# Interpretable COVID-19 Risk Assessment with Deep Learning on Chest X-Rays

Brendan Ashworth
MIT EECS/BCS + Physics
77 Mass Ave, Cambridge, MA 02139
brendana@mit.edu

Anna Landler
MIT CEE + EECS
77 Mass Ave, Cambridge, MA 02139
alandler@mit.edu

## Abstract

*While COVID-19 diagnosis is typically based on RT-PCR detection of SARS-CoV-2 viral fragments, assessing disease progression requires either a qualitative symptomatic analysis or an examination of chest radiographs, which can identify fluid in the lungs. Prior research has focused on achieving quantitatively similar COVID-19 diagnosis accuracy when applying computer vision techniques. In this paper, we replicate competitive accuracy in diagnosing COVID-19, before successfully classifying chest X-rays into severity categories of severe, moderate, mild, and negative for COVID-19. Finally, we interpret these diagnoses for practitioners using SHAP, highlighting pixels most important to the diagnosis.*

## 1. Introduction

Considerable work has been done regarding using machine learning for disease diagnosis. Automatic classification for numerous types of diseases is an ongoing area of research, with convolutional neural networks and autoencoders being two popular models to apply [8]. The applications of computer vision in healthcare range from early detection, diagnosis, severity classification and distinguishing between similar conditions, typically with augmenting the medical professional in mind. Success automatic disease detection would have implications for patient care and outcomes, such as enabling more at-risk patients to be prioritized [2].

For this work, we focused on severity classification for COVID-19. Research into the applications of computer vision on COVID-19 has shown great promise in augmenting practitioners that could be overwhelmed with high numbers of cases, considering the ongoing pandemic and slow vaccine rollout. Past researchers have made anonymized datasets available containing both chest X-rays (CXRs) of positively diagnosed COVID-19 patients, as well as those of healthy patients with no ongoing infection. These datasets can be used to train deep learning models.

## 2. Related Work

In evaluating related research, we focus on the scope of detecting fluid buildup in the lungs by deep learning on CXRs. We cover both diagnosis of COVID-19 and pneumonia. Originally, COVID-19 was considered a new type of pneumonia [16], due to their similarity in pathogenesis. We consider convolutional neural networks as applied to these tasks.

### 2.1. Pneumonia Identification

Because some severe cases of COVID-19 can lead to pneumonia, identifying pneumonia with deep learning is a signal that deep learning can be leveraged in COVID-19 diagnosis.

Pneumonia, an infection of the lungs, is identifiable in chest X-rays. Deep learning, as applied to the identification and diagnosis of pneumonia, has shown great accuracy: one team demonstrated 78.83% accuracy when diagnosing it through chest X-rays, leveraging a residual network architecture of CNNs [10].

### 2.2. COVID-19 Diagnosis

Computer vision has proven to be a powerful tool for the diagnosis of COVID-19 through convolutional neural networks. We consider diagnosis through chest CT scans and chest X-rays. Prior successful research contributions include 93.96% accuracy for binary classification and 83.89% accuracy for three-class classification when diagnosing COVID-19 through chest CT scans [1]. Another approach showed that deep transfer learning, when applied to CT scans, resulted in 85% prediction accuracy for COVID-19 [11].

Other approaches have also found success in diagnosing COVID-19 through chest X-rays (CXRs). When classifying between pneumonia, COVID-19, and healthy lungs, it was found that VGG19 ([12]) achieved an 89.3% accuracy against 860 CXRs [9]. Even with few-shot learning, classification accuracy as high as 96.4% has been reported, leveraging siamese networks [6].

## 3. Methods

### 3.1. Dataset

We built a unique dataset of CXRs by combining one database of COVID-19 diagnoses with another database of healthy lung CXRs.

