Classification of Medical Imagery using Convolutional Neural Networks

By: Mason Matthews, Josh Sullivan, Matthew Widjaja

Introduction

In this project report, we will detail the usage and application of Convolutional Neural Networks(CNNs) in classifying Medical Imagery. We utilized a dataset from Kaggle containing roughly 400 images of MRI scans of human lungs with either Covid-19, Viral Pneumonia or no major lung conditions to train a CNN to predict the health conditions(in relation to Covid-19 and Viral Pneumonia) of lungs from MRI scans. This project is also informative on the privacy and social concerns of large datasets containing potentially sensitive information. While the dataset we have used only contains ethically sourced information from a public institution, others may not have the same standard of ethics.

Project Goal

Our main goal for this project is for our Convolutional Neural Network to be able to accurately determine whether lungs are being affected by Covid-19, Viral Pneumonia, or neither. To further numericize our results, our intention was to have a recall of over 90%, an F1 Score of over 90%, a precision of over 90%, and an accuracy of over 90%, in order of importance. Beyond this, we emphasized learning the utilization of CNNs, as well as autoencoders, to best develop, train, test, and analyze datasets, specifically within the context of diagnostic medicine.

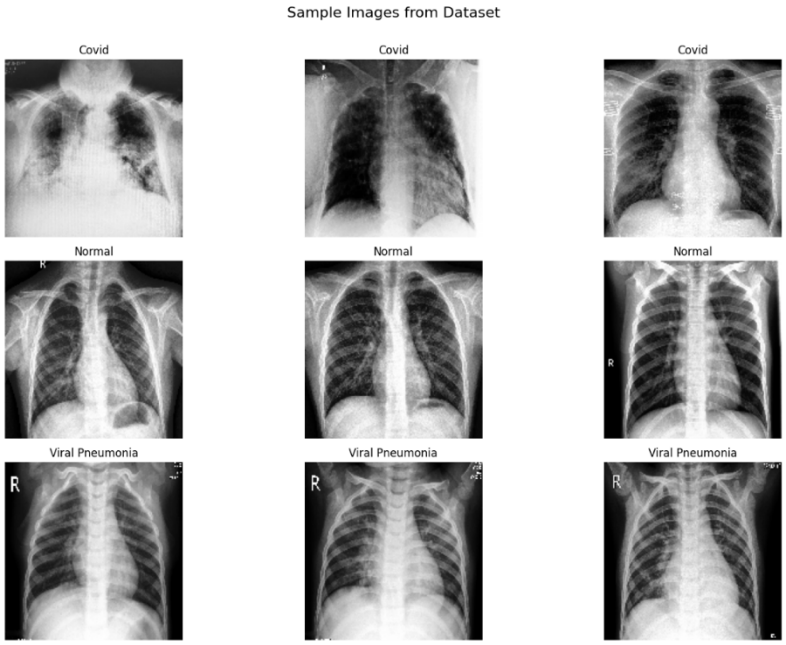
Dataset

The chosen dataset for this project and report comes from Kaggle.com, having been compiled and posted by a user by the name of Pranav Raikote. The dataset contains 317 images of chest x-rays of patients with Covid-19, Viral Pneumonia, and healthy, or “normal” lungs as they are described in the dataset. The images and patient data regarding said images were released by the University of Montreal in 2019.

A key challenge with this dataset is the inherent data imbalance. Specifically, the dataset contains more images of Covid-19 cases compared to the other two classes, Normal and Viral Pneumonia. This imbalance can lead to a model that is biased toward the majority class (Covid), while underperforming on the minority classes, particularly Viral Pneumonia. Additionally, the relatively small sample size of 317 images limits the ability of the model to generalize well, especially when deployed on unseen, real-world data. Small datasets can result in overfitting, where the model performs well on the training set but struggles with new data.

Despite these limitations, we aimed to optimize the model's performance based on the available dataset. Techniques such as class weighting were implemented to ensure that minority classes received appropriate attention during training. Moreover, the model was trained and validated carefully to balance its performance across all classes. While the current results demonstrate promising accuracy, further improvements could be achieved by using a larger and more diverse dataset. Such improvements would enhance the model's ability to generalize across various demographics and imaging conditions.

Below is a visual sample of the dataset, illustrating examples of Covid, Normal, and Viral Pneumonia chest X-ray images. These examples highlight the variability in image characteristics and the importance of ensuring robust performance for all categories.



