Accurate Classification of COVID-19 and Pneumonia from Chest X-ray Images

Emily Getzen, Revathi Vijayaraghavan, and Kyle Coleman

December 12, 2021

Abstract

SARS-CoV-2, the virus that causes the respiratory disease COVID-19, has infected hundreds of millions of people worldwide and caused millions of deaths. Early detection of COVID-19 through testing is critical, as prolonged infection without treatment and quarantine can result in serious illness and spread of the disease to other individuals. However, many standard COVID-19 tests, including the SARS-CoV-2 RT-PCR test, are in short supply in many rural areas. Also, some tests suffer from a relatively low sensitivity. Due to these shortcomings of molecular COVID-19 tests, a machine learning algorithm that can be used to accurately classify COVID-19 positive patients without standard testing is desirable. Here we train multiple machine learning models for the classification of COVID-19 patients using only chest X-ray imaging data. We utilized baseline models (logistic regression, decision trees, random forest, and bagging), neural networks (feed forward and CNNs), and transfer-learning-based methods. Our transfer learning model, consisting of the pre-trained Inception V3 network connected to a feed forward network, reported a sensitivity of 0.994 and AUC of 0.973. In addition to classifying patients as COVID-19 positive or negative, since pneumonia is a common symptom of COVID-19, we also trained each of our models to predict whether or not a patient has pneumonia using chest X-rays. The same transfer learning model described above attained a sensitivity of 0.915 and AUC of 0.989 when trained for pneumonia classification. In addition, our models are computationally fast, and since X-rays can be obtained within minutes, a COVID-19 diagnosis can be made far more quickly than if using a RT-PCR test. Given the impressive performance and computational efficiency of our models, we anticipate that they will become useful alternatives to molecular COVID-19 testing.

1 Motivation

The COVID-19 pandemic has affected the world population for almost two years now. Throughout this time there have been countless advances to fight the pandemic– improvements to diagnostic testing, life-saving medications and vaccines, and better public health knowledge to guide citizens on how to keep themselves and their loved ones safe. However, access to these resources remains inequitable in the United States, even two years later. Particularly, rural communities lack resources for testing compared to more populated regions of the country [1]. Rapid and widespread testing of COVID-19 is necessary in order to mitigate the disease and prevent hospitalizations and deaths. Additional mechanisms to identify COVID-19 in patients could be useful in the face of lack of resources for these communities. One such possibility is a chest X-ray. Additionally, a leading cause of mortality in COVID-19 patients is pneumonia caused by the disease [2].Pneumonia is a lung infection that causes inflammation to the air sacs inside the lungs. This can cause these sacs to fill with fluid, thus making it difficult for the patient to breathe. The majority of patients that account for morbidity and mortality with the infection develop severe pneumonia that requires mechanical ventilation. Earlier and more accurate detection of pneumonia is vastly important to save COVID-19 patients' lives and reduce suffering. To contribute to the fight against COVID-19, we are interested in developing and assessing various machine learning algorithms to detect COVID-19 and pneumonia in chest X-rays. Respectively, this would assist rural communities with less access to tests, and catch pneumonia in COVID-19 patients quickly such that providers can employ interventions faster. This endeavor has the potential to save countless lives.

2 Related Work

There are a variety of related works that inform our study. Nikolaou et. al. (2021) used the COVID-19 radiography database from Kaggle. They developed a CNN which was trained, validated, and tested on 15,000 X-ray images. It achieved 95 perent accuracy, 90 percent sensitivity, and 97 percent specificity with regard to differentiating COVID-19 from normal lungs. Drozdov et al. (2021) developed an AI algorithm called Covlx to differentiate normal, abnormal, non-COVID-19 pneumonia, and COVID-19 chest X-rays using a cohort of 293,143 chest x-rays. They achieved an AUC of 0.86, sensitivity of 0.83, and F1-score of 0.71 which performs on-par with board-certified radiologists. Osman et al. (2021) suggest a new COVID-19 identification technique based on locality-weighted learning and self-organization

