**PNEUMONIA DETECTION IN CHEST X-RAY IMAGES
USING DEEP CONVOLUTIONAL NEURAL NETWORK**

by

Hee Kyoung Nam, BSc, Ryerson University, 2022

A Major Research Project

presented to Ryerson University

in partial fulfillment of the requirements for the degree of

Master of Science

in the Program of

Data Science and Analytics

Toronto, Ontario, Canada, 2022

© Hee Kyoung Nam 2022

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A MAJOR RESEARCH PROJECT (MRP)

I hereby declare that I am the sole author of this Major Research Paper. This is a true copy of the MRP, including any required final revisions.

I authorize Ryerson University to lend this MRP to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this MRP by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my MRP may be made electronically available to the public.

Hee Kyoung Nam

PNEUMONIA DETECTION IN CHEST X-RAY IMAGES USING DEEP CONVOLUTIONAL NETWORK

Hee Kyoung Nam

Master of Science 2022

Data Science and Analytics

Ryerson University

ABSTRACT

Pneumonia is a life-threatening disease that leads to the death of individuals within a short period due to the flow of fluid in the lungs. Early diagnosis and drugs are very important to avoid the progress of the disease and chest X-ray is the most widely used method to diagnose pneumonia by radiologists. However, human-assisted approaches have some drawbacks such as expert availability, treatment cost, and availability of diagnostic tools. Hence, the need for an automated system that operates on chest X-ray images comes into place. This project proposes deep convolutional neural network and transfer learning approaches that automatically detect pneumonia from x-ray images. Moreover, the impact of image augmentation techniques on deep learning framework and visualization of feature maps and class activation maps are examined in this study. According to the findings of this project, deep convolutional neural network can detect pneumonia from chest x-ray images. The proposed Xception based CNN with transfer learning utilizing ImageNet parameters obtained the following results: accuracy (0.95), recall (0.94), precision (0.96), F1-score (0.95), and AUC score (0.95). These positive results allow us to consider the model as an alternative that can be useful in low-resource countries with a shortage of radiology experts and equipment.

Key words: Pneumonia, medical image classification, CNN, Transfer learning, Image augmentation, feature map visualization, class activation maps

ACKNOWLEDGEMENTS

Table of Contents

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

A. Problem definition

Pneumonia is a life-threatening disease that leads to the death of individuals within a short period due to the flow of fluid in the lungs. Almost a third of all victims were children and it is the leading cause of death for children under 5 [1]. The patients with pneumonia have their air sacs of the lungs called alveoli, filled with pus and fluid which makes them painful to breath and limit the oxygen intake. Therefore, early diagnosis and drug treatment at the right time is a key factor along the progress of the disease [2]. Today, one of the most conventional medical techniques used to diagnose pneumonia is chest x-ray. During diagnosis, expert radiologists look for white spots in the lungs, called infiltrates that identify an infection and white areas for the pneumonia fluid in the lungs. However, differentiating them in the x-ray images, especially in the early stage is difficult and it is where many radiologists fail to make the correct diagnosis. Either of false positive or false negative diagnosis has substantial impacts on patients. Hence, we need more reliable and consistent computational methods in the diagnosis steps that operates on chest x-ray images. The aim of this project is to utilize the power of machine learning and propose a DCNN model that diagnose pneumonia using chest x-ray radiographs with better accuracy.

B. Dataset

The chest X-ray pneumonia dataset is publicly available from Kaggle, and the image data is originally collected by Guangzhou women and children's medical center [3]. The dataset consists of 5932 chest x-ray images of one to five years old pediatric patients and images are classified into two categories, pneumonia and normal. The diagnoses for the images were graded by two specialized radiologists. Pneumonia x-rays include both bacterial and viral affected pneumonia as they share similar features. The train folder contains a total of 5292 images (3933 pneumonia images and 1359 normal images) as shown in Figure 1. The validation folder contains a total of 16 images, 8 images for each class. The test folder contains a total of 624 images (390 for pneumonia and 234 for normal classes). The number of pneumonia class images is almost three times greater than the number of normal class images, which shows that there is an imbalance in the distribution of the dataset.

C. Research Question

Since there is an imbalance of class sample size in the dataset, I examined how the imbalance affects the pneumonia classification performance of models. Also, prior studies have disclosed that augmenting image data for deep learning enhances the quality of training dataset and resolves overfitting problem [4][5]. In this study, I created three different versions of training dataset using the original dataset: 1) augmented image dataset, 2) balanced & augmented image dataset and 3) raw image dataset. With these three sets of training data, I found out which method and type of dataset performs well on detecting pneumonia from chest x-ray images.

In terms of creating a diagnosis model, different DCNN architectures are customized. I have customized my own CNN model with different parameters and made a comparison with transfer learning models to see which method are more effective for medical image classification. Five different pre-trained deep convolutional neural networks that are well known for medical image classification: ResNet50, DenseNet121, VGG16, Inception and Xception, are selected for transfer learning [6][7][8].

After the models are trained, I further explored what happens in the DCNN layers by visualizing feature maps of the transfer learning models and class activation maps to see which areas of images are used to interpret the prediction decision by the proposed model.

2. LITERATURE REVIEW

Razaak et al. discussed the challenges of AI-based methodology with regard to medical imaging [9]. Several biomedical image detection techniques have been proposed in the literature to help in diagnosing numerous diseases such as breast and renal cancers [10] [11].

For lung diseases in particular, some of proposed techniques used machine learning algorithms and others used deep learning methods for feature extraction and classification. The parameters of these methods are optimized to achieve high accuracy in disease classification. In Antin et al. study, the authors used logistic regression as their supervised machine learning method and

DenseNet as their deep learning method to diagnose Pneumonia [12]. From a relatively low scale of The Area Under the Curve (AUC) of logistic regression model, the study conclude that logistic regression does not adequately capture the complexity of X-ray images. Whereas densely connected network having 121 layers (DenseNet) achieves a better result. Deep learning is a better advancement over machine learning as it can easily operate on images and extract the features responsible to classify the disease.

There are also several recent works have proved the benefit of data augmentation in improving CNN performance for various deep learning applications [13]. Data Augmentation is a simple method to increase the dataset size by adding more invariant samples and thus reduce generalization error of the small dataset and avoid overfitting [14]. In Monshi et al. study, they demonstrated that the optimization of data augmentation and CNN hyperparameters is an effective tool to extract features from chest X-rays [15]. The CNN architectures used in this study were VGG-19 and ResNet50 and they proposed that such optimization increased the accuracy of both models. In addition, Nishio et al. also proposed a computer-aided diagnosis system for detection of COVID-19 pneumonia [16]. Their proposed model utilized VGG16 and studied the effect of conventional and mix-up method of chest X-ray image augmentation. Their results proved that the combinational conventional and mix-up augmentation methods were more effective than single type or no augmentation method. The conventional data augmentation method included rotation (+/- 15), x-axis and y-axis shift (+/- 15%), horizontal flipping, scaling, and shear transformation. The mix-up augmentation was set to 0.1.

