# Predicting Severity of COVID-19 from CT Scans

**Toral Shah** TORAL.SHAH1@NYULANGONE.ORG
*Department of Biomedical Informatics*
*NYU Langone Health*
*New York, NY, USA*

**Peter Hsu** PETER.HSU@NYULANGONE.ORG
*Department of Radiology*
*NYU Grossman School of Medicine*
*New York, NY, USA*

**Editor:**

## Abstract

The recent use of CT scans for detecting COVID-19 has shown promising results by providing more detailed diagnoses for patients. However, manual review by radiologists can be prone to bias which can result in inconsistent readings or disagreement between experts. Here, we propose a deep learning approach to provide consistent readings of CT scans to detect the presence and severity of COVID-19. We base our model architecture on the well-established ResNet and achieve an AUROC of 0.678 and 0.751 for detecting the presence and severity of COVID-19, respectively. With this framework, we believe that a more robust test result can be achieved than the gold-standard RT-PCR which can help hospitals more adequately allocate their resources in the event of another surge of COVID-19 cases.

**Keywords:** Deep Learning, COVID-19, Computed Tomography

## 1. Purpose

The purpose of our study was to build an automated method to predict both the presence and severe outcome of COVID-19 based on CT scans from the STOIC2021 challenge dataset. This dataset contains openly available CT scans from 2000 individuals suspected of being infected with SARS-CoV-2 during the first wave of the pandemic. We propose that a deep learning system trained and evaluated with this dataset will be able to accurately identify individuals who are likely to require additional hospital resources based on their CT scans.

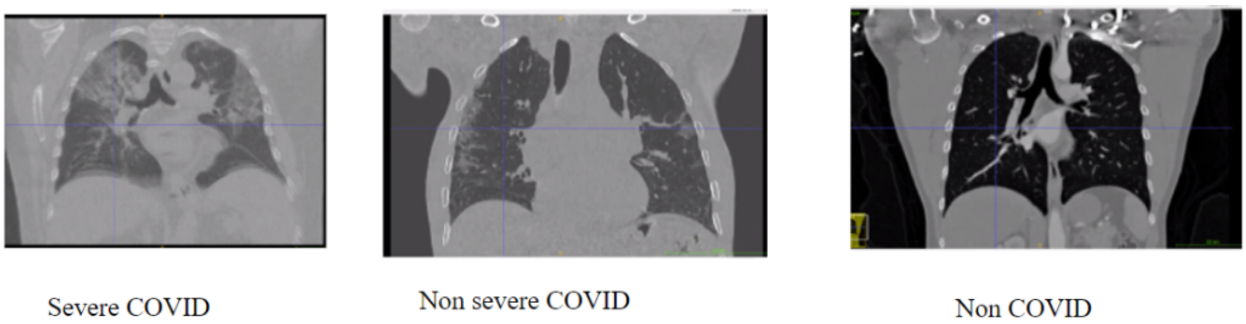
## 2. Introduction

The COVID-19 pandemic has been responsible for 1.6 million deaths worldwide by the end of 2020 and has overwhelmed healthcare resources (Simonsen and Viboud, 2021). The virus named SARS-CoV-2 infects the airway epithelial cells causing varied symptoms ranging from no or few symptoms to acute respiratory distress and death. While the reference standard for testing positivity is RT-PCR, this test does not provide any information on the severity of condition. The use of imaging methods such as Computed Tomography (CT) have been useful in providing visual indications of both the presence and severity of COVID-19 condition (M. Revel et al., 2021). However, to extract this information from CT scans, a review from a radiologist is needed. This manual review can be time-consuming process and is prone to inter-observer and intra-observer biases (N. Sushentsev et al., 2021). We aim to combat these issues with the proposal of an automated deep learning method which can provide both consistent and accurate predictions of COVID-19 presence and severity based on CT scans.

