X-Ray Detection of COVID-19

IDS 705 Final Project

Vanessa Tang, Sang-Jyh Lin, Sangseok Lee, Julio Portella
Team 9 Green

Abstract

Unavailability and inefficiency in COVID-19 diagnostic testing prevents millions from getting tested and increases risk to the healthy population. To help solve these issues, we propose a deep learning model to identify COVID-19 chest X-rays. We use transfer learning and tune a pre-trained model, ResNet50, to maximize overall accuracy. The resulting model has an overall accuracy of 92.4% on unseen data. This shows the potential for transfer learning and deep neural networks to distinguish amongst normal, pneumonia, and COVID-19 chest X-rays. In the current COVID-19 pandemic, this could help ease the pressure off traditional PCR diagnostic tests, while future implementation could use deep learning for diagnostic radiology.

Introduction

The novel coronavirus, COVID-19, has now affected over 200 countries worldwide, with almost 700,000 confirmed cases and over 30,000 deaths in the United States (*COVID-19 Map,* Johns Hopkins Coronavirus Resource Center). Thorough testing is important in slowing the spread of COVID-19 by preventing interactions between positive individuals and the healthy population as well as isolating positive individuals. Current diagnostic testing relies on reverse transcription polymerase chain reaction (RT-PCR) which can be time consuming and lead to delays in test results (Ai et al., 2020). Furthermore, these tests are not readily available, preventing many individuals from getting tested and thus increasing the risk to the healthy population.

The Centers for Disease Control (CDC) prioritizes testing to hospitalized patients, healthcare workers, elderly, immuno-compromised individuals, and those with mild symptoms (CDC, 2020). Despite the knowledge that people without symptoms may still be contagious, the CDC states that asymptomatic individuals are a non-priority (CDC, 2020). Making COVID-19 testing more accessible, available, and fast could help alleviate these problems.

This study proposes a solution to the shortage of tests through the development of a convolutional neural network (CNN) to identify COVID-19 amongst normal and pneumonia cases using chest X-rays. X-ray imaging is readily available in almost all hospitals and faster than running PCR, allowing more widespread availability and faster results.

Background

Our work with X-ray image detection of COVID-19 proposes one possible solution based on the fact that COVID-19 often presents as pneumonia and acute respiratory distress syndrome (CDC, 2020). Past research shows high accuracy in using CNNs to classify X-ray images for thorax diseases (Rajpurkar et al., 2017; L. Wang & Wong, 2020; X. Wang et al., 2017). Recently, X-ray imaging has also been shown to help detect respiratory distress in COVID-19 patients (Xu et al., 2020). Because COVID-19 acutely attacks the respiratory system, it is likely that X-rays can be used to help identify COVID-19.

As COVID-19 testing research continues, only a few other works have begun to delve into the idea of using machine learning models to diagnose COVID-19 based on X-ray images. Hall et al. and Narin et al. both used deep learning models to classify normal, pneumonia, and COVID-19 chest X-rays (Hall et al., 2020; Narin et al., 2020). However, both groups fail to recognize that their normal and pneumonia images are from pediatric patients aged 1 to 5, making their models inapplicable to adults.

In this paper, we aim to combat this issue by using a more representative dataset of normal, pneumonia, and COVID-19 chest X-rays from adults and develop a machine learning model using transfer learning to accurately distinguish amongst these three classes. Thus far, only one published paper has done similar work, building a deep CNN to classify chest X-rays into the same three classes (L. Wang & Wong, 2020). Overall, we aim to use transfer learning to tailor a CNN to distinguish amongst normal, pneumonia, and COVID-19 adult chest X-rays to propose a potential future solution to the shortage of RT-PCR diagnostic tests for COVID-19.

Data

The data for this project comes from two primary sources. First, COVID-19 chest X-rays are from a Github repository on COVID-19 image data collection (Cohen et al., 2020). Although this repository is an ongoing data collection project, at the time of our data collection stage, we had access to 115 total COVID-19 images. Second, pneumonia and normal chest X-rays were gathered from the NIH Clinical Center (X. Wang et al., 2017). We focused on pneumonia X-rays from the NIH dataset based on prior research that suggests that many COVID-19 patients develop pneumonia (CDC, 2020). The NIH dataset consists of 16,756 images, but due to computational limitations when training our model, we subsetted 1,360 normal and 1,360 pneumonia images.