For annotated COVID-19 CXRs, we used the RSNA International COVID-19 Open Radiology Database (RICORD) [13] [14] [3] to evaluate the severity and risk of a patient. This dataset contains 856 chest x-rays from 361 patients. Each patient is labelled with personally non-identifying information, including:

1. **Classification**: Typical, Indeterminate, Atypical, or Negative for Pneumonia
2. **Disease Grading**: Mild, Moderate, or Severe
3. **Clinical Variables**: Age, Study Date, Sex, Testing Result

We combine this with ChestX-ray8, a dataset that contains 108,948 CXRs of various conditions [15]. We filter for 1,000 CXRs taken of healthy patients to add to our negative classification.

This results in a dataset of 1,856 CXRs with a dataset slightly biased towards negative diagnoses, which is likely to be more representative of a real-world application where a positive diagnosis has not been made.

| Classification | # of CXRs | Relative % |
|---|---|---|
| Severe | 265 | 14.2% |
| Moderate | 308 | 16.6% |
| Mild | 164 | 8.8% |
| Negative for COVID-19 | 1,119 | 60.3% |

To understand the data, Figure 1 is an example of a CXR from a patient with a positive COVID-19 diagnosis with moderate severity. Note the clouding around the internal organs. The annotations for Figure 1 are below.

| | |
|---|---|
| **Sex** | Male |
| **Age** | 55 |
| **Disease Grading** | Moderate Opacities (3-4 lung zones) |
| **Classification** | Absence of typical findings |

### 3.2. Classification Problem

To formulate the classification problem, we performed two experiments. The first experiment was a 4-classification problem, where a CXR was classified according to its severity, one of: severe, moderate, mild, and negative for COVID-19. The second problem was binary classification, where severe, moderate, and mild were aggregated into a positive class, with 60.3% negative and 39.7% positive class imbalance in our training dataset.

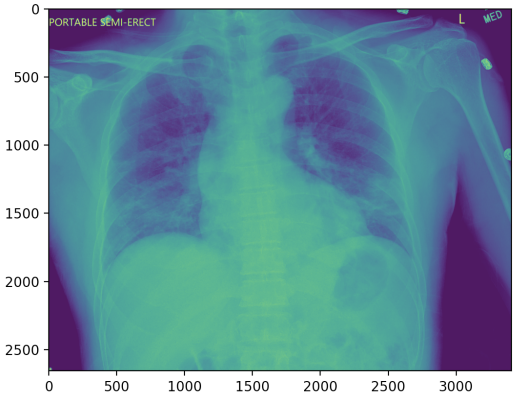


Figure 1. An example of a moderate COVID-19 case chest X-ray. The true X-ray is grayscale, here, converted to `viridis`.

Our dataset was split into a training, test and validation set, with an 85%, 7.5% and 7.5% split respectively. Accuracy is reported as the highest accuracy achieved on the validation set.

### 3.3. Model Architecture

In deciding the optimal model architecture to use, we compared multiple model architectures on the 4-classification problem. We found that, as compared to DenseNet and ResNet18, VGG demonstrated superior accuracy [5] [4] [12]. This architecture superiority on CXRs is also noted in other research.

| Architecture | Accuracy |
|---|---|
| VGG-11 | 86.5% |
| DenseNet-121 | 84.2% |
| ResNet-18 | 85.3% |

### 3.4. Data Augmentation

For training, we used multiple pre-processing steps to augment the size of our dataset and increase the generalizability of our model. We found that these preprocessing steps improved validation accuracy. One method we used was a random resized crop. Cropping the image is a common data augmentation task and will help the neural network perform well in x-rays that are not well centered, which we found to be relatively frequent in the dataset. We conducted the majority of trials with horizontal and vertical flips, as well as 20-30 degree rotations with probabilities that we varied between 0.2 and 0.5. Interestingly, there is a congenital condition, *situs inversus*, that results in mirrored internal organs (e.g., the heart on the right). Because the prevalence of this condition is approximately 0.01%, we can safely choose to ignore it. For the testing dataset, we chose to do a center crop. All images are grayscale.

| | Actual Positive | Actual Negative |
|---|---|---|
| **Predicted Positive** | 44 | 2 |
| **Predicted Negative** | 2 | 84 |

Table 1. Most of the images were correctly classified - hence 44 true positives and 84 true negatives. There were 2 false positives and 2 false negatives.

