# DISEASE DETECTION USING CHEST X-RAYS

**Manpreet Dhindsa**
mkd8vv@virginia.edu

**Gretchen Larrick**
jem3yb@virginia.edu

**Sarah Rodgers**
pjk2wq@virginia.edu

August 9, 2022

## 1 Abstract

A chest radiograph, often referred to as a chest X-ray, produces black and white images of the inside of the chest. The images capture various components such as the heart, lungs, blood vessels, airways, and bones. These images are used in detecting various lung conditions such as cancer, infections, or other chronic conditions. These images can be taken over time and allow for doctors to analyze the progression of specific diseases. Additionally, as shown through several articles in the literature review, chest x-rays recently have been used in detecting Covid-19 especially with the presence of pneumonia.

## 2 Motivation

In order to interpret and report the information displayed in the chest X-ray to the patient, a highly trained radiologist is required. As being the most commonly used film in medicine, there is a great opportunity to leverage this information to further research the impact of machine learning methods on chest X-rays. By leveraging machine learning algorithms, doctors can utilize the information reported to support diagnosing lung conditions. Through recent research, deep learning algorithms specifically have showed promising results in diagnosing lung conditions by incorporating transfer learning methods. We aim to explore how chest x-rays images can be used in detecting abnormal conditions and what those specific conditions are.

As discussed in the literature review below, a common theme seen across the published research papers is a lack of data available to train these machine learning algorithms, especially for more complicated conditions, and many of the algorithms we researched used the same existing data. Since the time and expertise of a highly trained radiologist is required to analyze the chest x-ray and classify the condition, the sparsity of the data is warranted. We compiled an extensive list of available chest x-ray images listed in the dataset section below that include a wide range of conditions and span many years. We aim to leverage data augmentation strategies to transform the existing images in our database in order to increase the number of images available. However, when doing so, we will pay careful consideration to not construe the meaning of the image to result in a different classification than the original image. By using these augmentation techniques, we aim to improve the generalization of the algorithm.

### 2.1 Literature Review

Image recognition and analyses have become very popular in healthcare research. [Castiglioni et al., 2021] describes the growth of machine learning and deep learning in biomedical research as well as the benefits and considerations for each approach. There is a focus on deep learning models as the multi-layered neural network framework has proven to be successful in image processing. There are existing image datasets such as ImageNet that have been continuously leveraged in deploying neural networks for medical image classification. However, when specifically focusing on the research performed using chest X-ray images to automatically detect illnesses, the existing research is immense. There are a few categories of analyses that were seen that include algorithms to detect specific illnesses, to detect general illnesses, and to detect Covid-19. Analyses span deep neural networks to convolutional neural networks and more traditional machine learning techniques such as support vector machines.

In [Seah et al., 2021] the authors obtained over 800,000 chest x-ray images from the United States, Europe, and Australian where the patients were adults 16 years of age and older. The deep learning models used to assess these images were three convolutional neural networks (CNNs) using two architecture types, EfficentNet for the attributes and classification model, and a U-Net architecture with an EfficientNet backbone for the segmentation model. In addition to deep learning, the x-rays were assessed by human radiologists to understand their effectiveness at identifying clinical findings. The scope of these clinical findings is large, and there are 127 clinical findings that are possible for identification in the dataset. Findings for the model show that radiologists had an average AUC of 0.713 (95 % CI 0.645-0.785) where the deep learning model had an AUC of 0.957 (95 % CI 0.954-0.959) over all the findings. This indicates X-ray detection via deep learning can provide high accuracy over multiple illness (clinical findings).

Additional deep learning models have been developed with a focus on a specific condition. The [Ashhar et al., 2021] study focuses on lung cancer and detecting it earlier than current methods to prevent the number of fatalities. Multiple CNN architectures were evaluated that included GoogleNet, SqueezeNet, DenseNet, SuffleNet and MobileNetV2 to classify lung tumors into malignant and benign categories. The GoogleNet produced the best results for tumor classification with an accuracy of 94.5% and specificity of 99.1%. Another study [Jaiswal et al., 2019] focuses on easing the process of pneumonia identification. Pneumonia has been a focus for many deep learning algorithms as radiologists find it beneficial to identify if a patient has pneumonia or not. This study uses an approach that incorporates global and local features for pixel-wise segmentation called Mask-RCNN. The model ingests the chest X-ray image as an input and predicts the bounding box of the image. The evaluation metric for pneumonia identification was the intersection over union (IoU) and suggested that this model is effective in identifying pneumonia in chest X-rays.

Google Research recognized that many existing algorithms that leverage deep learning methods on chest X-ray images were for specific clinical conditions such as cancer, pneumonia, tuberculosis, etc. These algorithms are helpful in diagnosing if a patient has a specific condition, but then if the patient does not have that illness, doctors are left with running multiple algorithms to determine if a patient has a specific disease. Using this process can leave patients and doctors with inconclusive results as algorithms are not built for every possible abnormality and not every abnormality is medically known. As seen in the [Nabulsi et al., 2021] study, a classifier was built to determine if a chest X-ray contains any abnormality. The goal of this algorithm is to serve as a general purpose algorithm to be used in the clinical workflow to first identify if a patient has an abnormal condition then perform additional tests to determine what that abnormality specifically is. The study used a CNN and EfficientNet-B7 as a feature extractor which was pre-trained on ImageNet. The CNN used a cross-entropy loss and momentum optimizer with a learning rate of 0.00004 and momentum value of 0.9. Regularization methods were also used that included dropout with a keep probability rate of 0.5. The algorithm was evaluated on two test datasets that result in 0.87 - 0.94 accuracy on detecting abnormalities.