In Sharma et al. built different deep convolutional neural network models to extract characteristics from chest X-rays and categorise them to detect pneumonia [17]. In addition, this study trained their proposed CNN models using both the original and augmented data to evaluate the effect of dataset size on the model performances. The proposed two CNN architecture designs, one with and one without a dropout layer. The principle of dropout layer is to randomly zero out the edges of hidden units at each step of the training phase. The CNN models had convolution layers varied from 32 units to 128 units, max pooling size from 2 x 2 from 3 x 3 and 0.5 for the dropout layer with 0.5. The augmentation techniques included rescale, rotation range, zoom range, width and height shift, shear, horizontal flip, and fill mode to nearest. The models were trained for 20 cycles with adam optimizer with its learning rate set to 1e-4 and the batch

size of 32. Their results of proposed CNN models showed that CNN with dropout layer trained on augmented images outperformed the other models.

In Hashmi et al.'s work [18], pre-trained ResNet50 multilayer architecture was used to identify pneumonia on chest X-ray images. The dataset used in this study is identical to the dataset I used for this project. As the dataset contain insufficient number of X-ray images, data augmentation techniques were deployed to increase the size of the training dataset. The authors further scale up ResNet50 model by compound scaling. Training images were resized to 224*224 and images of healthy chest X-rays were augmented twice with augmentation settings (crop and pad = 0.25, Horizontal shift =0.15, Rotation = 35, Vertical shift = 0.2). The most optimum results were obtained with the learning rate of 0.001 and Stochastic gradient descent optimizer. The proposed compound scaled ResNet50 attained a test accuracy of 98.14%, an AUC score of 99.71 and an F1 score of 98.3 on the test data from the Guangzhou Women and Children's Medical Center pneumonia dataset.

Similarly, the authors in [19] used the same publicly available Pneumonia dataset as my dataset in Kaggle to implement four different architectures including two pre-trained ResNet152V2, MobileNetV2, a Convolutional Neural Network, and a Long Short-Term Memory. Among the proposed four CNN architectures, ResNet152V2 architecture achieved the best accuracy result of 99.22% and the other architectures achieved more than 91% in accuracy, recall, F1-score, precision, and AUC.

Rajaraman et al. addressed the challenges of interpreting CNN model behavior that could adversely affect the clinical decision [20]. CNNs are often described as black boxes as their performance do not give adequate explanations and insight on the form of function. The authors present a visualization strategy for localizing the region of interest in the Pediatric chest X-rays and customized VGG16 model and attain 96.2% and 93.6% of accuracy values in detecting bacterial and viral pneumonia classification respectively. They embedded a CAM-compatible architecture to the customized VGG16 model to visualize the model predictions.

The authors of [21] presented several transfer learning methods as feature extractors to classify normal and pneumonia-infected chest X-rays. Training a deep Convolutional neural network

models from scratch requires a lot of inputs and efforts because there are millions of trainable parameters. Instead, there are baseline models such as ResNet, Xception, or DenseNet available for researchers to reuse their pre-trained weights using transfer learning technique. In this study, Lujan-Garcia et al. use several convolution Neural Network pretrained weights on ImageNet as an initialization for the proposed model to automatically classify healthy and pneumonia chest X-ray images of people. Among the evaluated models, Xception outperformed VGG-16, ResNet and Inception-V3. This network contains 36 convolutional layers including a global average pooling, dropout, layers and adam optimizer is used. The transfer learning technique support numerous studies to resolve a small insufficient dataset problem and help in achieving a good generalization of models. It is reported that the outperformed Xception model achieved a precision value of 0.843, a recall value of 0.992, a F-1 score of 0.912 and an AUC value of 0.962 for the ROC curve. The dataset used in this study also had imbalance issue where pneumonia class had almost three times more samples than normal class samples. To solve the imbalance, the authors chose to use Random Under sampling. The Random Under sampling is a non-heuristic method to help to combat the imbalance problem [22]. The method eliminates a portion of samples from a majority class to balance out the training data.

In Rahman et al. they proposed four different pre-trained deep convolutional neural networks [23]. AlexNet, RsNet18, DenseNet201, and SqueezeNet were used for transfer learning. This study reported three schemes of classifications which were normal and pneumonia, bacterial and viral pneumonia, and normal, bacterial, and viral pneumonia. Based on their research, DenseNet201 outperformed other three deep CNN architectures in three difference classification schemes. The classification accuracy, precision, and recall of normal and pneumonia images, bacterial, and viral pneumonia images, and normal, bacterial, and viral pneumonia were 98%, 97%, and 99%; 95%, 95% and 96%; and 93.3%, 93.7% and 93.2%, respectively.

3. EXPLARATORY DATA ANALYSIS

The following section is conducted to analyse the entire dataset and summarize its main characteristics which would guide the modeling process and help answering preliminary questions. For such image classification problem, we can observe the distributions of variables like class distribution, pixel intensity distribution and more.

A. Distribution of class in dataset
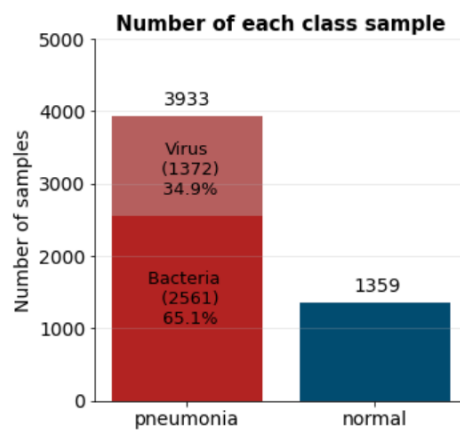
Figure 1. class distribution in the training set

Figure 1 shows the distribution of class in dataset. This data is highly imbalanced. There is almost three times more pneumonia samples than the normal sample size. Such unequal ratio is very normal when it comes to medical data [24]. The problem of class imbalance is that it usually leads a learning bias to the majority class. To deal with the class imbalance problem, many prior studies have used sampling technique to adjust the size of training dataset and this technique is also selected for this project [25] [26].

B. Visualizing samples of each class
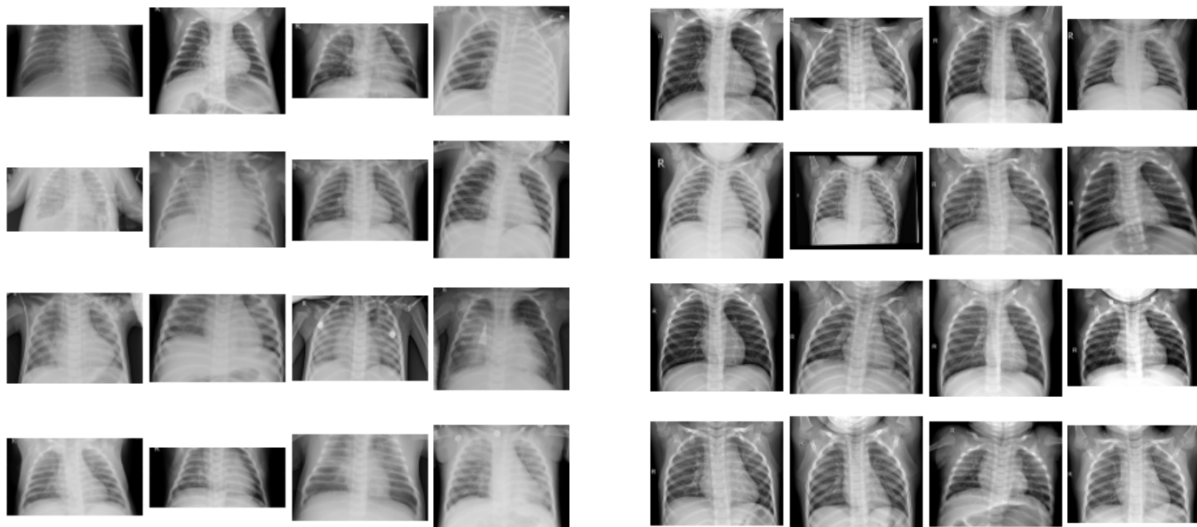
Figure 2. 16 randomly selected images from Pneumonia (left) and Normal (right) training set.

