Segmentation of lung CT scans to identify COVID-19 infection using U-Net CNN architecture

Mackenzie A. Looney
Department of Electrical and Computer Engineering
Duke University
Durham, NC 27708
mackenzie.looney@duke.edu

Abstract

The rapid spread of COVID-19 has led to a variety of tools being developed and utilized for detection, prognosis and management of the disease. One such tool is the use of lung CT scans to identify infection. This may be most beneficial for early detection and disease management. This paper focuses on the design of a convolutional neural network model to predict the location of COVID-19 infection on lung CT scans. The model is developed with and without a physical layer, and both versions of the model are evaluated with an intersection over mean metric. A goal of the project is to allow for reduced radiation exposure by creating a model that could accurately predict the location of COVID-19 infection on images with simulated noise artifacts. Twenty CT scans of COVID-19 patients were used to develop and evaluate the model.

1 Introduction

After first being identified in December of 2019, the novel coronavirus disease was classified as a global pandemic in March, 2020. The disease is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Transmission of COVID-19 was rapid, as it can be spread through direct contact, human-to-human transmission through droplets, and indirect contact, as an airborne illness [1]. Over the course of the pandemic, COVID-19 has spread to over 210 countries and territories [2], with over 500 million confirmed cases at the time of this paper [3].

Reverse transcriptase polymerase chain reaction (RT-PCR, or PCR) is the leading method of COVID-19 diagnosis. PT-PCR tests work by putting a sample in contact with a primer that matches segments of COVID-19 genetic material, targeting specific genes such as the ribonucleic acid (RNA)-dependent RNA polymerase, and envelope genes, or the open reading frame 1b and nucleoside regions. Tests have been developed to look for the genetic markers of the disease in mucus, saliva, feces, and blood [4]. With the nature of COVID-19, the need for high sensitivity is much larger than the need for high specificity in test, meaning that it would be better for a test to have more false positives than false negatives [5]. Studies have shown that the average specificity of COVID-19 PCR tests is 98 percent, while the sensitivity of those tests is ranges from 50 percent to 80 percent [2]. Other studies found that PCR sensitivity is much higher, with the range in result coming from the variety of PCR tests available, the differences in the portions of genome being replicated by each, and the times at which tests were taken [4]. As such, secondary verification of negative PCR results may be beneficial for symptomatic patients.

One method that may be used to verify PCR results is computed tomography (CT). CT may offer higher sensitivity than PCR, with studies reporting sensitivity between 80 and 90 percent and specificity between 92.8 and 96 percent [2]. In a study of 1014 patients who underwent both PCR test and CT scan, 888 showed abnormalities in the CT consistent with COVID-19 infection while only 601 initially had positive results from the PCR test. Seventy-five percent of the 413 patients with

negative PCR test results had positive CT scans. Within six days, between 60 percent and 93 percent of the patients with abnormalities in the CT returned positive PCR results. Twenty-four patients showed improvement on CT before PCR test results became negative [6]. This led researchers to believe CT was a good indicator of COVID-19 infection.

Particularly in countries with limited access to PCR tests and limited resources for laboratories, CT scans may be an option to supplement. One country where this is the case is Ghana, where there are many CT machines but low access to PCR testing. In areas such as this, it has been shown that CT scans may allow for rapid diagnosis of COVID-19 [2]. In other countries, with more readily available PCR testing, CT scans may still aid in the prognosis and management of COVID-19 [7].

1.1 Computed Tomography

The machines used for CT scans consist primarily of a large tube called a gantry, and a bed for the patient to lie on. Continuous gantry rotation allows for narrow x-ray beams to shoot through the patient's body from 360 degrees of emitter location. At the same time as the gantry rotates, the bed passes through the machine, allowing for scans of various body parts to take place as quickly as possible, which minimizes motion blur [8]. Unlike traditional x-ray, which uses film, CT utilizes a digital receiver that is located opposite the x-ray emitter. This allows for scan data to be passed directly to a computer, which takes the raw form of the data, a sinogram, and generates slice images. These slices show cross-sectional 2D images of the body [9]. Through the use of an array of emitters and detectors, three-dimensional visualization is also possible, making CT useful for a wide variety of diagnostic and prognostic tasks [8].