We found that applying a Gaussian blur introduced a regression in validation accuracy.

### 3.5. Hyperparameter Tuning

We found that the following hyperparameters gave us the best results:

| Hyperparameter | Initial Value |
|---|---|
| Epochs | 75 |
| Learning Rate (SGD) | 0.0001 |
| Momentum (SGD) | 0.9 |
| Batch Size | 32 |

## 4. Results

Our method of applying a VGG-11 architecture to an augmented dataset of CXRs gave us accuracy comparable to other previous works of research. For 4-classification of a CXR into severe, moderate, mild, and negative for COVID-19, we achieved a maximum validation accuracy of 86.5%. The confusion matrix for the validation set can be found in Table 2. Only 3.5% of negative COVID-19 cases were misdiagnosed as being positive, while all severe cases were correctly classified as being so. The model comparably struggled on distinguishing mild from moderate cases. Given how qualitative these categories are, future work might investigate how a quantitative metric of case severity could be predicted from a CXR. The model was trained over 75 epochs, with training, validation, and testing accuracy plotted in Figure 2.

When applied to binary classification of negative and positive cases of COVID-19, our model achieved a heldout validation accuracy of 97.1%, with a precision of 95.6% and a recall of 95.6%. The confusion matrix for binary classification can be found in Table 1. The model was trained over 75 epochs, with training, validation, and testing accuracy plotted in Figure 3. This accuracy is competitive with prior research: a recently published paper achieved 96.4% with siamese networks [6].

## 5. Interpreting Diagnoses with SHAP

When applying computer vision to medical diagnoses, it is important to be transparent and allow for medical practitioners to clearly interpret the diagnosis provided. To achieve this goal, we take advantage of SHAP (SHapley

| | Actual | | | |
|---|---|---|---|---|
| | **Negative** | **Mild** | **Moderate** | **Severe** |
| **Pred Negative** | 84 | 2 | 0 | 0 |
| **Pred Mild** | 3 | 5 | 4 | 0 |
| **Pred Moderate** | 0 | 0 | 5 | 0 |
| **Pred Severe** | 0 | 5 | 13 | 12 |

Table 2. There are 108 correct classifications out of a total of 133. Severe classification was perfect. Negative was did relatively well. Mild showed a range of classifications, including 2 false negatives out of 12 total images. Moderate tended to be classified as severe, but with no false negatives.

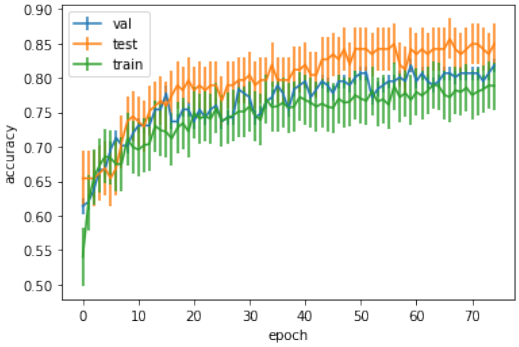


Figure 2. Accuracy over epochs for training with 4 classes.

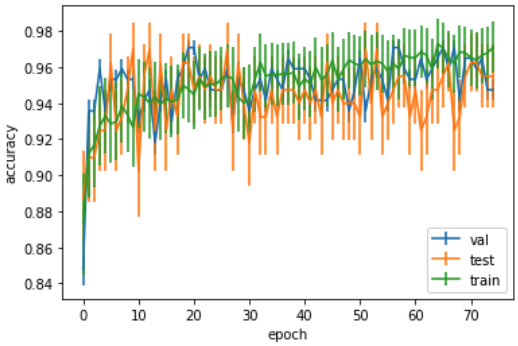


Figure 3. Accuracy over epochs for training with 2 classes.