Figure 2 shows samples from each class. Pneumonia images tend to have larger white regions which represent fluid. However in some cases, it is hard to differentiate between an image with pneumonia and an image of a healthy patient with naked eye. We can also see that the width and height of the training images vary.

C. Average image of each class

The average pixel intensity value of each pixel across all images of each class is computed. We see from the average image that pneumonia class image has more widely spreaded white regions around the lung area.
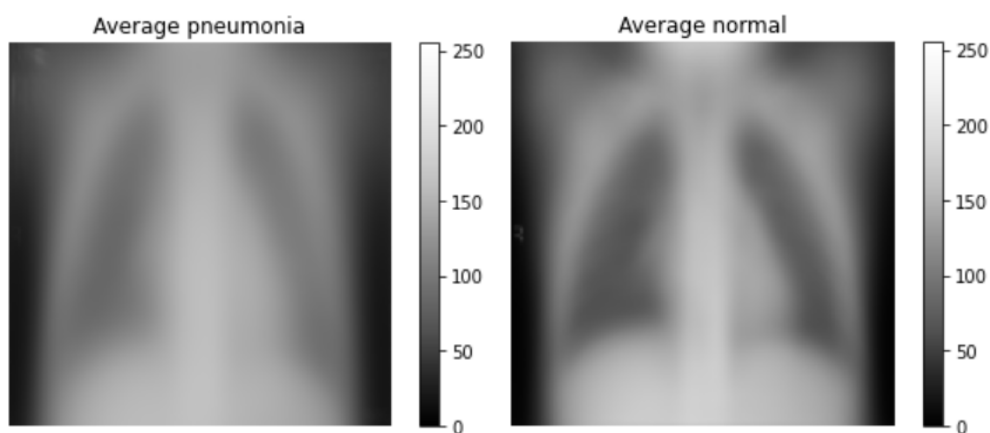
Figure 3. Average image of pneumonia images (left) and normal images (right)

D. Contrast between normal and pneumonia average images

The contrast between normal and pneumonia average images is computed, and the red intensity indicates the areas where difference is shown in the dark areas of the average image and the blue intensity indicates the areas in light colors in the average image. Since the normal lung size and the lung position overlaps between normal and pneumonia samples, the lung itself is shown in light red color. However, as pneumonia samples have larger lungs filled with fluid, the outer part of the lung shows the greatest red intensity. This indicates that the lung size of the pneumonia patients is more spread out compared to the healthy lung.
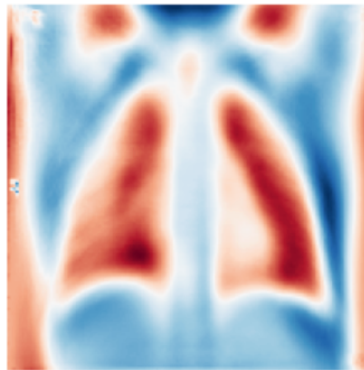
Figure 4. Contrast between the average image of pneumonia and average image of normal class. Red intensity indicates the difference in the dark areas of the average image and blue intensity for the bright areas.

E. Standard deviation (variability)

To observe which area is most variable in each class, I have computed standard deviation of each class images. Bright colors represent higher variability. For the normal case, lung size doesn't change much so there is a border around the outer part of the lung. Whereas for the pneumonia case, there is more variability within the lungs.
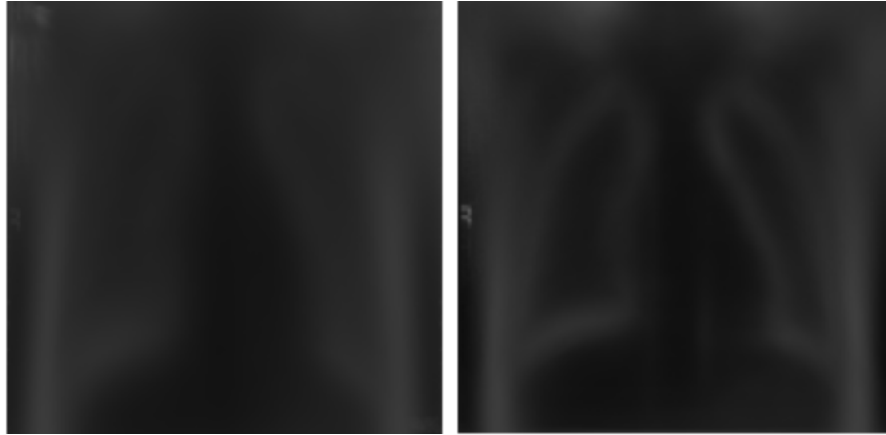
Figure 5. Standard deviation of Pneumonia images (left) and normal images (right)

4. METHODOLOGY AND EXPERIMENTS

This section demonstrates how experiments were conducted and the assessment steps taken to evaluate the overall validity and reliability of proposed architectures for pneumonia detection using chest X-rays. In this image classification problem, customized CNN architecture and transfer learning methods are implemented. All the experiments were performed under the pro-version of Google Colab where 32GB of RAM is available.

A. Image pre-processing

Before training the model, we need image preprocessing steps to improve model performance and reduce generalization errors. Since the width and height of the images vary, I have resized all images to 128 x 128 x 3 for consistency and all the pixel values are divided by 255 for normalization. Then, I have applied two techniques: oversampling and image augmentation to training set before feeding them into models.

I.   Oversampling: Method to improve class imbalance for image data

Through EDA, we have seen that the sample sizes of two classes are noticeably different. The exact imbalance ratio of the training set is as follows:

$$\text{Imbalance ratio} = \frac{Majority\ class\ sample\ size}{Minority\ class\ sample\ size} = \frac{3933}{1359} = 2.894$$

Imbalanced datasets pose a challenge for classification modeling as most of the artificial intelligent based classification algorithms are designed around the assumption of an equal number of class sizes [27]. Unequal distribution of classes in the training set leads to poor predictive performance, specifically for the infrequent class. To resolve this imbalance issue, I have chosen to use over-sampling technique. Over-sampling augmentations re-sample imbalanced class distribution by creating new minority class observations and add them to the training set.

II.   Data Augmentation: avoid overfitting

Image data augmentation is a technique used to artificially generate new transformed images from existing data. It performs as a regularization mechanism that helps in expanding limited datasets and avoiding immediate overfitting at real time over the training set [4].

III.   Data groups

To test that the effectiveness of the augmentation and oversampling methods on medical image classification problem, I have generated three different data groups using the original dataset:

1) Augmented data
   The image augmentation settings selected for in this study are listed in the table 1. The augmentation technique is applied to training set only. Figure 6 shows augmented samples of a pneumonia image.