For the use in COVID-19 diagnosis, prognosis and management, CT scans specifically look for COVID-19 pneumonia in the lungs, which is characterized by "peripheral, bilateral, or multifocal rounded foci of ground-glass opacity (GGO)", along with "airspace consolidation, with or without an associated crazy-paving pattern and a reverse halo abnormality" [7]. Scans that show atypical lung appearances are not expected for COVID-19 pneumonia, and would lead providers to a different diagnosis such as bacterial pneumonia [7].

1.2 Radiation Exposure

As CT imaging utilizes x-ray, it involves radiation exposure which may pose health risks to patients. It is estimated that around two percent of cancers in the United States are related to CT imaging [10]. To better understand the risk that radiation exposure poses, the following section will be a brief literature review on radiation exposure, dosage, and the relationship to cancer risk.

In medical imaging, dose equivalence is measured in units of sieverts (Sv), rather than the more standard grays (Gy) used for absorbed dose. This is because different forms of radiation have different biological effects on tissue. Sieverts are calculated as the product between absorbed dose and the radiation weighting factor. The radiation weighting factor used for x-rays is 1.0, thus for CT imaging radiation dose and dose equivalence are equivalent [10].

Annually, it is estimated that people are exposed to 3 mSv in their homes. This primarily occurs due to radon gas. From studies on survivors of atomic bombs in Japan, it has been shown that exposure to over 100 mSv poses a large risk of causing cancer, in particular thyroid cancer. The medical risk associated with radiation exposure between 10 and 100 mSv is less documented. The average CT results in a radiation dose of 10 mSv. For patients that require more than one CT study in a year, the risk associated with radiation exposure due to CT becomes relevant. However, in many cases, reducing CT radiation exposure results in noise artifacts on the CT image. For this reason, looking for opportunities to reduce radiation exposure while maintaining the benefits of the CT for medical imaging becomes important [10].

2 Methodology

To evaluate the effects that a simulated physical layer has on the model, this project was done in two steps. In the first, a convolutional neural network (CNN) model was created to train and test on a set of CT scans. In the second, a physical layer was added to the CNN before a set of CTs scan with

simulated noise were input to the model. These steps were then repeated with the input of a scan with a 0.9 signal to noise ratio. This allows for a comparison between the predicted masks for each case.

2.1 The COVID-19 segmentation data

The data set used throughout this project consisted over twenty CT scans obtained in 2020 [11], [12] and corresponding annotation masks made by experts to highlight the location of COVID-19 pneumonia infection in the lungs [13]. Each CT scan in the data set contains between 39 and 418 slices. In total there are over 3,000 available slices, and matching annotation masks, in the data set. Fig. 1 shows an example slice from the CT scan database, on the left, and the corresponding expert annotation mask overlaid on that slice, on the right.

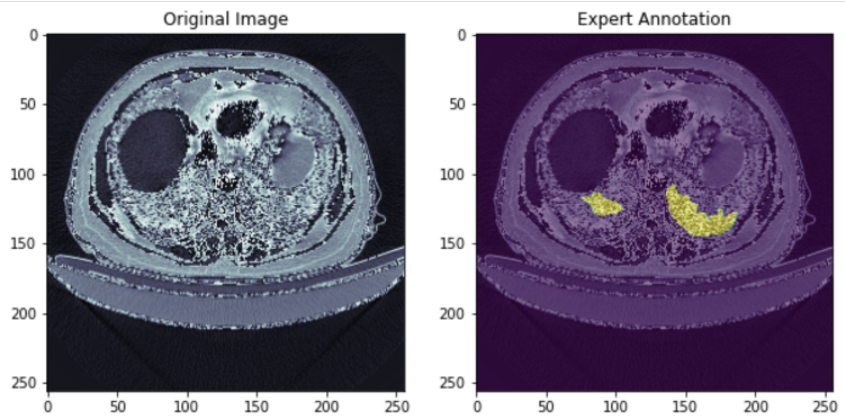

Figure 1: Example CT slice and expert annotation

To process this large data set the Python library NiBabel, which has been designed to give easy access to a variety of medical imaging file formats [14], was utilized to access the CT data as Numpy arrays [15]. As the images varied in size, they were all sized-down to 256 by 256 pixels. When randomly splitting the data into training and testing subsets, ten percent of the data was used for testing and ten percent of the data was used for validation.