The success of deep learning in the imaging field has been showcased in a variety of settings. In particular, convolutional neural networks (CNNs) have shown the ability to accurately classify medical images of varying conditions (L. Cai and Zhao, 2020). Here, we base our method, STOIC21, on the ResNet architecture which is a subtype of CNN that has been successfully been implemented in the classification of disease (K. He et al., 2015). We trained this model to predict the presence and severity of COVID-19 based on the largest dataset of CT images of COVID-19 suspects and patients collected to date, the STOIC2021 challenge dataset (Boulogne et al., 2021). We hypothesize that this model to has the capability to reliably return a more robust test result than RT-PCR and aid hospitals in allocating their resources for high-risk patients.

## 3. Data

We utilized the STOIC2021 Challenge dataset which is part of the STOIC Project, a multi-center observational study of the diagnostic and prognostic value of CT in COVID-19 (M. Revel et al., 2021). The total dataset contains Computed Tomography (CT) scans from 10,735 patients, but only a fraction is accessible for public use. We used the openly available set of 2000 CT images which was available through Amazon Web Services (AWS). This dataset contains chest CT scans from individuals suspected of being infected with SARS-CoV-2 during the first wave of the pandemic in France from March 2020 to April 2020 (Figure 1).The ground truth for these images is defined as an RT-PCR Test Result and 1-month patient outcome follow up. The labels are broken up as follows: 795 Healthy, 1205 COVID; 1699 Non-Severe, 301 Severe.



Severe COVID

Non severe COVID

Non COVID

Figure 1: CT scans of Severe COVID, Non-Severe COVID, and Non-COVID patients from the STOIC2021 dataset. The first image (left) indicates a need for additional hospital resources for treatment while the second and third (middle and right) images likely would not require additional hospital resources.

## 4. Methods

### 4.1 Data Preprocessing

Images from the STOIC2021 dataset contained pixel values ranging from -2000 to 2000 and varied dimensions ranging from 512x512x124 to 512x512x1199. Images underwent resampling to provide our network with consistent input shape and to significantly speed up computation time. Images were also converted from .MHA files to .NII files for convenience.

The conversion of images from .MHA to .NII file format took place in Python using the SimpleITK library (R. Beare and Yaniv, 2018) (Z. Yaniv and Beare, 2018) (B. C. Lowekamp and Blezek, 2013). The converted NifTi images were resampled to an isotropic 256x256x256 if their dimensions were all greater than 256 using the mrgrid feature of MrTrix3 (Tournier et al., 2019). If images had dimensions that were smaller than 256, they were zero padded to 256 before being used as input to the network to avoid artificial upsampling of image values. In addition, the pixels in the images were rescaled to values from 0-255 to provide more consistent values. Other pixel rescaling techniques such as 0-1 scaling or 0 mean, 1 variance scaling were also tested.

### 4.2 Training, Validation, and Testing Sets

Our dataset suffered from imbalances both for healthy cases vs COVID-19 cases and for non-severe vs severe cases. However, since the class imbalance for non-severe and severe was significantly worse than healthy vs COVID-19, we focused on balancing the split of data for non-severe and severe cases.

To ensure that none of the datasets were lacking in severe COVID-19 cases, the images were first split into non-severe (1699) and severe (301) sets. Each dataset was then randomly split into training and testing sets following an 80:20 split. The resultant training sets were then split again to form training and validation sets which also followed an 80:20 split. The non-severe and severe sets for training, validation, and testing were then combined to form the complete datasets. This resulted in 1279 images for training, 320 images for validation, and 401 images for testing. There is no patient overlap between the sets.

### 4.3 Model Architecture

Our model is based on the Residual Neural Network (ResNet) architecture. ResNets improve flow of information and gradients through a network compared to a traditional CNN through the use of residual connections (K. He et al., 2015). These residual connections add identity mappings into the layers of the model which increase computational efficiency by providing dimensionality reduction of feature maps. This also proposed to be easier to optimize than standard deep model connections.

The original ResNet architecture is modified so that it can take chest CT images as input and return predictions for both the presence and severity of COVID-19 (Figure 2). We changed the first convolutional layer to accept 256 channels as input and the fully connected layer from 1000 to 4 output values. The first two outputs correspond to [0,1] predictions of COVID-19 and the second two outputs correspond to [0,1] predictions of severity.

