

AUTOMATIC CLASSIFICATION OF PNEUMONIA USING DEEP LEARNING

JEHANZEB MOBARIK



A REPORT SUBMITTED AS PART OF THE REQUIREMENTS FOR THE DEGREE
OF BSc IN COMPUTER SCIENCE
AT THE SCHOOL OF COMPUTING
ROBERT GORDON UNIVERSITY
ABERDEEN, SCOTLAND

May 2020

Supervisor Dr. Eyad Elyan

Abstract

Deep learning over the years has shown to produce human level performance across different computer vision tasks. In this project we explore the possibility of applying deep learning techniques to diagnose chest X-rays that show pneumonia infected lungs. We developed a deep learning pipeline using techniques researched in the literature review and pushed the model performance via an iterative process to maximise metrics like recall, specificity, precision and F1-score. We ultimately found that there is a long way to go before deep learning methods are used in the clinical settings, but new frontiers in AI and medicine are rapidly developing and ultimately will overcome a lot of the challenges discussed in this paper.

Acknowledgements

Special thanks must go to my Honours supervisor Dr. Eyad Elyan for his inspiration, guidance and advice during this project. Thanks also goes to all the RGU staff from lecturers, lab assistants and the school office members. I am particularly grateful to my family for their love and support for the past four years, especially my mum.

Declaration

I confirm that the work contained in this BSc project report has been composed solely by myself and has not been accepted in any previous application for a degree. All sources of information have been specifically acknowledged and all verbatim extracts are distinguished by quotation marks.

Signed: Jehanzeb Mobarik

Date: 25/05/2020

Contents

Abstract	2
Acknowledgements	3
Declaration	4
1 Introduction	1
1.1 Overview	1
1.2 Aims and Objectives	1
1.3 Chapter List	2
2 Literature Review	3
2.1 Overview	3
2.1.1 Faster Diagnosis	3
2.2 Traditional Methods of Machine Learning	4
2.3 State-of-the-Art	5
2.4 Related Work	8
2.4.1 CheXNet	8
2.4.2 Multi-task Learning for Chest X-ray Abnormality	9
2.4.3 Abnormality Detection and Localization in Chest X-Rays using Deep Convolutional Neural Networks	12
2.5 Analysis of Related Work	15
2.6 Medical Images	16
2.6.1 Challenges	16
2.7 Summary of Review	17
2.8 Moving Forward	17
3 Requirements Analysis	18
3.1 Deep Learning Model	18
3.1.1 Functional Requirements	18

3.1.2	Non-Functional Requirements	19
3.2	Web App for visualisation	20
3.2.1	Functional Requirements	20
3.2.2	Non-Functional Requirements	20
3.3	Methodology	21
4	Experiment Setup	22
4.1	Hardware Implementation	22
4.2	Software Implementation	23
4.3	Dataset	23
4.3.1	Dataset Preprocessing	25
4.3.2	Normalising Images	26
4.4	Going Beyond Accuracy	26
5	Experiments & Results	28
5.1	Experiment 1	28
5.1.1	Improving Experiment 1	31
5.2	Hyper-Parameter Tuning	33
5.2.1	Learning rate	33
5.2.2	Adding Convolution and Fully Connected Layers	36
5.3	Transfer Learning	39
5.4	Augmentation	42
5.5	Summary of Results	45
5.6	Model deployment & Interface	46
6	Evaluation	48
6.1	Overview	48
6.2	Implementation Discussion	49
6.3	Legal, Social, Ethical, Security and Professional Considerations	50
6.3.1	Legal	50
6.3.2	Security	51
6.3.3	Ethical	51
7	Conclusion	51
7.1	Overview	52
7.2	Future Work	52
7.2.1	Extra patient information	52
7.2.2	Generative Models	53
7.2.3	Classification and Localisation	54

References	56
A Project Poster	59
B Project Proposal and Ethics Form	61
C Project Log	71

Chapter 1

Introduction

1.1 Overview

Chest X-rays are one of the most popular medical imaging techniques that are used to look for abnormalities within a patient ([Organization et al., 2001](#)). One of these abnormalities includes pneumonia, which is an infection of the lungs and in the X-ray image, the infection appears as faint white spots depending on the severity of the infection. Radiologists diagnose chest X-rays; however, over the past few years, there has been a drop in the number of experienced radiologists, leading to increase backlogs for the National Health Service(NHS). With the advent of public datasets and compute power, the possibility of a real-time diagnosis tool of X-rays using deep learning to classify as either normal or infected lungs , is needed today more than ever. According to the Center of Disease Control and Prevention(CDC), more than one million people are placed in hospitals due to pneumonia infected lungs and accounts for over 50,000 deaths in the USA ([Rajpurkar et al., 2017](#)). In this paper, we investigate some of the key techniques of deep learning and apply some of those techniques in the implementation stage as well as investigating the new frontiers of deep learning and the impact these frontiers can have on the automatic diagnosis of medical images.

1.2 Aims and Objectives

The main aim for this project is to build deep learning pipeline and a simple interface which allows for chest X-rays to be uploaded to a model, and return a prediction with a confidence score of the result. The aim of this project will be undertaken by taking on the following objectives:

Objective 1: Study existing techniques that have already been published during the

literature review stage

Objective 2: Research relevant literature in deep learning and medical imaging

Objective 3: Grasp a solid understand of Convolutional Neural Networks(CNN)

Objective 4: Implement a deep learning pipeline using the correct open source technologies like Keras and Tensorflow

Objective 5: Apply an iterative approach to examine and improve the model performance

Objective 6: Conduct testing to evaluate the model performance using appropriate performance metrics

Objective 7: Research how a simple interface can be created so that an X-ray image is uploaded and passed through a model which returns a prediction.

1.3 Chapter List

Chapter 2 Literature Review. This chapter reviews the relevant literature associated to the project aims and objectives.

Chapter 3 Requirements Analysis. In this chapter we define the project spec to understand what requirements must be met and those that are beneficial to the project.

Chapter 4 Experiment Setup. This chapter deals with the hardware and software setup as well as a discussion of the dataset used for the experiments.

Chapter 5 Experiments & Results. This chapter deals with the detailed experiments carried out, as well as a summary of the results comparing different experiments across a range of metrics. We also detail the steps taken to create the simple interface.

Chapter 6 Evaluation. This chapter deals with the overall evaluation of the project by reflecting on the implementation stage and detailing the outcomes and lessons learned during the project. Here important considerations are also made such that legal, ethical and security issues related to project are discussed.

Chapter 7 Conclusion. Finally we conclude by summarising the project and the key findings as well as looking at the next frontier of deep learning in medical imaging .

Chapter 2

Literature Review

This chapter is a review of the relevant literature around deep learning and how this type of machine learning has advanced in the past years . In particular, the report will discuss how deep learning can be used in the diagnosis of pathologies in chest X-rays. As well as this we will review successful projects from the past that are closely linked to this thesis.

2.1 Overview

Radiology is a branch of medicine which uses medical imaging technology to generate images so that radiologists can view structures within the human body. The types of images range from X-rays, CT scan(computed tomography scan) and ultrasounds. ([Bradley, 2008](#)). For the past few years, radiology departments in the UK have seen disruption in the National Health Service(NHS). A reoccurring issue amongst radiology departments is lack of staff to interpret scans. According to the Royal College of Radiologists, the NHS does not have enough radiologists to meet diagnostic imaging demands leaving patients at risk. ([Rimmer, 2017](#)).

2.1.1 Faster Diagnosis

One area which can augment the diagnosis of medical scans is image classification. This process of image classification involves teaching a model to learn features from images and then test on unseen images to compute accuracy. In recent times, image classification through deep learning methods has produced superhuman performance on several image-based classification tasks([Krizhevsky, Sutskever, & Hinton, 2012](#)).

2.2 Traditional Methods of Machine Learning

Before the advancements in deep learning, less sophisticated methods of image classification were employed. This method of machine learning required the crucial step of feature extraction before inputting data into a classifier. In order to visualise the process that needs to take place when using traditional methods of machine learning please refer to figure 2.1.

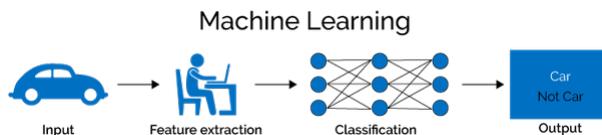


Figure 2.1: Figure showing the process of traditional image classification using feature extraction (Image taken from TowardsDataScience)

Computers see images as pixel values, and each image can be viewed as a 2D matrix of pixel values. When building a machine learning pipeline to classify an image, we need to be able to tell what is unique about each image so that the classifier shown in figure 2.1 can correctly classify. This uniqueness in an image can also be viewed as the features. Classification can then be thought of as detection of features in an image, and if features associated with a class are present in the input, we can predict if the input image is of a specific class ([Alexander, 2019](#))

In order to build a classification pipeline, the model needs to know what features it is looking for in an input image. These features it is looking for are a predefined set of features. In order to apply this rule, we can leverage human knowledge about a picture in a given domain and extract the most important features. We can then detect those manual features and use that detection to make a classification prediction ([Popescu & Sasu, 2014](#))

However, images can have lots of variation. If we want to build a robust pipeline, our model needs to be invariant to image variation, while still being sensitive to the differences in individual classifications. Manual features defined by humans can be used, but where this manual detection breaks down is in the detection task itself, and this is merely based on the fact that one class can have incredible amount of variation. This can be illustrated with the [2.2](#) provided by ([Szegedy et al., 2015](#))



Figure 2.2: Two distinct class from 1000 classes of the ImageNet database. Domain knowledge is required to distinguish between separate classes

So, how can we do better? We want a way to both extract features and detect their presence in an image automatically. This is where deep learning can be harnessed to extract features, no matter how the image is presented, automatically.

2.3 State-of-the-Art

Convolutional Neural Networks are a special kind of Multi-layered Perceptron(MLP). Like almost every other neural network they are trained with a version of back-propagation algorithm. Where they differ is in the architecture. This contrast can be viewed in figure 2.3 and figure 2.4. CNN's are designed to recognise visual patterns directly from pixel values with minimal preprocessing. They can recognise patterns with extreme variability(such as handwritten characters) and with robustness to distortions and simple geometric transformations. ([LeCun et al., 1989](#)) CNN's differ from MLP's in that they take advantage of spatial information of an image by convolving a patch of the input image with a filter weight of the same dimension. This results in a 2D matrix called a feature map. These filter weights are applied over the whole image and are iteratively refined in the training process , so that the correct features can be extracted. Different weight filter can be used to extract different features. This is achieved by passing the input across multiple convolution layers. In order to visualise this please refer to figure 2.4. From the diagrams it can be appreciated that MLP's lose all spatial information about an image by flattening every pixel and connecting these pixel values to a neuron in the hidden layer. As well as this we can see that MLP's have many more parameters in the network compared to CNN which can lead to over-fitting especially if the number of training samples is limited. ([Szegedy et al., 2015](#)) Starting with LeNet-5 , CNNS's have typically had a standard stacked structured of convolutional layers, followed by ReLU operation, max-pooling and then a fully connected layer. There are variants of this basic design in different image classification

tasks and they have all yielded respective results on classification tasks such as MNIST and CIFAR. Although, the most notable includes the ImageNet classification challenge where 1.2 million images were classified into 1000 different classes. In the past years CNN architectures have dominated this challenge and the first publication using CNN architecture was produced by Alex Krizhevsky and Geoffery Hinton in 2012([Krizhevsky et al., 2012](#)). The architecture was given the name Alex-Net and achieved a top-5 error rate of 15.3% outperforming the previous state-of-the-art, SIFT ([Lowe, 2004](#)) which achieved a 26.2% error rate using traditional methods. Since then, new CNN architectures have been published with improved results and this can be seen in the table below.

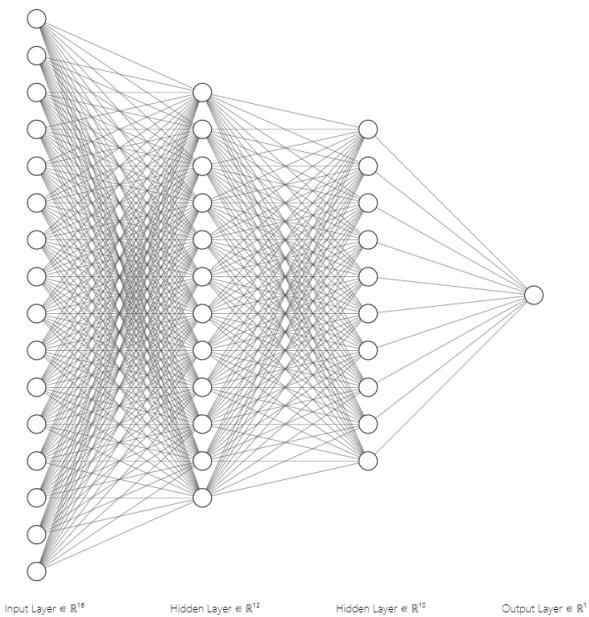


Figure 2.3: Architecture for MLP. Shows how all pixels value in image are flattened losing all spatial information(<https://alexlenail.me/NN-SVG/index.html>)

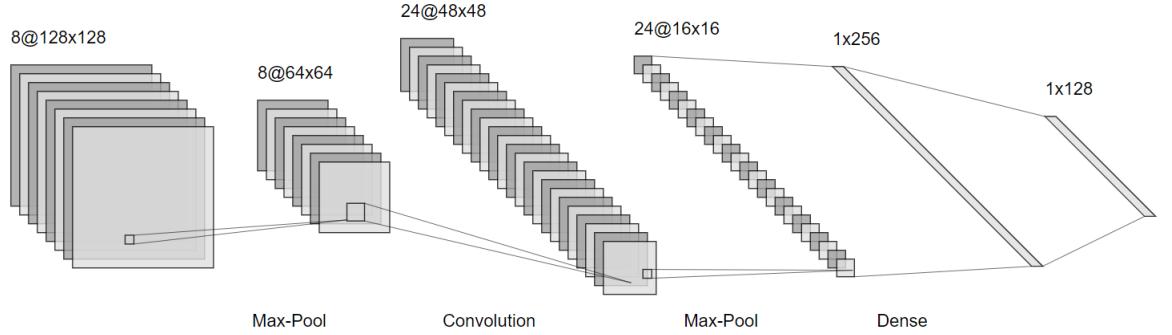


Figure 2.4: Architecture for Le-Cun CNN. Shows how spatial information is preserved as well different operations applied (<https://alexlenail.me/NN-SVG/LeNet.html>)

Model	Top-5 error rate	Number of Layers
AlexNet(2012)	15.3%	8 layers
VGG16(2014)	7.3%	19 layers
GoogleNet(2014)	6.7%	50 layers
InceptionV3(2015)	3.58%	50 layers
ResNet(2015)	3.57%	152 layers

Table 2.1: Results of different models with Human error rate at 5.1%

From the table we can see that for large datasets such as ImageNet, there is seems to be a trend developing. That is that the number of layers as well as layer size is increasing. A bigger size means that the model holds more parameters and can make it more prone to over-fitting. To address this problem, a technique developed by Google called “dropout” was developed ([Hinton, Srivastava, Krizhevsky, Sutskever, & Salakhutdinov, 2012](#))

2.4 Related Work

Many papers have been published in the domain of using deep learning for medical image analysis. This has been spurred in part due to large carefully curated labelled datasets which have led to papers showing expert-level performance on many medical image classification tasks (Irvin et al., 2019). In this section we will review work related to the thesis title.