Methodology

Data Preprocessing:

The dataset used for this project consists of medical images categorized into three classes: Covid, Normal, and Viral Pneumonia. These images are divided into training and testing sets, stored in respective directories, and loaded using TensorFlow's ImageDataGenerator. To prepare the images for the neural network, a preprocessing step rescales the pixel values to a range between 0 and 1 using rescale=1/255. This normalization ensures that the network can learn effectively and avoids large gradient updates that might hinder convergence.

A significant challenge with this dataset is the inherent data imbalance. Specifically, there are more images of Covid-19 cases compared to the other two classes, Normal and Viral Pneumonia. This imbalance can cause the model to be biased toward the majority class (Covid), leading to poorer performance on the underrepresented classes. Additionally, the relatively small sample size of 317 images limits the model's ability to generalize well to new data, increasing the risk of overfitting.

To address the data imbalance, class weights were calculated and applied during model training to ensure that the minority classes received appropriate attention. The calculated class weights are as follows:

- Class 0 (Covid): 0.7537
- Class 1 (Normal): 1.1952
- Class 2 (Viral Pneumonia): 1.1952

During initial experiments, we also explored using autoencoders to improve feature extraction and model generalization. However, given the relatively clean and small dataset, we observed diminishing returns with autoencoders. As a result, we focused solely on the CNN approach to optimize performance based on the available data.

Model Design:

Model: "sequential"

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 256, 256, 32)	896
max_pooling2d (MaxPooling2D)	(None, 128, 128, 32)	0
conv2d_1 (Conv2D)	(None, 128, 128, 64)	18,496
max_pooling2d_1 (MaxPooling2D)	(None, 64, 64, 64)	0
conv2d_2 (Conv2D)	(None, 64, 64, 128)	73,856
max_pooling2d_2 (MaxPooling2D)	(None, 32, 32, 128)	0
dropout (Dropout)	(None, 32, 32, 128)	0
flatten (Flatten)	(None, 131072)	0
dense (Dense)	(None, 128)	16,777,344
dropout_1 (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 128)	16,512
dense_2 (Dense)	(None, 3)	387

Total params: 16,887,491 (64.42 MB)
Trainable params: 16,887,491 (64.42 MB)
Non-trainable params: 0 (0.00 B)

The Convolutional Neural Network (CNN) used in this project consists of multiple convolutional, pooling, and fully connected layers. The input images are resized to 256x256 pixels with three channels (RGB). The model begins with a Conv2D layer containing 32 filters with a kernel size of 3x3, followed by a MaxPooling2D layer to reduce spatial dimensions. Two additional convolutional layers are added with 64 and 128 filters, respectively, each followed by pooling operations. To prevent overfitting, Dropout layers with a rate of 30% are applied after the third convolutional layer and the first dense layer.

The fully connected section of the model flattens the feature maps and passes them through two dense layers with 128 neurons and ReLU activation functions. Finally, a softmax output layer with 3 neurons is added to classify the images into one of the three classes: Covid, Normal, or Viral Pneumonia. The model is compiled using the Adam optimizer, with categorical crossentropy as the loss function and accuracy as the evaluation metric. The model is trained for 10 epochs with a batch size of 16, balancing efficiency and performance.

Although we experimented with autoencoders for feature extraction and dimensionality reduction, their effectiveness was limited due to the relatively clean and small dataset. Ultimately, the CNN-based architecture provided the best trade-off between complexity and performance, making it the most suitable choice for this task.

Ethical Considerations:

In medical image analysis, several ethical considerations must be addressed to ensure the responsible use of neural networks. Data privacy is a primary concern, as medical images often contain sensitive patient information. Compliance with data protection regulations like GDPR and HIPAA is essential to ensure anonymity and confidentiality. Removing identifying markers and metadata from the images is a necessary step to protect patient privacy.

Another critical ethical issue is bias and fairness. A model trained on imbalanced or biased datasets may underperform for underrepresented patient groups, leading to misdiagnosis. For example, if the dataset contains more images for Covid cases than Viral Pneumonia, the model may struggle to detect the minority class accurately. Additionally, deploying a model with poor generalization can result in incorrect classifications, potentially delaying life-saving treatments. Ensuring transparency in predictions is equally important, as clinicians must trust the decisions made by AI systems.