For a comparison, we also constructed a 3D CNN that is based on the work of T. Kim et al. (2021) with 3 convolutional layers, 3 max pooling layers, and 2 fully connected layers with the number of output features being 4, the same as the modified ResNet models. Each convolutional layer included batch normalization and Exponential Linear Unit activation functions (ELU).

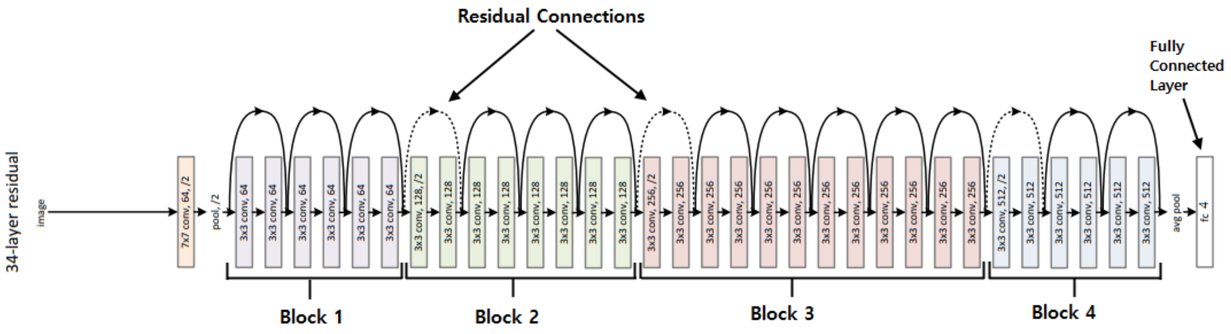


Figure 2: Architecture of a modified ResNet-34 for the classification of COVID diagnosis and Severe COVID diagnosis from CT images. This model consists of 4 Convolutional Blocks and ends in a fully connected layer of size 4 which indicates predictions for presence and severity of COVID-19.

### 4.4 Model Training, Validation, and Testing

To select a final model for our project, we trained and evaluated several versions of ResNet that are initialized with the pretrained weights on the ImageNet dataset. In particular, we compare ResNet-18, ResNet-34, and ResNet-50 using the same optimizer, loss function, batch size, and number of epochs.

Weighted Cross Entropy Loss as the loss function for training each model. The loss function weights were determined to be [0.6, 0.4] for presence of COVID-19 and [0.15, 0.85] for severity of COVID-19 based on the class imbalances in the dataset. Models were trained using the Adam optimizer with standard parameters (B= 0.9, B2 = 0.999). We used an initial learning rate of 1e-5 that is decayed by a factor of 10 each time the validation loss plateaus after 5 epochs, and selected the model with the lowest validation loss. Models were trained across 100 epochs using a batch size of 4. The random seed was set to 2022 so our results would be reproducibile.

Model comparisons were made based on the performance metrics of each on the test set of CT images. The primary metric for evaluation was the Area Under the Receiver Operating Curve (AUROC). This metric was selected as an indicator for the discrimination ability of our model based on the True Positive Rate (TPR) and the False Positive Rate (FPR). Our secondary metric for comparison is accuracy to get a sense of the percentage of correct predictions that our models make. Initial predictions were made with a decision threshold of 0.5 for both presence and severity of COVID-19. Alternative decision thresholds were determined from the ROC Curve which provided an optimal prediction criteria.

The 3D CNN was trained for 25 epochs with weighted Cross Entropy Loss. We tried optimizing the choice of weights here by utilizing the compute_class_weights function from sklearn. This yielded weights of [1.2, 0.86, 0.59, 3.3]. This model was initialized with random weights and we used the Adam optimizer with a starting learning rate of 1e-3 since the weights were not pretrained.

