CNN model was determined to have the best performance and behavior when classifying chest X-ray images. It is important to note that the 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 and obtaining more data. In addition, it is important to use regularization in deep learning models, such as Dropout, in order to reduce the chance of overfitting. The final step for this project would be to deploy this model into a user-friendly application. The application could then be used to aid medical staff in diagnosing pneumonia. It is important to note that this application should not replace medical advice.

V. IMPLEMENTATION AND CODE

The code and data needed to replicate this study can be found in the following GitHub repository https://github.com/cako-dalphy/deep_learning.

REFERENCES

[1] Pneumonia: Overview. (2018). Retrieved March 16, 2021, from https://www.ncbi.nlm.nih.gov/books/NBK525774/

[2] Mattila, J. T., Fine, M. J., Limper, A. H., Murray, P. R., Chen, B. B., & Lin, P. L. (2014). Pneumonia. Treatment and diagnosis. *Annals of the American Thoracic Society*, *11 Suppl 4*(Suppl 4), S189–S192. https://doi.org/10.1513/AnnalsATS.201401-027PL

[3] Coronavirus cases:. (n.d.). Retrieved March 16, 2021, from https://www.worldometers.info/coronavirus/

[4] Yamashita, R., Nishio, M., Do, R., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, *9*(4), 611–629. https://doi.org/10.1007/s13244-018-0639-9

[5] Hashmi, M. F., Katiyar, S., Keskar, A. G., Bokde, N. D., & Geem, Z. W. (2020). Efficient Pneumonia Detection in Chest Xray Images Using Deep Transfer Learning. *Diagnostics (Basel, Switzerland)*, *10*(6), 417. https://doi.org/10.3390/diagnostics10060417

[6] Mooney, P. (2018, March 24). Chest x-ray Images (Pneumonia). Retrieved March 16, 2021, from https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

[7] Manasson, A. (2020, February 25). *Why using CRISP-DM will make you a better Data Scientist*. Medium. https://towardsdatascience.com/why-using-crisp-dm-will-make-you-a-better-data-scientist-66efe5b72686

[8] Byeon, E. (2020, October 15). *Exploratory Data Analysis Ideas for Image Classification*. Medium. https://towardsdatascience.com/exploratory-data-analysis-ideas-for-image-classification-d3fc6bbfb2d2

[9] *Image Size*. (2017, February 9). Canon Australia. https://www.canon.com.au/explore/glossary/image-size

[10] *Normalization*. (n.d.). Codecademy. https://www.codecademy.com/articles/normalization

[11] Brownlee, J. (2019, July 5). *How to Configure Image Data Augmentation in Keras*. Machine Learning Mastery. https://machinelearningmastery.com/how-to-configure-image-data-augmentation-when-training-deep-learning-neural-networks/

[12] Mohebban, S. (2020, July 20). *Image Classification: Using AI to Detect Pneumonia - Towards Data Science*. Medium. https://towardsdatascience.com/using-ai-to-detect-pneumonia-3ec4601acd07

[13] Brownlee, J. (2019, August 6). *How to use Learning Curves to Diagnose Machine Learning Model Performance*. Machine Learning Mastery. https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/