map (LWL-SOM) based on images from chest-X-rays which had better results for distinguishing COVID-19 and non-COVID-19 patients than current machine learning baselines. Lastly, Kwon et al. (2021) recognizes that detecting pneumonia from chest X-rays is challenging even for experienced radiologists. Thus, they develop an ensemble model for predicting pneumonia that used 157,016 chest X-ray images. It achieved an AUC of 0.983. As one can see, many of the recent works surrounding this problem were published within the last year, thus displaying the fact that diagnosis of COVID-19 and identification of pneumonia via chest X-ray images is a pertinent problem requiring a solution. Although the previous works achieved impressive performance with regard to their models, they also require many images to achieve such results. We believe we can beat their results with less images, which would also reduce computation time.

3 Data

We use chest X-rays from the COVID-19 Radiography Database from Kaggle. This database contains the chest X-rays of 3616 COVID-19 patients, 10192 normal patients and 1345 viral pneumonia patients. It is the result of a collaboration between researchers from Qatar University and the University of Dhaka, along with collaborators from Pakistan and Malaysia. All images are in Portable Network Graphics (PNG) file format and the resolution of the images are 299x299 pixels. The link to the database can be found here: Kaggle COVID-19 Radiography Database. For our models, we randomly sample 500 normal images, 500 COVID-19 images, and 500 pneumonia images to ensure balanced cases and controls in the models. Since one of our main aims was to learn a better model using less data, this will also help us achieve and validate that. Another reason for doing so was to decrease the computational intensity for models like the transfer learning model.

Figure 1: X-ray of Normal Lungs

Figure 2: X-ray of COVID-19 Lungs

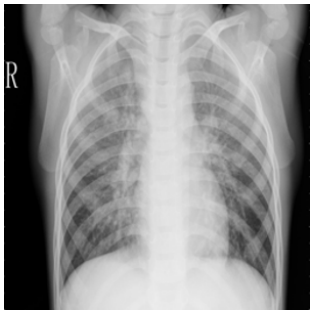
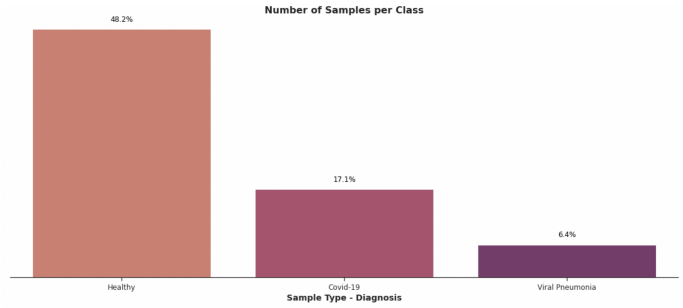


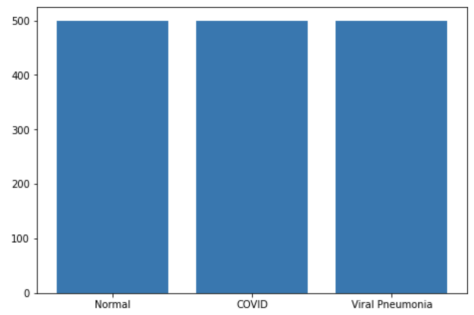
Figure 3: X-ray of Pneumonia Lungs

3.1 Dataset Visualization

We see that the initial distribution of data is not even, the dataset is unbalanced. Around 48% of the images are normal images, with 17% COVID-19 images and 6% pneumonia images. As mentioned before, we randomly sampled 500 images from each category for our purposes. The original distribution and that used for training are shown in the following figure.