Augmentation settings	
Augmentation procedure	Parameter value
Rotation range	10°
Zoom range	0.1
Width Shift	0.1
Height Shift	0.1

Table 1. Image augmentation settings
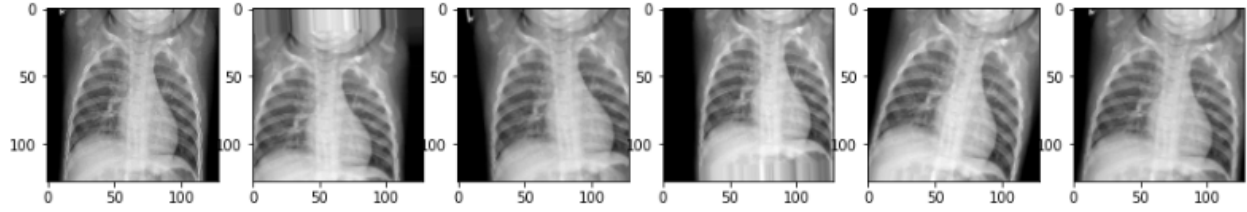
Figure 6. A pneumonia image in the first index (first image) is augmented to five different transformed images, shown in the second to sixth indexes.

2) Balanced (oversampled) and augmented data

Although training set images are augmented to avoid overfitting, the first augmented data group still has a class imbalance issue. Thus, I have used oversampling technique to create synthetic minority class (Normal) data and added the created data to the normal class training set. Downside of oversampling is usually the overfitting of the oversampled classes, so image augmentation is applied together to handle the overfitting problem. Figure 7 shows the change in the sample size before and after the oversampling.

a)   b)
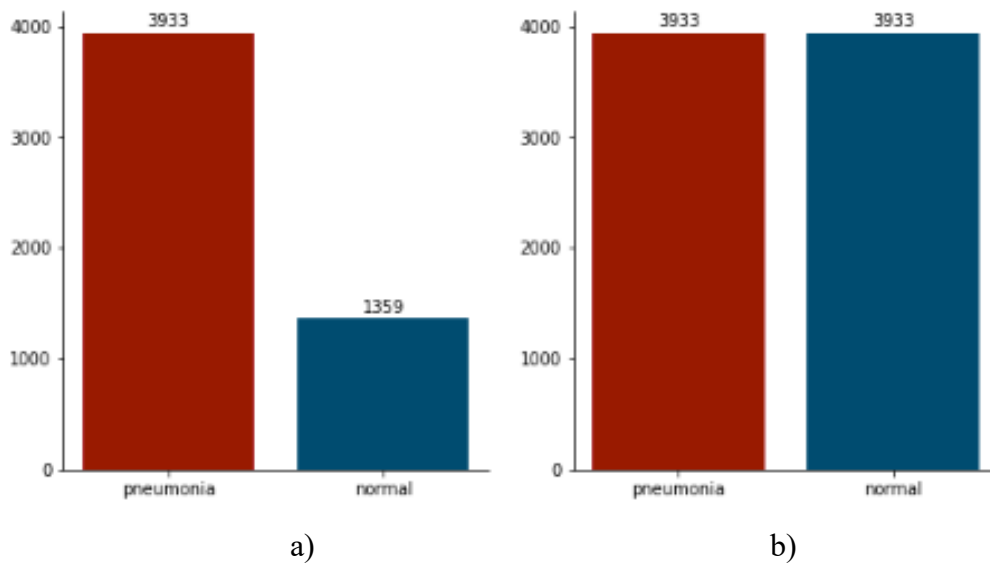
Figure 7. imbalanced dataset before oversampling (a) and balanced dataset after oversampling (b)

3) Without augmentation (raw data)

To evaluate the effectiveness of data augmentation in image classification, neither of the augmentation nor oversampling technique is applied to the training set for the control group.

B. Customizing Convolutional Neural Network

Customized convolutional neural network (CNN) is employed to explore patterns in pneumonia affected chest x-ray images. The proposed customized CNN architecture is demonstrated in Figure 8. The architecture of the CNN model has three main layers, input, feature extraction and classification layers. The first input layer has an input image size of 128 x 128 x 3. For feature extraction, 5 convolutional layers blocks with 3x3 kernel size are built and the convolutional layers are followed by batch normalization with ReLU activation functions. Furthermore, I have added max-pooling layers with 2x2 kernel size and drop out layers in between blocks. After the feature extraction layers, the output is passed to the flattened layer to flat the three-dimensional convolutional layers to one dimensional shape for a dense layer of the classification layer. The dense layer is deeply connected with its preceding layer which allows dense layer neurons to be connected to every neuron of its preceding layer. Then, the classification output layer consists of a dense layer with one node as it is a binary classification problem so sigmoid activation function is used. Also, Adam optimizer is used to reduce the cross-entropy loss. The total number of parameters of the customized CNN is 558,849 where the trainable parameters are 557,761 and non-trainable are 1088. The epoch size is set to 15 and batch size of 32 is selected for CNN architecture.
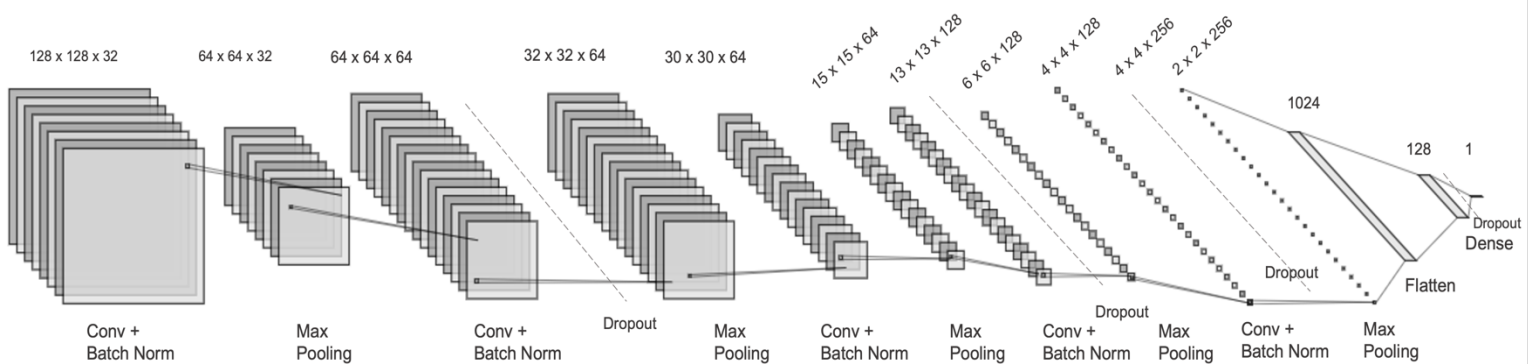

Figure 8. customized CNN architecture

C. Transfer learning

The pretrained deep convolutional neural networks, ResNet50, DenseNet121, VGG16, Inception and Xception were considered for this work. Pretrained weights on ImageNet were used as the Transfer learning method to train the input [28]. For pre-trained models, the original architecture followed by a dense layer with 128 units with ReLU activation function and a final output dense layer with sigmoid function were used to classify images. While instantiating pre-trained models, the 'include_top' argument is set to false, in which case the fully connected output layers to make prediction is not loaded and I have manually added new output layers which include one dense layer with 128 units and ReLU activation function and a dense layer with single node for binary classification problem. A pre-trained model without a top will output activations from the last convolutional block directly. One common approach to summarize the descending activations is to add a pooling layer. Thus, the 'pooling' argument is set to max for global max pooling application which replace the flatten layers and generates one feature map for each corresponding category of the classification task. Each model is run for 15 epochs with a batch size of 32. Also, I have set all the model's layers to be trainable.