2.4.1 CheXNet

In this project, led by Stanford ML Group, an algorithm was developed to interpret chest X-ray images and detect if Pneumonia is present while showing a heat map of the area in the image that had the highest activations (Rajpurkar et al., 2017). Example output can be see in figure 2.5

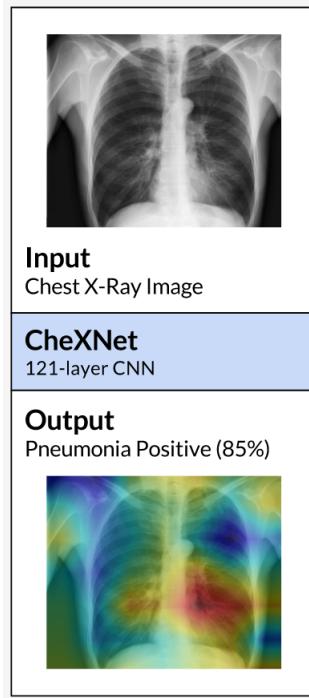


Figure 2.5: Example of CheXNet output of class label along with probability from softmax layer and heat-map

Results showed that the algorithm can detect the pathology at a level exceeding radiologists. The algorithm used a 121-layer CNN which was trained on ChestX-ray14 (Wang et al., 2017) which is currently the largest publicly available chest X-ray dataset, containing over 100,000 frontal view images and 14 different pathologies. The CNN used is based on the DenseNet architecture (Huang, Liu, Van Der Maaten, & Weinberger,

2017) but modifications were made for single output in the fully connected layer by applying a non-linearity sigmoid function, so that only Pneumonia would have a probability. Pre-trained weights were initialised from a model trained on ImageNet (Deng et al., 2009) to speed up the process of training.

2.4.1.1 Results

Precautions were put in place to make sure that patient images between training and testing sets never overlap. In order to test the accuracy of the model, 420 chest X-rays were collected and 4 practising radiologists annotated where they think the Pneumonia is present. Through the CheXNet tests it showed the algorithm exceeds average radiologist performance on the **F1 metric**. The model was comparative with expert radiologists on metrics such as **accuracy**, **sensitivity** and **specificity**. The biggest difference was in terms of speed as radiologists took around 4 hours on average to interpret the images whereas CheXNet algorithm took less than 2 minutes (Rajpurkar et al., 2017). Results from figure 2.6 present the F1 score for algorithm and average of the 4 clinicians.

	F1 Score (95% CI)
Radiologist 1	0.383 (0.309, 0.453)
Radiologist 2	0.356 (0.282, 0.428)
Radiologist 3	0.365 (0.291, 0.435)
Radiologist 4	0.442 (0.390, 0.492)
Radiologist Avg.	0.387 (0.330, 0.442)
CheXNet	0.435 (0.387, 0.481)

Figure 2.6: F1 Score for CheXNet and average F1 score for 4 radiologists

2.4.2 Multi-task Learning for Chest X-ray Abnormality

In this paper, researchers built a model trained on 297,541 chest X-rays. The system is based on a novel multi-task deep learning architecture that in addition to classifying and showing a heat-map of the abnormalities also supports the segmentation of the lungs and heart (Guendel et al., 2019). The research demonstrated that by training the model concurrently on these tasks, one can increase the classification performance. This research produced state of the art performance of 0.883 AUC on average across 12 different abnormalities. ChestX-Ray 14 (Wang et al., 2017) and PLCO datasets were combined to increase the amount of variability of images and this also showed increase in performance.

As mentioned before techniques were applied to increase accuracy. One of the methods

employed was to normalise the images. A challenge in processing chest radiographs is that there can be large variabilities in the appearance of the image and this is due to the acquisition source, radiation dose ([Guendel et al., 2019](#)). The paper proposes to dynamically window each image by adjusting the brightness and contrast via a linear transformation of the image intensities. Example of output is shown in figure [2.7](#)



Figure 2.7: We can see that certain parts of the X-ray become more visible hence making abnormalities more clear for the model

In addition , the paper also harnesses spatial knowledge related to individual pathologies as well as underlying structure i.e the heart and lungs which can be exploited to increase classification performance. The paper proposes that once can focus the learning task to the heart and lung region as information outside of these regions may be regarded as irrelevant for the diagnosis of heart/lung abnormalities.

In order to achieve this segmentation, a DenseNet model shown in figure [2.8](#) ([Huang et al., 2017](#)) has been used and figure [2.9](#) presents the architecture as well as an example showing the segmentation technique. Another technique that was used to take advantage of spatial information was the use of several approximate spatial labels provided in the PLCO dataset. For five abnormalities(Nodule,Mass,Infiltrate,Atelectasis,Hilar Abnormality) there are rough location information available to help the model locate where the abnormality is usually located.

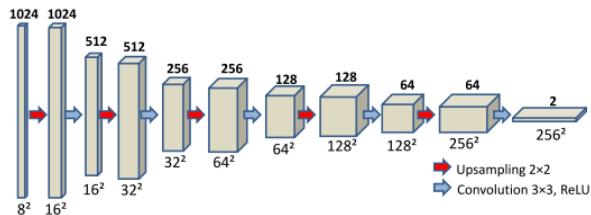


Figure 2.8: Architecture for DenseNet

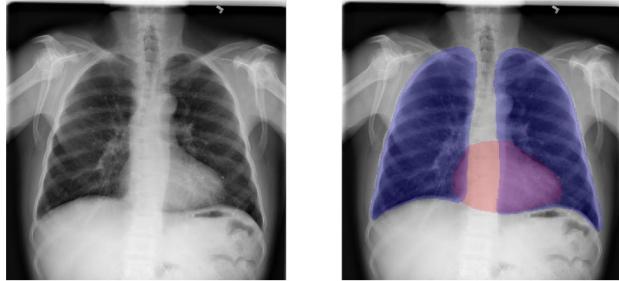


Figure 2.9: Distinct segmentation of lungs and heart

2.4.2.1 Results

For this paper, different experiments were undertaken to show the performance of the model. As a baseline, performance was measured by only training the classification part on the PLCO dataset. The test was evaluated with an average **AUC** score across 12 abnormalities of **0.859**.

Results also showed improved classification scores when the network was additionally trained to generate lung and heart segmentation masks (as shown in figure 2.10). Performance increased to 0.866 on average across the abnormalities.

As mentioned another technique was to use spatial knowledge and the impact of this showed an average improvement of 0.011

This paper also used 2 datasets and by combining both, the average AUC score reaches 0.870. This increase in score is due to more variability in the learning process. As well as this the normalisation technique previously shown in figure 2.7 was applied when training the model and this had a 2 fold benefit. The first was that the training process was reduced on average 2-3 times. The researchers hypothesised that this is due to the images being more aligned in terms of brightness and contrast. Another reason was due to the generalisation of the model parameters and this lead to a performance gain of 0.876.

Finally, the researchers upscaled the input image size to 512 in each dimension and adjusted the DenseNet layer to take in 16x16 patch size at each layer. This final network architecture change also added all of the previous techniques mentioned before. The following image shows the results obtained across all techniques and finally the result of combining all the techniques which produces a state-of-the-art performance of **0.883**

Dim. Size	256 × 256	256 × 256	256 × 256	256 × 256	256 × 256	512 × 512
Data	PLCO	PLCO	PLCO	PLCO+NIH	PLCO+NIH	PLCO+NIH
Features	-	Seg.	Loc.	Loc.	Loc.+Norm.	Loc.+Norm.+Seg
Nodule	0.810	0.815	0.830	0.832	0.831	0.881
Mass	0.829	0.839	0.840	0.867	0.869	0.884
Granuloma	0.884	0.886	0.887	0.887	0.893	0.912
Infiltrate	0.865	0.863	0.864	0.877	0.882	0.891
Scaring	0.841	0.842	0.843	0.850	0.848	0.861
Fibrosis	0.870	0.875	0.863	0.875	0.871	0.884
Bone/Soft Tissue Lesion	0.841	0.846	0.834	0.840	0.850	0.848
Cardiac Abnormality	0.926	0.928	0.922	0.923	0.927	0.926
COPD	0.882	0.883	0.877	0.880	0.882	0.874
Effusion	0.909	0.938	0.927	0.932	0.949	0.939
Atelectasis	0.849	0.858	0.858	0.867	0.855	0.832
Hilar Abnormality	0.796	0.815	0.812	0.815	0.850	0.859
Mean (Location Labels)	0.830	0.838	0.841	0.852	0.857	0.869
Mean	0.859	0.866	0.863	0.870	0.876	0.883

Figure 2.10: Results showing increasing mean AUC score

2.4.3 Abnormality Detection and Localization in Chest X-Rays using Deep Convolutional Neural Networks

For this paper, researchers conducted experiments on three datasets. From the datasets, researchers explored the performance of various deep CNN's for detection of heart disease from chest X-rays. Researchers used binary classification on diseases such as Cardiomegaly against normal images ([Islam, Aowal, Minhaz, & Ashraf, 2017](#)). The paper explores several CNN models such as AlexNet, VGG-Net and ResNet. These models have different layers and typically achieve higher accuracy as layers increase as noted in State-of-the-Art section.

Another part of the research, similar to previous papers reviewed was to produce a heat-map showing which areas of the X-ray the model has the highest activation on. This paper used the sensitivity of softmax scores of occlusion on a certain region in the chest X-ray to find which region in the image is responsible for the classification decision.

2.4.3.1 Results

The first experiment used single model CNN's on the Indiana dataset, and from the tests, it was shown that deeper models like VGG-19, ResNet improve the accuracy significantly. It was found that Cardiomegaly detection improves 6% compared to using shallower models like AlexNet. Although results showed high specificity for ResNet-101 and high sensitivity for VGG-16, the VGG-19 gave the highest AUC score of 0.94. This at least 1% higher AUC for other models. As well as this, dropout was used, which increased performance on the shallower networks but degraded deeper ones. For all experiments, it was noted that extracting features from earlier layers lead to improved accuracy of 2-4%. It was concluded that shallower networks were better for detecting smaller objects, and as an example, it was found that shallower layers from ResNet-15 trained on Cardiomegaly had better performance than later layers across different

metrics. figure 2.11 shows a more in depth picture of findings.

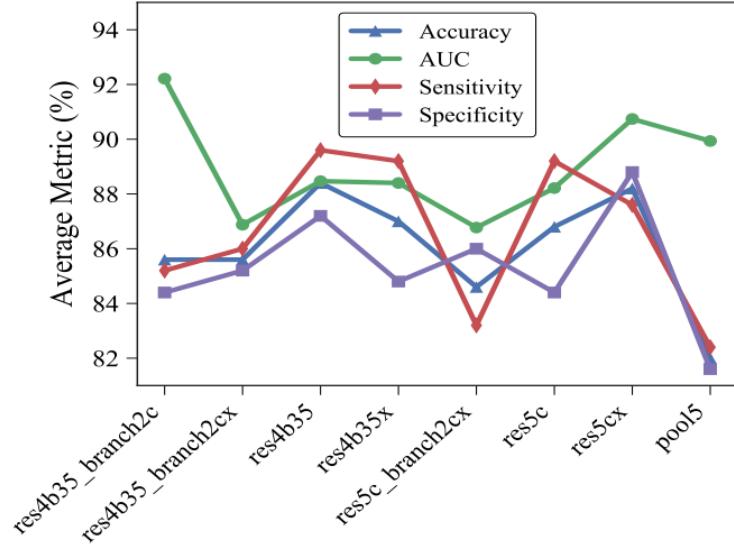


Figure 2.11: Researchers showed that shallower layers in models had a higher accuracy on Cardiomegaly detection compared to deep layers

Another fascinating result was on the localisation of Cardiomegaly shown in the image below. It can be observed that from the figure 2.13, the model has higher activations from the heart region, which is expected as Cardiomegaly concerns abnormal enlargement of the heart. What was revealing was that other than an enlarged heart for a patient with Cardiomegaly, there is not much difference in features between a normal image and an image classified as Cardiomegaly. But the model can still differentiate between a normal heart which is enlarged due to age or physical attributes of the patient.



Figure 2.12: Occlusion sensitivity used to create heap map of region where it thinks Cardiomegaly exists

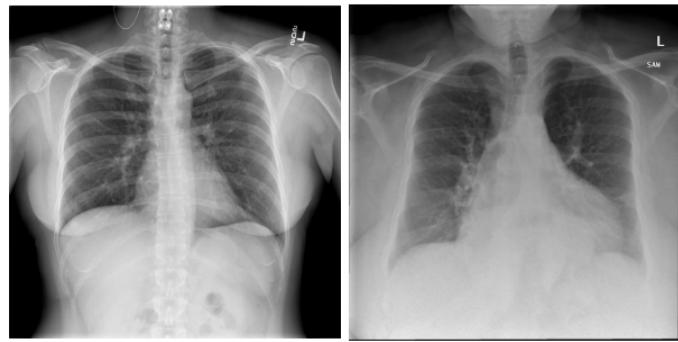


Figure 2.13: Example of normal heart on the left and Cardiomegaly diagnosed heart on the right

Finally, another point raised in the paper was the model used for localisation of cardiomegaly were counter-intuitive to traditional methods of cardiomegaly diagnosis, which consider the relative size of heart and lung. But, most of the signals contributing to the softmax score come from the heart alone indicating features in the shape of the heart and its surrounding region alone help in detection of the pathology and perhaps the lung and relative size to the heart are less critical in the diagnosis for the model.

2.5 Analysis of Related Work

In light of reviewing the related work, we can see a wide array of methods have been applied to produce several State-of-the-Art results on pathologies such as Pneumonia, Lung Nodules and Cardiomegaly. The paper from Stanford ML Group was one of the first to be tested on such a large scale dataset and made a mark in the application of deep learning on X-rays. Results from this paper show a 121-layered network produce results comparable to expert radiologists by using an architecture similar to that of DenseNet with some tweaks. As stated the ChexNet model has a higher F1 score than the average F1 score for four practising radiologists. After the success of ChexNet, more papers were released in the same domain trying to improve on accuracy metrics. The article discussed in section 1.4.2 uses more advanced features to locate the pathology more accurately. Given these additional features, it allows for the model to disregard certain regions of the images hence reducing noise in training process. From the results presented. It shows a clear improvement in accuracy across different pathologies once these features are introduced to the model. Lastly, the paper from X demonstrates a clear difference of performance on know CNN architectures. In section X, it showed that deeper layers led to better performance on the ImageNet classification task. However contrary to most image classification tasks, paper shown in section 1.4.3 shows how even though we see a better accuracy on deeper models, we see a higher AUC score for models like VGG-19 compared to ResNet-101, which has more layers, and this was due to earlier layers being better at detecting smaller objects. We also see the use of an ensemble of models to increase accuracy, which was something that was not experimented with in the previous papers reviewed. As well as this, we see a model trained on Cardiomegaly questioning the current ways of detecting this pathology such that it disregards traditional methods of looking at the relative size of the lung and heart in comparison. Lastly, from the three papers reviewed, we have seen the constant evolution which has allowed for more sophisticated methods of pathology detection and perhaps a combination of techniques in the future can increase detection amongst heart and lung pathologies.

2.6 Medical Images

As mentioned in the Overview section, radiology departments in the UK have struggled to meet imaging diagnostic demands. The survey published by the Royal College of Radiology reported 97% of radiology departments were unable to do so within their working hours. As mentioned in the paper, “It points to an insufficient number of radiologists to meet the increasing demand for imaging and diagnostic services”. According to the report it cited that radiology has the second lowest proportion of trainees to consultants: 26 trainees to every 74 consultants. This is compared to an average in all specialities of 40 trainees for every 60 consultants. The report also cites that there was a particular workforce shortage prominent in Scotland, where the consultant workforce grew by 7% from 2010 to 2016 but demand for CT,MRI scans increased by 10%. One of the profound impacts of such shortage is the risk to patient care as well as economical effects. The NHS paid nearly £88m in 2016 for backlogs of radiology examinations to cover backlogs of radiology examinations to be reported. To cover these backlogs 92% of radiology departments paid staff overtime, 78% outsourced to private companies and 52% employed ad hoc locums. This amount could have paid for at least 1028 full time radiology consultants.([Rimmer, 2017](#))

Given this what has been done to augment the analysis of medical images in the NHS and reduce extreme workload and large expenses ? One particular software which has bee used in the clinical setting is Computer Aided Diagnosis(CAD). CAD is technology which is designed to decrease observational oversights and false negative rates for physicians interpreting medical images such as mammographies ([Castellino, 2005](#)). According Dr. Paul Chang, the reason this type of software did not receive as much attention, compared to the possibility of using deep learning, was that it used traditional methods of machine learning. (*Video from RSNA 2018: Dr. Paul Chang on AI and radiology*, n.d.)