This article from The Lancet Digital Health highlights that radiologists can improve their performance by reviewing results from a deep-learning model called EfficientNet [Takahashi and Usuzaki, 2022]. However, in this deep learning model, it is not able to take into account results from a chest CT like a radiologist can, which can put the model performance at a disadvantage [Takahashi and Usuzaki, 2022]. Therefore, this article emphasizes a transfer-learning strategy that transfers data from CT scans to x-ray images in order to better train models with these two different methods combined [Takahashi and Usuzaki, 2022]. Even though the current model is very helpful for helping radiologists already, this article explains that the introduction of transfer-learning could help from a resources perspective as well in case some parts of the world do not have access to as many imaging methods for their radiologists [Takahashi and Usuzaki, 2022].

As the Covid-19 pandemic began, researchers and data scientists initiated the use of artificial intelligence and machine learning methods to diagnosis patients with Covid-19. There have been many research papers published related to Covid-19 that utilize deep learning techniques and we will highlight a few that were most related to our project goals. As seen in [Erdaw and Tachbele, 2021], chest X-ray images were used to automatically classify Covid-19 pneumonia compared to other pneumonia cases along with normal images. From our research, there were many similar studies performed that classified the different types of pneumonia with a focus on Covid-19. This study leveraged 1,100 images and used a support vector machine (SVM) classifier algorithm. The SVM algorithm was distinctly chosen over deep learning methods because of the known well performance on small sets of data. This study used a pyramid histogram of oriented gradients (PHOG) method to extract 630 features from the images. The SVM algorithm resulted in 99.3% accuracy for binary classification and 97.3% accuracy in multi-level classification for automatically detecting Covid-19 pneumonia.

We now shift the focus to studies that specifically leverage deep learning techniques related to Covid-19 chest X-ray images. A study of Covid-19 and X-ray image analysis using deep learning is found in [Jain et al., 2021]. This study had a narrow scope of detection, analyzing for Covid-19 positive patients through the use of X-ray images. The convolutional neural networks used in this analysis were Inception V3, Xception, and ResNeXt. Data was sourced through publicly available data, via Kaggle with a total number of images in the dataset at 6432 images. The key measurement in this study was accuracy, with Xception giving the highest accuracy at 97.97%, and the other models all

resulting with accuracies in the high 90%. A concern within this study deals with overfitting, with ResNeXt having a tendency to over-fit models. This is to be address in future studies on newer publicly available datasets.

Daniel Moses in his literature review [Moses, 2021] provides and extremely comprehensive understating of deep learning and X-ray imaging. The datasets used in this study cover multiple areas of diagnosis including lung nodules, pneumonia detection, Covid-19 pneumonia, tuberculosis, and a generalized abnormality detection. Each of these had various model performance, but for the purpose of this study we will focus in on the Covid-19 pneumonia detection. The datasets used were publicly available on 11 different models. The best performing model was found to be DenseNet-121 with an accuracy of 98.69 %. To achieve this high value, Moses augmented the dataset by rotating the images 120 - 140 degrees.

In this next study, the authors created a classification model to determine if a chest x-ray is normal, bacterial, or viral with the assumption that if the results are viral, that will indicate a strong probability that that individual has Covid-19 [Hammoudi et al., 2021]. Their data is from a Chest X-Ray Images (Pneumonia) dataset on Kaggle, which only consists of chest x-rays from children, however, in their results it seems that this model can also work for adult x-rays [Hammoudi et al., 2021].Currently, the model is able to discern between the three classes, however, there is not as strong of a performance when it comes to distinguishing between the Non-Covid-19 viral class and the Covid-19 viral class [Hammoudi et al., 2021]. The best performing model was found to be DenseNet169 with an average accuracy of 95.72% for detecting all three classes but individually had accuracies of 97.97%, 96.62%, and 92.57% for detecting the bacteria, virus, and normal classes, respectively [Hammoudi et al., 2021]. They mention in their conclusion that they may have future work to better discern more accurately between the Non-Covid-19 viral class and the Covid-19 viral class [Hammoudi et al., 2021].

This study from the International Journal of Environmental Research and Public Health classified x-ray images into two classes: Covid-19 and not Covid-19 as well as into three classes: Covid-19, Normal, Pneumonia with three different architectures: Inception-V3, ResNet-50, and VGG-19, and the two-class accuracies for both architectures were 100% while the three-class accuracies were 97%, 98.55%, and 98.55% for Inception-V3, ResNet-50, and VGG-19, respectively [Awan et al., 2021]. This study used Apache Spark to help with big data as well as transfer learning as shown in previous studies mentioned, as well. Along with the three models mentioned, logistic regression was also used to help train the models. With this combination, this experiment was able to perform very strongly and show confident results. However, the authors list some limitations with their models with one of them being that the model performs well on similar images, but if the images are tilted or rotated, the model does not perform as strongly [Awan et al., 2021].