D. Training optimization

For training optimization, ReduceLROnPlateau is added as a callback which is a function to be applied at training stages. ReduceLROnPlateau is a scheduling technique that monitors a validation accuracy and if no improvement is seen for a 'patience' number of epochs, it decays the learning rate. I chose the minimum learning rate as 1e-6, the patience value as 2 and the factor value as 0.3 which means the system reduces the learning rate by 0.3 when the validation accuracy stops improving.

E. Performance measures

To evaluate the performance of a binary classification model, specific metrics measures such as loss, accuracy, precision (specificity), recall (sensitivity) and f1-score are examined with the help of classification report and confusion matrix. In addition, the reception operating characteristic curve (ROC curve) and the Area under the curve (AUC) which is the measure associated with it were computed to measure the efficiency of the proposed models.

F. Interpretation of visualization of intermediate layers and class activation maps

Although deep neural networks produce notable outcomes in recognizing images, our understanding of how these models work, especially at the intermediate layers remains unclear [29]. In this section, one sample image is fed into deep neural networks of pretrained models and feature maps of the image will be examined by observing which areas of image are activated by convolutional layers. Similarly, class activation map (CAM) is a powerful technique to get visual explanation of the discriminative image region a CNN is used to identify a specific class in the image [30]. In this process, an image is fed to the network and classification output of the network is computed. The weights of the final convolution layer and the output are stored, and the CAM is computed by multiplying each depth from the output of the final convolutional layer and sum them all. The computed CAM is then extended to the size of the input image and produce a coarse localization map that is highlighting the important regions in the image of the prediction conception.

5. RESULTS AND DISCUSSION

The performances of customized CNN model and pre-trained models were examined with three different data inputs having different augmentation and sampling settings. For CNN architectures, models were built from the scratch. For transfer learning, pretrained weights of models on ImageNet were used as initialization and last few layers were customized. This project aims to find out which method is more effective in classifying pneumonia using chest x-ray images and visualize what happens in the intermediate layers and which region in the images is most relevant to the pneumonia and normal category.

I.   Customized CNN

Training and validation accuracy and loss charts & confusion matrix for the test set

1) Augmented data group:

2) Balanced and augmented data group:
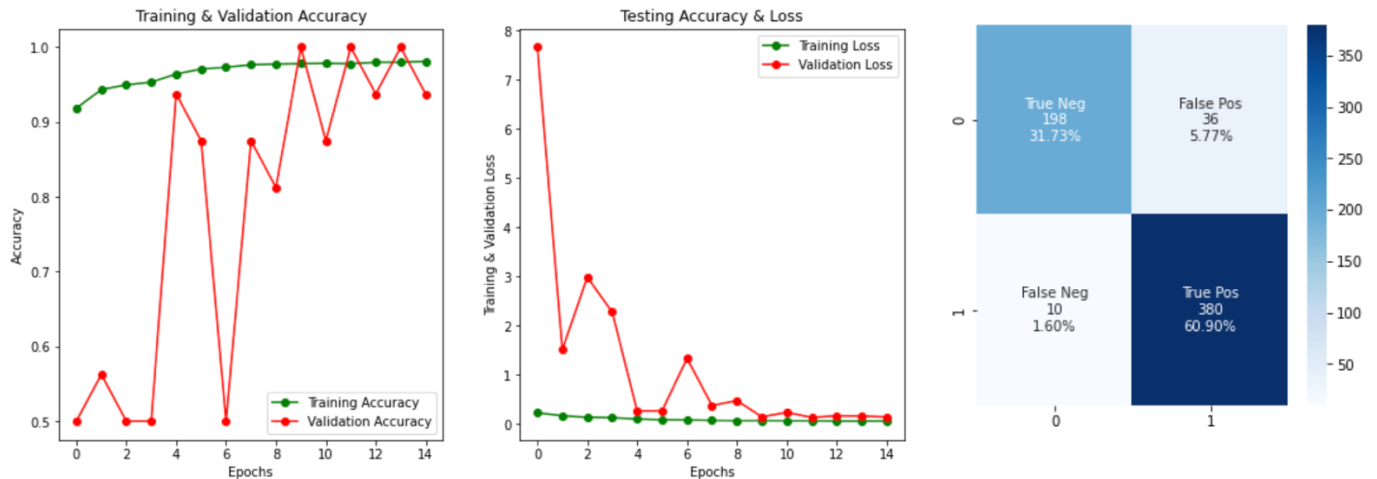
3) Non-augmented data group:
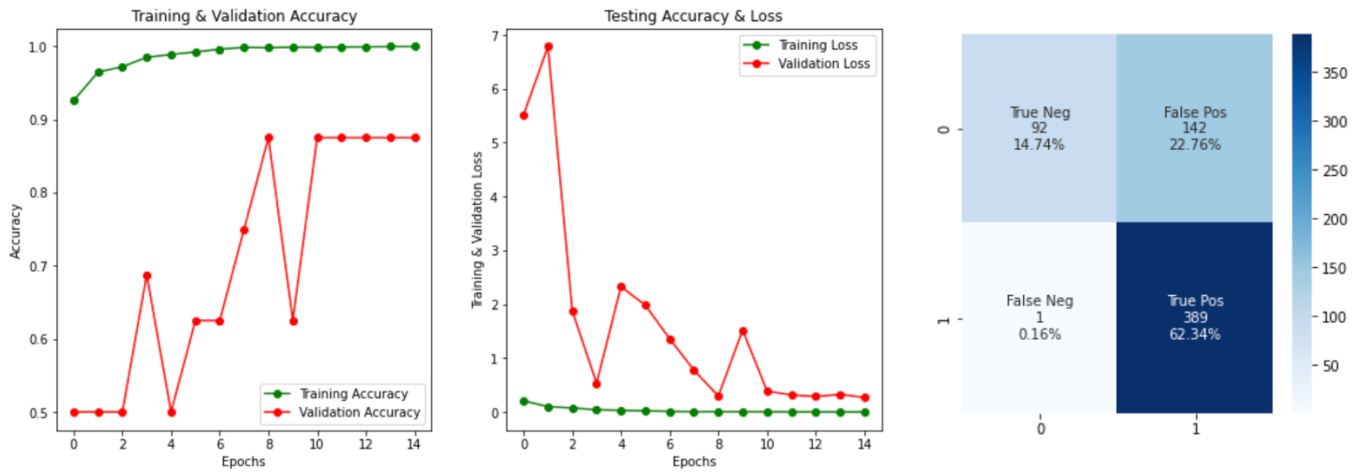
Figure 9. Charts of accuracy, loss and confusion matrix for the customized CNN model trained with three different data groups

From figure 9, we see that all the data groups were trained well. The accuracy increases, and the loss decreases with epochs. From the confusion matrices which were obtained using test set, we see that the model trained with balanced and augmented dataset performed well in distinguishing both normal and pneumonia samples (correctly classified 380 out of 390 pneumonia and 198 out of 234 normal samples). The model trained with raw data only missed one pneumonia sample out of 390 pneumonia samples. However, it performed very poorly on classifying normal samples. It correctly classified only 92 out of 234 normal samples. The CNN model trained with augmented dataset on the other hand, correctly classified 5 more pneumonia cases (385 out of 390 pneumonia) than balanced and augmented data group. However, it classified 15 less normal cases (183 out of 234 normal). I assumed that reducing error is the top priority in this study, so I conclude that CNN architecture trained with balanced class sample size and augmented images resulted in the least number of errors and performed the best among the other data group in classifying pneumonia and normal chest x-ray images.