2.6.1 Challenges

There many challenges associated when applying new technologies in a clinical setting. Medical images are very large and complicated and for most of the image classification tasks in the ImageNet example , distinct features or pixels can be learned by a model. Dr. Matthew Lungren , radiologist at Stanford Medical Center, mentioned how most of the pixels in a chest X-Ray from a healthy person who happens to have lunge cancer , will tell you that the image is of a healthy person. But there is only a small proportion of the image which indicates the finding of lung cancer fo the classifier. Dr Lungren also noticed how medical images may have, what is known as artefacts or tokens in the image. Doctors are able to deal with context when looking at an image , but

such artefacts or tokens may lead to a model making a decision with higher degree of confidence but in actual fact the classification is incorrect ([DrMatthew, 2018](#))

2.7 Summary of Review

From the review, it is quite evident that with the ever-growing concern in the NHS of a shortage of radiologists ([Rimmer, 2017](#)), solutions need to be found in order to solve the issue of increased backlogs. Image classification, as mentioned, can provide augmentation of diagnosis of X-rays. Great leaps and bounds have been made in the field of deep learning and particularly well known CNN architectures show performance exceeding that of human intelligence in the case of ImageNet challenge ([Krizhevsky et al., 2012](#)). As well as this more sophisticated example of applying deep CNN's to medical images has shown outstanding results on pathologies such as Pneumonia and Cardiomegaly ([Rajpurkar et al., 2017](#)). This problem of diagnosing X-rays also comes with its challenges which have been outlined in the challenges section. Other hurdles must be kept in mind for implementation purposes, and these include making sure that pixel values are not lost. Given the nature of X-rays, pixel loss can lead to pathologies being removed from an image such as lung nodules and according to the Royal College of Radiologists certain rules need to be adhered to so that images do not lose quality when it comes to diagnosing them. Lastly, when training a model on medical images, the dataset chosen needs to be balanced in class labels. As well as this one of the big obstacles is lack of datasets of reliable labelling of datasets as noise in class labels can lead to reduced performance.

2.8 Moving Forward

In this section, we will conclude how the rest of this paper will be implemented and how the proposed problem in the introduction will be approached. From the findings above, the aim is to apply CNN architectures to build an accurate image classification pipeline on predicting the presence of Pneumonia in chest X-rays. In order to train the model, a public dataset will need to be used for the project, which has been used in previous papers as this solidifies the quality of the dataset. The aim is to build an accurate model which can predict and localise the Pneumonia. Another objective would be to build an interface which allows the entry of a chest X-ray and a predicted class label will be given back as well as a heat-map indicating which parts of the image had the highest activation. Lastly, if results from the initial pipeline show high performance across a number of metrics, then the model could be extended to classify more than one pathology contained in the dataset.

Chapter 3

Requirements Analysis

In order to answer the question proposed in the literature review, several objectives must be met. As this project cover two stages: Application of deep learning model on medical images and visualisation, this requirements analysis is split up into two sections: first section relating to requirement analysis of the implementation of deep learning model and the second being visualisation.

These requirements have been gathered based on the MoSCoW rule, which prioritises different requirements by stating the project must, should or could achieve the detailed requirements by the end of the implementation stage. A requirement that the project “must” achieve has the highest priority, where if this requirement is not met, the developed product will not be fit for purpose. The ones that “should” be completed are next in the priority hierarchy, whereby the project will still be functional without them, but their addition to the project would improve the end product. The requirements that “could” be completed are those that are not necessary to the overall functionality. However, these would benefit the project.

3.1 Deep Learning Model

3.1.1 Functional Requirements

3.1.1.1 Dataset

Must pre-process images from the dataset. Since input into the neural network, used in the implementation phase, requires pixel values, every image will need to be converted to its respected pixel value. Moreover, any resizing of the images will also need to be done as input to the neural net needs to be of a specific shape. Hence, this needs to be investigated in the implementation stage.

3.1.1.2 Convolutional Neural Network

Must build a model using image classification algorithms like CNN.

3.1.1.3 Improve Performance of model

Must look into ways of tweaking model parameters, and evaluating the results of these improvements, as per the ones described in the literature review.

3.1.1.4 Comparison of different algorithms

A report showing a comparison of different algorithm employed to build a model and predict could be created depending on the time.

3.1.1.5 Extending classification

The model's architecture could be further extended to diagnose more than one pathology.

3.1.2 Non-Functional Requirements

3.1.2.1 Programming Language and Libraries

Must use Python for rapid development and prototyping of CNN model via libraries such as Keras and sci-kit learn for visualisation and explanation of results.

3.1.2.2 Proof of Concept

Must be fully functional by the proof of concept stage

3.1.2.3 Anonymised Date

Must use anonymised patient data from publicly available datasets in order to conform to GDPR rule.

3.1.2.4 Accuracy

Model should diagnose images with an acceptable accuracy rate

3.2 Web App for visualisation

3.2.1 Functional Requirements

3.2.1.1 Visualisation

If the model evaluation shows promising performance, visualisation and explainability of results must be used in the form of a web application:

- Must show a classification probability to the user when a chest X-ray is tested on the deployed model
- Should show a localisation of where the model believes the Pneumonia is present in the X-ray.

3.2.1.2 User Input

User must be allowed to upload an image through the interface and get back results from the image being tested on the model.

3.2.2 Non-Functional Requirements

3.2.2.1 Programming Language and Libraries

Must use Javascript,HTML,CSS and cloud services for web application development:

- Should use Amazon Web Services(AWS) Sage Machine and AWS Lambda to deploy trained and tested model. The model should be accessed via the deployed RESTful API

3.2.2.2 Completion Date

Must be completed by the project demonstration day or could be operational by proof of concept deadline

3.2.2.3 Compatibility

Should be compatible with all major web browsers

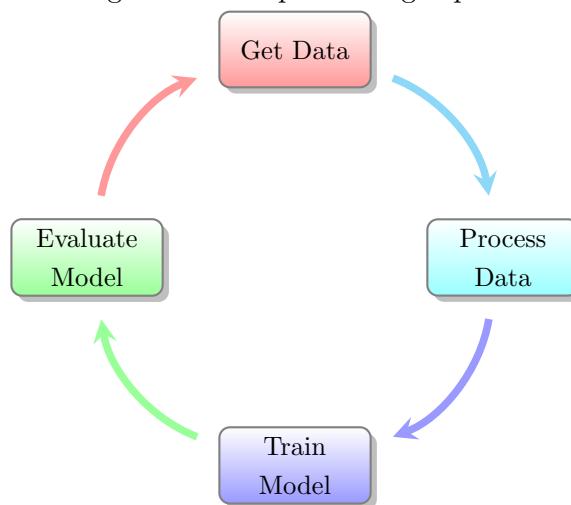
3.2.2.4 Image quality

Must not reduce image quality when passing image to cloud service where the model has been deployed upon

3.3 Methodology

The requirements in the previous section will be delivered in an agile manner. Using an agile approach consists of implementing the model, testing it against the test set, evaluating and reporting on the results of the performance. This process will be iterative until satisfactory results from the model are achieved, then leading to the deployment of the model onto a cloud service. The image shown in figure 3.1 demonstrates the steps that will be taken to build a deep learning pipeline that allows us to apply algorithms on the dataset and analyse performance.

Figure 3.1: Deep Learning Pipeline



Chapter 4

Experiment Setup

Before moving onto explaining process of applying the machine learning models for this paper, the following chapter explains the experimental setup that took place before starting the experiments. In this chapter the hardware and software setup is discussed. As well as this, a general overview of the dataset that is used for the experiments is discussed and more importantly the preprocessing that had to occur before running experiments is explained. Lastly, evaluation metrics will be discussed so that experiments in the implementation chapters can be analysed against metrics that determine the success or failure of an experiment.

4.1 Hardware Implementation

A computer system with the specifications shown in Table 4.1 will be used to carry out the experiments in this paper. Through the help of Robert Gordon University's Research Hub for Artificial Intelligence, this project makes use of the Research Hub's Nvidia DGX-1 to accelerate the deep learning experiments carried out in this paper.

Table 4.1: Computer System Specification

Operating System	Ubuntu 18.04 Linux Host OS
CPU	2x Intel Xeon E5-2698 v3 (16 core, Haswell-EP)
System Memory	512GB DDR4-2133 (LRDIMM)
GPU	8x NVIDIA Tesla P100 (3584 CUDA Cores)
GPU Memory	128GB HBM2 (8x 16GB)

4.2 Software Implementation

The setup for the experiments in this paper involves the use of **Python 3**. The solution heavily relies on **Keras**, which is an open-source neural network library written in python and is capable of running on top of **TensorFlow** as its back-end. Keras is designed to enable fast experimentation with deep neural networks and contains the basic building blocks to build one such as adding layers, objective functions, activation functions and optimisers. Moreover, Keras through the use of TensorFlow allows for the use of distributed training of deep learning models on multiple Graphics Processing Units - GPUs. This is further achieved by using CUDA, which is a parallel computing platform and Application Programming Interface(API) created by Nvidia. The CUDA platform is a software layer that gives direct access to the GPU allowing for highly parallel manipulation of large blocks of data.

4.3 Dataset

The dataset that will be used for training and testing of the learning algorithm has been taken from a public repository ([?](#), [?](#)). The dataset includes X-ray images of both normal patients and patients with pneumonia infected lungs. Based on the information given by the authors that published the dataset, for use in the paper titled “Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning” ([Kermany et al., 2018](#)), the data has been obtained from the Chest X-ray images selected from a respective cohort of pediatric patients aged from one to five years old from Guangzhou Women and Children’s Medical Center.

The dataset comes with two folders which include training and validation. Each folder contains subfolders for each type of classification, normal and pneumonia. The training folder contains 5218 images split into 1342 normal images and 3876 pneumonia images. The validation folder contains 624 images split into 234 normal images and 390 pneumonia images. Figures [4.1](#) and [4.2](#) help to visualise the distribution of the training and testing set respectively. The distribution of both training and testing are important to notice as this will be an important consideration when evaluating the model performance.

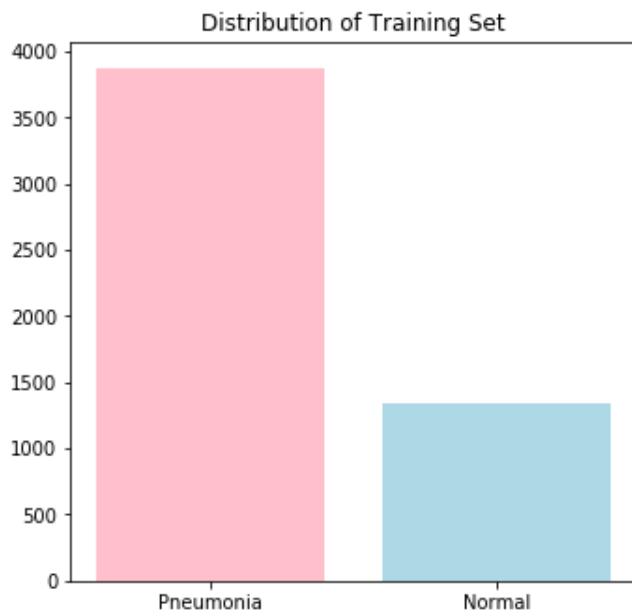


Figure 4.1: Class Distribution of Training Set

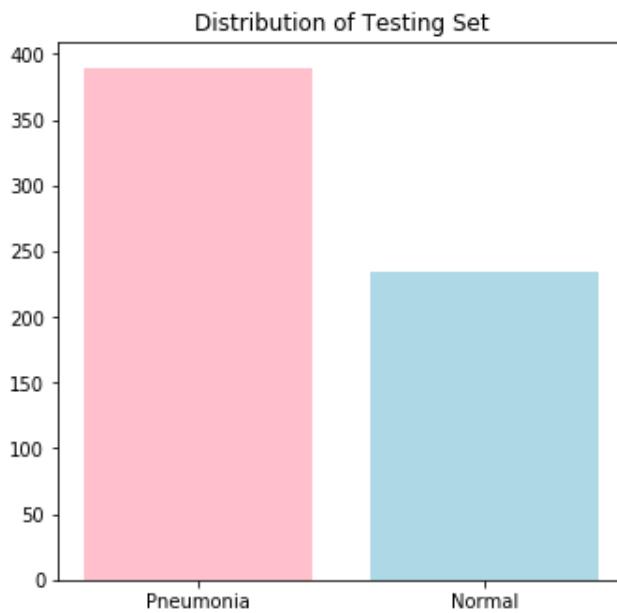


Figure 4.2: Class Distribution of Testing Set

To visualise the images used to train the learning algorithm, figure 4.3 and 4.4 shows an example of both a normal patient and a patient suffering from pneumonia. It might be noticeable in figure 4.3 that a healthy X-ray would normally be clearer, and the general area of the chest is more visible compared to cases where pneumonia is present, as

shown in figure 4.4. However, the onset of pneumonia in patients can differ depending on the severity of the infection. As such, there may only be a few pixels that denote the presence of pneumonia which makes the task of detecting the disease a challenge.

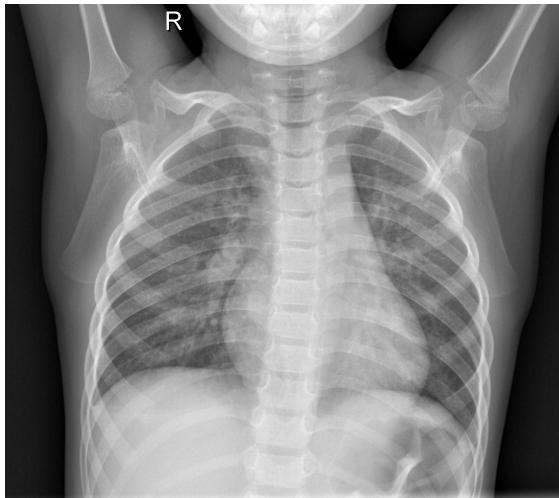


Figure 4.3: Normal Chest X-ray

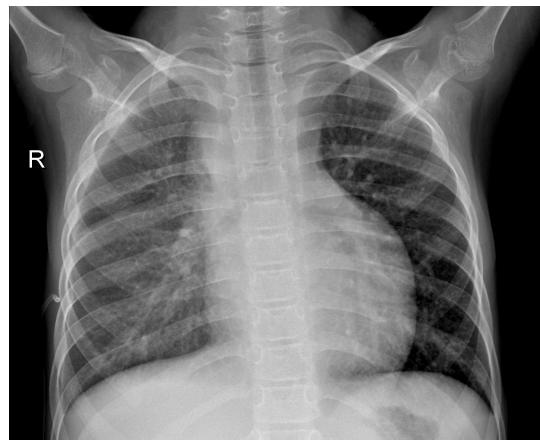


Figure 4.4: Pneumonia Chest X-ray

4.3.1 Dataset Preprocessing

During exploration of the dataset it was shown that the X-ray images provided were of different sizes. This most likely due to different X-ray imaging machines used during collection of the dataset. Before applying the learning algorithm on the X-ray images we needed to make sure that each input image was of the same dimension. This is due to the way a CNN is configured via libraries like Keras, where the size of the input layer is constant. All images were resized from their original size to 250 x 250 which meant that some image were scaled up and other were scaled down.