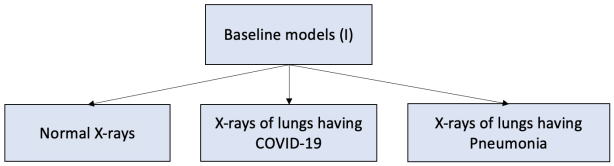
(a) Distribution of original dataset



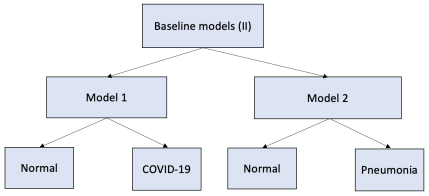
(b) Distribution of dataset used for training

4 Problem Formulation

We model this problem as a classification problem. We have three categories of images - Normal Lung, Lung with Pneumonia and Lung with COVID-19. Typically, this would be considered as a multi-class classification problem and we did treat it as such. However we see that there is no data in the dataset that tells us whether pneumonia and COVID-19 are completely independent. The one thing we know for sure are that the images labeled "Normal" are healthy and will be the controls for our use case. Since we don't know the dependence of COVID-19 on pneumonia and vice-versa, we first test the baseline models to detect the three classes, following which we convert the problem into a binary classification problem. We do so by building two models in each of the baseline methods - one that is trained to detect Normal and COVID-19 images, and one that is trained to detect Normal and Pneumonia images. We then compare which of the two (multiclass or binary) methods gives better recall and F1-scores, and continue to use that method of classification for any further models.



(a) Multiclass classification using baseline methods



(b) Binary classification with 2 models using baseline methods

We develop at least two models for lung infection classification using chest X-ray imaging data. We start by testing the baseline methods - which includes Logistic Regression, Decision Trees, K-NN, RandomForests, and Bagging. Based on these results, the next models we will build are neural networks. The parameters will be decided using cross validation. Finally, we also investigate the method of Transfer Learning, and compare the output of that to the custom neural networks and CNNs. The testing of each of the models will be done from a random sampling of the dataset that wasn't included in the training dataset. The models will be evaluated using AUC, recall and F-1 scores. An interesting method we explore later is to try and predict Pneumonia using a model trained to identify COVID-19 and vice-versa. This will help us understand the correlation between the two diseases, if it exists.

5 Methods

5.1 No-model classification

Here, we calculate the frequencies of each of the classes and always output the class that had the highest frequency in the training set. If we were to look at this for the initial dataset (before sampling), the most frequently occuring class if the Normal lung. By default, any and every image will be classified as a normal image. If we consider the sampled images dataset, for multiclass classification, the model will randomly choose one of the classes as the output, with each having a probability of **33%**, since there are equal number of images in each class. Finally, if we consider splitting the model into two binary classifications, we will get the output to be Normal images in **50%** of the cases, COVID-19 in **25%** of the cases and Pneumonia in **25%** of the cases. This method is not accurate, since we are not using the data given to us at all.

5.2 Baseline Models

5.2.1 Pre-Processing

We simply convert the images of size 299x299x1 into a flattened vector and pass this as the input to the models. Thus, the effective input was a vector of size (98401,). We directly discuss the results obtained on varying different parameters for the baseline models.

5.2.2 Logistic Regression

As shown in Tables 1, 2 and 3, the binary classification models perform much better than the multiclass classification method. We also notice that even in cases when the accuracy is high, sometimes the recall is not as good. The regularized model did not perform much better than the non-regularized model. Finally, in all cases, we note that feature scaling tends to give the best results for the Logistic Regression model.

Table 1: Multi-class Classification using Logistic Regression

Method	Accuracy	Recall	F1 Score
Basic Model	0.864	0.863	0.863
L2 Regression	0.856	0.856	0.85
Feature Scaling (0-1)	0.862	0.863	0.861

Table 2: Binary Classification using Logistic Regression (COVID-19)

Method	Accuracy	Recall	F1 Score
Basic Model	0.925	0.932	0.928
L2 Regression	0.925	0.908	0.929
Feature Scaling (0-1)	0.93	0.932	0.929

Table 3: Binary Classification using Logistic Regression (Pneumonia)

Method	Accuracy	Recall	F1 Score
Basic Model	0.91	0.894	0.911
L2 Regression	0.91	0.903	0.912
Feature Scaling (0-1)	0.935	0.932	0.937

5.2.3 Decision Trees

Table 4: Multi-class Classification using Decision Trees

Method	Accuracy	Recall	F1 Score
Basic Model	0.774	0.774	0.778
Max Depth = 7	0.787	0.787	0.781
Criterion=Entropy	0.78	0.783	0.784