## 5. Results

Our best performing model converged at the 10th Epoch with a Validation AUROC of 0.715 for COVID prediction and 0.696 for severity prediction. We found that the ResNet-34 architecture outperformed the other model configurations on the testing set of images (Table 1). This model achieved an AUROC of 0.678 and 0.751 for COVID-19 prediction and severity prediction, respectively, on the test set. It also achieved respective accuracy of 65.34% and 84.29%. The pixel rescaling technique that performed the best for this model was the 0-255 scaling. To highlight the True Positive Rate compared to the False Positive Rate, the AUROC across the test set was plotted (Figure 3). Additionally, Confusion Matrices were generated to highlight the exact amounts of True Positives, True Negatives, False Positives, and False Negatives for each (Figures 4 & 5).

| **Model** | **AUROC COV.** | **AUROC Sev.** | **Accuracy COV.** | **Accuracy Sev.** |
|---|---|---|---|---|
| ResNet-18 | 0.617 | 0.733 | 64.09 | **85.54** |
| ResNet-34 | **0.678** | **0.751** | 65.34 | 84.29 |
| ResNet-50 | 0.654 | 0.742 | **66.33** | 85.31 |
| 3D CNN | 0.526 | 0.533 | 54.61 | 81.55 |

Table 1: Comparison between ResNet-18, ResNet-34, and ResNet-50 on the Test set of CT images. Metrics for Comparison include AUROC of COVID-19 predictions, AUROC of Severity predictions, Accuracy of COVID-19 predictions, and Accuracy of Severity predictions. The best performance is indicated with bold.

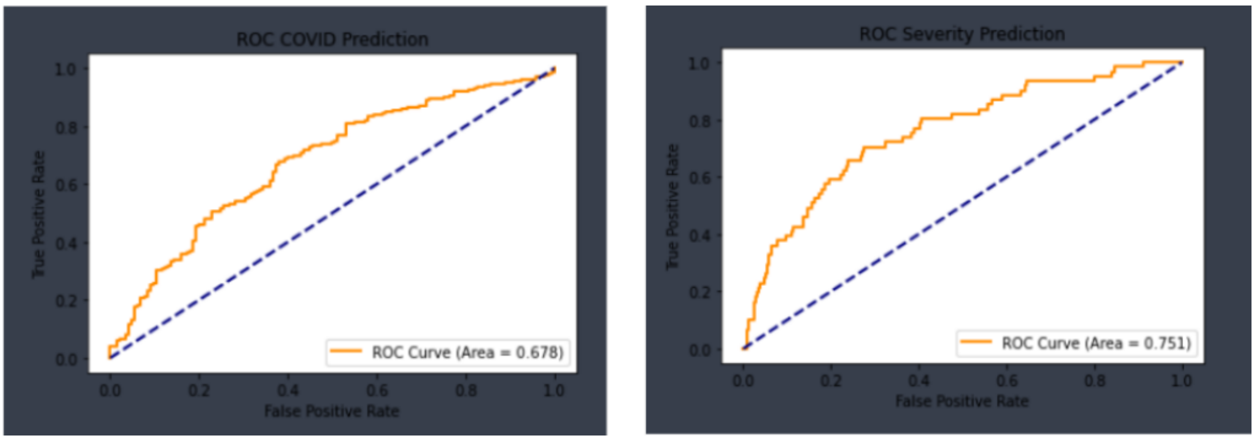


Figure 3: Area Under the Receiver Operating Curve (AUROC) performance on the test set. Performance for COVID-19 detection (left) and performance for severity detection (right).