4.3.2 Normalising Images

Data normalisation is an important step which ensures each pixel value for an image has a similar data distribution. This makes convergence faster while training a neural network. Although non-normalised data can be presented to the neural network, this can result in challenges during training, such as slower training of the model.

As discussed in section 2.3 “State of the Art”, neural networks process inputs using small weight values which are multiplied by the input value. If these inputs are on a large unsigned integer scale, like pixel values, this may disrupt or slow down convergence of the training process. A pixel value of 0(white) compared to 255(black) are both equally as important but the 255 value will have a greater effect on the final prediction. This large value could cause the output to be wrong and lead to the training process taking longer.

For the purposes of this paper, data normalisation is applied by dividing each pixel value by 255 and as a result all inputs to the model have a range between 0 and 1.

To better understand the stages of normalisation, figure 4.5 shows the steps taken to normalise each image.¹

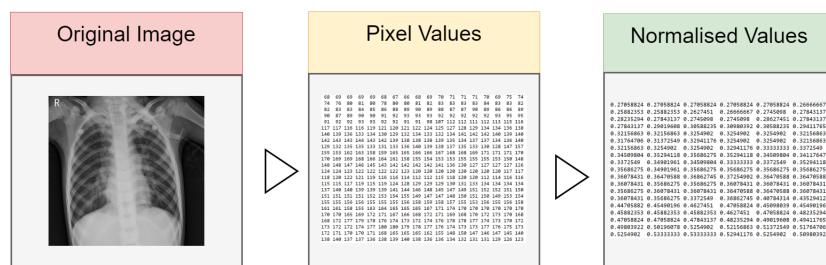


Figure 4.5: Stages of the Preprocessing and Normalisation

4.4 Going Beyond Accuracy

As shown in figure 4.2, the class distribution of the testing set is imbalanced. Due to this imbalance, it is important to look beyond accuracy to validate a model performance. For example, since the test set is imbalanced, if all the test cases were labelled positive, i.e. a class label of pneumonia, then the accuracy on the test set would be 63%. However, the accuracy does not show the underlying performance of the model, which indicates that it is classifying every patient as having the infection.

Due to the domain of this paper and its aim: Automatic classification of X-ray Images

¹ Python Notebook:Preprocessing

as either normal or abnormal, a confusion matrix needs to be considered when evaluating results. Figure 4.6 shows an example of a confusion matrix that will be used to analyse results. The number of pathological samples that are *correctly* identified as a pathological sample by the model is called a true positive (TP). The number of pathological samples that are *incorrectly* classified as normal by the model is called a false negative(FN). The number of normal samples that are *correctly* identified as normal is called a true negative(TN) and in a similar fashion, the number of normal samples that are *incorrectly* identified as pathological samples is called false positive(FP). Given the domain of this study, the cost of a false negative is higher than a false positive

		Predicted Class	
		N	P
		True Negative (TN)	False Positive (FP)
Actual Class	N		
	P	False Negative (FN)	True Positive (TP)

Figure 4.6: Confusion Matrix

Although accuracy is still considered, this study will also examine the values of precision, recall/sensitivity and specificity. Accuracy, specificity, recall and precision are computed as follows:

$$Accuracy = \frac{TN + TP}{TN + FP + FN + TP}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Recall/Sensitivity = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

- **Accuracy** is a basic measure shown as a percentage of the total correct classification out of all the samples in the testing set.

- **Specificity** shows the degree to which the model correctly identifies normal samples as normal.
- **Recall/Sensitivity** is a measure which shows the degree to which the model does not miss a pathological sample
- **Precision** is measure of the exactness or quality of the model when it returns results for a specific class

Chapter 5

Experiments & Results

In this chapter, the experiments that were performed during the duration of this project will be discussed. This discussion will include the successful experiments as well as the experiments that failed due to a flaw in the initial setup. For each experiment the general setup will be discussed. This includes the CNN architecture, epochs, batch size, learning rate and training and testing splits. Each of the experiment results will be discussed in detail where the evaluation metrics from 4.4 will be applied to the model results.

5.1 Experiment 1

Model Architecture: For the first experiment that was carried out a simple CNN architecture was used to classify the images. The setup of the CNN involved the use of two convolutional layers each with ReLU activation. After the final convolutional layer, Max Pooling is applied to reduce the dimensionality of the network. After applying Max Pooling the features extracted are flattened and sent through a fully connected layer consisting of 128 neurons. ReLU activation is applied once more before adding a further 2 neurons and finally a Soft-max layer. Figure 5.1 shows the layers that have been used in Experiment 1



Figure 5.1: Model Architecture for Experiment 1

Hyper-parameters

Hyper-parameter setup for the experiment is the following:

- Imbalanced training set (*Non-Shuffled*)
- Balanced testing set
- Batch Size = 64
- Epochs = 1
- Adam optimiser (lr=0.01)
- Categorical Cross Entropy Loss Function

Results

For the initial experiment, the following confusion matrix shows the models performance:

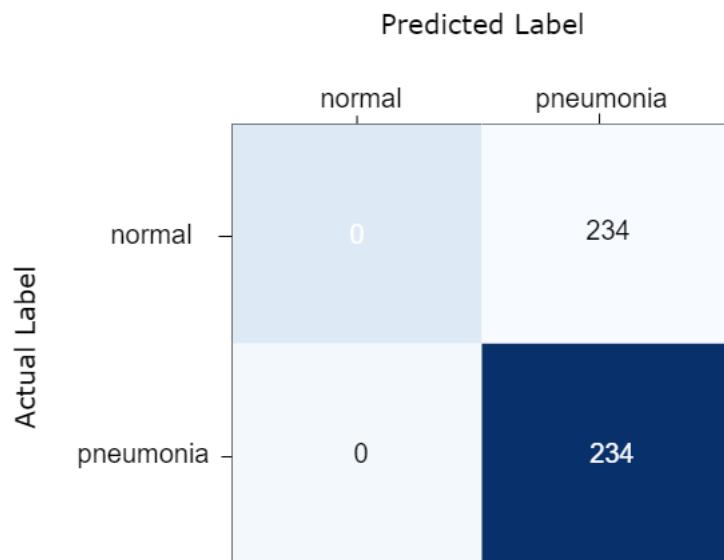


Figure 5.2: Experiment 1 Confusion Matrix

After inspecting the results above, it can be observed that the model trained on the above hyper-parameters has an accuracy of 50% on the test set shown in the confusion matrix above. From the matrix, it shows the model has a Recall/Specificity score of 1 since the model was able to identify all patients with pneumonia. However, although the model can classify pneumonia cases correctly, the opposite is true for the normal cases as the model classifies them as positive cases, hence giving a Specificity score of 0.

Observations were made to understand why the model produced the results in figure 5.2. Upon reviewing the setup for experiment 1, it was noted that since the training data was not shuffled it meant that for each epoch the model was learning 74% of the pneumonia cases first and a further 26% of the normal cases. It meant that the model was learning in a sequential manner which affects the weight updates. It is important that moving forward the training set is shuffled so that batches of input contain a mixture of both positive and negative samples.

Another observation noted when reviewing results is that the number of epochs was too low. The number of epochs determines the number of passes over the entire training set. It means the number of epochs must be higher so that the model has more chances to learn from the training set.

Lastly, in this experiment, it was decided that the testing should be balanced. However, after reflection experiments going forward should not balance out the testing set as this could remove vital images which could include difficult samples and may test the model's capability further. Balancing of the provided data would not reflect the medical environment where it is common to have imbalanced datasets.

5.1.1 Improving Experiment 1

Model Architecture In this experiment, the aim is to see the effect higher epochs and shuffled training data has on the model performance. The model architecture is the same as the previous experiment which is illustrated in figure 5.1.

Hyper-parameters

- Imbalanced training set (*Shuffled per epoch*)
- Validation set (*10% of training*)
- Imbalanced testing set
- Batch Size = 64
- Epochs = 10
- Adam optimiser (lr=0.01)
- Categorical Cross Entropy Loss Function

The main differences in hyper-parameters settings include increased epoch to allow the model to have more passes to learn from the training data. The training set has also been shuffled and the testing set has been changed to the unbalanced distribution mentioned in the previous experiment. To track the model performance over the entire training process this experiment used 10% of the training set to act as a validation set. By doing so we can track the model accuracy and loss over 10 epochs.

Note: Keras allows us to shuffle the data via the *shuffle* parameter which shuffles the order of the data at every epoch. This allows the model to learn in a varied manner since the sequence of the training data is different per epoch. By using this feature in Keras it means reproducibility is not guaranteed but experiments ran multiple times should show similar results.

Results

For this experiment, the following confusion matrix shows the model performance after changing hyper-parameters. Since the model has been run for more than one epoch, an accuracy and loss graph can be used to track the model performance on the training and validation set through out the training process.

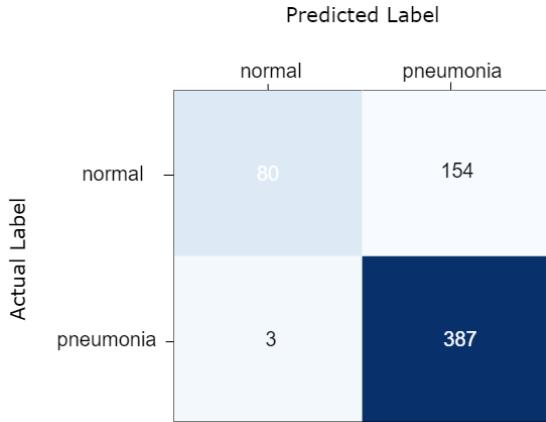


Figure 5.3: Experiment 1.1 Confusion Matrix

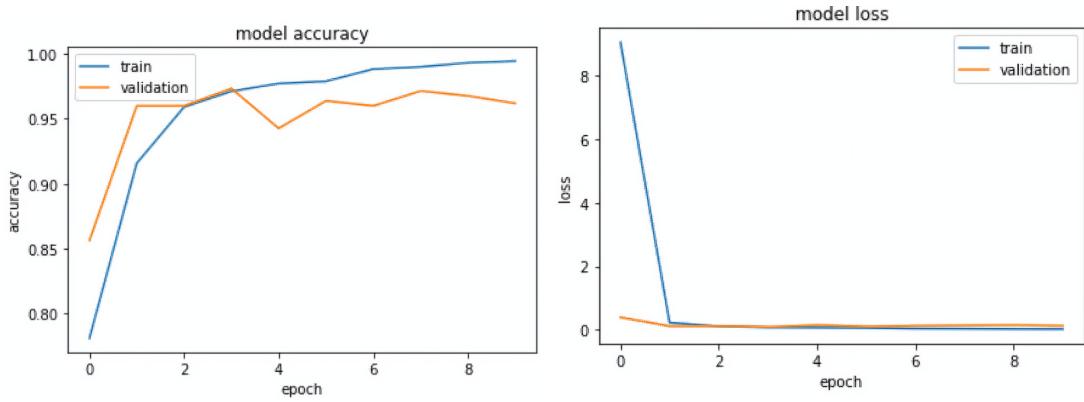


Figure 5.4: Model Accuracy

Figure 5.5: Model Loss

Results Analysis

The confusion matrix in figure 5.3 shows that after shuffling the training data and increasing the number of epochs the model achieves a 75% **Accuracy** on the test set. **Recall/Sensitivity** score is 0.99 and the **Specificity** score is 0.3. From figure 5.4 we can track the model accuracy on both the training and testing set. The graph shows us the model is able to quickly learn the training set close to 100% accuracy between the first and third epoch(Python epoch begins from 0th epoch) and continues to increase which suggests the model could have been trained for longer. Figure 5.5 shows a similar story but measures the models loss which shows the training loss drops quickly in the first 2 epochs. The validation line in figure 5.4 shows the accuracy increases a lot between the first and second epoch, however, later on in the training process validation accuracy tends to fluctuate before it starts to drop.

Although the model has a reduced Recall/Specificity score, which is expected since the

model has learned in a varied manner, the number of false positives has gone down which suggests increased epochs and a shuffled training set improves model performance. However, it still evident that due to the imbalance in the training set the model is classifying almost double the amount of false positives compared to true negatives. To diagnose the issue of why the validation data is suffering from a case of over-fitting the same experiment was ran for four additional runs where the test accuracy ranged from 71% to 75%.

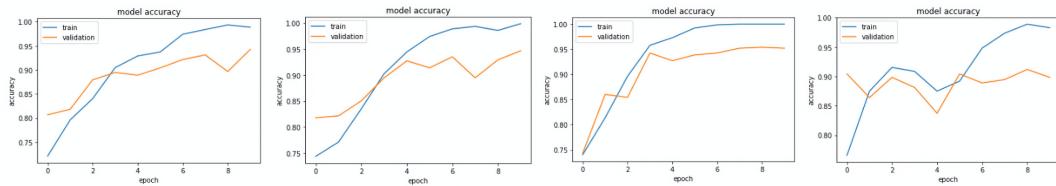


Figure 5.6: Identical experiments with different outcomes

Figure 5.6 shows that the same experiment ran multiple times results in a different training process, where even though training accuracy for all five experiments ends in a similar trajectory, the validation accuracy shows a different story as all five validation accuracies are different. On reflection the reason the same model produced different accuracies is due to the value of the learning rate being too high for this dataset.

The learning rate is a hyper-parameter that controls how much to change the model in response to the estimated error each time the model weights are updated. In this specific situation we hypothesis that since the data is shuffled each of the experiments sees a different sequence of data which when coupled with a high learning rate causes drastic changes in weight updates which effects the training and validation accuracy.

5.2 Hyper-Parameter Tuning

In this experiment(s) the aim is to apply hyper-parameter tuning to improve performance of the previous experiment and stabilise the training and validation accuracy and loss. There are a range of hyper-parameters to experiment with however, the experiments carried out in this section will focus mainly on adjusting Adam learning rate, increasing epochs, changing the number of convolution operations in the convolutional layers and number of dense connections in the fully connected layers.

5.2.1 Learning rate

It was noted in the experiments carried out in section 5.1.1 that after running the same experiment five times the overall testing accuracy varied as well as the validation

accuracy and loss. The aim of this experiment is to reduce the learning rate to see the effect it has on stabilising results.

Model architecture

The model architecture is the same as before and is illustrated in figure 5.1

Hyper-Parameters

The model hyper-parameters are largely the same with learning rate changing from 0.01 to 0.0001.