Additive exPlanations) [7]. SHAP assigns pixels in an image importance in the final diagnosis: a medical professional can look at the proposed label alongside the pixels that are most important in the diagnosis. Practically, this might help a medical practitioner identify areas of the CXR that weren't identified prior, or even help a radiologist identify discrepancies in the deep learning model.

To review the contributions of [7], SHAP assigns features in a deep learning model importance according to how influential the feature is in the final classification. That is, for every feature $i$, it has a Shapley value $\phi_i$ that is positive if it influences the model towards a prediction, and negative if it influences away. This leads to a formulation for the individual Shapley values in terms of $x$, the features of

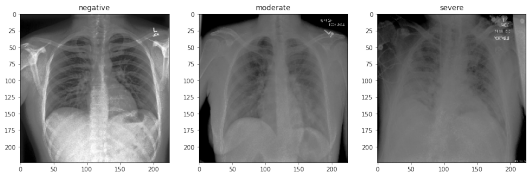


Figure 4. Chest X-rays for the interpretable data.

a model, $f$, the model, and $M$, the count of all features.

$$\phi_i(f,x) = \sum_{z' \in x'} \frac{\|z'\|!(M - \|z'\| - 1)!}{M!}[f_x(z') - f_x(z'\ i)]$$

This formulation follows from combined cooperative game theory results [7].

To investigate how we can use SHAP to interpret a diagnosis, we consider three images taken from our dataset, shown in Figure 4. The left image is a negative, healthy CXR, while the middle is a moderate case of COVID-19, and the right is a severe case. All were correctly classified by the model. By leveraging an open source SHAP library to calculate the pixelwise-Shapley values for each image, we can plot $\phi_i$ for each classification and each image on a coordinate grid. These can be found in Figure 5.

Each row represents an interpretation of a given CXR. The left image is the ground truth image from the dataset, in grayscale. The four labeled images to the right are the images with Shapley values plotted on top according to the classification labeled at top. For example, the top row is a negative CXR. One can tell that the negative Shapley values are particularly pink, especially around the areas of the lungs not obstructed by the ribcage. This is indicative that the model has identified a lack of clouding in the area. Comparatively, the next two rows have blue dots in the negative column, indicating that there is a positive COVID-19 diagnosis to be made. The model strongly indicates a severe case on the bottom row by coloring in the severe image with many high (pink) Shapley values, signaling a high importance on pixels around the internal organs in driving the classification to severe.

Shapley values make this deep learning more interpretable for medical practitioners, who may be overwhelmed by many COVID-19 cases, or have difficulty in distinguishing them. Comparatively, Shapley values are advantageous to activation mappings, which can blur together the focus on an image, losing granularity in interpretation.

## 6. Conclusion

We synthesized a unique dataset of CXRs, which amounted to 1,856 images in total. We trained a number of models across various neural network architectures, data augmentations, and hyperparameters to extract the most accurate automated detection model for COVID-19. Based on our analysis, we used a VGG11 architecture, with a number of data augmentations. We conducted experiments with 2-class positive-negative classifications and a more nuanced 4-class model with labels: negative, mild, moderate, and severe. The 2-class and 4-class experiments achieved best heldout validation accuracies of 97.1% and 86.5%, respectively. These results are comparable to the state-of-the-art.

## 7. Discussion

Better accuracy could likely be achieved by incorporating more patient information, such as age or symptoms, into the model. Generally speaking, it is beneficial to have more training data, so results could be improved with an expanded dataset. Finally, the 4-way classification ground truth is based on a panel of experts; these assessments often disagreed. It could be useful to have more expert assessments to better average out the variability between these ground-truth classifications.