The final dataset consists of 2,835 X-rays images from the posteroanterior view, sized 224 by 224 pixels (Figure 1). This dataset was split into train and test datasets with 2,625 and 210 images in each, ensuring that patients in the training and test sets are mutually exclusive to prevent multiple X-rays from one patient appearing in both the training and test sets. To ensure generalization, the training data was split into 60% training and 40% validation sets. In comparison to other splitting patterns (Figure 2), a 60-40 split showed the best generalization performance in test accuracy. This test data is the same dataset as that of Wang et al. to enable performance comparisons. The distribution of the three classes are shown in Figure 3.

COVID-19 Pneumonia Normal

Figure 1. Example of X-ray images for COVID-19,Pneumonia, and Normal patients

Train-Validation Split	Train Accuracy	Validation Accuracy	Test Accuracy
70% - 30%	96.01%	91.32%	86.73%
60% - 40%	96.45%	91.24%	92.38%
50% - 50%	97.60%	95.20%	86.84%

Figure 2. Table of train, validation, and test accuracies for three different training data splits into train and validation subsets.

	Train	Validation	Test
Normal	749	511	100
Pneumonia	749	511	100
COVID-19	64	41	10

Figure 3. Table of distribution of data for each class in the training, validation, and test datasets.

One major challenge is the lack of COVID-19 X-rays, consisting of only 4.4% of the training and 0.47% of the test datasets. Second, every patient's body can react differently to infections, which may be apparent in chest infection localization. Furthermore, X-rays are best suited for detecting dense objects, such as bones, and may not provide as high-quality imaging on soft tissue as other techniques such as CT scans or MRIs. This is potentially problematic if X-rays are unable to detect smaller or less aggressive infections.

Methods

After developing a baseline SVM, we used transfer learning to leverage architecture and pre-trained ImageNet weights, then tuned hyperparameters, added layers, and trained on our X-ray images to better fit our project (Figure 4). We selected the final model based on highest overall accuracy on the test data, and analyzed performance using a variety of metrics including class precision and recall.

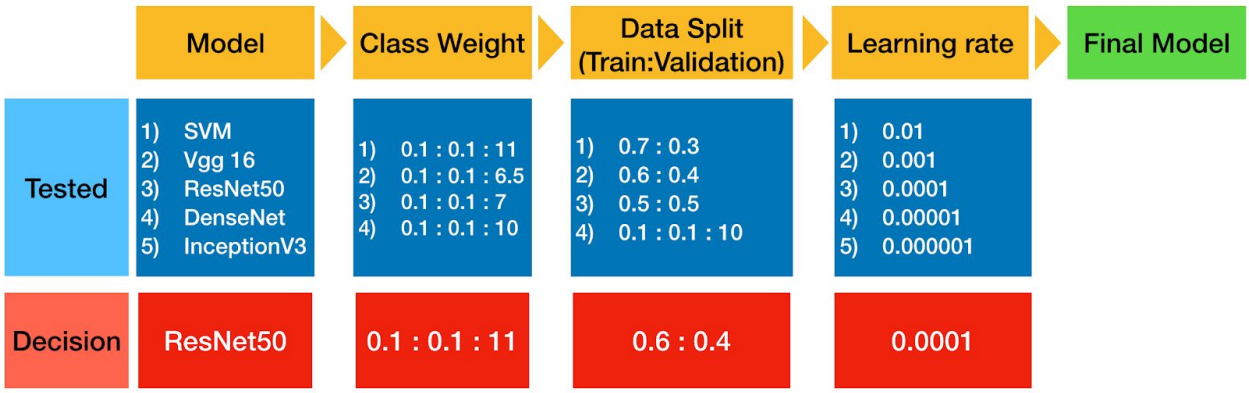


Figure 4. Model selection and testing flowchart. 1) We first developed a baseline SVM and selected the pre-trained model with the highest accuracy, ResNet50. 2) We then chose the class weights based on class ratios within the training data. For example, we applied 0.1 : 0.1 : 0.7 and 0.1 : 0.1 : 0.65 (normal : pneumonia : COVID-19) for a dataset with 700 normal, 700 pneumonia, and 105 COVID-19. 3) We tested different training and validation splits and found that a 0.6 training and 0.4 validation split had the best generalization performance. 4) We tested different learning rates and found that a learning rate of 0.0001 produced a smoother learning curve than higher learning rates and enabled faster convergence than lower learning rates (Appendix).