Results

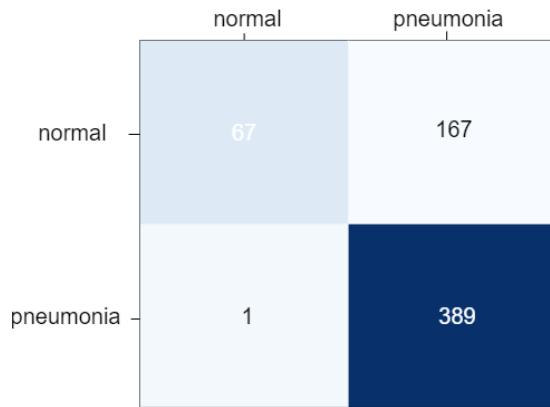


Figure 5.7: Confusion Matrix after adjusting learning rate

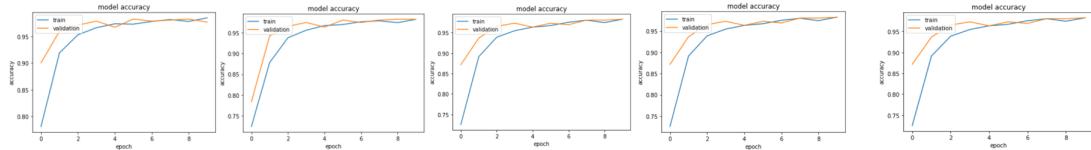


Figure 5.8: Model accuracy after changing learning rate

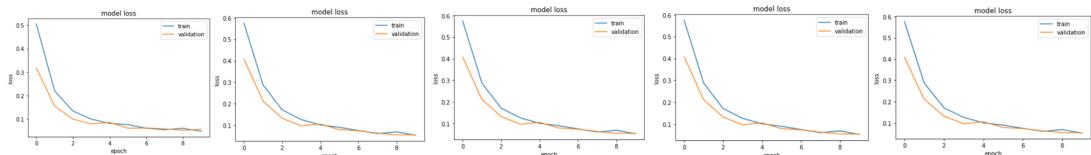


Figure 5.9: Model loss after changing learning rate

In this particular experiment, although we see a decrease in performance, the same model ran five times produces results clustered closer together with testing accuracy across five models being the same except one which differs by one percent. In figure

[5.8](#) and [5.9](#) we see that for the five models ran, all of them show a similar trajectory with both training and validation accuracy closely follow each other. A change in learning rate also appears to reduce the discrepancy between the training and validation accuracy acting as a regularisation effect which stopped over-fitting. The same can be said for the loss graph show in figure [5.9](#) which shows the training loss is still lowering with the validation loss starting to flatten out. This suggests that the model could have been trained for longer.

Comparing figure [5.6](#) and [5.8](#), we can see that the model after tuning the learning rate is stable across multiple runs. As a result of reducing the learning rate, although it stabilised results, the model has smaller weight updates which slows down the process of convergence.

Therefore, we ran the same experiment with 20 epochs which resulted in test results shown in the confusion matrix in figure [5.10](#) and the model accuracy loss shown in figure [5.11](#) and [5.12](#) respectively.

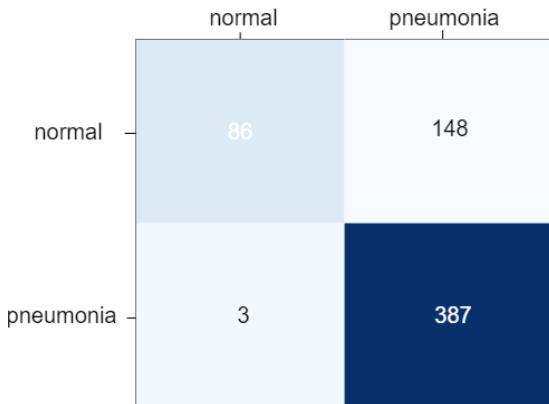


Figure 5.10: Results after 20 epochs

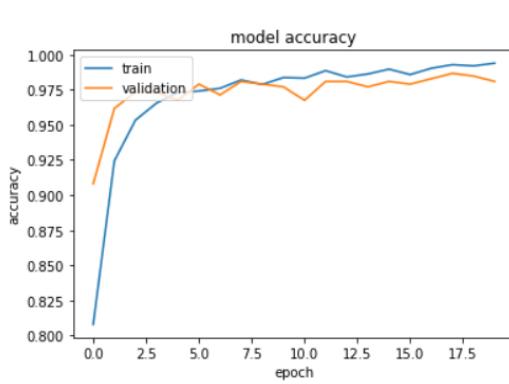


Figure 5.11: Model Accuracy

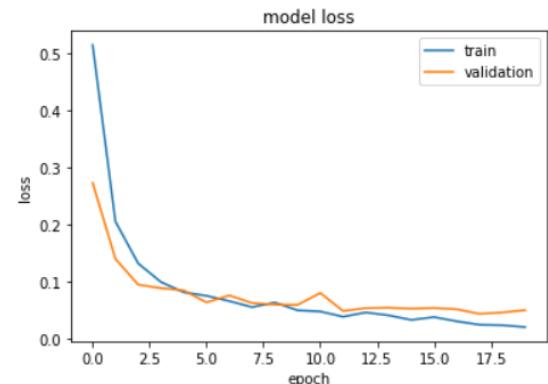


Figure 5.12: Model Loss

After running the experiment over 20 epochs, the overall testing **Accuracy** is 76% with a **Recall** and **Specificity** score of 0.99 and 0.37, respectively. From the model accuracy graph shown in figure 5.11, the training process is generally stable when comparing both the training and validation lines, which suggests that a reduced learning rate and increased epoch value has kept stability during training. There are slight deviations after the 9th epoch until the 20th where the training and validation accuracy and loss start to deviate from each other. When comparing the result from the experiment in section 5.1.1, the results in this experiment after reducing the learning rate and increasing epochs shows a slight increase in **Accuracy** and **Specificity**.

5.2.2 Adding Convolution and Fully Connected Layers

Previous experiments showed that we were able to stabilise results over many runs and increase specificity. Another approach that we took to further reduce the higher number of false positives, is to alter the convolutional and fully connected layer in the model architecture. Figure 5.13 and 5.14 shows a comparison of the model architecture between the previous experiments and the one used in this experiment.



Figure 5.13: Model Architecture for previous experiments

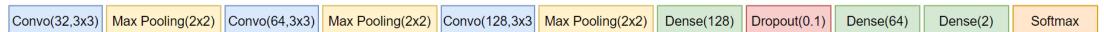


Figure 5.14: New Model architecture

What is noticeable is the addition of a convolution layer and the number filters in each of those layers. These layers are responsible for extracting the relevant features of an image, before passing them through a fully connected(FC) layer, where the classification process occurs. Previous experiments maintained the same number of filters in the two convolutional layers, whereas in this experiment we doubled the amount of filters after every layer. The intuition behind this choice is that CNN's typically detect lower level features via the filters from the first convolutional layer, such as edges and lines. Once the lower level features are extracted, the second layer typically detects motifs by spotting particular arrangements of edges, regardless of small variations. Finally, the third layer may assemble motifs into larger combinations that correspond to parts of familiar objects (LeCun, Bengio, & Hinton, 2015). By increasing the number of filters, a higher number of inputs are passed to the FC layer, and as such, it makes sense to increase the number of neurons in the FC layer which is why an additional layer of 64 neurons is added.

Moreover, the addition of Max Pooling helps with reducing the dimensionality of the network after every convolutional operation and lastly, a technique known as dropout was also used in this experiment to help combat over-fitting. Previous experiments showed signs of over-fitting towards the end of the training process and so increasing the number of convolutional layers could increase over-fitting and hence drop performance on the test set. Dropout is a regularisation technique where randomly selected neurons are ignored during training, which means that their contribution to the activation of downstream neurons is temporarily stopped during the forward propagation, and subsequently any weight updates during the back-propagation stage are not applied. Figure 5.15 shows an example of how dropout works.

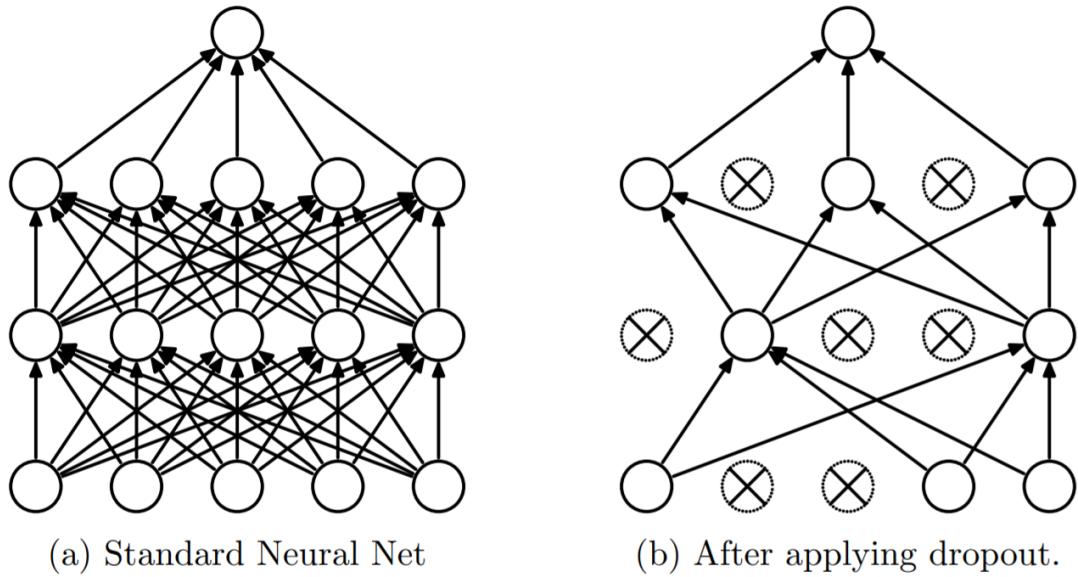


Figure 5.15

In the paper which introduced the concept of dropout (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014), the recommended value for randomly dropping out neurons is 0.5, however for this experiment we will use a value of 0.1 since the FC layer is much smaller in this experiment compared to the original paper and the previous experiments did not show a huge deviation between the training and validation lines.

Model Architecture

As mentioned above the model architecture is shown in figure 5.14.

Hyper-Parameters

The model hyper-parameters are unchanged from the previous experiment.

Results

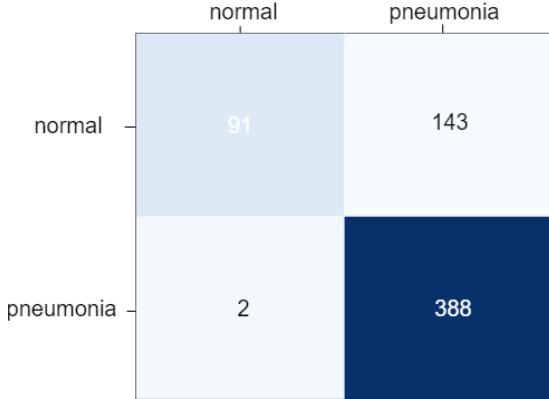


Figure 5.16: Testing results after adjusting architecture

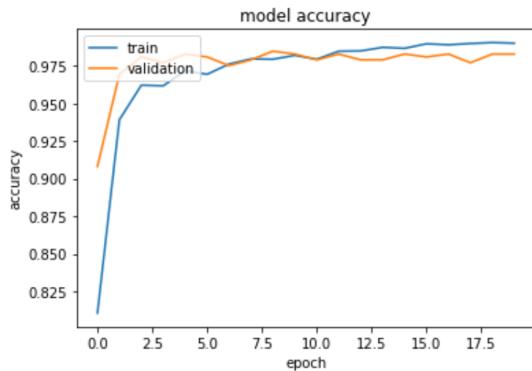


Figure 5.17: Model Accuracy

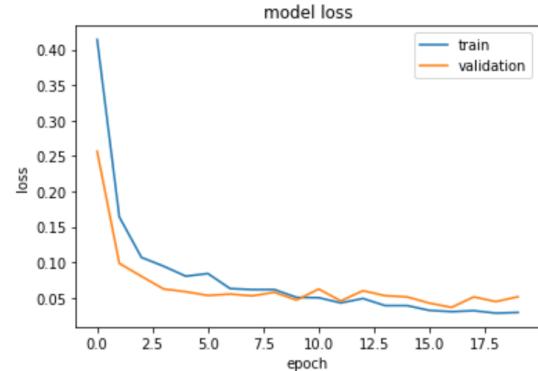


Figure 5.18: Model Loss

Figure 5.16 shows the results when testing the trained model on the test set. The results show a small increase in **accuracy** from the previous experiment of 76% to 77% and **specificity** has increased a small amount from 0.37 to 0.38, hence an improvement in the number of true negatives and false positives . **Recall** drops slightly, but remains high due to the imbalance which favours the positive class. The model accuracy and loss for this experiment shows that the training and validation lines are stable during the training process and does not show signs of over-fitting.

The changes made in this experiment did not improve performance dramatically since the aim was to increase the number of true negatives to a point where there is at least more true negatives than false positives. However, so far this experiment with model architecture and hyper-parameters, has produced the best results on specificity which is the value we are trying to increase.

5.3 Transfer Learning

A novel technique used in deep learning, is to harness pre-configured models that have been optimised for feature extraction. This approach is called transfer learning, where we take advantage of an existing CNN’s convolutional layers as a feature extractor to output feature maps used as inputs to the FC layer. This type of learning has shown to overcome challenges such as limited dataset size and training from scratch (Rajpurkar et al., 2017). Examples of pre-trained models include VGG16 (Simonyan & Zisserman, 2014), which is the model that will be used in this experiment to see if transfer learning can improve results. Figure 5.19 shows the model architecture for VGG16, where a total of 13 convolutional layers are applied to the input image.

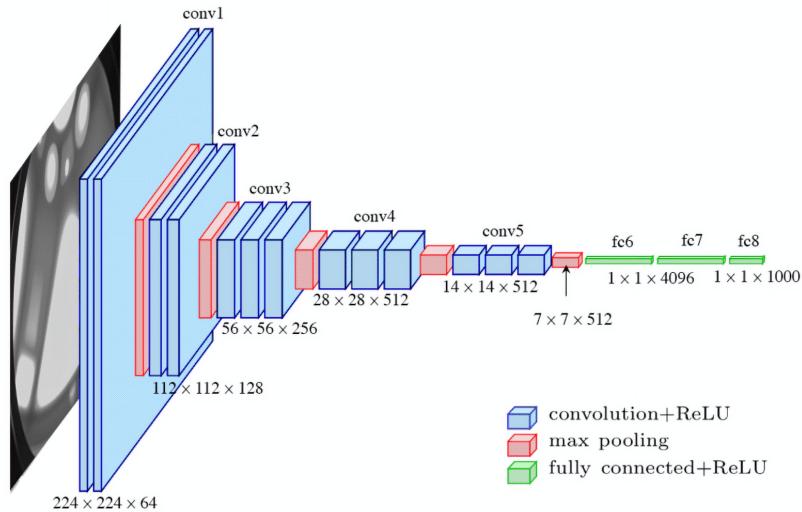


Figure 5.19: VGG16 Architecture

There are different ways in which transfer learning can be used, however in this experiment we simply use VGG16 as a feature extractor and attach our own FC layer at the end for the classification process. Figure 5.20 is a visualisation of how transfer learning is used in this experiment. To use VGG16’s convolutional layers exclusively, as feature extractor, we froze the convolutional base and only trained the FC layer. Another change that had to be made to use VGG16, was to use a 250x250 input size instead of a 224x224. Lastly, VGG16 is trained on three dimensional data(RGB) so we had to increase the dimensions of the X-ray images from one to three.



Figure 5.20: Transfer Learning technique

Model Architecture

The model architecture used for this experiment, uses the convolutional layers in figure 5.19, and the FC layer from the previous experiment as it showed the best performance.

Hyper-Parameters

The hyper-parameters remain unchanged from the previous experiment.

Results

		normal	pneumonia
normal	normal	75	159
	pneumonia	1	389

Figure 5.21: Testing results using VGG16

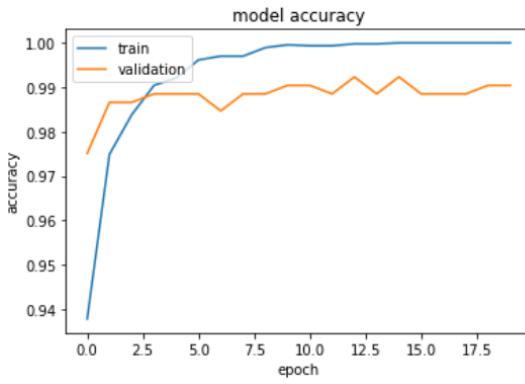


Figure 5.22: Model Accuracy

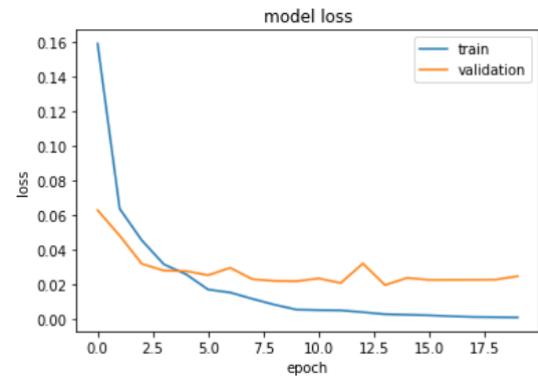


Figure 5.23: Model Loss

After reviewing the test results we can see this experiment produced sub-optimal results in comparison to the previous experiment. Analysis of the confusion matrix in figure 5.21, shows that accuracy is 74% compared to 77% from the previous experiment. The aim of this experiment was to increase the specificity score and hence increase the true negatives. However, this experiment was not able to suffice that aim.