Table 5: Binary Classification using Decision Trees (COVID-19)

Method	Accuracy	Recall	F1 Score
Basic Model	0.89	0.894	0.894
Max Depth = 7	0.895	0.913	0.90
Criterion=Entropy	0.87	0.855	0.872

Table 6: Binary Classification using Decision Trees (Pneumonia)

Method	Accuracy	Recall	F1 Score
Basic Model	0.81	0.865	0.825
Max Depth = 7	0.78	0.82	0.796
Criterion=Entropy	0.78	0.767	0.778

As shown in Tables 4, 5 and 6, we observe here too that the binary classification models outperform the multi-class model. A few notable observations are that the model perform better when the criterion is set to "gini" (split based on gini index) as compared to "entropy" (split based on entropy). We also observe that setting the max_depth to 7 gave a much better result in the COVID-19 binary model as compared to the Pneumonia model.

5.2.4 Ensemble Methods

We use two similar methods here - Extra Tree Classifier and Random Forest Classifier. They essentially perform similar computations, except in the way they perform splits and how they sample the data. Random forests perform bootstrapping (drawing samples without replacement), whereas extra trees don't. Random Forests choose the best split and extra tree chooses a random split.

A. Extra Tree Classifier

Table 7: Classification using Extra Tree Classifier

Method	Accuracy	Recall	F1 Score
Multi-class classification	0.899	0.899	0.897
Binary Classification (COVID-19)	0.95	0.951	0.953
Binary Classification (Pneumonia)	0.943	0.961	0.952

B. Random Forests

Table 8: Classification using Random Forest Classifier

Method	Accuracy	Recall	F1 Score
Multi-class classification	0.892	0.892	0.890
Binary Classification (COVID-19)	0.94	0.98	0.944
Binary Classification (Pneumonia)	0.93	0.951	0.933

C. Bagging

Table 9: Classification using Bagging

Method	Accuracy	Recall	F1 Score
Multi-class classification	0.863	0.863	0.861
Binary Classification (COVID-19)	0.933	0.935	0.930
Binary Classification (Pneumonia)	0.88	0.855	0.881

We observe that the binary models perform better compared to the multi-class model. We see that the Extra Tree Classifier and the Random Forest classifier give extremely good scores. Overall we see that the ensemble methods outperform Logistic Regression and Decision Trees thus far.

An additional conclusion after testing all the models with different hyperparameters is that in general, the binary models on two sets of classes tended to perform better with higher recall and F1-scores. Since that is our main aim, we focus our following neural network models and transfer learning model to perform binary classification of COVID-19 and Pneumonia.

5.3 Neural Network Architectures

5.3.1 Overview

We constructed and trained four different neural network models for COVID-19 and pneumonia classification. Each model was trained separately for each classification task. Two of models are feed forward neural networks, and two are convolutional neural networks (CNNs). We utilized CNNS due to their well-known efficacy in image classification tasks, and we wanted to compare their performance to feed forward neural networks, which are generally simpler models that require less training time. The parameters of each model were optimized using the Adam optimizer (learning rate 10^{-3}) with a binary cross-entropy (BCE) loss function. The number of epochs used for training a given model was determined using the validation loss over a maximum of 50 epochs. All neural networks were trained using PyTorch.

5.3.2 Feed Forward Neural Networks

We constructed two feed forward neural networks for disease classification. The first network (Net1) consists of two fully connected layers, and the output of dimension 1 is used as input for the sigmoid activation function. Specifically, Net1 has the following architecture:

Fully connected layer (Input size: 299*299*1, Output size: 32)→ ReLU activation function→ Fully connected layer (Input size: 32, Output size: 1)→ Sigmoid activation function

The second feed forward neural network (Net2) is similar to Net1, but has an additional fully connected layer and uses dropout of 50% after the first fully connected layer for regularization. The architecture of Net2 is shown here:

Fully connected layer (Input size: 299*299*1, Output size: 256) → ReLU activation function → Dropout (50%)→ Fully connected layer (Input size: 256, Output size: 32)→ ReLU activation function→ Fully connected layer (Input size: 32, Output size: 1)→ Sigmoid activation function

Plots of the Net1 and Net2 training and validation loss for the COVID-19 classification problem are shown in Figure 6. As evidenced by the lack of fluctuations in the validation loss when approaching the maximum number of epochs, the models do not appear to overfit after 50 epochs. Hence, when calculating the evaluation metrics for the validation set, we used Net1 and Net2 models trained for 50 epochs.