## 8. Individual Contribution

I led the model architecture search, hyperparameter tuning, and also investigated which preprocessing and data augmentation tasks resulted in the highest heldout validation accuracy. I maintained a table of results for each iteration. I determined how to host our datasets and import them into our code. I also developed a script to generate the confusion tables. Brendan and I contributed equally to the research into state-of-the-art in COVID-19 severity classification and the final paper writing.

## References

[1] Hammam Alshazly, Christoph Linse, Mohamed Abdalla, Erhardt Barth, and Thomas Martinetz. Covid-nets: Deep cnn architectures for detecting covid-19 using chest ct scans. *medRxiv*, 2021. 1

[2] Min Chen, Yixue Hao, Kai Hwang, Lu Wang, and Lin Wang. Disease prediction by machine learning over big data from healthcare communities. *IEEE Access*, 5:8869–8879, 2017. 1

[3] K Clark, B Vendt, K Smith, J Freymann, J Kirby, P Koppel, S Moore, S Phillips, D Maffitt, M Pringle, L Tarbox, and F Prior. The cancer imaging archive (tcia): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057, 2013. 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 2

[5] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018. 2

[6] Shruti Jadon. Covid-19 detection from scarce chest x-ray image data using few-shot deep learning approach. *Medi-*

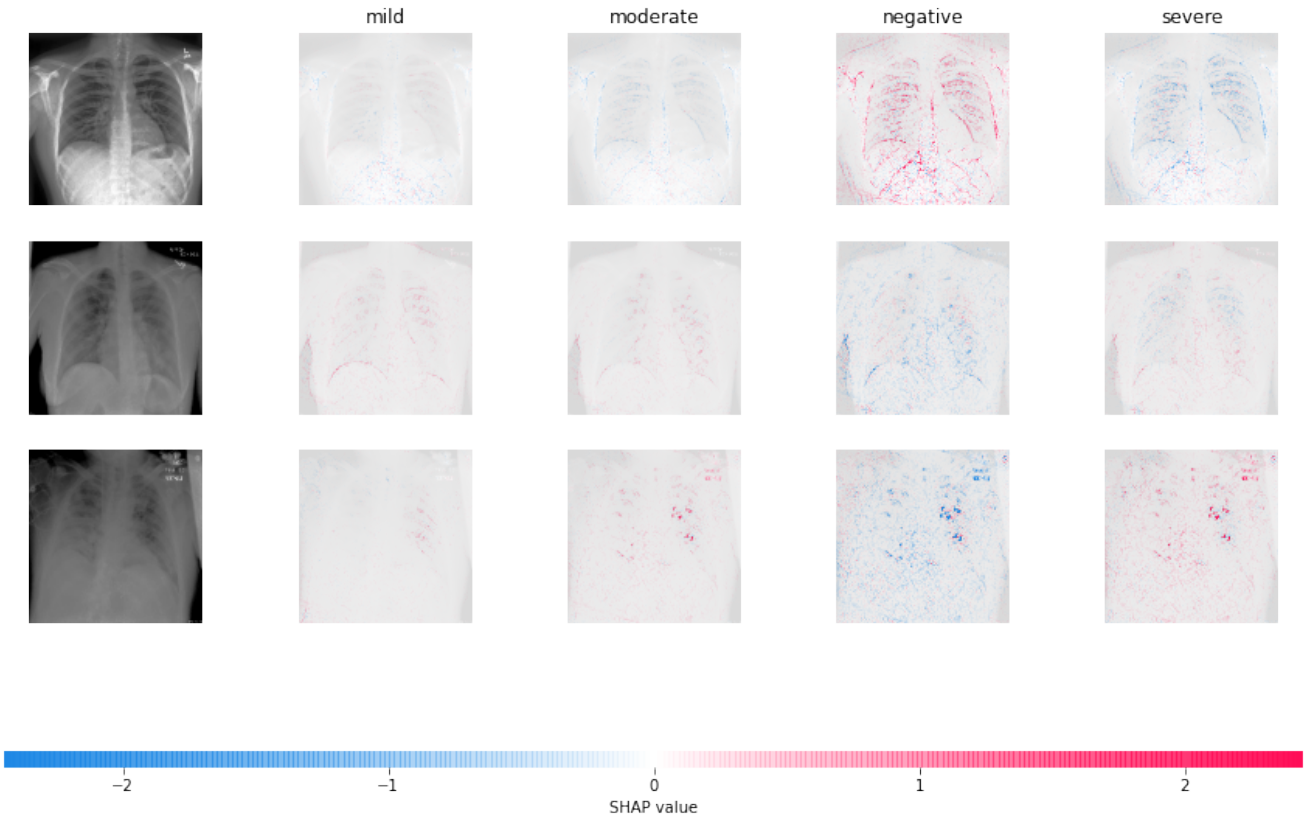


Figure 5. SHAP values plotted on top of three CXRs. Top row: negative, middle row: moderate, lower row: severe. Pink represents a strong positive classification towards the top category, and blue represents a negative classification towards any other category.

*cal Imaging 2021: Imaging Informatics for Healthcare, Research, and Applications*, Feb 2021. 1, 3

[7] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. 3, 4

[8] Chengsheng Mao, Yiheng Pan, Zexian Zeng, Liang Yao, and Yuan Luo. Deep generative classifiers for thoracic disease diagnosis with chest x-ray images, 2018. 1