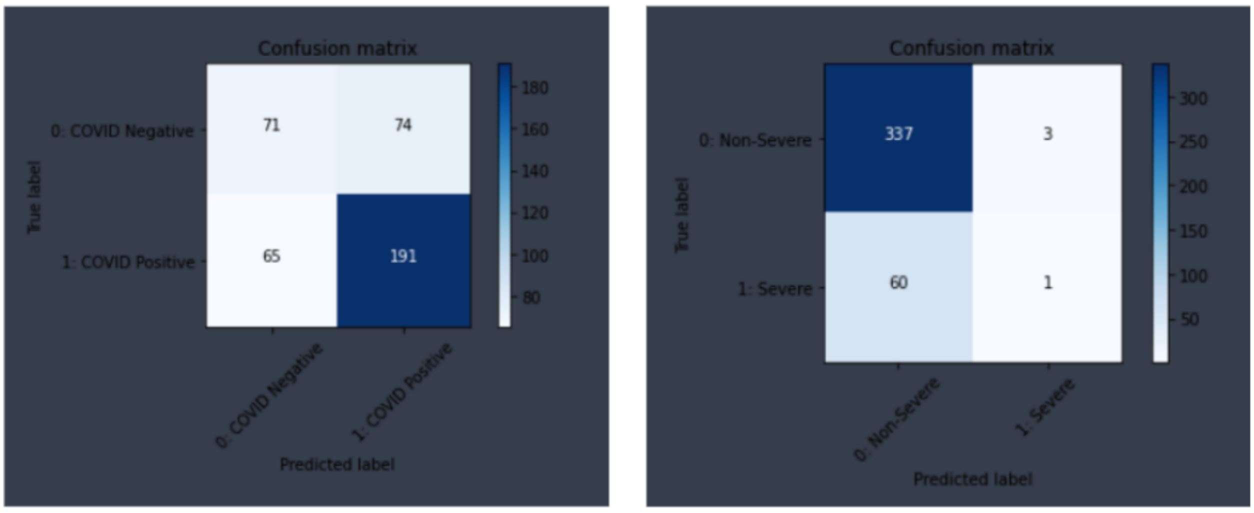


Figure 4: Confusion Matrices corresponding to the performance of ResNet-34 on the test set with a default decision threshold of 0.5 for predicting the presence of COVID-19 (right) and predicting the severity of COVID-19 (right)

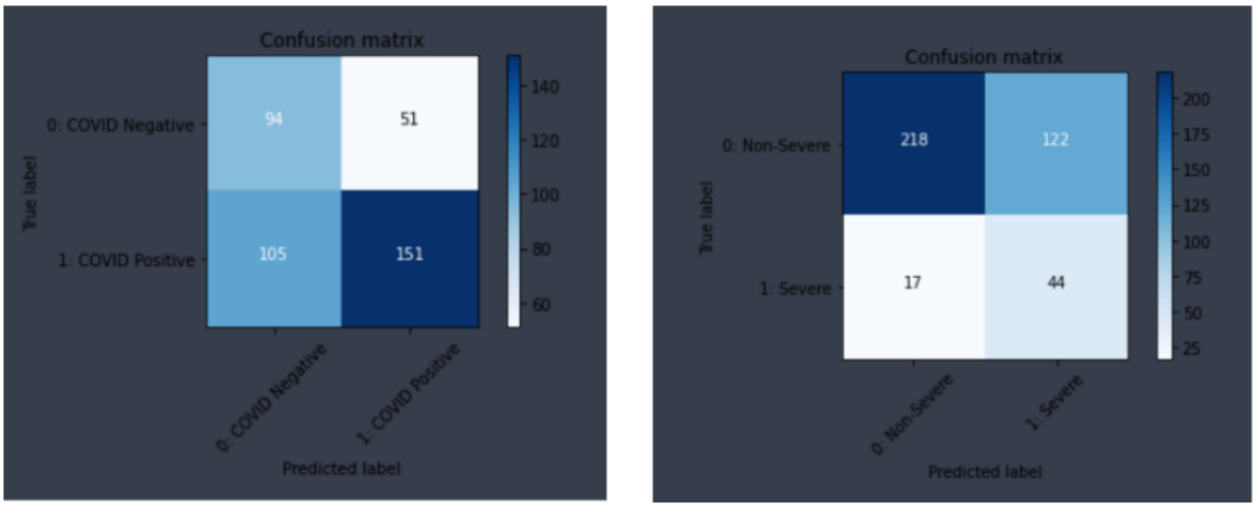


Figure 5: Confusion Matrices for Performance on test set with optimized decision thresholds for prediction of presence of COVID-19 (threshold of 0.59, left) and severity of COVID-19 (threshold of 0.19, right).

## 6. Discussion

Our goal for this project was to create an automated deep learning system that would provide consistent and accurate predictions based on the visual information gathered from the input images. We hypothesized that a ResNet architecture would be capable of providing a robust test result to potentially help hospitals allocate their resources for COVID-19 patients more effectively. We trained several configurations of ResNet to predict the presence and severity of COVID-19 based on CT scans from the STOIC2021 dataset to test this hypothesis. We found that the ResNet-34 architecture provided the best results based on its AUROC and accuracy performance. However, while the model has the ability to make distinctions between the types of CT scans in the STOIC dataset, it is clearly not a reliable or accurate enough system in its current state. In the base configuration, it only correctly predicted a single instance of severe COVID-19 (Figure 4). Even with an optimized threshold from the ROC Curve, the increased number of correct severity predictions came with a significantly higher number of false positives (Figure 5). This "optimized" threshold also reduces the performance of the model on predicting presence of COVID-19.