Winning the ImageNet challenge in 2015, ResNet uses residual learning, which decreases the computational complexity and expense of training deep neural networks while solving the issue of degrading training accuracy (He et al., 2015). Rather than simply stacking layers, ResNet implements shortcut connections, where layer outputs skip one or more layers and then are added to stacked layer outputs (He et al., 2015). These shortcut connections perform identity mapping without adding parameters or increasing computational complexity (He et al., 2015). In addition, based on the research results done by X. Wang et al., ResNet50 is the most efficient pre-trained model for thorax disease identification (X. Wang et al., 2017). After testing various available pre-trained models through Keras, we found that ResNet50 had the best performance in overall accuracy.

Architecture

The architecture of our network is shown in Figure 5. We implemented an Average Pooling layer between ResNet50 output and fully connected layer to reduce the dimensionality of the ResNet50 output and retain high feature variation. Since X-ray images include many high value pixels representing white colors, applying a Max Pooling layer may cause loss of variation in the image, especially for pneumonia and COVID-19 patients, leading to poor performance. Hence, the Average Pooling layer is a better choice. After reducing the dimensionality, we decreased the output shape from (7, 7, 2048) to (3, 3, 2048). This was still high dimensional data compared to our output of 3 classes and can increase the chance of overfitting. Therefore, we

implemented a Dropout layer as a regularization method to avoid overfitting and reduce computation time.

Our final model architecture has almost 24 million parameters, so we augmented our data by randomly rotating, shifting, flipping, cropping, and shearing the training data. This augmentation can gauge model accuracy and prevent overfitting to a relatively small number of images in our dataset without sacrificing computation ability or choosing a smaller network (Scott et al., 2017). These operations help our model converge within 30 epochs without overfitting.

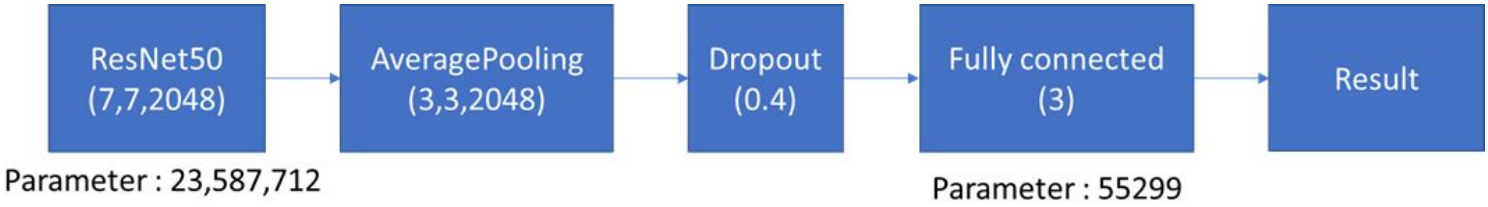


Figure 5. Architecture of CNN classifier with number of parameters. The (0.4) under Dropout layer is the dropout parameter setting. Other numbers under layer name represent the output shape for that layer.

Results

The final model was selected based on highest overall accuracy, with train, validation, and test dataset accuracies at 96.5%, 91.2%, and 92.4%, respectively (Figure 6). We also gauged model performance on class precision and recall (Figure 6). On the test data, these results showed that the final model performed well in identifying all three classes, with both precision and recall for normal, pneumonia, and COVID-19 classes at 0.93, 0.92 and 0.90, respectively.

	Train		Validate		Test	
	Precision	Recall	Precision	Recall	Precision	Recall
Normal	0.98	0.96	0.94	0.89	0.93	0.93
Pneumonia	0.96	0.97	0.89	0.92	0.92	0.92
COVID-19	0.93	0.95	0.82	0.71	0.90	0.90
Accuracy	0.965		0.912		0.924	

Figure 6. Class precision, recall and overall accuracy for train, validate, and test data.

We evaluated model performance by plotting the precision-recall (PR) and receiver operating characteristic (ROC) curves (Figure 7). In the PR curves (Figures 7a, 7b), the curves for all three classes (0: normal, 1: pneumonia, 2: COVID-19) hug the upper right corner, confirming that the final model performs well in terms of precision and recall for all three

classes. In the ROC curves (Figures 7c, 7d), the curves for all three classes hug the upper left corner, indicating that the model performs well in terms of true and false positive rates. Overall, these curves by class confirm that the model performs well in distinguishing amongst the three classes.