This next study works to classify chest x-ray images into a similar set of three classes as compared to the previous article: Covid-19 pneumonia, no-Covid-19 pneumonia, and non-pneumonia and resulted in a 94.5% overall accuracy [Heidari et al., 2020]. This improves upon a previously mentioned study that couldn't as accurately classify Covid-19 pneumonia and no-Covid-19 pneumonia because this experiment uses two image pre-processing steps to remove most of the diaphragm region, remove image noise and improve image contrast as well as creates a pseudo color image to feed into an existing deep learning model that is already a strong performer for color images in a transfer learning strategy [Heidari et al., 2020]. This particular VGG16 model for the color images uses Softmax as the activation with an Adam optimizer [Heidari et al., 2020]. Additionally, data augmentation was performed in this study in order to increase the training size by using shearing, zooming in on some images, randomly rotating images, shifting images left, right, up and down, and finally, flipped horizontally [Heidari et al., 2020]. This study confirmed that with the data augmentation, the image pre-processing, and the transfer learning method, there can be a strong performance to distinguish Covid-19 pneumonia from no-Covid-19 pneumonia.

## 3 Methods

Machine Learning algorithms have the ability to learn patterns seen in images to support diagnosing a patient. Since a highly trained radiologist is needed to interpret a chest X-ray to determine the underlying meaning of the image, these algorithms can support providing insights that a human may have missed or supplement their findings. The methods in this paper aim to continue the research of developing deep learning algorithms to detect diseases with a primary focus on Convolutional Neural Networks (CNNs).

The modeling process focused on using transfer learning techniques. Transfer learning is a machine learning method where a model developed for a task is reused as the starting point for another model. The existing frameworks available through the Keras machine learning library were explored including ResNet, VGG, Xception, and DenseNet. These methods are used in the study of a binary class model that classifies chest X-rays as normal and not normal and a multi-class model that focuses only on non-normal images and classifies chest X-rays as Covid-19, Pneumonia, and Other. A sample of images in both sets of classifications are displayed below.

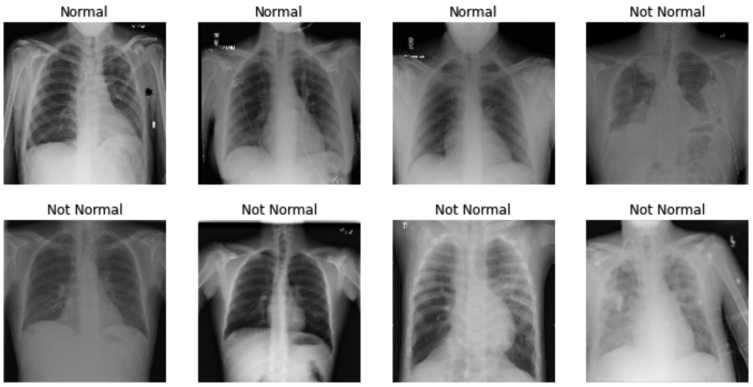


Figure 1: Binary data of normal and not normal images

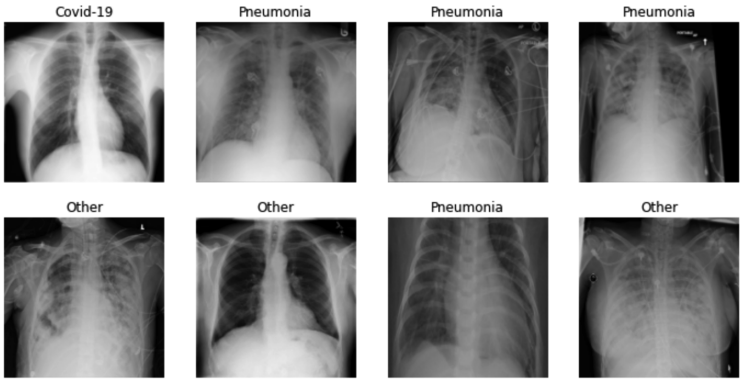


Figure 2: Multi-class data of covid-19, pneumonia, and other images

### 3.1 Data

To create our deep learning networks, we are utilizing datasets across various sources. It has been seen that many studies use the same images provided by the National Institute of Health as it is the largest known dataset of chest X-ray images. Below is a listing of the datasets:

- Mendeley Data: Large Dataset of Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images
- Kaggle: COVID-19 chest xray
- Kaggle: COVID-QU-Ex Dataset
- NIH: ChestX-ray8 Database

The distributions of the disease types from the combined datasets were similar for the binary classification (normal and not normal) as seen in the figure below.

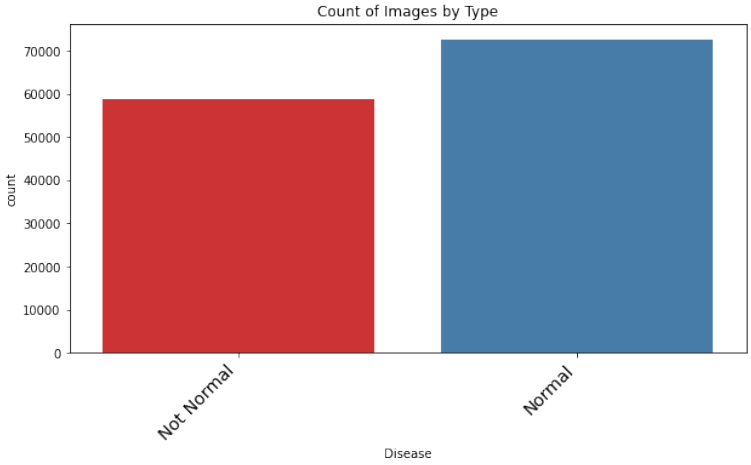


Figure 3: Distribution of disease type

When breaking down the specific diseases of the not normal images, Pneumonia and Covid-19 were the primary diseases found in these images. To create a more robust model that classifies the specific diseases, the normal images were not included in the data pool for the multi-class model. Images that were not for Pneumonia or Covid-19 were grouped together to reduce the number of classes.