2.2 U-Net CNN

As a CNN architecture, U-Net was designed to be used for the segmentation of biomedical images [16]. It consists of two symmetric paths connected together by skip connections that copy the output of a portion of the first path to the input of the corresponding portion on the second path. The first path, called the encoding or compressing path, using a series of convolutions, compresses the spatial features of an image into learned filters. The second path, called the decoding or decompressing path, using a symmetrical a series of convolutions, decompresses those learned filters so that they can be represented in the original spatial dimensions [17].

For this project the compressing path of the U-Net architecture used four layers. Each layer utilized two 3 by 3 two-dimensional convolutions with ReLu activation, followed by batch normalization, which aims to re-center and re-scale features to create more robust CNNs. After which the results are saved to later be copied to the corresponding level in the decompression path. Finally, the result is max-pooled. In the last layer of the compression path, drop out is also performed on the result to help with over fitting of the model. At the bottleneck, two 3 by 3 convolutions are performed, with the same ReLu activation, followed by drop out. On the decompression path there are once again four layers. The first step in each layer is up sampling to a higher spatial dimension followed by adding in the saved result of the compression path. Then two 3 by 3 convolutions are performed followed by batch normalization. Finally, to find the output of the model a final 1 by 1 convolution is done with Sigmoid activation. Fig. 2 shows a visualization of these steps.

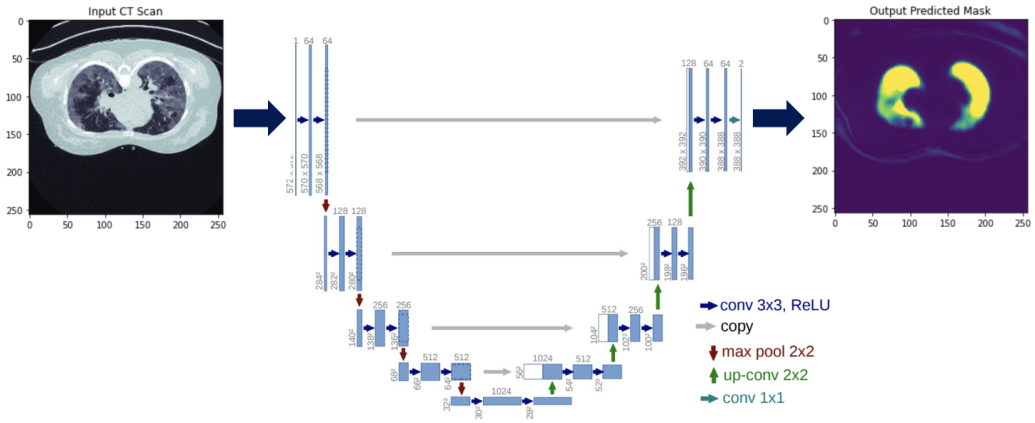

Figure 2: Visualization of U-Net CNN architecture [17]

2.3 Physical layer

With the goal of simulating noise produced by a reduction in radiation dosage, noise was added to the CT images with various values for signal to noise ratio (SNR). The SNRs chosen were 0.1, 0.3, 0.5, 0.7, and 0.9. To obtain this desired level of noise a function was written to randomly select a portion of total pixels, as determined by the SNR, and change them to 0. This method allowed for the level of noise to be determined, where as many functions to add noise do not allow for an SNR input. The five levels of noise, which can be seen in Fig. 3, were created for each training and testing image. After creation they were put into the CNN. The first layer of the CNN is the physical layer, which trained a weighted sum of the noise. This was done through the creation of a custom TensorFlow Keras layer that utilized initialized weights with 0 mean and 0.05 standard deviation, and a 1 by 1 convolution with 1 filter, and trainable weights [18]. Fig. 4 shows the output of this layer compared to the original image.