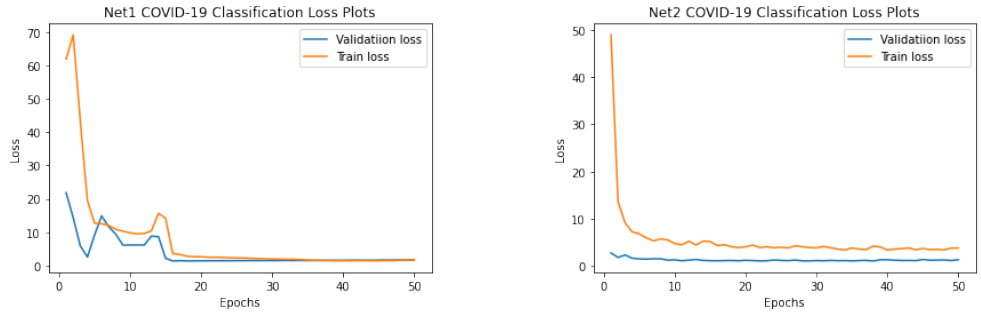


Figure 6: Net1 and Net2 training and validation loss plots for COVID-19 classification

Plots of the Net1 and Net2 training and validation loss for the pneumonia classification problem are shown in Figure 7. Similar to the loss plots for the COVID-19 classification problem, the Net1 and Net2 validation loss remains relatively low when approaching the maximum number of epochs, and so the final Net1 and Net2 models were trained for 50 epochs.

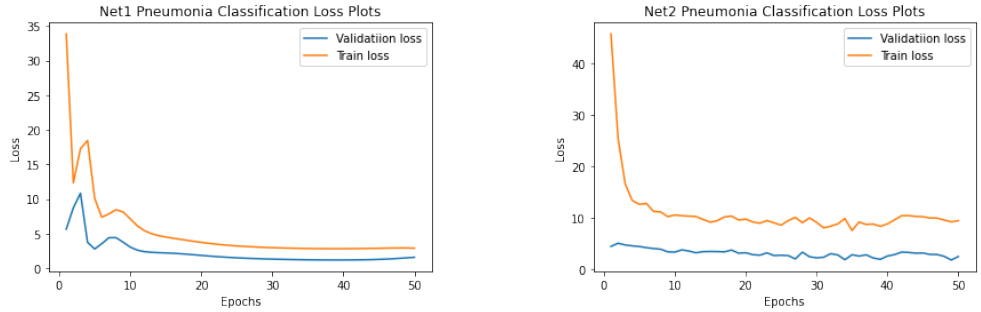


Figure 7: Net1 and Net2 training and validation loss plots for pneumonia classification

5.3.3 Convolutional Neural Networks

In addition to the feed forward neural networks described above, we also constructed two different neural network architectures consisting of convolutional layers. The first CNN (Net3) has the following architecture:

Convolutional layer (in channels:1, out channels:16, kernel size:4, stride:3, padding:1) → ReLU activation function → Fully connected layer (Input size: 100*100*16, Output size: 128) → ReLU activation function → Dropout (50%) → Fully connected layer (Input size: 128, Output size: 1) → Sigmoid activation function

The second CNN (Net4) uses three convolutional layers and a max pooling layer for regularization. Specifically, Net4 has the following architecture:

Convolutional layer (in channels:1, out channels:8, kernel size:4, stride:3, padding:1) → ReLU activation function → Convolutional layer (in channels:8, out channels:16, kernel size:6, stride:3, padding:1) → ReLU activation function → Convolutional layer (in channels:16, out channels:32, kernel size:2, stride:1, padding:1)→ Max pooling layer (Input size:34x34x32, Output size: 17x17x32)→ Fully connected layer (Input size: 17*17*32, Output size: 128)→ ReLU activation function→ Dropout (50%)→ Fully connected layer (Input size: 128, Output size: 1)→ Sigmoid activation function

During training for the COVID-19 classification task for 50 epochs, visualization of the validation loss indicated that both models overfit the data. As a result, we trained Net3 and Net4 for 25 and 20 epochs, respectively. The plots of the training and validation losses for Net3 and Net4, which are shown in Figure 8, do not show signs of overfit when reducing the numbers of epochs for training to 25 and 20, respectively.