From this transfer learning experiment, we concluded that perhaps transfer learning 'out of the box' as a feature extractor did not perform as well compared to a model with convolutional layers optimised for this specific dataset. However, further work could be done into adjusting the VGG16 model, such that a select few convolutional layers are allowed to train. The intuition behind this is based on a previous comment, where the earlier filters detect lines and edges compared to later filters which deal with the high level feature, such as the area of the infection. Hence, deeper filters could be fine-tuned to build a model with better results.

5.4 Augmentation

After optimising results via hyper-parameter tuning, the number of normal cases diagnosed as pneumonia by the model is larger than the number of cases correctly identified as normal. This issue is mainly due to the reason that the model is trained predominantly on the pneumonia class, hence, the normal class is under-represented. A common approach to overcoming the issue of low **specificity** is to find ways of oversampling the under-represented class. Since this project could not acquire new X-ray images from patients, data augmentation was used which is a common approach to improve the performance of computer vision systems. Data augmentation can be used in various ways such flipping, rotating and shearing. For this specific experiment, we chose to over-sample the normal class by rotating images randomly between -10° and $+10^\circ$. Figure 5.24 shows a healthy X-ray taken from the dataset which can be rotated to create a new sample as shown in figure 5.25

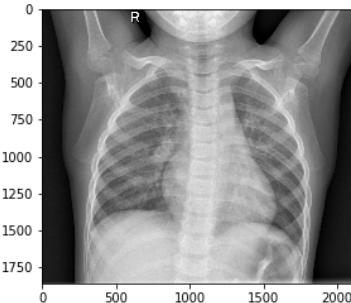


Figure 5.24: Before Augmentation



Figure 5.25: After Augmentation

The next step was to decide how much the normal images were to be oversampled by. We decided to double the amount of normal images in the training set to counteract the dominance of the pneumonia class. The validation data used to track the performance of the model against unseen data over the 20 epochs was not given the augmented images and uses 10% from the original training set.

Model Architecture

Since the best performance from the previous experiment showed that six layered CNN outperformed all other model architectures including transfer learning, the model architecture from section 5.2.2 will be used. Figure 5.14 shows a detailed look of the model architecture

Hyper-Parameters

Hyper-parameters are unchanged from the previous experiment.

Results

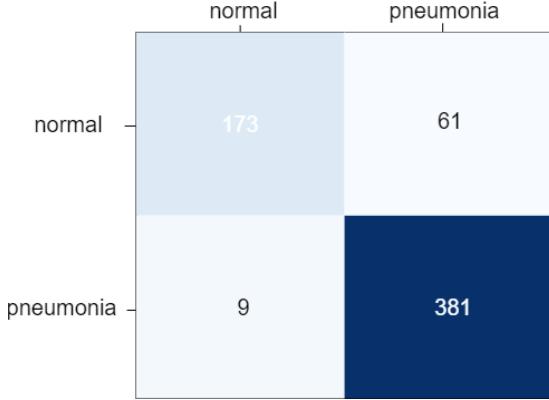


Figure 5.26: Testing results after augmentations

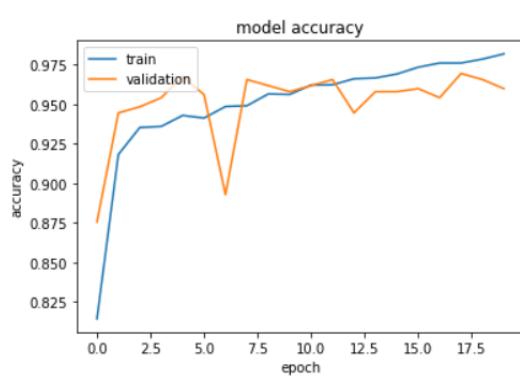


Figure 5.27: Model Accuracy

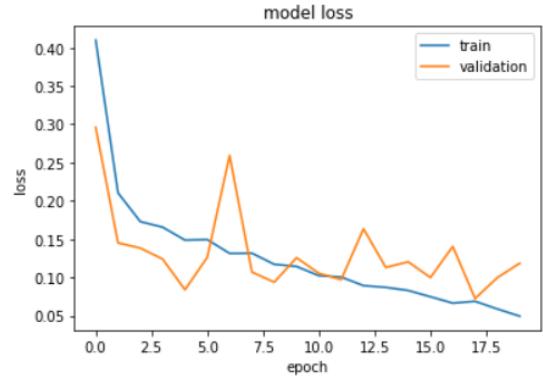


Figure 5.28: Model Loss

Figure 5.26 shows the model result on the test set after applying the augmentation techniques to increase the number of normal X-ray images. What can be observed is that there is a substantial increase in true negatives and drop on false positives. The overall accuracy of this experiment was 89% compared to the previous optimal results in section 5.2.2, which was 77%. As a result of applying augmentation, specificity score nearly doubled from 0.38 to 0.74. After applying the augmentation it was noted that recall decreased, however, this was mainly due to an increase in the normal samples which ultimately has an affect on the models weights. Figure 5.27 and 5.28 explain the model accuracy and loss after 20 epochs, which shows that the training accuracy steadily increases and could be ran for more epochs until it starts to plateau. The validation loss generally decreases from start to end, but there seems to be fluctuation and ultimately indicates a case of over-fitting. We ran the same experiment again, but

this time with 50 epochs to allow the training accuracy to flatten out and observe the affect it has on the test score.

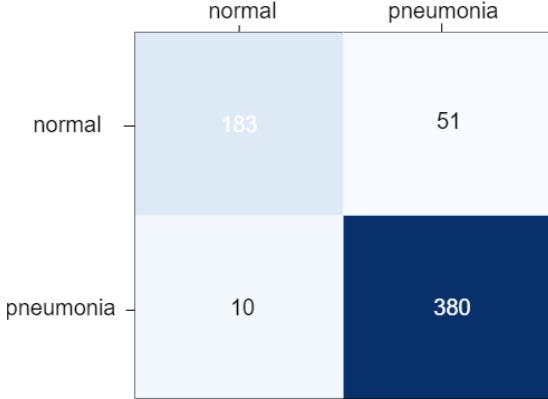


Figure 5.29: Testing results after increasing epochs

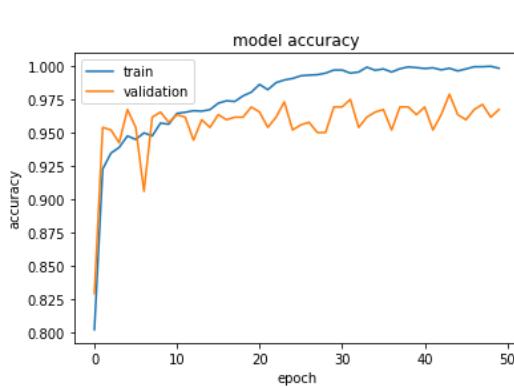


Figure 5.30: Model Accuracy

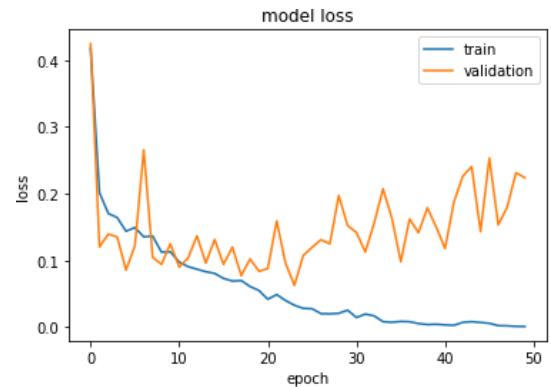


Figure 5.31: Model Loss

From figure 5.30 and 5.31, we can see that running the same experiment for 50 epochs generally flattens out the training process and this seems to increase the number of true negatives, which ultimately leads to accuracy increasing to 90% and specificity increasing from 0.74 to 0.78. Again recall is slightly lower at 0.97 compared to 0.98, but this is expected as the model is trained with a combination of higher epochs and increased number of normal X-ray images. The model accuracy and loss shows a discrepancy between the training and validation even though the testing accuracy increases compared to the previous experiment. This could suggest that the training and testing images are quite similar.

5.5 Summary of Results

Experiment	Accuracy(%)	Recall	Specificity	Precision	<i>F1-Score</i>
3Conv+3FC	77%	0.99	0.38	0.73	0.84
VGG16	74%	0.99	0.32	0.71	0.83
Augmentation	89%	0.97	0.74	0.88	0.93

Table 5.1: Summary of Experiment Results

Table 5.1 is a summary of the experiment results which shows the most important metrics when evaluating the model performance. The table also summarises the most important experiments as a result of tuning certain parameters. An additional performance metric used to summarise a models performance is the F1-score which provides a harmonic average of precision and recall. The F1-score is represented with the following formula:

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

What can be observed from the table above is that although the six-layered network and the transfer learning approach resulted in a very high recall score, both experiments appeared to show a bias towards the positive class, which is why specificity is low for both experiments. The final F1-Score for the two experiments were quite similar. However, a much smaller network with optimised convolutional layers achieved better results across all the metrics(except recall)compared to the transfer learning model.

After finding that a simpler model produced a higher F1-score and specificity, we needed to counteract the bias towards the positive class by introducing data augmentation to double the number of normal samples via random rotations which ultimately increased performance across all metrics, except recall as the model was fed a more balanced distribution of positive and negative samples. However, recall largely remained close to one, showing that the model did very well in recognising the infection in the images while increasing in specificity and ultimately boosting the F1-score.

5.6 Model deployment & Interface

One of the requirements for the project was to ultimately show how a model could be trained and deployed, demonstrating the applications of deep learning through an interface. To do so we created a front-end application using **Streamlit**, which is a open-source Python library that allows construction of web-apps for machine learning and data science purposes.

The web-apps purpose is to show how deep learning can be used to provide a fast diagnosis of medical images efficiently. A core principle of the web-app was to provide users with a diagnosis and a probability of how confident the model believes the infection is present, or equally, how confident it is in diagnosing the patient as healthy. To extract the model confidence we can use the last layer of the model which includes 2 neurons and a Softmax activation function, as shown in figure 5.14. Softmax squashes a vector in the range of 0-1, and all resulting elements add up to one and is applied to the output score of the two neurons, s . The Softmax function cannot be applied independently to each s_i since it depends on all elements of s . For a given class s_i , the Softmax function is computed as follows:

$$f(s)_i = \frac{e^{s_i}}{\sum_j^C e^{s_j}}$$

Where s_j are the scores inferred by the model for each class in C . Note that the Softmax activation for a class s_i depends on all the scores in s .

The process of taking a user input of an X-ray image follows the same principles that is required when training the model, where the input image must be pre-processed before passing it through the model and extracting the Softmax score. This process is shown in figure 5.32, and figure 5.33 shows an example of the interface where a user uploads an X-ray image and is provided with the model prediction and confidence for each classification.

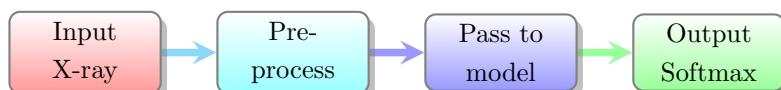


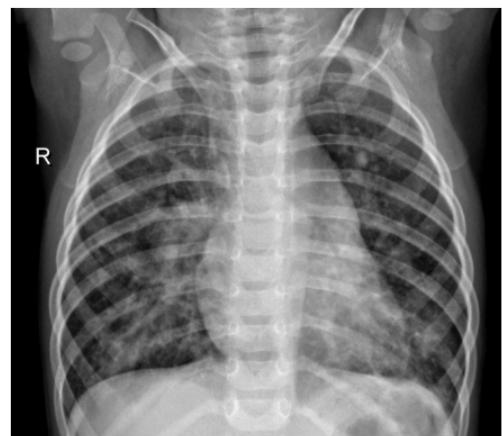
Figure 5.32: Process of diagnosis

Pneumonia classification task

Honours Project Demonstration.

Choose an image...

person15_virus_46.jpeg
[browse files](#)



Uploaded Image,

Classifying...

Model Prediction:Pneumonia

Pneumonia:90%

Normal:10%

Figure 5.33: Web-app

Chapter 6

Evaluation

In this chapter, we will evaluate the project success on the requirements that were defined before the implementation stage, and provide a discussion on the results and key findings that were observed during implementation.

6.1 Overview

The project requirements laid out in chapter 3, was created using the MoSCoW framework which allows us to categorise certain requirements based on their level of importance. In the end, all of the “must” haves for this project were implemented, since this paper applied well known deep learning models like ConvNets. Fine-tuning of the hyper-parameters in order to improve results was carried out, and that was a “must” have for the project, as it showed continual process of improving results as depicted in figure 3.1. The project made use of common neural network libraries such as Keras and Tensorflow for fast prototyping which was also a “must” have. We met the criteria of complying with GDPR rules, as X-rays from patients did not include any Protected Health Information(PHI).

If results from the implementation stage showed good performance across various metrics, which it did , it was a “must” that a simple interface was created to demonstrate deployment of a model. This was achieved using a Python framework called Streamlit, which allowed us to create a web-app accessed through all major browsers which aims to provide a model prediction and confidence score via a Softmax output as percentage, when a user inputs an X-ray that needs to be diagnosed by the model.

Some requirements that were categorised as a “should” have, were not achieved. This was either due to time, or, limited dataset access. The impact of not implementing the

following requirements does not affect the outcome of the project, but meeting those requirements would have improved the overall research.

The first was being able to extend the classification pipeline to other pathologies. Since most of the development time was spent pushing the model performance for pneumonia classification, it was not possible to do so.

Another requirement that would have added to the overall research was applying pneumonia detection as well as classification, which allows a person to understand the area of lungs, which is most indicative of the model prediction. However, because the time spent increasing the classification metrics was the primary purpose, this was not possible.

The last requirement, which was classified as a “could” have, was to deploy the model to a cloud web service. However, upon reflection, we decided this was not part of the requirements list, as the deployment of the model would need to be rigorously examined before letting others access it as a potential diagnostic tool which is not verified by governing a body.

Overall, all of the essential requirements were met for this project given its limitations and it is also important to reflect on the key takeaways from the implementation stage.

6.2 Implementation Discussion

While exploring the dataset, although we did not have much knowledge of where the pathology was present, it was evident that each X-ray image was of different size and quality. As mentioned in the dataset section 4.3, we hypothesised that different X-ray machines could have been used when collecting the images and the amount of radiation applied to perform the X-ray, may vary (Bradley, 2008). This presented some challenges when training the model as resizing of images can sometimes cause a reduction in performance. Although this is not definitive, future research could be performed to examine the effect image sizes and radiation dose has on the model performance.

Moreover, although chest X-rays are the most common form of image modality, a public dataset might not be representative of either class(normal or infection). The impact class imbalance has on the training phase of a model can cause a bias towards one class, and this is evident in figure 5.16 and 5.21 , which shows a large number of false positives in the initial experiments carried out. In this body of work, the main aim was to automate the classification of pneumonia in chest X-rays, and the dataset supplied had a higher number of positive cases of the infection, compared to negative cases which

helped with the aim of this paper, but introduced problems like low specificity for some experiments.