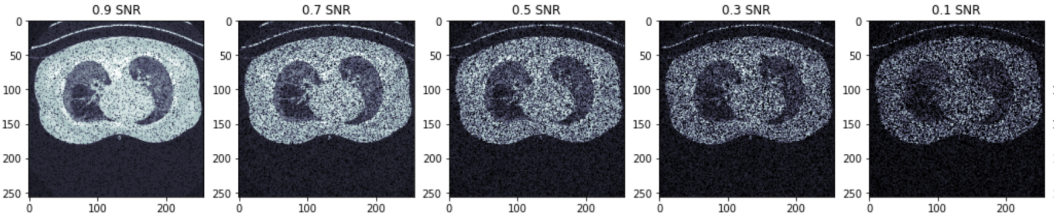

Figure 3: The five levels of noise, from 0.9 SNR to 0.1 SNR

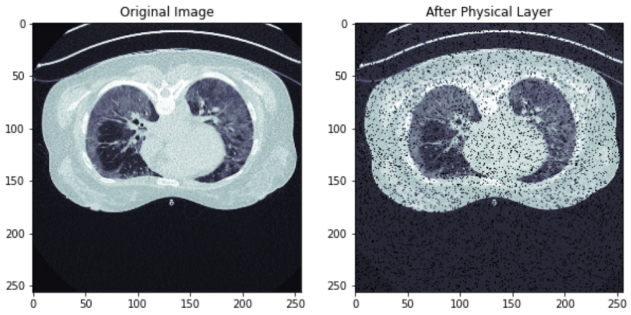

Figure 4: Original CT scan compared to output of physical layer

For the second run performed on a CT scan with 0.9 SNR, the physical layer was modified to use a sum of five simulated noisy images with SNR of 0.85, 0.8, 0.75, 0.7 and 0.65.

2.4 Metrics

To measure the performance of the models, both with and without the physical layer, several metrics were used. The first metric is the intersection over union (IOU). IOU can be understood as the number of pixels that can be found in both the ground truth mask and the predicted mask (the intersection of the two masks) over the number of pixels that are in the either the ground truth mask or predicted mask (the union of the two masks). To create a graph to show this metric a series of thresholds were defined, from zero to one with a step size of 0.05. For each threshold, the number of pixels that were classified as true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) were found using the SciKit learn confusion matrix function [19]. True positives are the pixels that are identified as infected on the ground truth mask and the predicted mask. True negatives are the pixels that are identified as not infected on the ground truth mask and the predicted mask. False positives are the pixels that are identified as not infected on the ground truth mask but are identified as infected on the predicted mask. False negatives are the pixels that are identified as infected on the ground truth mask but identified as not infected on the predicted mask. Using these values, IOU can be calculated using the equation:

$$IOU = TP/(TP + TN + FN)$$

Receiver-operator curve (ROC) was used as a second metric for measuring the performance of the model. ROC is shown as the true positive rate over the false positive rate. The true positive rate is 1-specificity. The equation for specificity is $TN/(TN + TP)$. The false positive rate is sensitivity, which can be found with $TP/(TP + FN)$. By finding these values for each of the thresholds, one can plot true positive rate against false positive rate to obtain an ROC.

The final metric used to evaluate model performance was a precision-recall curve (PRC). Using the same thresholds as above, one can calculate the precision with $TP/(TP + FP)$. Recall is equivalent to sensitivity, for which the equation has already been given. These values can be plotted against one-another to obtain the PRC for the model.

To evaluate the training process and the number of epochs used for model fitting, plots were made for the training loss, validation loss, training accuracy, and validation accuracy, according to TensorFlow model metrics. The loss is defined as the summation of errors made in each subset of data. Accuracy is the percentage of correct decisions that the model makes during model fitting. Neither loss nor accuracy are considered good metrics for segmentation, thus these plots only serve to show that an adequate number of epochs were used for model fitting.

3 Results

3.1 Clean Run

Qualitatively, the results of the CNN with and without the physical layer were very similar. Fig. 5 shows the original scan on the left, the expert annotated mask in the center, and the predicted infection on the right. The top row is the CNN without the physical layer and the bottom row is the CNN with the physical layer. The physical layer is only applicable to the predicted infection. Looking carefully there are small differences between the two predicted infections, with the most noticeable difference being a light blue false positive region on the left lung of the image without the physical layer.

Quantitatively, the results of the CNN with and without the physical layer are displayed on graphs in Fig. 6. These graphs show the metrics that were previously described in the section 2.4. The results of the CNN without the physical layer are on the top row. The results of the CNN with the physical layer are on the bottom row.