Biases:

Bias in training data can significantly impact the performance and fairness of a medical neural network. In this project, the class distribution is imbalanced, with Covid images being more frequent compared to the Normal and Viral Pneumonia classes. This imbalance can cause the model to overfit to the majority class (Covid), leading to poorer recall for minority classes such as Viral Pneumonia. For instance, the confusion matrix highlights that Viral Pneumonia images are misclassified more often, indicating that the model struggles to correctly identify this underrepresented class.

The imbalance can have serious implications for medical diagnostics. If a model is biased toward detecting Covid, it might fail to identify other critical conditions like Viral Pneumonia, delaying appropriate treatment. This issue is compounded when the dataset lacks diversity, as the model may not generalize well across different demographic groups or imaging devices. Over-reliance on biased models could lead to inaccurate predictions, diminishing trust among clinicians and harming patient outcomes. Addressing this imbalance is critical to ensure that all conditions are equally and reliably detected.

Proposed Solutions to Mitigate Bias

To mitigate the effects of bias, several strategies are proposed. First, data augmentation can be used to artificially increase the representation of minority classes by applying transformations such as rotation, flipping, scaling, and zooming. This technique helps improve generalization and prevents overfitting to dominant classes. Additionally, using class weights, as implemented in this project, ensures that the model pays equal attention to all classes during training. By assigning higher weights to minority classes, the network can improve recall for underrepresented conditions.

For further improvement, ensemble models can be employed, combining predictions from multiple architectures such as ResNet or VGG to increase robustness and reduce errors. Incorporating Explainable AI (XAI) techniques like Grad-CAM can help visualize which regions of the image influence the model's predictions, providing transparency and building trust among medical professionals.

A critical step for improving model generalization is expanding the dataset to include more diverse and representative samples. Collaborating with medical professionals and institutions can facilitate access to a wider range of data, including images from different demographics, imaging devices, and clinical settings. This will help address dataset biases and ensure that the model performs well across all patient groups. Additionally, regular audits and performance evaluations on these diverse datasets are essential to identify and mitigate biases that may arise over time.

Results

Overall these two models provided similar results, however there are differences between the specific accuracies of the models.

The first model, the model including an autoencoder, gave solid results that most certainly provided good insight into the performance of the model and saw reliable outputs. The scores of the precision, recall, F1 Score, and support are as follows:

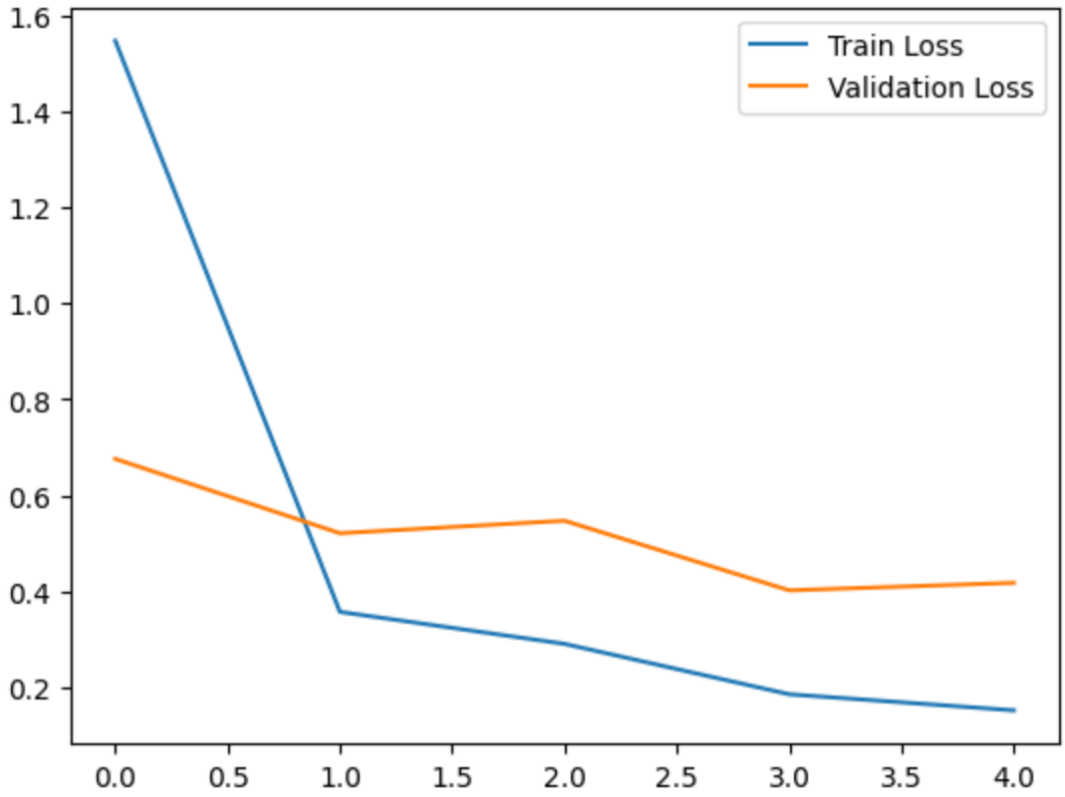
	Precision	Recall	F1 Score	Support
Covid	0.83	0.96	0.89	26
Normal	0.84	0.80	0.82	20
Viral Pneumonia	0.82	0.70	0.76	20