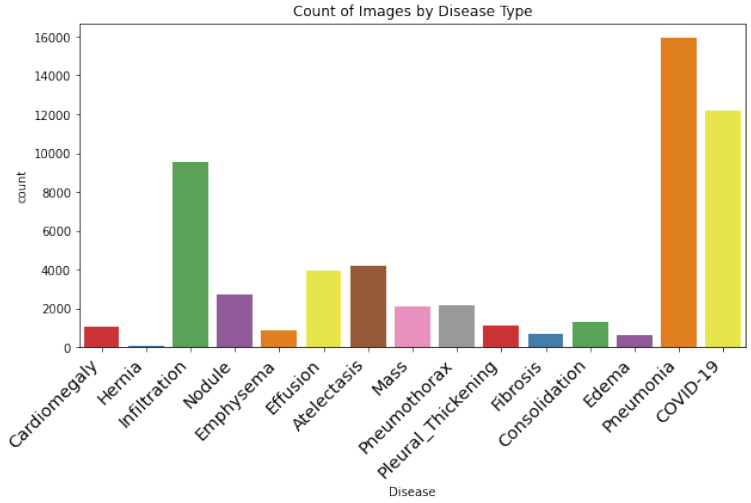


Figure 4: Distribution of not normal images

Data Augmentation uses various techniques to increase the volume and diversity of the data during the training process [Mikołajczyk and Grochowski, 2018]. The chest X-ray images are transformed by adding slight modifications to create new images and then are used to train the model. This creates a large pool of images and helps minimize over fitting when training the model. The images were transformed by being horizontally flipped and also rotated as seen in the Figure below. The augmentation techniques were applied to the training datasets for both the binary and multi-class models.

These transformations aim to simulate other chest X-ray images that would be seen from different patients, noise from various imaging technicians, and other factors from the environment such as movement in the patient. The augmented images were used in both the binary and multi-class approaches for select transfer learning models. This strategy was implemented only on a portion of the models so that the effectiveness of the augmentation can be assessed.

Data augmentation strategies were also utilized to strengthen the generalization of the models. To further improve the generalization, other regularization strategies were tested. Regularization strategies aim to reduce the generalization error while not changing the training error [Moradi et al., 2020]. In addition to data augmentation, early stopping and

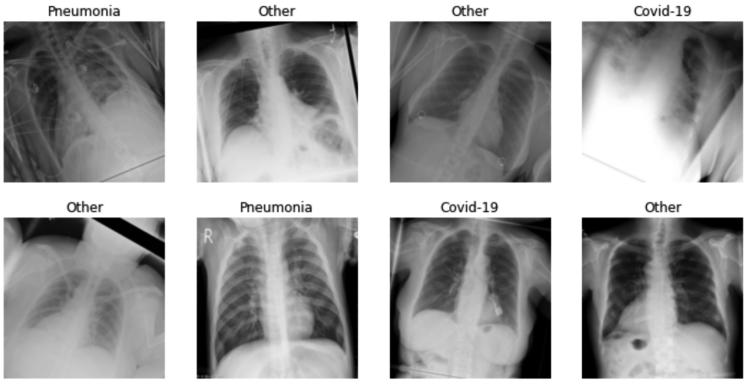


Figure 5: Augmented multi-class data

dropout techniques were used. While training each of the transfer learning models, the hyperparameters were tuned to increase the accuracy and minimize the loss. This included altering the learning rate, optimizers, number of epochs, activation functions, and more in order to produce the highest performing models.

## 4 Experiments

Data from the sources listed above were downloaded and partitioned into specific class structures. The overall size of these files when merged became cumbersome for the machine learning programming space. Google Colab had memory allocation errors which impacted speed of analysis. To mitigate this, the data was uploaded to Kaggle and accessed directly from a Kaggle Notebook to perform machine learning. Additionally, data was reduced to 80% of the total file volume for the binary classification study and the multi-class study at a 70% reduction after all the non-disease files. Each of the classifications were then divided again into a 80/20 split of training data/validation data through the use of the built in preprocessing class in Keras.

### 4.1 Binary

The first step in binary analysis was to identify via the literature review which models returned a high accuracy result for the datasets when run individually. Three architectures were identified for the first analysis, Inception-V3, DenseNet-201, and VGG-16. The table below shows the accuracy results for this data. The optimizers used in each architecture was RMSprop with a learning rate of 0.0001 and a momentum of 0.9. The additional architectures were analyzed at various optimization, including changing drop out, adjusting layers, and changing learning rates. The specific models and their hyperparameters can be found in the appendix. In addition to accuracy, other metrics were captured including Precision, AUC, and Recall. These values allowed for a more granular look at how the model was performing. All of the models were run using the same data set at a set seed to ensure the values would remain constant if a rerun was necessary.

Table 1: Binary Analysis Results

| Number | Architecture | Accuracy(%) |
|---|---|---|
| 1 | InceptionV3 | 55 |
| 2 | DenseNet201 | 65 |
| 3 | VGG16 | 77 |

### 4.2 Multi-Class

We also wanted to create a more specific classification model that classified diseases into 3 groups: COVID-19, Pneumonia, and Other.