After running multiple experiments show in chapter 4, we observed that after end-to-end training and hyper-parameter tuning of a six-layered CNN, it outperformed an 'out-of-the-box' transfer learning approach using VGG16. The VGG16 model was used exclusively as a feature extractor by freezing the convolutional layers and attaching the same fully connected architecture as the six-layered model. The six-layered model outperformed the transfer learning approach across all metrics, except recall, including accuracy, specificity, precision and F1-score. **Results suggest that training a model end-to-end was better suited compared to the transfer learning approach, for the specific dataset used in this paper.**

We also noted that a novel approach to combatting issues of low specificity was to oversample the under-represented healthy X-ray images. After applying random rotations on the healthy images to create new samples, we retrained the optimised six-layered CNN with an increased number of normal X-ray images and observed that specificity almost doubled, with all other metrics increasing except recall, which dropped slightly but remained close to a value of one which is desirable.

6.3 Legal, Social, Ethical, Security and Professional Considerations

This project is an integral part of the Honours degree accreditation by the British Computer Society(BCS). Therefore, the author(s) must observe the BCS code of conduct that can be found at the following address: <https://www.bcs.org/membership/become-a-member/bcs-code-of-conduct/>. To meet the requirements of the code of conduct it is important to full engage with aspects that are impacted a result of this project. These aspects include the **legal, social, ethical, security and professional issues**. In this section, we will cover the aspects which could potentially be impacted as a result of the outcome of this project.

6.3.1 Legal

Firstly, the legal aspects of this project need to be considered as this project uses patient X-ray images. It is important to note that the dataset used does not include any Patient Health Information(PHI), therefore there is no way to identify a patient from his/hers X-ray image. This aligns with the GDPR rules which were mentioned in Chapter 3, as the patients details are not provided in the public dataset, hence GDPR right for the patients are protected.

Another legal issue that needs to be considered, is if the software created as a result of this project is used for purposes outside of the research realm. The software created is only intended as a proof of concept and created for the purposes of the BSc Honours Degree. Therefore, if the software is used for other purposes, then blame of any outcomes of using the software shall not lay with the author(s) of this paper .

6.3.2 Security

Security issues also need to be considered as the software created for this project accepts a patient X-ray image. It is important to note that no data is saved externally when users submit X-ray for classification. As a result of the web-app running locally no information is stored permanently.

6.3.3 Ethical

Ethical issues are considered for this project as the possibility of the software being used in a clinical environment as a diagnostic tool on real patients is possible. It should be stressed that the project and its associated software components should only be used for research purposes and not as a diagnostic tool on real patients. The consequence of doing so could lead to misdiagnosis and blame for the incorrect diagnosis will not lay on the authors of this paper, as the software is not intended to diagnose patients in a clinical setting.

Chapter 7

Conclusion

In this chapter, we will conclude the project by taking into account the things that were learnt while trying to achieve the aims and objectives of this paper. We will further end by considering the limitations of the paper and discuss some of the future work.

7.1 Overview

In this project, we were able to apply deep learning methods to classify highly complexed X-ray images. We looked at several techniques to push the model performance on the test set by iteratively fine-tuning the model via key hyper-parameters which control the training process and evaluated the performance of each result on several key metrics.

Two-thirds of the global population lacks access to radiology diagnostics ([Mollura et al., 2010](#)), even when medical imaging resources are available. After pushing the performance metrics on the test set, we demonstrated a simple interface that could be used to provide quick diagnostic capabilities through a trained and tested deep learning model.

7.2 Future Work

Although there is a long way to go before deep learning models for pneumonia classification can be deployed in the wild, it is also important to identify some of the limitations of this body of work and look at the types of innovations that could be added to improve performance.

7.2.1 Extra patient information

One of the limitations that was identified during the research of relevant literature and the implementation stage, is that firstly, the only image given per patient to the model was a **frontal** view chest X-ray. Typically, **lateral** view chest X-rays are used by radiologists to further enhance their diagnosis choice and research has shown that up to 15% of accurate diagnoses require a lateral view image ([Raoof et al., 2012](#)). The following figure demonstrate a research paper that uses both the frontal and lateral view X-rays together as input to a DualNet architecture where two CNNs are trained in parallel for each input and their outputs are concatenated together before a final FC layer makes a prediction ([Rubin et al., 2018](#)).

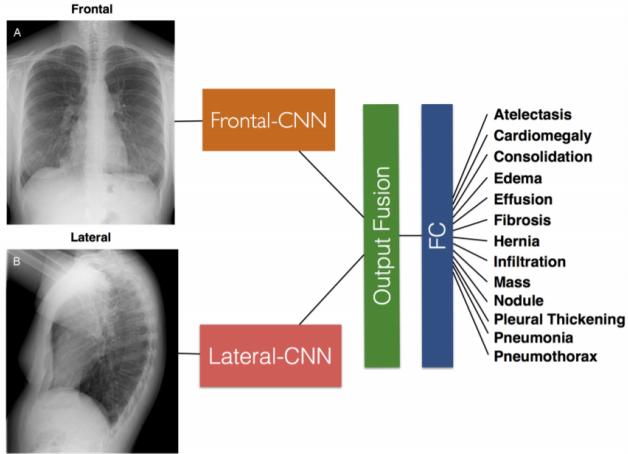


Figure 7.1: Deep learning architecture for two X-ray images as input

Secondly, radiologists are able to review patient history and physical appearance when coming to a conclusion for the diagnosis. It has been shown that when radiologists are not provided this information, their performance decreases when diagnosing the patient solely on the X-ray image ([Berbaum, Franken, & Smith, 1985](#)). This evidence shows that there could be an argument to develop a deep learning architecture where both frontal, lateral X-ray images and patient description such as their breathing state, could be used to further enhance performance on key metrics.

7.2.2 Generative Models

Imbalanced datasets, like the one used in this study, that contain one or more classes that are under-represented, can lead to the model during training becoming bias towards the positive class. Therefore solutions are needed to combat the bias. In this paper, we used simple data augmentation via affine transformations by rotating the X-ray image between -10° and $+10^\circ$ to create new samples. However, these transformations might not introduce enough variation to the training set and although after applying data augmentation in section [5.4](#) resulted in a performance gain, this approach to over-sampling has shown to overfit in specific domains ([Buda, Maki, & Mazurowski, 2018](#))

In some datasets augmenting complex images such as X-rays can prove to be challenging as there might not be too much of a shift in the original image such as flipping and shearing, which are conventional methods.

One approach is to train a generative model that can capture the subtle details in an X-ray image and introduce more variability in the images produced compared to simple data augmentation ([Ali-Gombe, Elyan, & Jayne, 2019](#)). Recent improvements

in generative models have shown the capability to produce high-quality images that are indistinguishable from the real samples. The following figure represents work carried out by Nvidia, which shows their generative model producing high-quality images when training a model on real images of celebrities ([Karras, Aila, Laine, & Lehtinen, 2017](#)).



Figure 7.2: 1024 x1024 facial images generated using generative models

7.2.3 Classification and Localisation

One of the requirements that was discussed as being a “should” have for the project and that was not met, was the ability to perform detection on the X-ray so that the model can point to the areas in the image that are most indicative to the models prediction. One of these methods is to use class activation mappings (CAMS) ([Zhou, Khosla, Lapedriza, Oliva, & Torralba, 2016](#)) which allows an input image to feed into the fully trained network and extract feature maps that are outputted by the final convolution layer filters. CheXNet, a popular CNN reviewed in chapter 2 uses CAMS to output a ‘heatmap’ which shows where the model is looking when making its prediction ([Rajpurkar et al., 2017](#)). This can be highly useful, as certain artefacts or tokens with the X-ray image could cause the model to output a prediction that is totally wrong, due to the model focusing on the wrong parts of the image. The following figure is an example from the CheXNet paper which implements CAMS as a way of relaying to a doctor the areas that they may have missed due oversight when examining the X-ray or indeed when the doctor needs to review the models performance.

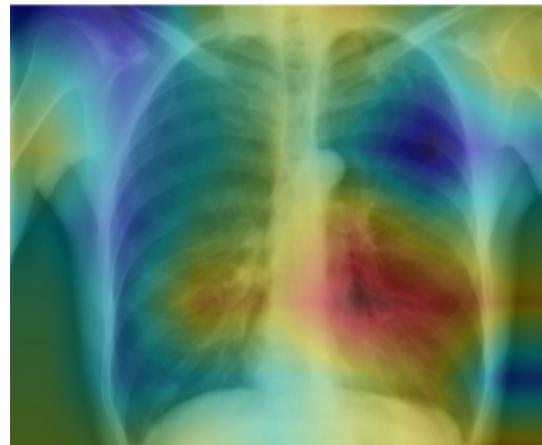


Figure 7.3: Class Activation Mappings indicating where the infection is present

The research in this paper definitely warrants further development before deep learning models can be used in a clinical setting and the advancements in deep learning described above, could definitely be applied and tested to see if key performance metrics are increases.

References

- Alexander, A. (2019). *Mit 6.s191: Convolutional neural networks*. Retrieved from <https://youtu.be/H-HVZJ7kGI0?t=6>
- Ali-Gombe, A., Elyan, E., & Jayne, C. (2019). Multiple fake classes gan for data augmentation in face image dataset. In *2019 international joint conference on neural networks (ijcnn)* (pp. 1–8).
- Berbaum, K., Franken, J. E., & Smith, W. (1985). The effect of comparison films upon resident interpretation of pediatric chest radiographs. *Investigative radiology*, *20*(2), 124–128.
- Bradley, W. G. (2008). History of medical imaging. *Proceedings of the American Philosophical Society*, *152*(3), 349–361.
- Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, *106*, 249–259.
- Castellino, R. A. (2005). Computer aided detection (cad): an overview. *Cancer Imaging*, *5*(1), 17.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (pp. 248–255).
- DrMatthew, L. (2018). *Ai in radiology at stanford: Rise of the machines*. Retrieved from <https://youtu.be/Gigd1rkZTSE>
- Guendel, S., Ghesu, F. C., Grbic, S., Gibson, E., Georgescu, B., Maier, A., & Comaniciu, D. (2019). Multi-task learning for chest x-ray abnormality classification on noisy labels. *arXiv preprint arXiv:1905.06362*.
- Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.
- Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4700–4708).

- Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-IIcus, S., Chute, C., ... others (2019). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *arXiv preprint arXiv:1901.07031*.
- Islam, M. T., Aowal, M. A., Minhaz, A. T., & Ashraf, K. (2017). Abnormality detection and localization in chest x-rays using deep convolutional neural networks. *arXiv preprint arXiv:1705.09850*.
- Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.
- Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., ... others (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5), 1122–1131.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436–444.
- LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4), 541–551.
- Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2), 91–110.
- Mollura, D. J., Azene, E. M., Starikovsky, A., Thelwell, A., Iosifescu, S., Kimble, C., ... others (2010). White paper report of the rad-aid conference on international radiology for developing countries: identifying challenges, opportunities, and strategies for imaging services in the developing world. *Journal of the American College of Radiology*, 7(7), 495–500.
- Organization, W. H., et al. (2001). *Standardization of interpretation of chest radiographs for the diagnosis of pneumonia in children* (Tech. Rep.). World Health Organization.
- Popescu, M. C., & Sasu, L. M. (2014). Feature extraction, feature selection and machine learning for image classification: A case study. In *2014 international conference on optimization of electrical and electronic equipment (optim)* (pp. 968–973).
- Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... others (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*.
- Raoof, S., Feigin, D., Sung, A., Raoof, S., Irugulpati, L., & Rosenow III, E. C. (2012). Interpretation of plain chest roentgenogram. *Chest*, 141(2), 545–558.
- Rimmer, A. (2017). Radiologist shortage leaves patient care at risk, warns royal college.

Bmj, 359, j4683.

- Rubin, J., Sanghavi, D., Zhao, C., Lee, K., Qadir, A., & Xu-Wilson, M. (2018). Large scale automated reading of frontal and lateral chest x-rays using dual convolutional neural networks. *arXiv preprint arXiv:1804.07839*.
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1), 1929–1958.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1–9).
- Video from rsna 2018: Dr. paul chang on ai and radiology.* (n.d.). Retrieved from <https://youtu.be/o2-CrrtuJI?t=300>
- 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. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2097–2106).
- Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2921–2929).

Appendix A

Project Poster

Automatic Classification of Pneumonia Using Deep Learning

Jehanzeb Mobarik & Dr Eyad Elyan

Introduction

Chest X-rays are one of the most popular medical imaging techniques that are used to look for abnormalities within a patient(WHO,2001). One of these abnormalities includes pneumonia infected lungs which appear upon the X-ray as obscure white spots. Radiologists diagnose chest X-rays, however, over the past few years, there have been a drop radiologists leading to increase backlog for the National Health Service(NHS). With the advent of public datasets and compute power, the possibility of real-time diagnosis tool of X-rays showing either healthy or pneumonia infected lungs is needed today more than ever.

Project Aim

This project aims to implement a pneumonia-classification pipeline using various deep learning techniques. Also, this project aims to apply techniques such as data augmentation to deal with class imbalance and transfer learning to help with the challenges that come with a limited dataset

Methods

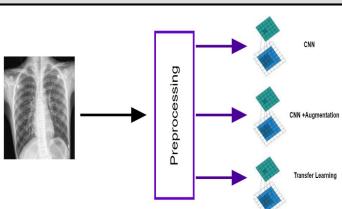


Figure 1: Methods of deep learning

The dataset contained various sizes of images due to different devices used for scanning which proved challenging since all images need to be scaled to the same size in order to classify with a Convolutional Neural Net(CNN). Different methods of using CNNs is illustrated in figure 1. The CNN algorithm is one of the well known deep learning techniques used to extract features automatically without human assistance (Krizhevsky et al., 2012) .

Figures and Results

normal		pneumonia	
normal	91	143	pneumonia
pneumonia	2	388	pneumonia
normal	183	51	pneumonia
pneumonia	10	380	pneumonia

Figure 2 : Confusion Matrix before(left) and after(right) data augmentation



Figure 3 : Example of Data augmentation

By running several experiments and fine tuning **hyper-parameters**, our initial model of 3 convolutional layers and 3 fully connected layers achieved an **accuracy** of **77%** on the test set. However, the model suffered from low **specificity** and **precision** as the dataset was imbalanced with normal X-rays under represented. From the confusion matrix in figure 2, it can be observed the model is highly biased towards the positive class resulting in a high false positive count.

To combat this under representation, data augmentation is used to increase the number of normal X-rays via affine transformations. Figure 3 shows a case of augmentation where a single patient X-ray can be rotated in various degrees to create new samples.

Our model was then trained on the original and augmented images which resulted in **accuracy** increasing to **90%** with both **precision** and **specificity** increasing. Figure 2 shows a confusion matrix(right) of the model evaluated on the test set showing less false positives and increase in false negatives

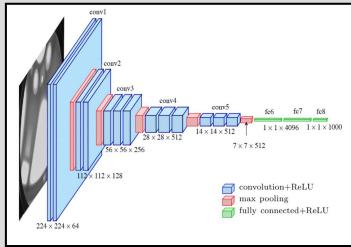


Figure 4: VGG16 Architecture

Algorithm	Accuracy	F1-Score
CNN	77%	0.84
CNN+Aug	90%	0.93
VGG16	74%	0.83

Table 1: Summary of Algorithm Performance

A common approach to overcoming limited dataset size and long training time is to utilize transfer learning. This approach allowed us to use a pretrained VGG16 model, shown in figure 4, whose convolutional layers are trained to extract low to high level features of an image. We adapted the VGG16 model to work on X-ray images by freezing the convolutional layers and training the fully connected layers which allows us to take advantage of VGG16's feature extraction without training from scratch (Simonyan et al, 2015)

We evaluated the performance of each of the algorithms on accuracy and F1-score, which is a measure of the harmonic average between precision and recall. Three different algorithms were compared where it was noted that transfer learning produced sub-optimal scores compared to a smaller network. The table above shows a CNN without data augmentation producing higher accuracy and F1-score compared to the transfer learning approach. Finally, the same CNN model with data augmentation outperformed all other algorithms on accuracy and F1-score

Conclusion

In this project, we provided an overview of the deep learning techniques applied on medical images to diagnose pneumonia in chest X-rays. Different CNN models were trained and tested on chest X-rays where it was observed that novel methods such as transfer learning via the VGG16 model performed worse compared to a much simpler model with only 3 convolutional layers. We also demonstrated that techniques such as data augmentation to help with class imbalance greatly helped in reducing the number of false positives. Ultimately this project proved there is a place where deep learning can be utilised in healthcare to automate diagnosis. In future work we hope to build a generalised classification pipeline on more than one pathology.