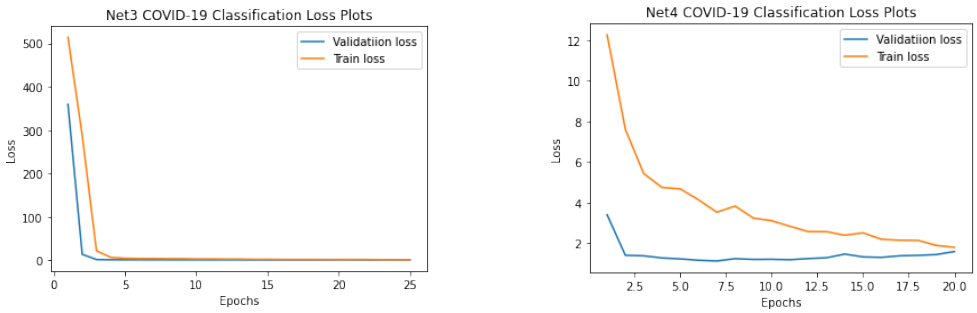


Figure 8: Net3 and Net4 training and validation loss plots for COVID-19 classification

When being trained for the pneumonia classification task, Net3 and Net4 do not overfit after 50 epochs, as evidenced by the plots in Figure 9.

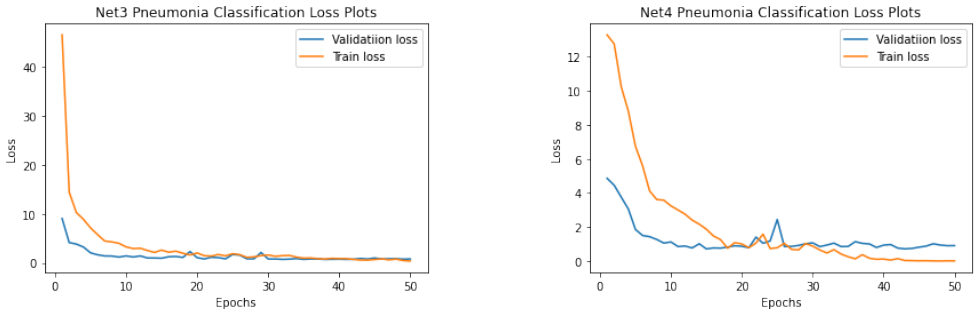


Figure 9: Net3 and Net4 training and validation loss plots for pneumonia classification

In addition to the four neural network architectures described above, we also constructed an ensemble method in which the output is the prediction using the mean of the output of the sigmoid function across the four neural networks. When evaluating performance on the validation set for COVID-19 classification, Net4 resulted in the highest AUC and F1 Score, but the sensitivity was slightly lower than those given by Net1, Net2, and the ensemble method. Since Net3 had the highest measures for a majority of metrics when evaluated on the validation set, we chose this as the model to use for test set evaluation. For pneumonia classification, Net3 performed the best across all metrics (AUC, Sensitivity, F1 Score) when evaluated on the validation set, and so this model was chosen for test set evaluation.

5.4 Transfer Learning

For the transfer learning portion of this work, we focus on the InceptionV3 architecture.

5.4.1 Pre-Processing

The InceptionV3 algorithm requires images to be in RGB format (input shape of (299,299,3). The vast majority of the images in our dataset are greyscale, with an input shape of (299,299,1). Thus, what we did to pre-process is repeat the greyscale torches 3x to create a false RGB image of the desired input shape. Additionally according to the algorithm, the images needed to be loaded into a range of [0,1] and normalized using mean = [0.485, 0.456, 0.406] and std = [0.299, 0.224, 0.225].