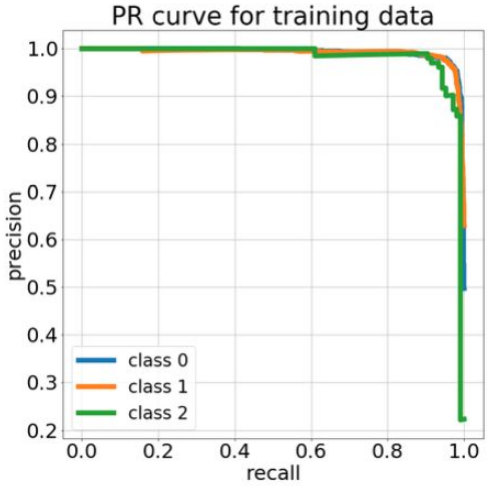


Figure 7a. PR curve for training data.

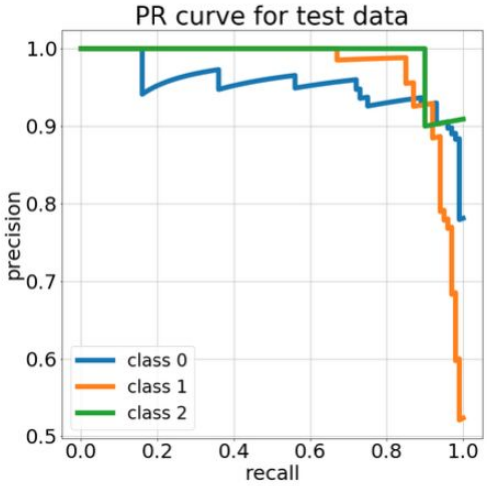


Figure 7b. PR curve for test data.

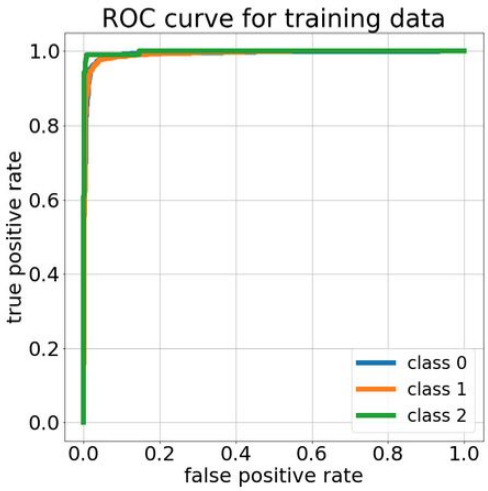


Figure 7c. ROC curve for training data.

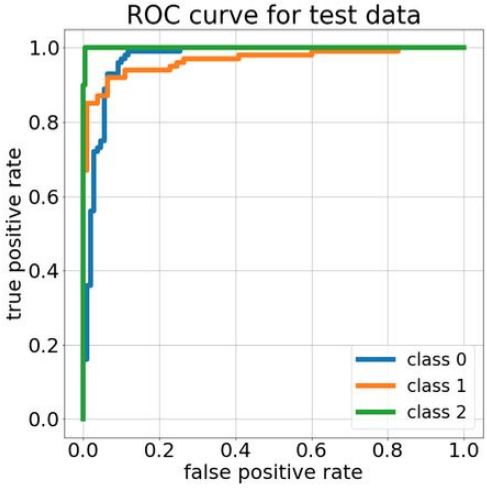


Figure 7d. ROC curve for test data.

Figure 7. PR (a, b) and ROC (c, d) curves on the training (a, c) and test (b, d) datasets, where class 0: normal, class 1: pneumonia, and class 2: COVID-19.

Based on the confusion matrices (Figure 8), our model performs well, with low counts on the off-diagonal indicating few incorrect predictions. On the test data, only 9 images were falsely classified as having pneumonia or COVID-19 (Figure 8), which is a relatively low false positive rate at 9%.