Looking first at the IOU vs. Thresholding Value graphs on the right, the peak value of the CNN without the physical layer is about 0.6 at a threshold of 0.2 and decreases steadily as threshold values increase. With the physical layer the peak IOU value is close to 0.7 at a threshold of 0.5. It does not decrease as dramatically as threshold values increase. This suggests that the confidence of the prediction was higher for the CNN with the physical layer.

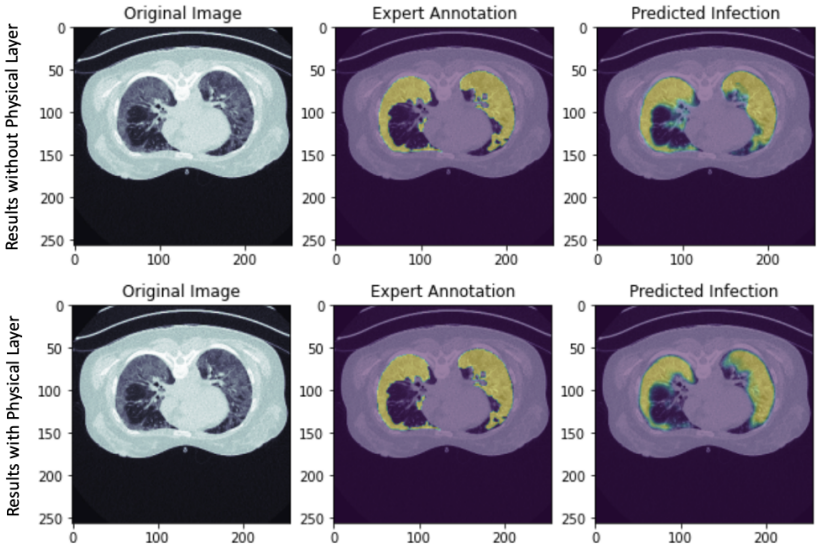

Figure 5: Mask results of CNN with and without physical layer

The ROC graphs in the middle show a similar result. Ideally, an ROC reaches a value of 1.0 true positive rate at a value of 0.0 false positive rate, and then maintains the 1.0 true positive rate as false positive rate increase to 1.0. Both ROCs are nearly ideal, however, the ROC produced by the CNN with the physical layer is more ideal than the ROC produced by the CNN without the physical layer.

The PRC graphs on the right are not as conclusive. Ideally, a PRC mirrors the ideal ROC, precision starts at 1.0 when sensitivity is 0.0 and remains at 1.0 until sensitivity is 1.0, at which point precision drops to 0.0. For the CNN without the physical layer, precision starts slightly lower than 1.0, and drops steadily until sensitivity reaches 0.9, at which point precision drops more rapidly to 0.0. For the CNN with the physical layer, precision starts at a similar level, decreasing steadily until sensitivity passes 0.95, at which point sensitivity rapidly drops to 0.0.