For initial baseline models, we experimented with ResNet-50, and this model included a layer with 512 hidden units and a ReLU activation. In the final layer, softmax activation was used, and RMS prop optimizer was used when compiling the model. ResNet-50 was attempted first because we saw in the literature review that others had seen success with transfer learning architectures especially with Inception-V3, ResNet-50, and VGG-19. This initial model finished with an accuracy of .8936 and a loss of .4677.

We also experimented with VGG-16, which resulted in a 0.8002 accuracy. This model included a layer with 512 hidden units and a ReLU activation. In the final layer, softmax activation was used, and RMS prop optimizer was used when compiling the model.

We also experimented with Inception-V3, which resulted in a 0.7869 accuracy and 9.1972 loss against the validation set. This involved a layer with 1024 hidden units and a ReLU activation. Similarly, in the final layer, softmax activation was used, and RMSProp optimizer was used when compiling the model.

| Number | Architecture | Accuracy(%) |
|---|---|---|
| 1 | ResNet50 | 89 |
| 2 | VGG16 | 80 |
| 3 | InceptionV3 | 79 |

Moving forward, we experimented with various models such as VGG-16, VGG-19, InceptionResNet-V2, Inception V3, ResNet-50, ResNet-152V2, DenseNet-121, DenseNet-201, and Xception using data augmentation and as part of that, some augmentation included images that were flipped while some included images that were flipped and rotated in order to compare all results. We compared different optimizers such as RMSprop, SGD, and Adam with different learning rates, and we experimented with dropout, momentum, and decay rates along with the number of epochs and steps per epoch.

## 5 Results

The table below contains the best performing model for each of the model types analyzed. DenseNet-121 classified the chest X-rays of COVID-19, Pneumonia, and Other with 95% accuracy, and InceptionResNet-V2 classified chest X-rays as normal and abnormal with 71%. The best-performing DenseNet-121 model did not include any data augmentation, used RMSprop as the optimizer with a learning rate of 0.0001, had a dropout of 0.7, and used 5 epochs with 50 steps per epoch. The best-performing InceptionResNet-V2 model also did not include data augmentation, used RMSprop as the optimizer with a learning rate of 0.0001, and had a dropout of 0.5.

| Type | Multiclass | Binary |
|---|---|---|
| DenseNet-121 | 95% | 67% |
| DenseNet-201 | 81% | 58% |
| Inception-V3 | 89% | 63% |
| InceptionResNet-V2 | 89% | 71% |
| ResNet-50 | 89% | 55% |
| ResNet-152V2 | 70% | 53% |
| VGG-16 | 87% | 68% |
| VGG-19 | 89% | 57% |
| Xception | 65% | 56% |

Figure 6: Results from Multi-Class and Binary Classification

The architecture for the best performing models are listed below. For binary we see InceptionResNet-V2's architecture.

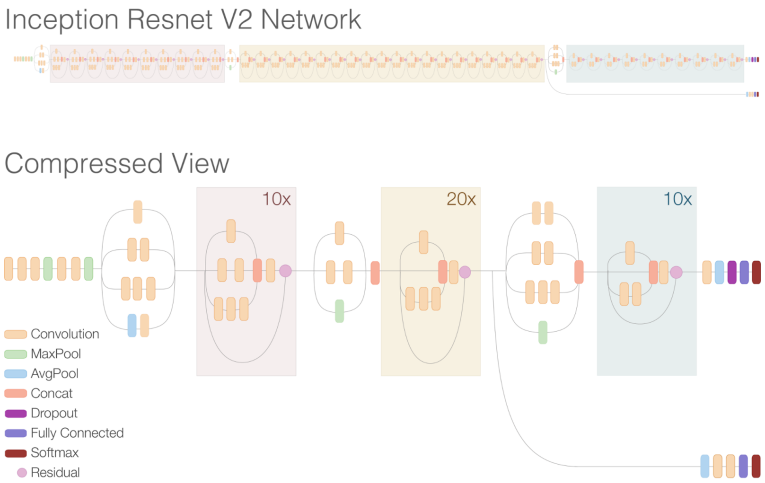


Figure 7: Architecture for InceptionResNet-V2. Sourced from https://ai.googleblog.com/2016/08/improving-inception-and-image.html

For multi-class, the best-performing model is DenseNet-121. Below is an image to show how this architecture functions.

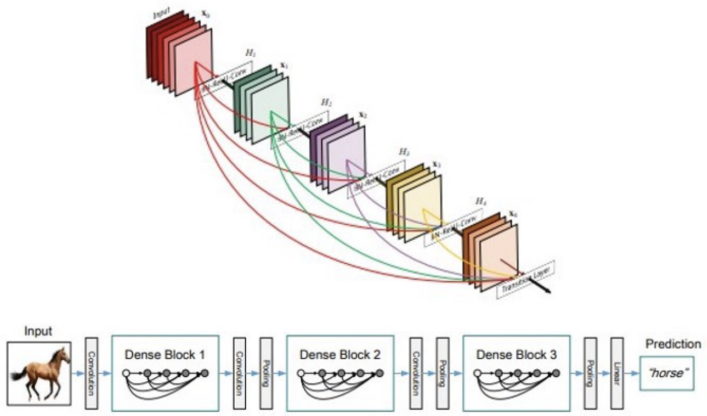


Figure 8: DenseNet-121 Architecture. Sourced from https://arxiv.org/pdf/1608.06993.pdf