Accuracy	N/A	N/A	0.83	66
Macro Average	0.83	0.82	0.82	66
Weighted Average	0.83	0.83	0.83	66

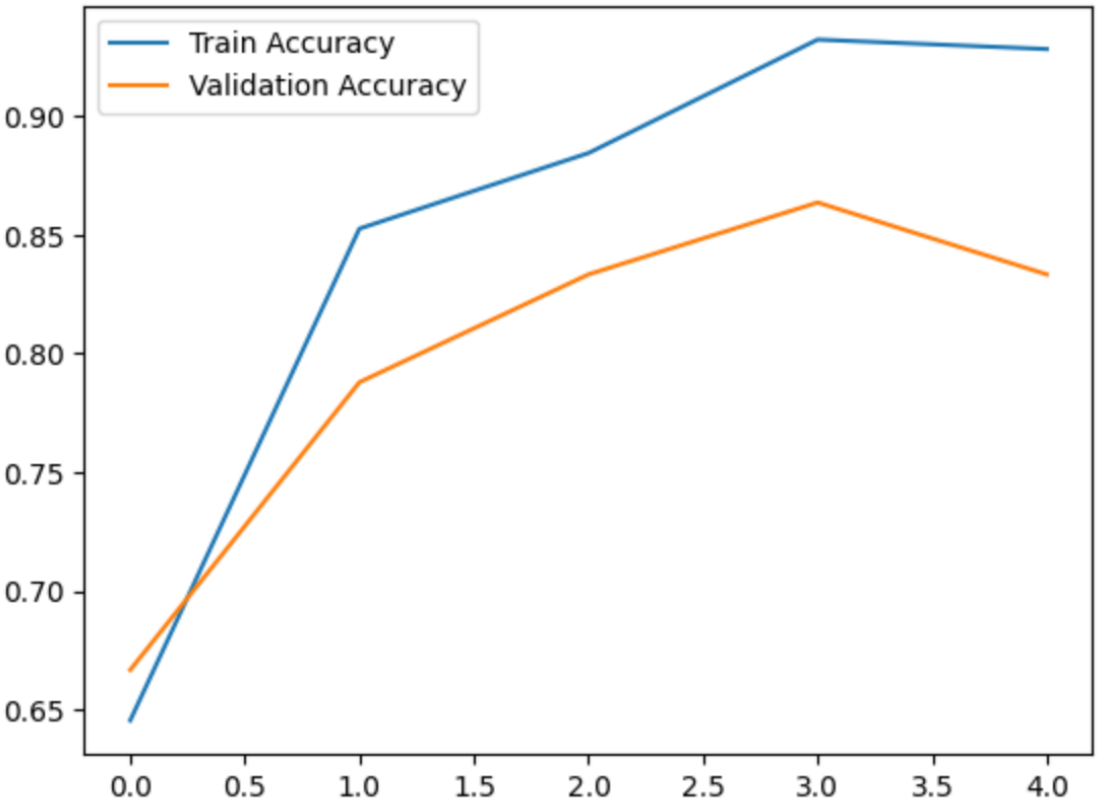
This table displays the values from the various methods of analysis and the multiple averages of the scores collected. This accuracy is consistent across the board, and while it may not be exceedingly high, it is consistent across the board and shows a lack of overfitting and ease of application. The general consistency also shows a uniformity in the ability of the neural network to process a range of images. The use of an autoencoder means there is a chance that some data could be skewed due to poor image denoising. However, this is likely a small issue based on the generally high accuracy. The only large deviances are the lower scores of viral pneumonia according to the Recall and F1 Scores. These outliers are likely due to Pneumonia being slightly more difficult to identify simply from images.