ROC Curve analysis of CNN model
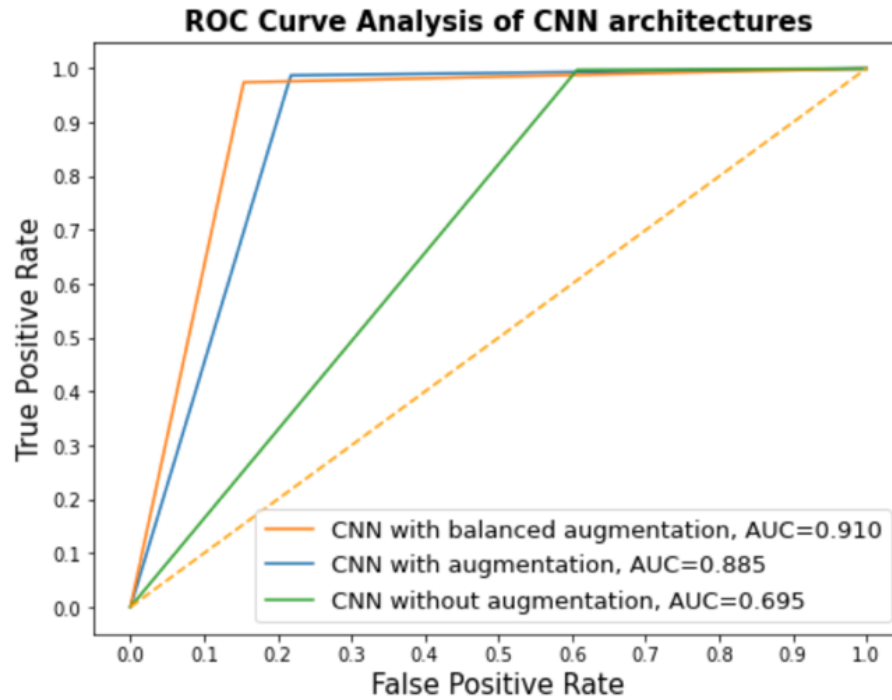
Figure 10. ROC curve and AUC score of the customized CNN architectures for the test set

Evaluation metric values for the customized CNN model using test set

Model trained with	Accuracy	Loss	Recall Normal| Pneumonia		Precision Normal| Pneumonia		F1-score Normal| Pneumonia		AUC score
augmented	0.9102	0.2555	0.78	0.99	0.97	0.88	0.87	0.93	0.885
Balanced & augmented	0.9263	0.24	0.85	0.97	0.95	0.91	0.90	0.94	0.91
Non-augmented	0.7708	1.223	0.39	1.00	0.99	0.73	0.56	0.84	0.695

Table 2. Evaluation metric values for the customized CNN model using the test set

II.   Transfer learning models

Confusion matrices for the transfer learning models trained with three data groups and their performance were evaluated using the test set

1. ResNet50

	Augmented data group		Balanced and augmented		Non-augmented data group

| | True Neg 205 32.85% | False Pos 29 4.65% | True Neg 216 34.62% | False Pos 18 2.88% | True Neg 119 19.07% | False Pos 115 18.43% |
| | False Neg 6 0.96% | True Pos 384 61.54% | False Neg 16 2.56% | True Pos 374 59.94% | False Neg 1 0.16% | True Pos 389 62.34% |

Augmented data group   Balanced and augmented   Non-augmented data group

2. DenseNet121

| | True Neg 203 32.53% | False Pos 31 4.97% | True Neg 205 32.85% | False Pos 29 4.65% | True Neg 143 22.92% | False Pos 91 14.58% |
| | False Neg 4 0.64% | True Pos 386 61.86% | False Neg 4 0.64% | True Pos 386 61.86% | False Neg 1 0.16% | True Pos 389 62.34% |

Augmented data group   Balanced and augmented   Non-augmented data group

3. VGG16

Augmented data group	Balanced and augmented	Non-augmented data group

4. Inception-V3

Augmented data group	Balanced and augmented	Non-augmented data group

5. Xception

Augmented data group	Balanced and augmented	Non-augmented data group

Figure 11. Confusion matrices for the transfer learning models trained with three data groups

Similar to the confusion matrix results of the customized CNN model, 'balance and augmented' data group trained models showed the least number of errors compared to the other two data groups. Among the transfer learning models, Xception model trained with the 'balanced and augmented' data group, showed the least number of false positive and false negatives when evaluated with the test set (correctly classified 383 out of 390 pneumonia cases and 211 out of 234 normal cases).

Evaluation metric values for the transfer learning models using test set

Transfer learning models	Data group	Acc.	Loss	Recall Normal| Pneumonia		Precision Normal| Pneumonia		F1-score Normal| Pneumonia		AUC score
ResNet50	augmented	0.9439	0.2152	0.88	0.98	0.97	0.93	0.92	0.96	
	Balanced & augmented	0.9455	0.1720	0.92	0.96	0.93	0.95	0.93	0.96	0.946
	Non-augmented	0.8141	1.4269	0.51	1.00	0.99	0.77	0.67	0.87	
DenseNet 121	augmented	0.9439	0.2586	0.87	0.99	0.98	0.93	0.92	0.96	
	Balanced & augmented	0.9471	0.3573	0.88	0.99	0.98	0.93	0.93	0.96	0.937
	Non-augmented	0.8526	1.1561	0.61	1.00	0.99	0.81	0.76	0.89	
VGG16	augmented	0.8686	0.4297	0.69	0.98	0.95	0.84	0.80	0.90	
	Balanced & augmented	0.9038	0.4140	0.76	0.99	0.97	0.88	0.86	0.93	0.807
	Non-augmented	0.7676	2.7321	0.39	0.99	0.98	0.73	0.56	0.84	
Inception-V3	augmented	0.9135	0.2419	0.78	0.99	0.99	0.88	0.87	0.93	
	Balanced & augmented	0.9439	0.2329	0.86	0.99	0.99	0.92	0.92	0.96	0.923
	Non-augmented	0.7949	0.8042	0.47	0.99	0.96	0.76	0.63	0.86	
Xception	augmented	0.9375	0.2294	0.85	0.99	0.98	0.92	0.91	0.95	
	Balanced & augmented	0.9519	0.2899	0.90	0.98	0.97	0.94	0.93	0.96	0.947
	Non-augmented	0.8093	1.5381	0.50	1.00	0.99	0.77	0.66	0.87	

Table 3.Evaluation metric values for the transfer learning models using the test set