The additional metrics associated with binary classification gives an insight on model performance at a more granular level. Building off the data in the above table, the metrics Precision, Recall, and AUC are added to the binary classification and show in the table below. The model that has the highest accuracy also needs high values in the other areas. In terms of AUC, DenseNet-121, InceptionResNet-V2, and VGG-16 all had high values. Precision had the highest values in VGG-16 and InceptionResNet-V2. And finally, Recall was highest in ResetNet-152V2 and DenseNet-201.

| Type | Accuracy | AUC | Precision | Recall |
|---|---|---|---|---|
| DenseNet-121 | 67% | 0.78 | 0.64 | 0.84 |
| DenseNet-201 | 58% | 0.71 | 0.56 | 0.97 |
| Inception V3 | 63% | 0.67 | 0.66 | 0.75 |
| InceptionResNetV2 | 71% | 0.78 | 0.75 | 0.67 |
| ResNet-50 | 55% | 0.55 | 0.62 | 0.91 |
| ResNet-152V2 | 53% | 0.5 | 0.53 | 0.9975 |
| VGG-16 | 68% | 0.76 | 0.77 | 0.57 |
| VGG-19 | 57% | 0.66 | 0.55 | 0.92 |
| Xception | 56% | 0.6 | 0.8 | 0.22 |

Figure 9: Results of Binary Classification with additional metrics

Overall, multi-class was able to have much stronger results than binary in terms of accuracy. This was expected from previous research though because trends are more easily identified for specific diseases rather than generalized for all types of anomalies.

## 6 Conclusion

Generalizing images as normal and abnormal was more difficult than classifying images as specific diseases. This is intuitive as there are several types of anomalies that can cause an image to be considered abnormal whereas a specific disease exhibits particular patterns. The data augmentation strategies proved to help generalize the multi-class approach, however did not perform as well for the binary approach. When using only horizontal flip transformations, the models produced the highest validation accuracy compared to including rotational transformations.

There are many additional explorations that can be made to further optimize the results of classifying chest x-rays for disease detection. This includes further investigating the optimal data augmentation strategy for classifying images as abnormal and normal. Also, the more images that are publicly available of rare lung conditions can help improve the results of the models and support identifying rarer cases. Overall, this experiment proved successful in classifying specific diseases for chest x-ray images.

## 7 Member Contribution

| Member | Contribution(%) |
|---|---|
| Manpreet Dhindsa | 33.33% |
| Gretchen Larrick | 33.33% |
| Sarah Rodgers | 33.33% |

### 7.1 Manpreet Dhindsa

- Multi-class classification
- Supported binary classification
- Wrote report and prepared presentation slides

### 7.2 Gretchen Larrick

- Data handling and preparation
- Binary classification
- Wrote report and prepared presentation slides

### 7.3 Sarah Rodgers

- Data augmentation
- Supported classifications
- Wrote report and prepared presentation slides

## References


[Ashhar et al., 2021] Ashhar, S. M., Mokri, S. S., Abd Rahni, A. A., Huddin, A. B., Zulkarnain, N., Azmi, N. A., and Mahaletchumy, T. (2021). Comparison of deep learning convolutional neural network (cnn) architectures for ct lung cancer classification. *International Journal of Advanced Technology and Engineering Exploration*, 8(74):126.

[Awan et al., 2021] Awan, M. J., Bilal, M. H., Yasin, A., Nobanee, H., Khan, N. S., and Zain, A. M. (2021). Detection of covid-19 in chest x-ray images: A big data enabled deep learning approach. *International journal of environmental research and public health*, 18(19):10147.

[Castiglioni et al., 2021] Castiglioni, I., Rundo, L., Codari, M., Di Leo, G., Salvatore, C., Interlenghi, M., Gallivanone, F., Cozzi, A., D'Amico, N. C., and Sardanelli, F. (2021). Ai applications to medical images: From machine learning to deep learning. *Physica Medica*, 83:9–24.

[Erdaw and Tachbele, 2021] Erdaw, Y. and Tachbele, E. (2021). Machine learning model applied on chest x-ray images enables automatic detection of covid-19 cases with high accuracy. *International Journal of General Medicine*, 14:4923.

[Hammoudi et al., 2021] Hammoudi, K., Benhabiles, H., Melkemi, M., Dornaika, F., Arganda-Carreras, I., Collard, D., and Scherpereel, A. (2021). Deep learning on chest x-ray images to detect and evaluate pneumonia cases at the era of covid-19. *Journal of medical systems*, 45(7):1–10.

[Heidari et al., 2020] Heidari, M., Mirniaharikandehei, S., Khuzani, A. Z., Danala, G., Qiu, Y., and Zheng, B. (2020). Improving the performance of cnn to predict the likelihood of covid-19 using chest x-ray images with preprocessing algorithms. *International journal of medical informatics*, 144:104284.

[Jain et al., 2021] Jain, R., Gupta, M., Taneja, S., and Hemanth, D. J. (2021). Deep learning based detection and analysis of covid-19 on chest x-ray images. *Applied Intelligence*, 51(3):1690–1700.

[Jaiswal et al., 2019] Jaiswal, A. K., Tiwari, P., Kumar, S., Gupta, D., Khanna, A., and Rodrigues, J. J. (2019). Identifying pneumonia in chest x-rays: A deep learning approach. *Measurement*, 145:511–518.