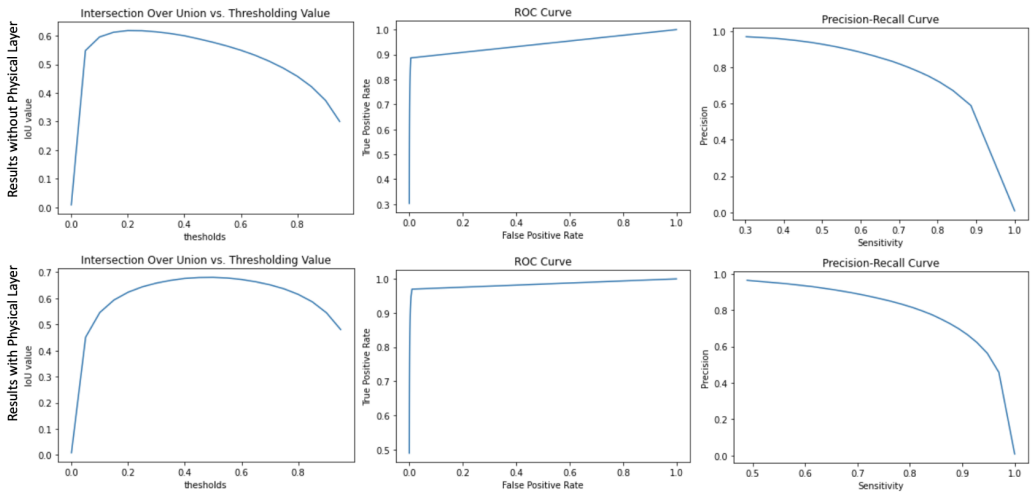

Figure 6: Graphical results of CNN with and without physical layer

3.2 Noise Run

When the models were run on a set of images with a SNR of 0.9, the predictions were not as accurate, but were, again, improved with the physical layer.

In the image comparison, the results of the CNN with and without the physical layer were still similar. Fig. 7 shows the original scan on the left, the expert annotated mask in the center, and the predicted infection on the right. The top row is the CNN without the physical layer and the bottom row is the CNN with the physical layer. There are small differences between the two predicted infections. Specifically, there are false positives on the right lung in the CNN without physical layer and false positives on the left lung in both the CNN with and without the physical layer. The widths of the predicted lesions also differ between models. It is difficult to determine from these images which model is more accurate.

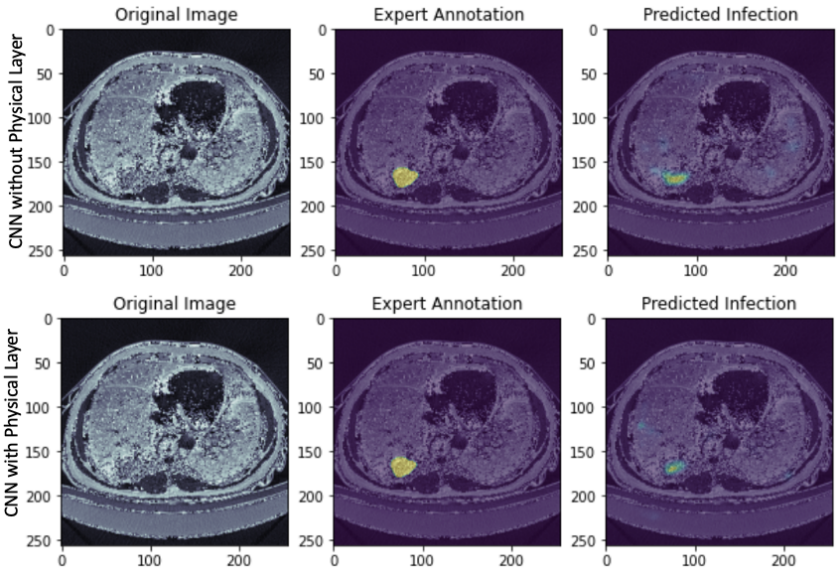

Figure 7: Mask results of CNN with and without physical layer for noisy CT input

Quantitatively, the results of the CNN with and without the physical layer are displayed on graphs in Fig. 8. The results of the CNN without the physical layer are on the top row. The results of the CNN with the physical layer are on the bottom row.

Looking first at the IOU vs. Thresholding Value graphs on the right, the peak value of the CNN without the physical layer is slightly greater than 0.2 at a threshold of approximately 0.3, and decreases rapidly as threshold values increase. With the physical layer the peak IOU value is close to 0.35 at a threshold of 0.15. It does not decrease as dramatically as threshold values increase, until 0.95, at which point it drops rapidly.

The ROC graphs in the middle show a marked improvement with the physical layer. The ROC produced by the CNN with the physical layer is more ideal than the ROC produced by the CNN without the physical layer.

For the CNN without the physical layer, precision starts slightly above 0.5, and drops rapidly until sensitivity reaches approximately 0.5, at which point precision drops more steadily to 0.0. This curve is worse than random selection. For the CNN with the physical layer, precision starts at approximately 0.9, decreasing steadily. This follows the line of random selection, for the most part, but that is an improvement over the CNN without the physical layer.