Acknowledgments

I would like to thank my supervisor, in particular, Dr Eyad Elyan for his supervision during this project. Without his guidance and encouragement this project would not have been possible. I would also like to thank my family for their support during this project.

References

Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

WHO. Standardization of interpretation of chest radiographs for the diagnosis of pneumonia in children. 2001

More Information

Email: j.mobarik@rgu.ac.uk



**ROBERT GORDON
UNIVERSITY ABERDEEN**

SCHOOL OF COMPUTING SCIENCE AND DIGITAL MEDIA

Computer Science

Appendix B

Project Proposal and Ethics Form

Detailed Project Proposal

First Name:	Jehanzeb
Last Name:	Mobarik
Student Number:	1612872
Supervisor:	Dr. Eyad Elyan

Defining your Project

1.1 Project title

Help: a brief statement about what you are actually going to do.

Using Deep Learning to Read Medical Images
Can Deep Learning diagnose chest X-rays?

1.2 Background

Help: Provide the background to your project. This section should highlight the main topics in the area you are going to research. Essentially what is the project about, what has been done before and why is this project important? ~500 words

For the past few years Radiology, a branch of medicine that uses medical imaging to diagnose patients, has been disrupted due to the advances in machine learning. The types of medical images range from X-rays, CT scan (computed tomography scan), ultrasound, etc. Existing solutions which exist in healthcare include CAD, computer aided diagnosis software which uses machine learning to detect for example Cancer in Mammograms and Cardiomegaly in chest X-rays. Although, these techniques have used older machine learning algorithms to diagnose and are not as accurate as some of the newer methods of learning.

With the advances in Convolutional Neural Networks more and more research has been published in the domain of image recognition for medical scans than ever before. The thesis proposed in this document involves applying the CNN algorithm to diagnose X-rays. Specifically, the project aims to diagnose whether a chest X-ray which is fed into a trained model can accurately predict an infection called Pneumonia. It is an infection of the lungs which has led to one million people being admitted into hospital in the USA. The project will be using a publicly available dataset of Chest - X rays which include 5,863 images of both normal patients and patients who have Pneumonia.

These images are both anterior and posterior and were selected from cohorts of paediatric patients of one to five-year olds from Guangzhou Women and Children's Medical Centre. These images were taken as part of a patients' routine clinical care. Furthermore, the diagnosis for the images were graded by two expert physicians before being cleared for training an AI system.

Currently there is a shortage of trained radiologists making it harder for patient X-rays to be diagnosed on time. With the advent of deep learning and availability of medical datasets several methods have been produced which have shown promising results and can speed up diagnosis. The most promising of all includes the ChexNet project led by researchers at Stanford ML Group. They developed a 121-layered CNN trained on Chest X-ray 14, a public dataset, holding 100,000 frontal view X-ray images with 14 different pathologies, which includes pneumonia. These images were first classified by trained radiologists, although the appearance of pneumonia in X-ray images is often difficult. This can cause discrepancies amongst radiologists' diagnosis. Results showed that ChexNet achieves an F1 score of 0.435, higher than the radiologist average of 0.387.

As mentioned previously, due to the shortage of trained radiologists, the National Health Service has seen waiting lists grow and patients waiting for more than 6 weeks to have medical imaging performed on them. The importance of this project is to demonstrate that with deep learning techniques, X-rays can be successfully diagnosed at an acceptable rate.

[Kermany, Daniel; Zhang, Kang; Goldblum, Michael (2018)," Labelled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification", Mendeley Data, v2]

1.3 Motivation

Help: To whom is this project important? A project must address a question/problem that generates a small piece of new knowledge/solution. This new knowledge/solution must be important to a named group or to a specific client (such as a company, an academic audience, policy makers, people with disabilities) to make it worthwhile carrying out. This is the **motivation** for your project. In this section you should address who will benefit from your findings and how they will benefit. ~300 words

The main beneficiaries of this project include Radiologists and other medical professionals (Since in some cases Radiologists are not the only experts to review X-rays).

The main motivation is to assist in the diagnosis of Medical X-rays. In scenarios where there may be a disagreement on whether a patient should be diagnosed with Pneumonia, an accurate deep learning model can lend another opinion to make a final decision. It should be stressed that even if an accurate model can be built it is important for the final decision to still be taken by the expert.

It is often asked whether artificial intelligence will replace Radiologists? In the eyes of Mayo Clinic electrophysiologist Paul Friedman, the answer is no. He believes that it will provide tools to help with diagnosis. The analogy given is that when using a flash light to see in the dark, the flashlight isn't replacing your eyes, rather it is letting you see further and use your eyes more effectively.

1.4 Aim & Objectives

Help: Outline what are the main things your project is going to do and what steps or milestones will be used to achieve this aim. The Aim is unlikely to change throughout your project; however, the objectives are likely to adapt to your ongoing research and development. In particular it is highly likely that you may wish to split objectives into sub-objectives as work progresses. A good clear set of objectives give you something to evaluate your final project against.

Aim: Build a deep learning model and an interface which allows for Chest X-rays to be uploaded and decide whether Pneumonia is present.

Objective 1: Study existing techniques that have already been published

Objective 2: Research relevant literature in deep learning and medical imaging

Objective 3: Grasp a solid understanding of Convolutional Neural Networks (CNN)

Objective 4: Implement correct technologies to build a machine learning pipeline

Objective 5: Try to look at how to improve on the existing technologies

Objective 6: Conduct appropriate testing to evaluate model performance

Objective 7: Research how the front-end interface should be built to allow for ease of access.

1.5 Key Techniques

Help: Perform some initial research into the area and outline what techniques you may research in further detail here. The techniques you cover here should include references to the papers where you have sourced the information. The techniques mentioned here are very likely to become the section headers in your literature review.

The main technique that will be used to analyse chest X-rays will be using deep learning. More specifically I will be using convolutional neural networks which is a class of deep neural networks and is mainly used for computer vision.

1.6 Legal, Social, Ethical, Professional and Security issues

Help: Here you should discuss any legal, social, profession and security issues that you believe may occur during the course of your project. It is not acceptable to write none in this box, all projects, regardless of focus will have to address issues in one, or more, of these categories. This is an extremely important part of your honours project to which there is no correct answer, this section must be fully discussed with your Honours Supervisor.

Since this project has the possibility of being used in a real-life environment where professionals can upload Chest X-rays and be given a diagnosis. There could be a possibility that the algorithm can give the wrong diagnosis and hence a wrong decision can be made on the patient treatment. In order to avoid this situation, this software should be used in accordance with diagnosis from other professionals so that another opinion can be given to validate the decision of the algorithm.

The responsibility in this situation lies with the healthcare professional to use the software in the correct manner.

1.7 Project Plan

Help: This is the project plan as to how you will go about achieving the objectives of the project.

1. *Gathering relevant research*
 - 1.1 Literature Review
 - 1.2 Review of technologies that can be used
2. *Review prior work*
 - 2.1 Gather examples of previous work where CNN has been used in Medical image diagnosis
 - 2.2 Report on these examples
3. *Design*
 - 3.1 Requirement gathering
 - 3.2 Non functional and functional requirements
 - 3.3 Choose libraries that will be used
4. *Implementation*
 - 4.1 Build model against the data provided
 - 4.2 Build interface for X-ray images to be uploaded
5. *Testing*
 - 5.1 Test model against test set
 - 5.2 Report on the finding
6. *Findings*
 - 6.1 Explain the findings when reporting test results
 - 6.2 Visualisation of results
7. *Evaluate project*
 - 7.1 Changes made from initial project plan
8. Conclusion

1.8 Ethics Form



STUDENT PROJECT ETHICAL REVIEW (SPER) FORM

The aim of the University's *Research Ethics Policy* is to establish and promote good ethical practice in the conduct of academic research. The questionnaire is intended to enable researchers to undertake an initial self-assessment of ethical issues in their research. Ethical conduct is not primarily a matter of following fixed rules; it depends on researchers developing a considered, flexible and thoughtful practice.

The questionnaire aims to engage researchers discursively with the ethical dimensions of their work and potential ethical issues, and the main focus of any subsequent review is not to 'approve' or 'disapprove' of a project but to make sure that this process has taken place.

The *Research Ethics Policy* is available at www.intranet.rgu.ac.uk/credo/staff/page.cfm?pg=7060

Student Name	Jehanzeb Mobarik
Supervisor	Dr Eyad Elyan
Project Title	Using Deep learning to Read Medical Image
Course of Study	BSc Computer Science
School/Department	School of Computing Science and Digital Media

Part 1 : Descriptive Questions			
1	Does the research involve, or does information in the research relate to:	Yes	No
	(a) individual human subjects		X
	(b) groups (e.g. families, communities, crowds)		X
	(c) organisations		X

	(d) animals?		X
	Please provide further details:		
2	Will the research deal with information which is private or confidential?	Yes	No
			X
	Please provide further details:		
	The dataset that the model will be run against is a public dataset of chest X-rays. It does not contain patient health information (PHI)		

Part 2: The Impact of the Research			
3	In the process of doing the research, is there any potential for harm to be done to, or costs to be imposed on	Yes	No
	(a) research participants?		X
	(b) research subjects?		X
	(c) you, as the researcher?		X
	(d) third parties?		X
Please state what you believe are the implications of the research:			
4	When the research is complete, could negative consequences follow:	Yes	No
	(a) for research subjects		X
	(b) or elsewhere?		X
Please state what you believe are the consequences of the research:			

Part 3: Ethical Procedures

5	Does the research require informed consent or approval from:	Yes	No
	(a) research participants?	<input checked="" type="checkbox"/> X	
	(b) research subjects	<input checked="" type="checkbox"/> X	
	(c) external bodies	<input checked="" type="checkbox"/> X	
	If you answered yes to any of the above, please explain your answer:		
6	Are there reasons why research subjects may need safeguards or protection?	Yes	No
		<input checked="" type="checkbox"/> X	
	If you answered yes to the above, please state the reasons and indicate the measures to be		
7	Has PVG membership status been considered?	Yes	X
	(a) PVG membership is not required.	<input checked="" type="checkbox"/> X	
	(b) PVG membership is required for working with children.		
	(c) PVG membership is required for working with protected adults.		
	(d) PVG membership is required for working with both children and protected		
	If you answered yes to (b), (c) or (d) above, please give details:		
8	Are specified procedures or safeguards required for recording, management, or storage of data?	Yes	No
		<input checked="" type="checkbox"/> X	
	If you answered yes to the above, please outline the likely undertakings:		
	No data will be stored on the software created as a result of this project.		

Part 4: The Research Relationship

9	Does the research require you to give or make undertakings to research participants or subjects about the use of data?	Yes	No
	If you answered yes to the above, please outline the likely undertakings:		
10	Is the research likely to be affected by the relationship with a sponsor, funder or employer?	Yes	No
	If you answered yes to the above, please identify how the research may be affected:		

Part 5: Other Issues

11	Are there any other ethical issues not covered by this form which you believe you should raise?	Yes	No

Statement by Student

I believe that the information I have given in this form is correct, and that I have addressed the ethical issues as fully as possible at this stage.

Signature	J. Mobarik	Date	29/09/2019
-----------	------------	------	------------

If any ethical issues arise during the course of the research, students should complete a further Student Project Ethical Review (SPER) form.

The Research Ethics Policy is available at
www.intranet.rgu.ac.uk/credo/staff/page.cfm?pge=7060

Part 6: To be completed by the supervisor

12	Does the research have potentially negative implications for the University?	Yes	No
	If you answered yes to the above, please explain your answer:		
13	Are any potential conflicts of interest likely to arise in the course of the research?		No
	If you answered yes to the above, please identify the potential conflicts:		
14	Are you satisfied that the student has engaged adequately with the ethical implications of the work? [In signifying agreement, supervisors are accepting part of the ethical responsibility for the project]	Yes	
	If you answered no to the above, please identify the potential issues:		
15	Appraisal: Please select one of the following		
	The research project should proceed in its present form – no further action is required	X	
	The research project requires ethical approval by the School Ethics Review Panel		
	The research project needs to be returned to the student for modification prior to further action		
	The research project requires ethical review by an external body. If this applies please give details		
	Title of External Body providing ethical review		
	Address of External Body		
	Anticipated date when External Body may consider project		

Affirmation by Supervisor

I have read the student's responses and have discussed ethical issues arising with the student. I can confirm that, to the best of my understanding, the information presented by the student is correct and appropriate to allow an informed judgement on whether further ethical approval is required.

Signature	Eyad Elyan	Date	29.09.2019
-----------	------------	------	------------

Appendix C

Project Log

Week	Tasks
16 th September	Project and supervisor chosen. Discussed the proposed project Eyad provided to go away and study regarding deep learning concepts
23 rd September	Meeting postponed due to Eyad attending an external meeting however work was done on the project
30 th September	Project spec and ethics form submitted
6 th October	Understand concepts of Latex and playing about with Latex template provided by Eyad.
13 th October	Meeting with Eyad where we discussed some of the deep learning concepts further to enhance my knowledge
17 th October	Started Lit Review by explaining the overview of the project and traditional methods of machine learning
24 th October	Continued to work on Lit review and other CW's
31 st October	Meeting with Eyad to give update on Lit review- He was disappointed in the progress as more content at this stage could have been added.
7 th November	State of the art and Future work section completed and Lit Review submitted
14 th November	AI and Cloud computing CW's were priority as well as Language & Compliers module-Mark Bartlett
21 st November	Requirements Spec created
29 th November	CW and Exam revision
6 th December	Prepping for exams
13 th December	Christmas Break
1 st January	Trello and GitHub created to track implementation changes
8 th January	Pre-processing
15 th January	Learning to use CL for DGX Machine and learn principles of Docker
22 nd January	Not much work as back at home in Glasgow
29 th January	First experiment ran and shown to Eyad during proof of concept meeting
5 th Feb	More work was done to improve results
12 th Feb	Issues with different results every time – neither me or Eyad could fix the solution
19 th Feb	Looked at other alternatives such as reducing learning rate which slowly stabilised results
26 th Feb	More experiments ran but failed
4 th March	CW for other modules was prioritised
11 th March	Further work was done on CW
18 th March	Hyper-parameter tuning was applied which improved results
25 th March	Work on Ethical hacking
1 April	Ethical hacking CW due but then extended due to Virus
8 th April	Online meeting scheduled with Eyad over Zoom where further advice was given on how to structure report
15 th April	Transfer Learning experiments completed
22 nd April	CW for modules was prioritised
29 th April	HCI CW submitted
6 th May	Moved back down to Glasgow due to extenuating circumstances
13 th May	Enterprise Web CW submitted and Report starting to come along
14 th May	Completed the Experiments section and augmentation experiments
21 st May	Most of the experiments tidied up with commenting of jupyter notebooks
24 th May	Project finished