[Mikołajczyk and Grochowski, 2018] Mikołajczyk, A. and Grochowski, M. (2018). Data augmentation for improving deep learning in image classification problem. In *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, pages 117–122.

[Moradi et al., 2020] Moradi, R., Berangi, R., and Minaei, B. (2020). A survey of regularization strategies for deep models. *Artificial Intelligence Review*, 53(6):3947–3986.

[Moses, 2021] Moses, D. A. (2021). Deep learning applied to automatic disease detection using chest x-rays. *Journal of Medical Imaging and Radiation Oncology*, 65(5):498–517.

[Nabulsi et al., 2021] Nabulsi, Z., Sellergren, A., Jamshy, S., Lau, C., Santos, E., Kiraly, A. P., Ye, W., Yang, J., Pilgrim, R., Kazemzadeh, S., et al. (2021). Deep learning for distinguishing normal versus abnormal chest radiographs and generalization to two unseen diseases tuberculosis and covid-19. *Scientific reports*, 11(1):1–15.

[Seah et al., 2021] Seah, J. C., Tang, C. H., Buchlak, Q. D., Holt, X. G., Wardman, J. B., Aimoldin, A., Esmaili, N., Ahmad, H., Pham, H., Lambert, J. F., et al. (2021). Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. *The Lancet Digital Health*, 3(8):e496–e506.

[Takahashi and Usuzaki, 2022] Takahashi, K. and Usuzaki, T. (2022). A comprehensive deep-learning model for interpreting chest x-rays. *The Lancet Digital Health*, 4(1):e6.

# 8 Appendix

| Binary Model Hyperparameters | | | | | | |
|---|---|---|---|---|---|---|
| Transfer Learning Model | Data Augmentation | Optimizer | Learning Rate | Steps per Epoch | Epochs | Other |
| **VGG-16** | | | | | | |
| Model 1 | N | RMSprop | 0.0001 | 50 | 5 | |
| Model 3 | N | RMSprop | 0.0001 | 50 | 5 | Dropout=0.5 |
| Model 4 | Y | RMSprop | 0.0001 | 50 | 5 | Dropout=0.5 |
| **VGG-19** | | | | | | |
| Model 3 | N | RMSprop | 0.0001 | 50 | 5 | |
| Model 4 | Y | RMSprop | 0.0001 | 50 | 5 | |
| **InceptionResNetV2** | | | | | | |
| Model 2 | N | RMSprop | 0.0001 | 50 | 5 | Dropout = 0.5 |
| Model 3 | N | RMSprop | 0.0001 | 50 | 7 | Dropout = 0.5 (More hidden layers than Model 2) |
| Model 4 | N | Adam | 0.0001 | 50 | 5 | Dropout = 0.5 (More hidden layers than Model 2) |
| Model 5 | N | RMSprop | 0.001 | 50 | 5 | Dropout = 0.5 (More hidden layers than Model 2) |
| Model 6 | N | RMSprop | 0.00001 | 50 | 5 | Dropout = 0.5 (More hidden layers than Model 2) |
| **Inception V3** | | | | | | |
| Model 1 | N | RMSprop | 0.0001 | 50 | 7 | Dropout = 0.55 |
| **ResNet-50** | | | | | | |
| Model 1 | N | SGD | 0.001 | 50 | 5 | |
| Model 3 | N | SGD | 0.1 | 50 | 5 | |
| **ResNet-152V2** | | | | | | |
| Model 1 | Y | SGD | 0.001 | 50 | 5 | |
| **DenseNet-121** | | | | | | |
| Model 4 | N | RMSprop | 0.0001 | 50 | 5 | Dropout = 0.7 |
| **DenseNet-201** | | | | | | |
| Model 1 | Y | RMSprop | 0.0001 | 50 | 5 | Dropout = 0.7 |
| Model 6 | N | RMSprop | 0.001 | 50 | 5 | Dropout = 0.5 |
| **Xception** | | | | | | |
| Model 1 | Y | RMSprop | 0.0001 | 50 | 5 | |
| Model 6 | N | RMSprop | 0.0001 | 50 | 5 | |
| **EfficentB7** | | | | | | |
| Model 1 | N | RMSprop | 0.0001 | 50 | 5 | |
| Model 2 | N | Adam | 0.0001 | 50 | 5 | Dropout = 0.5 |
| Model 3 | N | Adam | 0.001 | 50 | 5 | Dropout = 0.7 |