The following are graphs that visualize the training results.

Firstly, train loss and validation loss compared:

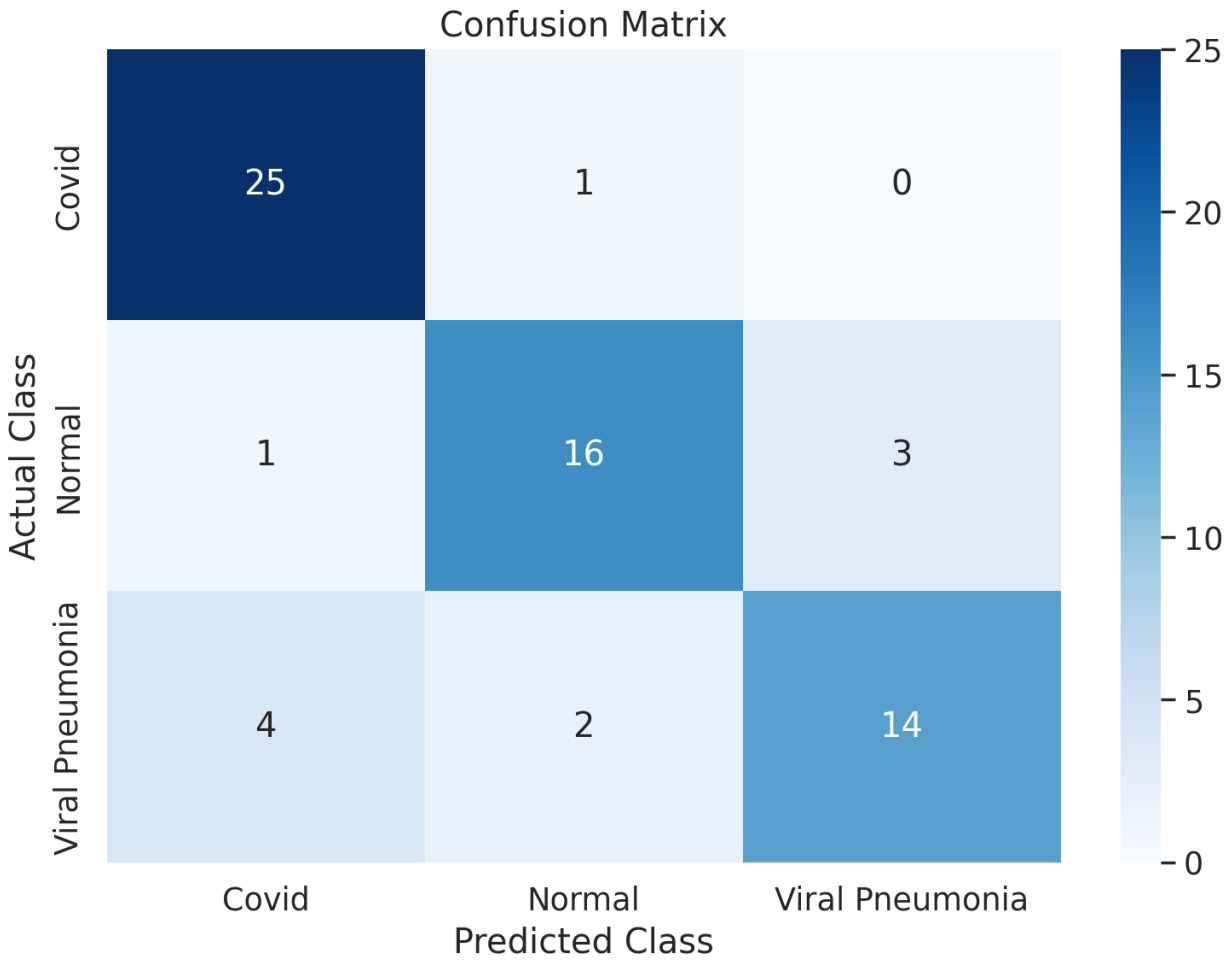

Train Loss
Validation Loss
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

And secondly, train accuracy and validation accuracy compared:



These graphs display a strong correlation between the two accuracies, however there are some easily identified differences. Firstly, the validation loss is much more horizontal and shows much less variance when compared to Train loss, however, it still improves regardless. This pattern is not as evident in the second graph when comparing Train accuracy and validation Accuracy. These slight differences likely arise due to the fact that validation accuracy is measured with a greater amount of updates when compared to train accuracy. These extra measurements can give the data a slightly different curve and adjust the overall look of the graph, however as is shown in the graphs this change is nothing entirely major.