ROC Curve analysis of CNN model
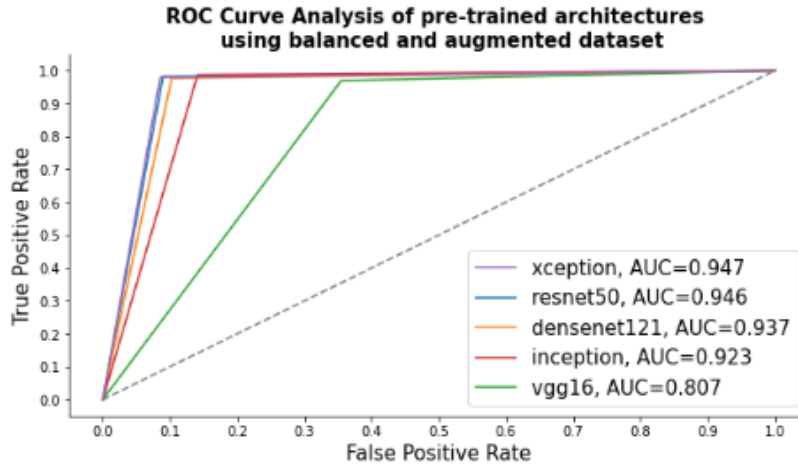
Figure 12. ROC curve and AUC score of the pre-trained architectures for the test set

Proposed Xception architecture

The Xception model trained with 'balanced and augmented' data group resulted in the least number of errors when testing is done using the test set. The experiment performance of proposed Xception model was assessed based on accuracy, recall, precision, F1-score and AUC, and the model showed values of 95.2%, 94%, 95.5%, 94.5%, and 94.7% respectively. It is evident that the proposed Xception model accomplished the highest results compared with the other architectures.

III.   Interpretation of visualization of intermediate layers

The feature maps of pre-trained models (ResNet, DenseNet, VGG16, Inception and Xception) are visualized in Table 4. The first feature map of the first nine layers and last three layers of each model are illustrated. We see that the initial layers of pre-trained models identify low-level features such as edges in the image and as it progresses deeper into the model, the model is learning more abstract features like specific regions (lung) of the image. The earlier features get combined and are passed onto the next subsequent layer inside network and the last layers of the network become complicated and we generally lose the ability to interpret the deeper feature maps as shown in the feature maps of the last layers.
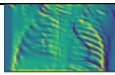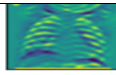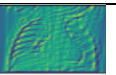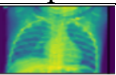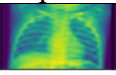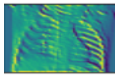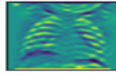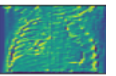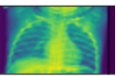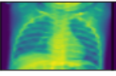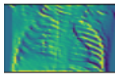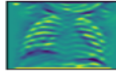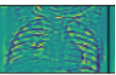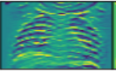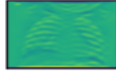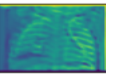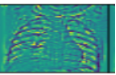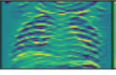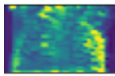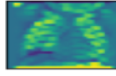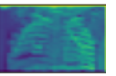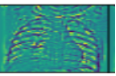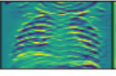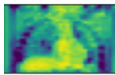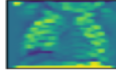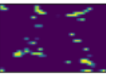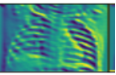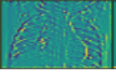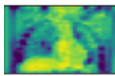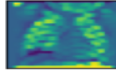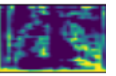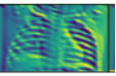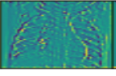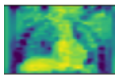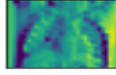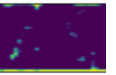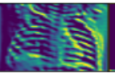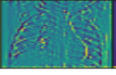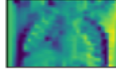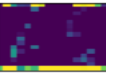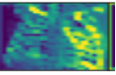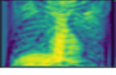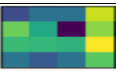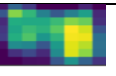
| ResNet	DenseNet	VGG16	Inception	Xception

Table 4. The first feature map of the first nine layers and last three layers of pre-trained models

IV.   Interpretation of visualization of class activation maps

I used the best performing Xception model to detect pneumonia in the samples and the weights of the last convolutional layer inside the 14th convolutional block are used to produce CAM. From figure 13, we see that there are good localization results, and the last convolutional layer sees that the lung regions are relevant when classifying the image.
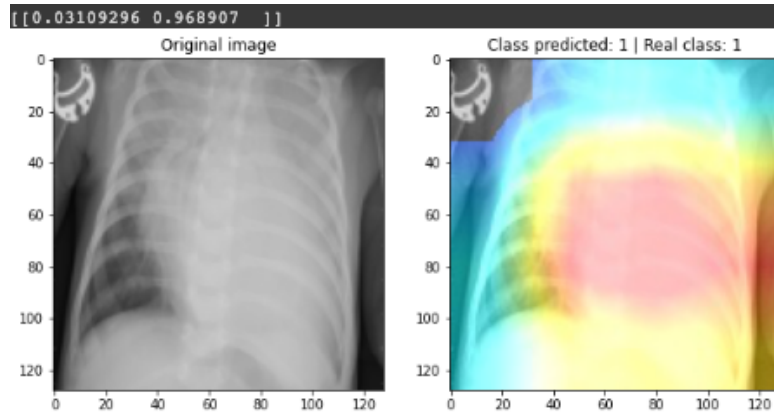
a) Original pneumonia sample image (left) and CAM applied sample image (right).
The model predicted that the image is pneumonia with 96.9% accuracy.
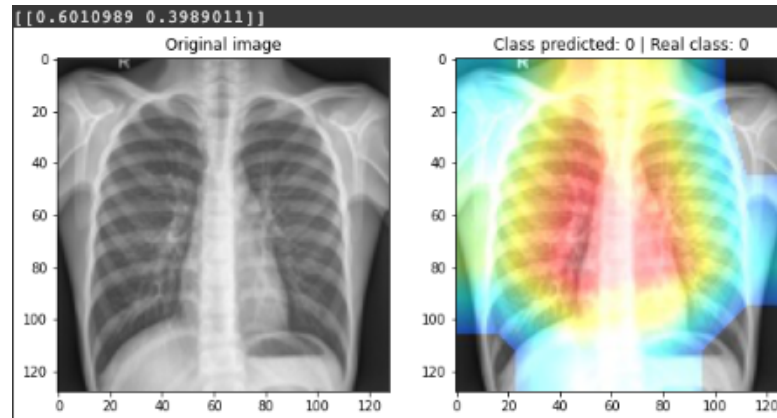
b) Original normal sample image (left) and CAM applied sample image (right).
The model predicted that the image is normal with 60.1% accuracy.

Figure 13. Heat maps for classified test examples using CAM

6. CONCLUSION AND FUTURE WORKS

This study presents a deep convolutional neural network-based models to detect pneumonia in pediatric chest x-ray images to expedite accurate diagnosis of the pathology. A customized CNN model and five different popular CNN-based deep learning algorithms for image classification were trained and tested. The Xception model demonstrated promising performance than the customized CNN model and the other four different deep CNN based transfer learning models under study. The model learned generic image features from ImageNet that served as a good initialization compared to learning from a random initialized weights and the model was trained to focus on learning current dataset specific features. The classification accuracy, recall, precision, F1-score, and AUC score were 95.2%, 94%, 95.5%, 94.5%, and 94.7% respectively. Using transfer learning approaches resulted in faster convergence with reduced bias, overfitting, and improved generalization, compared to customizing CNN model from scratch. A shallow customized CNN model worked pretty well in pneumonia detection, but it was challenging to deploy suitable and optimal CNN model with different configuration.