Figure 10: Model Hyper-parameters for Binary Classification

| Model Type | Transfer Learning Model | Accuracy (%) | AUC | Precision | Recall |
|---|---|---|---|---|---|
| VGG-16 | Model 1 | 54 | 0.64 | 0.54 | 0.95 |
| | Model 3 | 68 | 0.76 | 0.77 | 0.57 |
| | Model 4 | 62 | 0.73 | 0.79 | 0.38 |
| VGG-19 | Model 3 | 57 | 0.66 | 0.55 | 0.92 |
| | Model 4 | 54 | 0.6 | 0.53 | 0.97 |
| InceptionResNetV2 | Model 2 | 58 | 0.71 | 0.56 | 0.97 |
| | Model 3 | 62 | 0.65 | 0.59 | 0.91 |
| | Model 4 | 71 | 0.78 | 0.75 | 0.67 |
| | Model 5 | 59 | 0.67 | 0.57 | 0.91 |
| | Model 6 | 62 | 0.58 | 0.69 | 0.49 |
| Inception V3 | Model 1 | 63 | 0.67 | 0.66 | 0.75 |
| ResNet-50 | Model 1 | 52 | 0.52 | 0.53 | 0.75 |
| | Model 3 | 55 | 0.55 | 0.62 | 0.91 |
| ResNet-152V2 | Model 1 | 53 | 0.5 | 0.53 | 0.9975 |
| DenseNet-121 | Model 4 | 67 | 0.78 | 0.64 | 0.84 |
| DenseNet-201 | Model 1 | 58 | 0.71 | 0.56 | 0.97 |
| | Model 6 | 61 | 0.68 | 0.61 | 0.74 |
| Xception | Model 1 | 56 | 0.6 | 0.8 | 0.22 |
| | Model 3 | 53 | 50 | 53 | 0.9995 |
| | Model 6 | 53 | 53 | 53 | 0.996 |
| EfficentB7 | Model 1 | 58 | 0.62 | 0.6 | 0.64 |
| | Model 2 | 58 | 0.7 | 0.56 | 0.9 |
| | Model 3 | 53 | 0.5 | 0.53 | 0.99 |

Figure 11: Binary Classification Results with Additional Metrics

| MultiClass Model Hyperparameters | | | | | | | | Results |
|---|---|---|---|---|---|---|---|---|
| Transfer Learning Model | Data Augmentation | Early Stopping | Optimizer | Learning Rate | Steps per Epoch | Number of Epochs | Etc. | Accuracy |
| **VGG-16** | | | | | | | | |
| Model 1 | N | N | RMSprop | 0.0001 | 50 | 5 | | 80% |
| Model 2 | Y | N | RMSprop | 0.0001 | 50 | 5 | | 87% |
| **VGG-19** | | | | | | | | |
| Model 1 | Y | Y - Patience: 5 | RMSprop | 0.0001 | 50 | 5 | | 71% |
| Model 2 | Y | Y - Patience: 5 | RMSprop | 0.0001 | 50 | 5 | | 85% |
| Model 3 | N | N | RMSprop | 0.0001 | 50 | 5 | | 89% |
| **InceptionResNetV2** | | | | | | | | |
| Model 1 | N | N | RMSprop | 0.001 | | | Dropout = 0.7 | 54% |
| Model 2 | N | N | RMSprop | 0.0001 | | | Dropout = 0.5 | 89% |
| **Inception V3** | | | | | | | | |
| Model 1 | N | N | RMSprop | 0.0001 | 50 | 7 | Dropout = 0.55 | 79% |
| Model 2 | N | N | RMSprop | 0.0001 | 50 | 7 | Dropout = 0.7 | 89% |
| **ResNet-50** | | | | | | | | |
| Model 1 | N | N | SGD | 0.001 | 50 | 5 | | 89% |
| Model 2 | Y | N | SGD | 0.001 | 50 | 5 | | 66% |
| **ResNet-152V2** | | | | | | | | |
| Model 1 | Y | N | SGD | 0.001 | 50 | 5 | | 70% |
| **DenseNet-121** | | | | | | | | |
| Model 1 | N | N | Adam | 0.001 | 50 | 7 | Dropout | 31% |
| Model 2 | N | N | Adam | 0.001 | 50 | 7 | Dropout | 28% |
| Model 3 | N | N | RMSprop | 0.001 | 50 | 5 | Dropout = 0.5 | 56% |
| Model 4 | N | N | RMSprop | 0.0001 | 50 | 5 | Dropout = 0.7 | 95% |
| Model 5 | Y | N | RMSprop | 0.0001 | 50 | 5 | Dropout = 0.7 | 65% |
| Model 6 | Y | N | RMSprop | 0.0001 | 50 | 5 | Dropout = 0.7 | 88% |
| **DenseNet-201** | | | | | | | | |
| Model 1 | Y | N | RMSprop | 0.0001 | 50 | 5 | Dropout = 0.7 | 68% |
| Model 2 | Y | N | RMSprop | 0.0001 | 50 | 7 | Dropout = 0.8 | 55% |
| Model 3 | Y | N | RMSprop | 0.0001 | 50 | 7 | Dropout = 0.7 | 59% |
| Model 4 | Y | N | RMSprop | 0.0001 | 50 | 5 | Dropout = 0.7 | 81% |
| Model 5 | N | N | RMSprop | 0.0001 | 50 | 5 | Dropout = 0.7 | 76% |
| **Xception** | | | | | | | | |
| Model 1 | Y | N | RMSprop | 0.0001 | 50 | 5 | | 56% |
| Model 2 | Y | N | RMSprop | 0.0001 | 50 | 5 | Dropout = 0.7; Momentum 0.9; Decay 0.01 | 63% |
| Model 3 | Y | N | RMSprop | 0.0001 | 50 | 5 | Dropout = 0.5; Momentum 0.7; Decay 0.001 | 65% |
| Model 4 | Y | N | RMSprop | 0.001 | 50 | 5 | Dropout = 0.7; Momentum 0.7; Decay 0.001 | 53% |
| Model 5 | Y | N | RMSprop | 0.0001 | 50 | 5 | Dropout = 0.5; Momentum 0.7; Decay 0.001 | 60% |

Figure 12: Model Hyper-parameters for Multi-Class Classification