5.4.2 InceptionV3

The InceptionV3 model was chosen as our transfer learning algorithm due to the fact that it achieved the top position in the Conference on Computer Vision and Pattern Recognition (CVPR) in 2016. Inception was introduced in 2014 by Google (introduced as the InceptionV1 model). At the time, it had far fewer parameters compared to other prevalent models like VGG and AlexNet [5]. The innovation here is the Inception Module, which performs convolutions with differing filter sizes on the input, performs max pooling, and concatenates results for the next Inception module. Inception V2 increased accuracy and made the model less complex. Inception V3 included batch normalization, more

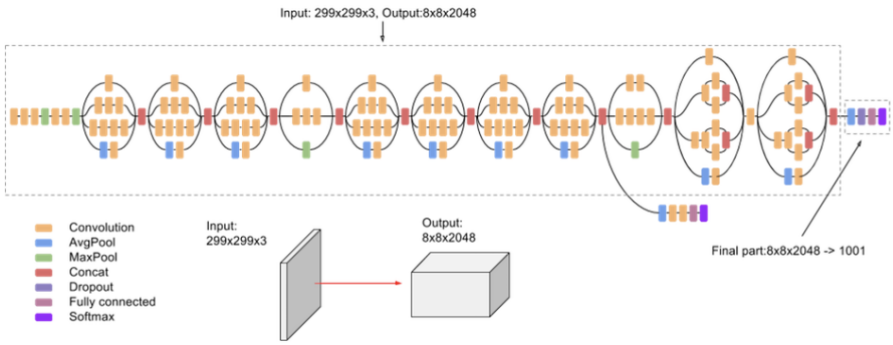


Figure 10: InceptionV3 architecture

factorization, and an RMSProp optimizer. The overall V3 model is made of asymmetric and symmetric building blocks that include convolutions, average pooling, max polling, dropouts, and fully connected layers [6].

The InceptionV3 algorithm is loaded with weights obtained from pre-training on ImageNet data. This data comes from a project aimed at categorizing images into almost 22,000 categories for computer vision research.

5.4.3 Feed Forward Neural Network Architectures

In transfer learning, the neural network model is first trained on a different problem, and then one or more layers from the trained model are used in a new model which is trained on the actual problem of interest. In this case, we do not load the fully-connected output layers of the InceptionV3 network, and instead design our own feed-forward neural network architectures. We try two different architectures and select one of them based on validation performance. The two models are:

- Feed forward neural network with one dense hidden layer of dimension = 1024, dropout = 0.1, and sigmoid activation function.
- Feed forward neural network with hidden layer 1 = 1024 nodes, hidden layer 2 = 500 nodes, dropout = 0.2, and sigmoid activation function.

For the COVID-19 model, through cross-validation it was determined that the second architecture yielded the best results. For the Pneumonia model, it was determined that the first architecture yielded the best results.

6 Results

6.1 Evaluation Metrics

The metrics we use for model evaluation are recall, F1 and AUC scores. When further testing our cross-comparison model (predict pneumonia using covid-19 model and vice versa), we will also use sensitivity and specificity to evaluate the results. We chose these metrics due to their importance in disease diagnosis. Recall takes a high priority since we want to limit the number of false negatives predicted.

6.2 Comparison of the Best Models

Tables 10 and 11 show, for all methods selected based on validation set performance, the AUC, recall, and F1 score when evaluated on the test sets for COVID-19 classification and pneumonia classification, respectively. The transfer learning algorithm provided the highest AUC and recall for COVID-19 classification, and the highest AUC for pneumonia classification.

Out of curiosity, we were interested to see if, given that the models are able to accurately predict COVID-19 and Pneumonia respectively, we can determine any similarities between COVID-19 and Pneumonia lung X-rays provided that the COVID-19 model does a good job of predicting pneumonia cases and vice versa. Thus, this is another attribute of the work that we assess. These results are shown in Tables 12 and 13.