Below is a confusion matrix of the first model (with autoencoder). This This shows how accurate our predictions were to the true values for extra context.

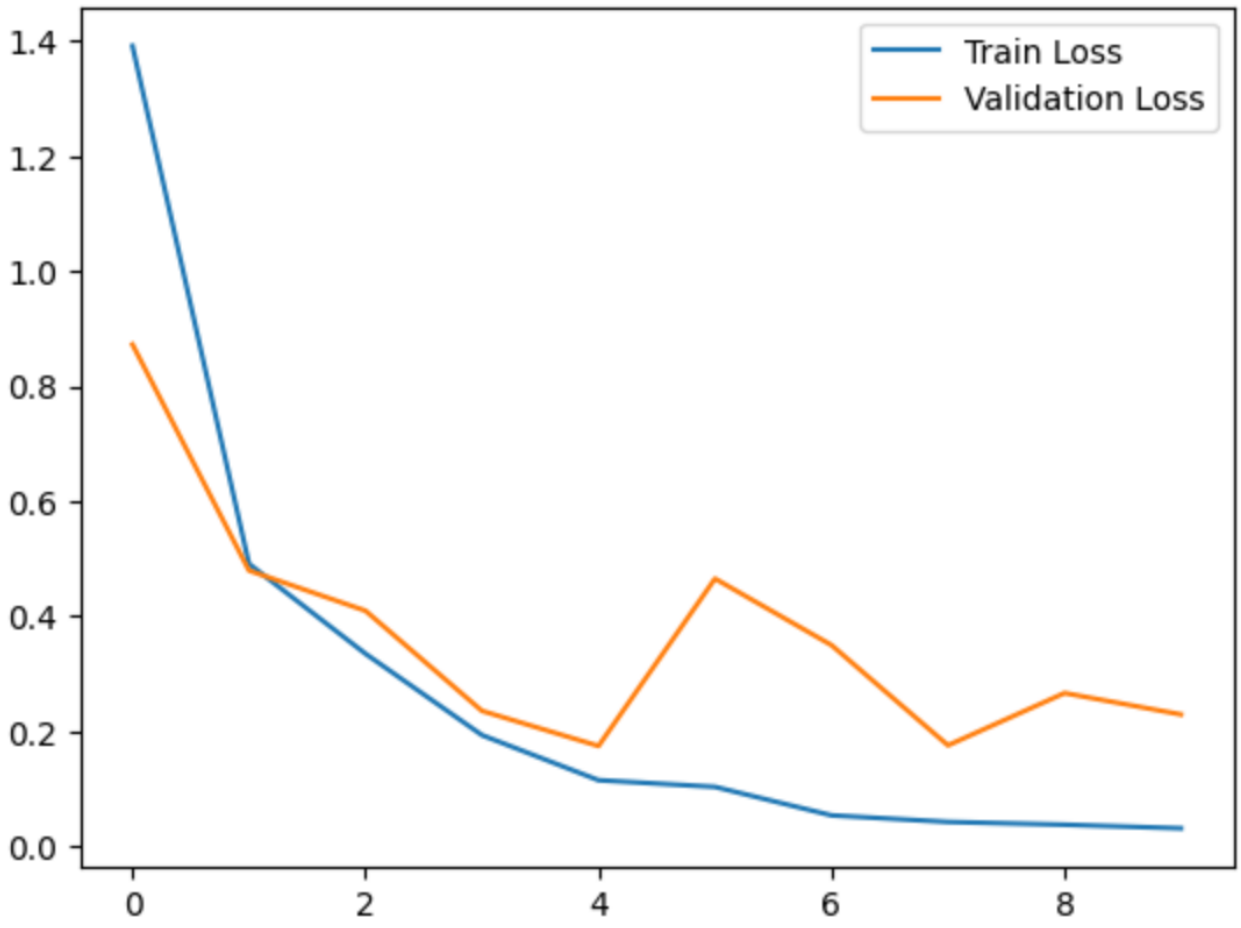


The second model is somewhat different from the first. Despite being similar at heart, there are some underlying key differences. Firstly, there is no autoencoder. While this might sound counterintuitive in most cases, in this specific use case with this particular dataset it made sense because the data is already relatively clean to begin with, therefore, adding an