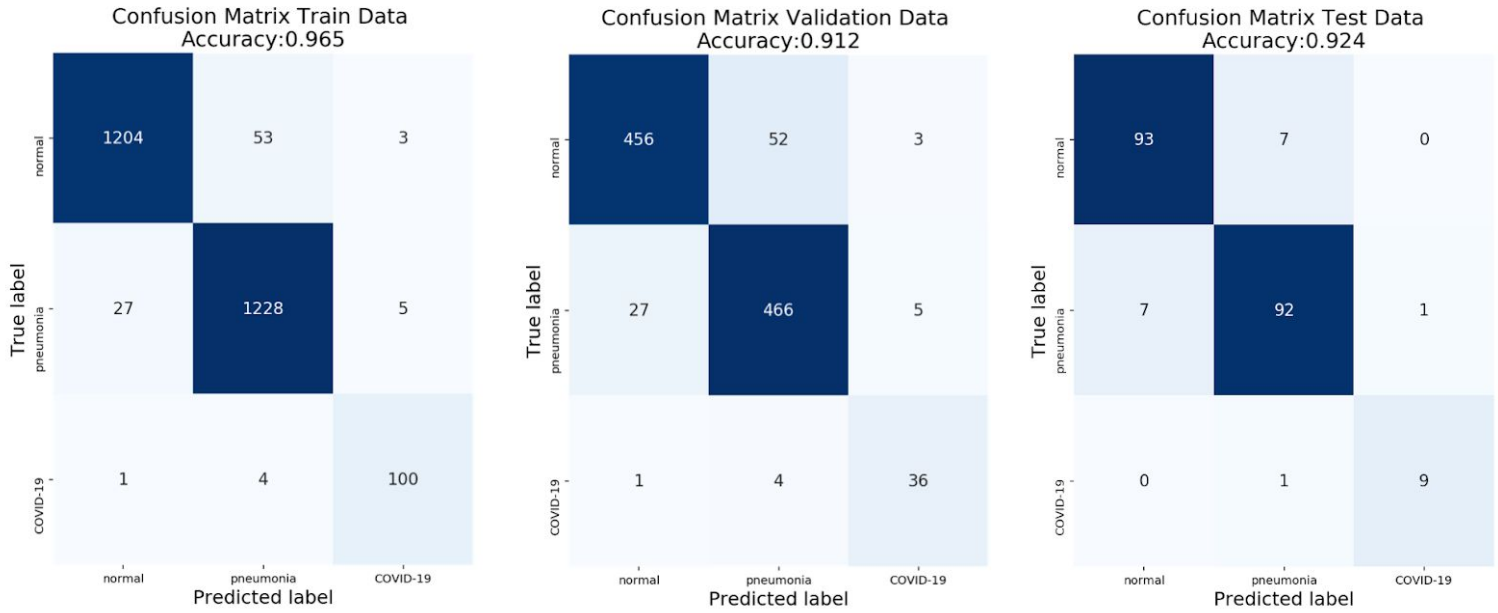


Figure 8. Confusion matrices for train, validation, and test datasets.(CNN)

We compared our final CNN to a support vector machine (SVM) trained on the same training data. We chose an SVM as a baseline model because SVMs have been shown to handle large amounts of data while being more flexible than other classification models (Hermes, 1999). Overall, SVMs are advantageous in image classification because they can handle nonlinear class boundaries and high-dimensional data while being flexible enough to yield high-performing models that can often generalize well in comparison to other more inflexible classifiers like logistic regression. The accuracy of the SVM is 99.0% on the training data and 83.8% on the test data, indicating that the SVM is overfitting to the training data.

We also compared our CNN to random guessing, and analyzed accuracy and confusion matrices on the test data (Figure 9). Based on overall accuracy, our CNN performs the best. Comparing class precision and recall (Figure 10), the CNN performs slightly better than the SVM for normal and pneumonia classes, but the SVM clearly struggles with COVID-19. In fact, the COVID-19 recall for the SVM and random guessing are the same. By analyzing class precision and recall, we can clearly see the strength of the CNN in identifying COVID-19.