After receiving some feedback on our 2D model implementation, we constructed a 3D CNN model to test if it could outperform the 2D ResNet implementations. Surprisingly, our 3D CNN performed significantly worse than our 2D implementations. A possible explanation for this is poor implementation given the time constraints of our project. Additionally, the model required a lot of dimensionality reduction to not reach memory limits of our computational hardware. We believe that additional time invested into 3D model architecture would yield better results than our best 2D models.

There are many other reasons for why our implementation might not have performed as well as we had hoped. In terms of our data, we had to reduce the resolution significantly to process the images in a reasonable amount of time. This downsampling of images could potentially have removed critical distinguishing details between the types of images in the set. It's also possible that other types of image normalization/pixel rescaling could have been more ideal for our problem. While some attempts for rescaling pixel values from 0-1 or with 0 mean and 1 variance were made, these did not appear to significantly improve model performance.

For future directions regarding this problem and dataset, we would recommend exploring more models with 3D architecture for classification. However, this requires strong computational hardware since 3D models generally use significantly more memory than 2D models. Stronger computational resources would also allow for a larger batch size which could speed up computation time and provide a smoother learning trajectory. Additionally, we would recommend downsampling less (if possible) to extract maximal information from high resolution scans.

Altogether, we present a 2D model that has some capability of distinguishing between COVID-19 and non-COVID-19 CT scans. While it was largely unsuccessful in returning accurate predictions for severity of COVID-19, we believe that it sets up a good framework for other, more experienced groups to follow up on. For now, a more reliable and robust test than RT-PCR is not widely available, but with our contributions to the field, we hope that other groups will be able to successfully develop one.

## References

L. Ibáñez B. C. Lowekamp, D. T. Chen and D. Blezek. The Design of SimpleITK. *Frontiers in Neuroinformatics*, 2013.

L. Boulogne et al. The STOIC2021 Challenge. *Kaggle*, https://stoic2021.grand-challenge.org/, 2021.

X. Zhang K. He et al. Deep Residual Learning for Image Recognition. *arXiv*, 2015.

J. Gao L. Cai and D. Zhao. A Review of the Application of Deep Learning in Medical Image Classification and Segmentation. *Annals of Translational Medicine*, 8(11), 2020.

C. Margerie-Mellon M. Revel, S. Boussouar et al. Study of Thoracic CT in COVID-19: The STOIC Project, 2021. *RSNA Radiology*, 301(1), 2021.

M. Kotnik-G. Shiryaev I. Caglic J. Weir-McCall N. Sushentsev, V. Bura et al. A Head-to-Head Comparison of the Intra- and Interobserver Agreement of COVID-RADS and CO-RADS Grading Systems in a Population with High Estimated Prevalance of COVID-19. *The British Journal of Radiology*, (2), 2021.

B. C. Lowekamp R. Beare and Z. Yaniv. Image Segmentation, Registration and Characterization in R with SimpleITK. *Journal of Statistical Software*, 86(8), 2018.

L. Simonsen and C. Viboud. A Comprehensive Look at the COVID-19 Pandemic Death Toll. *eLife*, (10), 2021.

T.T. Ho T. Kim, W.J. Kim et al. A 3D-CNN Model with CT-based Parametric Response Mapping for Classifying COPD Subjects. *Scientific Reports*, 11(34), 2021.

J. D. Tournier et al. MRtrix3: A fast, flexible and open software framework for medical image processing and visualisation. *NeuroImage*, 202:116–137, 2019.

H. J. Johnson Z. Yaniv, B. C. Lowekamp and R. Beare. SimpleITK Image-Analysis Notebooks: a Collaborative Environment for Education and Reproducible Research. *Journal of Digital Imaging*, 31(3):290–303, 2018.