autoencoder may introduce more overhead and increase complexity without improving metrics. This turned out to be true, hence the inclusion in this paper. In addition to this change, the classes were modified. In this case, the number of covid samples outnumbered the amount of normal or pneumonia cases, leading to an imbalance. To counteract this, weights were introduced. These helped to balance the inputs back out and give some variance to the model.

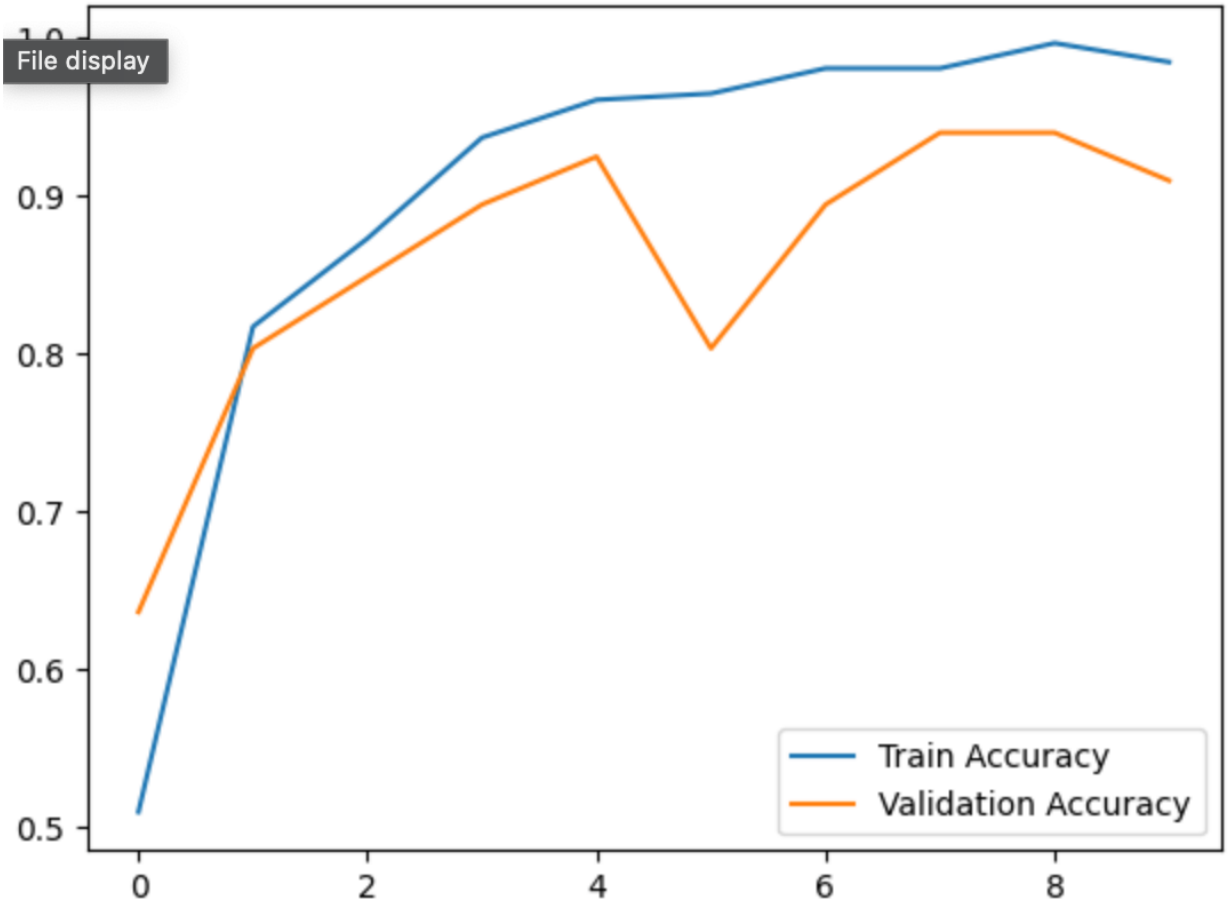
Here is a table of some of the metrics we observed with this model:

	Precision	Recall	F1 Score	Support
Covid	1.00	0.92	0.96	26
Normal	0.90	0.90	0.90	20
Viral Pneumonia	0.82	0.90	0.86	20
Accuracy	N/A	N/A	0.91	66
Macro Average	0.91	0.91	0.91	66
Weighted Average	0.91	0.91	0.91	66

Here is the training loss and validation loss graph:

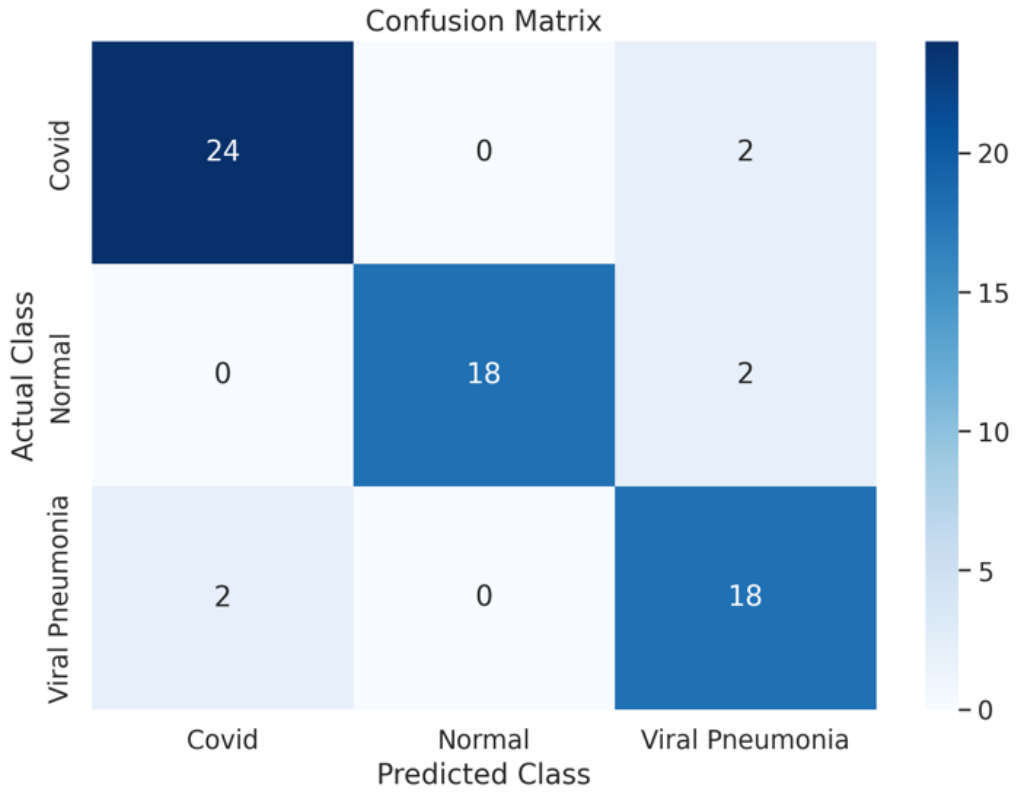


Here is the training Accuracy and validation accuracy graph:



These graphs are relatively unchanged from the previous ones however there is a key difference. The Accuracy is much better overall, however it has a strange jump in the validation portion. Overall this proved not to matter too much as the metrics showed large increases across the board, signifying that this was the better of the two models.

Finally, this Is the confusion matrix for this function:



Conclusions

Overall, the model was a great success, not only showing improvements from one revision to another but most importantly correctly identifying all cases in the lungs and proving to be successful and versatile across the board. In addition to this, we were able to create an efficient model that performs well while taking less resources.

Ethically there are some considerations we made. This dataset is public knowledge and is not infringing on anyone's rights and as such is not in violation of any ethical code.

To conclude, we created a model that successfully analyzes and identifies lungs infected with covid, pneumonia, and ones that are healthy, all while remaining ethical and non taxing to computer hardware with significant accuracy and high metric scores. This is a great success and displays how powerful Neural Networks truly can be when optimized and used properly.

References

https://www.kaggle.com/datasets/pranavraikokte/covid19-image-dataset