In this study, the impact of data augmentation and oversampling techniques are also evaluated. The performance difference between models trained with augmented data and non-augmented data was noticeable as they showed over 10% difference in accuracy. It shows that data augmentation is a necessary part of successful application of deep learning models on image data. In addition, over sampling infrequent class techniques were used to combat large class imbalance in the dataset. Class representation of the training set was adjusted by creating synthetic data of minority class. The downside of oversampling is the overfitting of the oversampled class, and it was handled with data augmentation which is known to prevent overfitting. As a result, models trained with 'balanced and augmented' data group showed the highest performance in detecting pneumonia among the three different data groups.

For the future work, we can extend the model usability to further classify the pneumonia class into bacterial and viral pneumonia. It is also possible to apply the proposed algorithms to several other diseases such as cancers, blockages and fractures which can be diagnosed by x-ray testing.

I hope that my findings are useful for developing clinically useful solutions like diagnosing pneumonia types in chest radiographs. I believe that the deep convolutional neural networks can be a good alternative to replace the traditional method of expert physicians interpreting medical images in many countries, particularly the low-resource regions.

7. APPENDIX

GitHub Link: https://github.com/katehee/pneumonia-detection-in-X-ray-images

8. REFERENCES

[1] Rudan I, Tomaskovic L, Boschi-Pinto C, Campbell H; WHO Child Health Epidemiology Reference Group. Global estimate of the incidence of clinical pneumonia among children under five years of age. Bull World Health Organ. 2004;82(12):895-903.

[2] Ayan E., Ünver H.M. Diagnosis of pneumonia from chest x-ray images using deep learning; Proceedings of the Scientific Meeting on Electrical-Electronics Biomedical Engineering and Computer Science (EBBT); Istanbul, Turkey. 24–26 April 2019; pp. 1–5.

[3] Chest X-Ray Images (Pneumonia). Kaggle. Accessed May 1, 2022. https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia

[4] Shorten C., Khoshgoftaar T.M. A survey on image data augmentation for deep learning.

[5] Yang, S, Xiao W, Zhang M, et al. Image Data Augmentation for Deep Learning: A Survey. arXiv: 2204.08610, 2022.

[6] D. Sarwinda, R. H. Paradisa, A. Bustamam, and P. Anggia, "Deep learning in image classification using residual network (ResNet) variants for detection of colorectal cancer," Procedia Computer Science 179, 423–431 (2021).

[7] T. Chauhan, H. Palivela, S. Tiwari. Optimization and fine-tuning of densenet model for classification of COVID-19 cases in medical imaging. International Journal of Information Management Data Insights (2021), Article 100020

[8] M. Rahimzadeh, A. Attar. A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray images based on the concatenation of Xception and ResNet50V2. Informatics in Medicine Unlocked, 19 (2020)

[9] Razzak MI, Naz S, Zaib A. Deep learning for medical image processing: overview, challenges and the future. 2017.

[10] A. Cruz-Roa, H. Gilmore, A. Basavanhally, M. Feldman, S. Ganesan, N.N.C. Shih, J. Tomaszewski, F.A. González, A. Madabhushi, Accurate and reproducible invasive breast cancer detection in whole-slide images: a deep learning approach for quantifying tumor extent, Sci. Rep. (2017) 7, doi:10.1038/srep46450.

[11] Xi IL, Zhao Y, Wang R, Chang M, Purkayastha S, Chang K, et al. Deep Learning to Distinguish Benign from Malignant Renal Lesions Based on Routine MR Imaging.

[12] Antin B., Kravitz J., Martayan E. Detecting Pneumonia in Chest X-rays with Supervised Learning. Semanticscholar Org.; Allen Institute for Artificial intelligence, Seattle, WA, USA: 2017. [Google Scholar]

[13] Calderon-Ramirez S. 2020. Correcting Data Imbalance for Semi-supervised Covid-19 Detection Using X-Ray Chest Images.

[14] Taylor L., Nitschke G. 2017. Improving Deep Learning Using Generic Data Augmentation.

[15] Monshi MMA, Poon J, Chung V, Monshi FM (2021) Covidxraynet: Optimizing data augmentation and cnn hyperparameters for improved covid-19 detection from cxr. Computers in Biology and Medicine 133:104375

[16] Nishio, M., Noguchi, S., Matsuo, H. et al. Automatic classification between COVID-19 pneumonia, non-COVID-19 pneumonia, and the healthy on chest X-ray image: combination of data augmentation methods. Sci Rep 10, 17532 (2020). https://doi.org/10.1038/s41598-020-74539-2

[17] H. Sharma, J. S. Jain, P. Bansal and S. Gupta, "Feature Extraction and Classification of Chest X-Ray Images Using CNN to Detect Pneumonia", 2020 10th International Conference on Cloud Computing Data Science & Engineering (Confluence), 2020.

[18] Pneumonia detection in chest X-ray images using compound scaled deep learning model https://www.tandfonline.com/doi/full/10.1080/00051144.2021.1973297

[19] Elshennawy, N. M., & Ibrahim, D. M. (2020). Deep-Pneumonia Framework Using Deep Learning Models Based on Chest X-Ray Images. Diagnostics (Basel, Switzerland), 10(9), 649. https://doi.org/10.3390/diagnostics10090649

[20] Rajaraman S, Candemir S, Kim I, et al. Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs. Appl Sci. 2018;8(10):1715.

[21] Luján-García J.E., Yáñez-Márquez C., Villuendas-Rey Y., Camacho-Nieto O. A Transfer Learning Method for Pneumonia Classification and Visualization. Appl. Sci. 2020;10:2908. doi: 10.3390/app10082908.

[22] Batista, G.E.A.P.A.; Prati, R.C.; Monard, M.C. A study of the behavior of several methods for balancing machine learning training data.

[23] Rahman, Tawsifur, et al. "Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray." Applied Sciences 10.9 (2020): 3233.

[24] D.-C. Li, C.-W. Liu, S.C. Hu. A learning method for the class imbalance problem with medical data sets. Comput. Biol. Med., 40 (2010), pp. 509-518

[25] Dai, Q., Chopp, H., Pouyet, E., Cossairt, O., Walton, M., & Katsaggelos, A. K. (2020). Adaptive Image Sampling Using Deep Learning and Its Application on X-Ray Fluorescence Image Reconstruction. IEEE Transactions on Multimedia, 22(10), 2564-2578. [8930037]. https://doi.org/10.1109/TMM.2019.2958760

[26] Dablain D, Krawczyk B, Chawla NV. DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data. IEEE Trans Neural Netw Learn Syst. 2022 Jan 27;PP. doi: 10.1109/TNNLS.2021.3136503. Epub ahead of print. PMID: 35085094.

[27] Ali, A., Shamsuddin, S. M., and Ralescu, A. L. (2013). Classification with class imbalance problem. A review. Int. J. Advance Soft. Compu. Appl. 7, 176–204.

[28] Standford vision lab. ImageNet. Updated Mar 11, 2021. Accessed May 1, 2022. https://www.image-net.org/

[29] V. Buhrmester, D. Munch and M. Arens. Analysis of explainers of black box deep neural networksfor computer vision: A survey. MAKE, 3(4):966–989, 2021

[30] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva and A. Torralba, "Learning Deep Features for Discriminative Localization," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2921-2929, doi: 10.1109/CVPR.2016.319.