Method	AUC	Recall	F1 Score
Logistic Regression	0.93	0.932	0.929
Decision Trees	0.895	0.913	0.90
Extra Tree	0.95	0.951	0.953
Random Forest	0.94	0.980	0.944
Bagging	0.933	0.935	0.930
CNN	0.914	0.942	0.920
Transfer Learning	0.973	0.994	0.932

Table 10: Comparison of model performance for predicting COVID-19 by best models (with parameters selected by cross-validation)

Method	AUC	Recall	F1 Score
Logistic Regression	0.935	0.933	0.937
Decision Trees	0.778	0.827	0.796
Extra Tree	0.943	0.961	0.952
Random Forest	0.929	0.951	0.933
Bagging	0.881	0.856	0.881
CNN	0.946	0.933	0.946
Transfer Learning	0.989	0.915	0.938

Table 11: Comparison of model performance for predicting Pneumonia by best models (with parameters selected by cross-validation)

AUC	0.812
Recall	0.157
Specificity	0.963
F1 Score	0.268

Table 12: Predicting COVID-19 using the pneumonia model

AUC	0.828
Recall	0.279
Specificity	0.964
F1 Score	0.423

Table 13: Predicting pneumonia using the COVID-19 model

7 Discussion

7.1 Conclusion

We see that across the metrics, transfer learning performs the best and the most consistently, which was an expected result. Primarily, because the Inception model is deep-CNN, and has been previosuly trained on millions of images. This helps the model identify simpler features and patterns in the images with more ease. However, we do see that CNN outperforms transfer learning in F1 score for the COVID-19 model, and Extra Tree Classifier outperforms transfer learning in Recall for the pneumonia model. An important observation here is that, even though the Extra Tree and Random Forest models perform well, they lack the flexibility for feature processing and modeling that convolutional neural net, or a transfer learning model can be structured to take advantage of the repetitive structure of images and the latent patterns in visual data. Also, we see that our best results outperform the results outlined in the Related Work section of the paper. Notably, Nikolaou et al. (2021) used the same COVID-19 Radiography Database that we did, and developed a CNN which was trained, validated, and tested on 15,000 X-ray images. We were able to outperform their results in terms of recall with just 1500 images. Thus, we achieved our goal of obtaining more accurate results with far less computation.

We also see that both models tested on alternate data (COVID-19 model tested on pneumonia cases and pneumonia model tested on COVID-19 cases) have low recall and high specificity. One possible reason for this could be the fact that both models were trained on normal images, which naturally makes it easier for the alternate model to identify the controls. Additionally, the normal chest X-rays seem to have better defined skeletal structures in the images. This could make the normal images easier to identify. Images with higher probabilities of being normal had very well-defined bone structures.

7.2 Limitations and Future Directions

We first enlist the limitations of this project. To begin with, images with different labels come from different sources which might have an impact on the findings. Secondly, the images in the dataset were in grayscale which could mean that there was potential information that was not included due to it not being in RGB. A future scope for this project would be the prediction of COVID-19 Pneumonia from regular pneumonia images.

8 References

Hendricks, Brian and Rajib, Paul and Smith, Cassie and Wen, Sijin, Kimble, Wes and Amjad, Ayne and Atkins Amy and Hodder, Sally. 2021. Coronavirus testing disparities associated with community level deprivation, racial inequalities, and food insecurity in West Virginia. *Annals of Epidemiology*

Grant, R. A. and Morales-Nebreda, L and Karnov, N. S. et al. 2021. Circuits between infected macrophages and T cells in SARS-CoV-2 pneumonia. *Nature*

Nikolaou, V and Massaro, S and Fakhimi, M and Stergioulas, L and Garn, W. 2021. COVID-19 diagnosis from chest x-rays: developing a simple, fast, and accurate neural network.*Health Inf Sci Syst*

Huilgol, Purva. 2020. Top 4 Pre-Trained Models for Image Classification with Python Code. *Analytics Vidhya*

Christian Szegedy and Vincent Vanhoucke and Sergey Ioffe and Jonathon Shlens and Zbigniew Wojna. 2015. Rethinking the Inception Architecture for Computer Vision. *arXiv*