[9] Md Mamunur Rahaman, Chen Li, Yudong Yao, Frank Kulwa, Mohammad Asadur Rahman, Qian Wang, Shouliang Qi, Fanjie Kong, Xuemin Zhu, and Xin Zhao. Identification of covid-19 samples from chest x-ray images using deep learning: A comparison of transfer learning approaches. *Journal of X-ray Science and Technology*, (Preprint):1–19, 2020. 1

[10] Can Jozef Saul, Deniz Yagmur Urey, and Can Doruk Taktakoglu. Early diagnosis of pneumonia with deep learning, 2019. 1

[11] Ahmad Shalbaf, Majid Vafaeezadeh, et al. Automated detection of covid-19 using ensemble of transfer learning with deep convolutional neural network based on ct scans. *International journal of computer assisted radiology and surgery*, 16(1):115–123, 2021. 1

[12] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. 1, 2

[13] E. Tsai, S. Simpson, M.P. Lungren, M. Hershman, L. Roshkovan, E. Colak, B.J. Erickson, G. Shih, A. Stein, J. Kalpathy-Cramer, J. Shen, M.A.F. Hafez, S. John, P. Rajiah, B.P Pogatchnik, J.T. Mongan, E. Altinmakas, E. Ranschaert, F.C. Kitamura, L. Topff, L. Moy, J.P. Kanne, and C. Wu. Medical imaging data resource center (midrc) - rsna international covid-19 open radiology database (ricord) release 1c - chest x-ray covid+ (midrc-ricord-1c). *The Cancer Imaging Archive*. 2

[14] E. Tsai, S. Simpson, M.P. Lungren, M. Hershman, L. Roshkovan, E. Colak, B.J. Erickson, G. Shih, A. Stein, J. Kalpathy-Cramer, J. Shen, M.A.F. Hafez, S. John, P. Rajiah, B.P Pogatchnik, J.T. Mongan, E. Altinmakas, E. Ranschaert, F.C. Kitamura, L. Topff, L. Moy, J.P. Kanne, and C. Wu. The rsna international covid-19 open annotated radiology database (ricord). *Radiology*. 2

[15] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. 2

[16] Jiangping Wei, Huaxiang Xu, Jingliang Xiong, Qinglin Shen, Bing Fan, Chenglong Ye, Wentao Dong, and Fangfang Hu. 2019 novel coronavirus (covid-19) pneumonia: serial computed tomography findings. *Korean journal of radiology*, 21(4):501, 2020. 1