3.3 Convergence Verification

To verify that both of the models had been given enough epochs during model fitting, accuracy and loss graphs were produced. These graphs can be seen in Fig. 9. The graphs that correspond to the

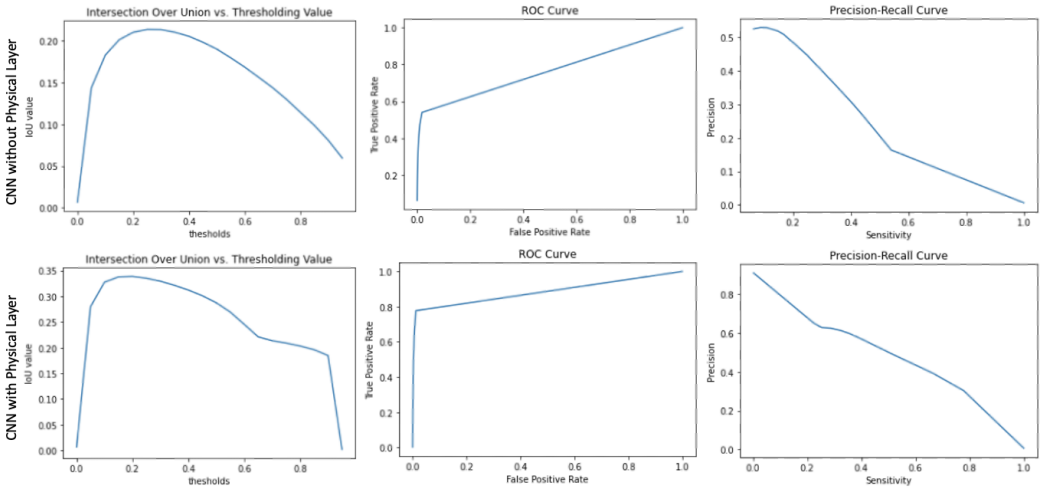

Figure 8: Graphical results of CNN with and without physical layer for noisy CT input

CNN without the physical layer are on the left and the graphs corresponding to the CNN with the physical layer are on the right.

In the model accuracy graphs, for both models, training and testing accuracy converge near 1.0 at the end of 30 epochs. Similarly, in the model loss graphs, for both models, training and testing loss converge near 0.0 at the end of 30 epochs. This shows that both models were given adequate epochs for model fitting.

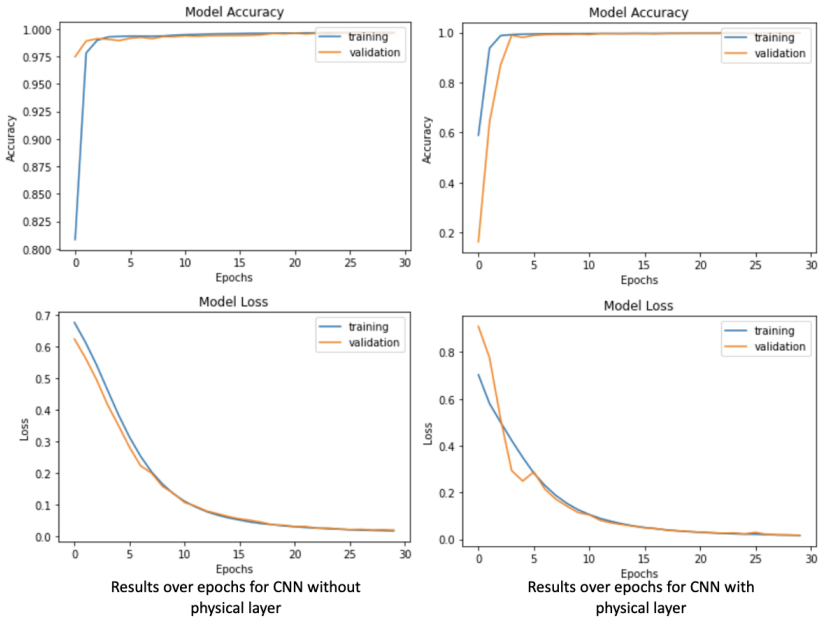

Figure 9: Results of CNN with and without physical layer

4 Discussion and Conclusion

The goal of this project was to find if radiation exposure could be reduced while maintaining the ability of predictive segmentation. Based on the results, it is clear that the models made in this project

did not succeed in that goal. Future experimentation with different models, and different physical layers, may lead to a better result. Rather than seeking to add noise on the images, it may be better to experiment with optimizing sampling frequencies.

Surprisingly, this project did find that using a weighted sum of noise as the physical layer improved the accuracy of the model. This was true for both the clean run and the run with inputted noise. Future experimentation may indicate why this is the case by comparing this proposed physical model to others.

Acknowledgments

I would like to thank Dr. Roarke Horstmeyer for his advice and guidance throughout this project. Additionally, thank you to Kanghyum Kim and Shiqi Xu for their help on understanding and implementing U-Net architecture.

References

[1] M. Lofti, M.R. Hamblin and N. Rezaei, "COVID-19: Transmission, prevention, and potential therapeutic opportunities," *Clin Chim Acta*, vol. 508, pp. 254–266, 2020.

[2] B.D. Sarkodie, and Y.B. Mensah, "CT scan chest findings in symptomatic COVID-19 patients: A reliable alternative for diagnosis," *Ghana Medical Journal*, vol. 54, no. 4, pp. 97–99, 2020.

[3] World Health Organization, "WHO Coronavirus (COVID-19) Dashboard," *World Health Organization*, 2022. [Online]. Available: https://covid19.who.int. [Accessed: Apr. 28, 2022].

[4] D. Shyu, et al., "Labratory Tests for COVID-19: A Review of Peer-Reviewed Publications and Implications for Clinical Use," *Missouri Medicine*, vol. 117, no. 3, pp. 184–195, 2020.

[5] A. Swift, R. Heale and A. Twycross, "What are sensitivity and specificity?," *Evidence Based Nursing*, vol. 23, pp. 2–4, 2020.

[6] T. Ai, et al., "Correlation of Chest CT and RT-PCR Testing for Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases," *Radiology*, vol. 296, no. 2, pp. E32–E40, 2020.

[7] S. Machnicki, et al., "The usefulness of chest CT imaging in patients with suspected or diagnosed COVID-19," *Chest Journal*, vol. 160, no. 2, pp. 652–670, 2021.

[8] G.D. Rubin, "Computed Tomography: Revolutionizing the Practice of Medicine for 40 Years," *Radiology*, vol. 273, no. 25, pp. S45–S74, 2014.

[9] National Institute of Biomedical Imaging and Bioengineering, "Computed Tomography (CT)," *National Institutes of Health: National Institute of Biomedical Imaging and Bioengineering*, 2022. [Online]. Available: https://www.nibib.nih.gov/science-education/science-topics/computed-tomography-ct. [Accessed: Apr. 28, 2022].

[10] E. C. Lin, "Radiation risk from medical imaging," *Mayo Clinic Proceedings*, vol. 85, no. 12, pp. 1142–1146, 2010.

[11] O. Paiva, *Helping Radiologists to Help People in More Than 100 Countries! Coronavirus Cases* (2020). Distributed by Coronacases.org. Accessed: April 4, 2022. [Online]. Available: https://coronacases.org

[12] Y. Glick, *COVID-19 pneumonia*. (2020). Distributed by Radiopaedia.org. Accessed: April 4, 2022. [Online]. Available: https://radiopaedia.org/playlists/25887

[13] M. Jun, et al., *COVID-19 CT Lung and Infection Segmentation Dataset*. (April 20, 2020). Distributed by Zenodo. Accessed: April 4, 2022. doi: 10.5281/zenodo.3757476.

[14] M. Brett, et al., *nipy/nibabel: 3.2.1*. (November 28, 2020). Distributed by Zenodo. Accessed: April 4, 2022. doi: 10.5281/zenodo.4295521.

[15] T. E. Oliphant, *Guide to numpy*, 1st ed. 2006. https://numpy.org/

[16] P. Kalane, "Automatic detection of COVID-19 disease using U-Net architecture based fully convolutional network," *Biomed Signal Process Control*, vol. 67, 2021.

[17] R. Horstmeyer, BME 548L, Class Lecture, Topic: "Lecture 15: Beyond classification – segmentation and autoencoders." Duke University, Durham, NC, Mar. 14, 2022.

[18] M. Abadi, et al., "TensorFlow: Large-scale machine learning on heterogeneous systems", *Google Research*, Nov. 9, 2015. https://www.tensorflow.org

[19] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011. https://scikit-learn.org/