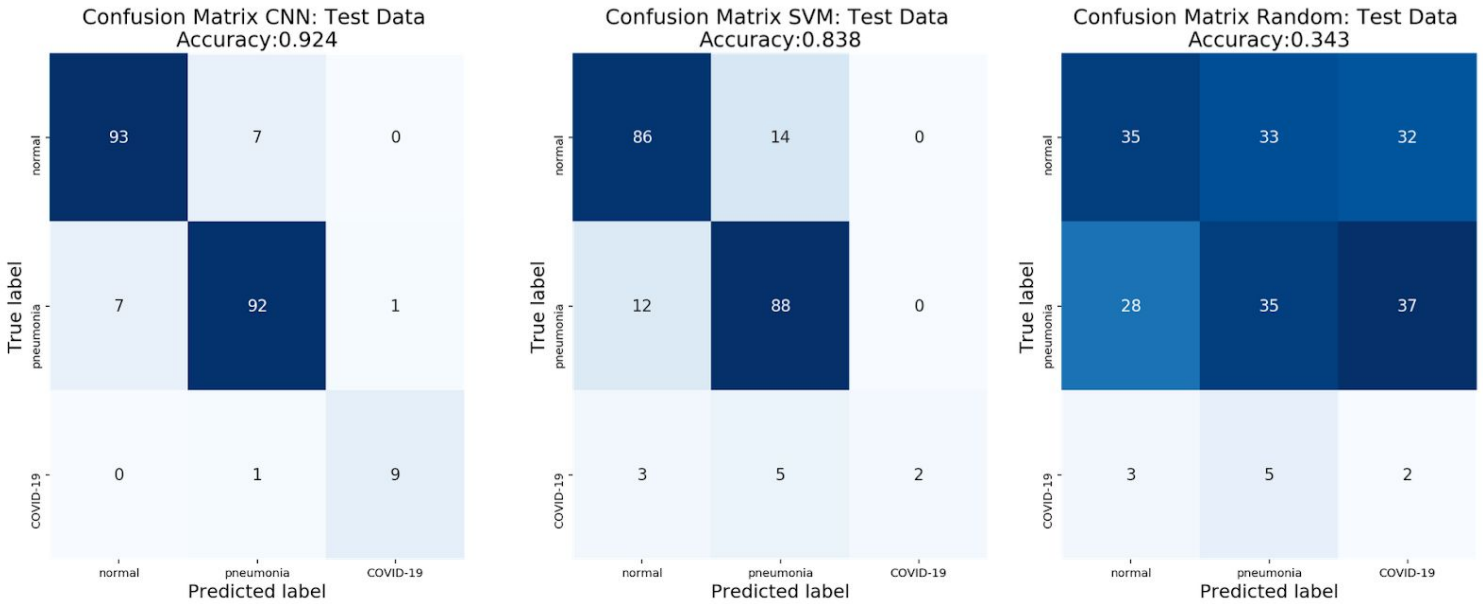
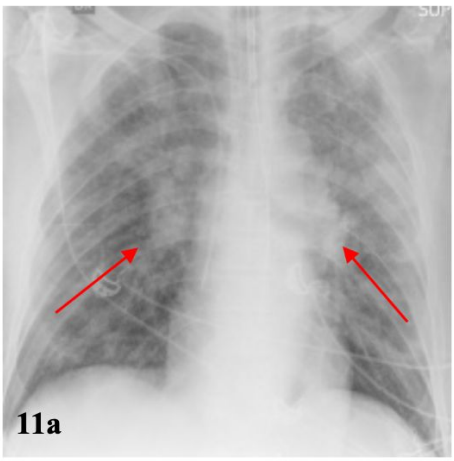


Figure 9. Confusion matrices and overall accuracy for CNN, SVM, and random guessing on the test data.

	CNN		SVM		Random	
	Precision	Recall	Precision	Recall	Precision	Recall
Normal	0.93	0.93	0.85	0.86	0.53	0.35
Pneumonia	0.92	0.92	0.82	0.88	0.48	0.35
COVID-19	0.90	0.90	1.00	0.20	0.03	0.20
Accuracy	0.924		0.838		0.343	

Figure 10. Precision, recall, and overall accuracy on the test data for CNN, SVM, and random guessing. The CNN and SVM use the same training dataset. COVID-19 recall for the SVM and random guessing are identical.

We also examined classification patterns to identify areas where the model succeeded and struggled. Specifically, we selected two misclassified images in the test data: Figure 11a) predicted pneumonia but actually COVID-19 and Figure 11b) predicted COVID-19 but actually pneumonia. First, Figure 11a shows white areas of inflammation similar to those seen in pneumonia X-rays(Simpson et al., 2020), indicated by the red arrows. Second, it is possible that the model struggles when images contain a large amount of white matter, as in Figure 11b, leading to misclassification as COVID-19. Overall, in looking at these two specific misclassified images from the test data, we hypothesize that the model struggles when images contain a lot of white matter in less clearly defined regions. Further research and advice from radiologists could clarify areas where our model struggles.



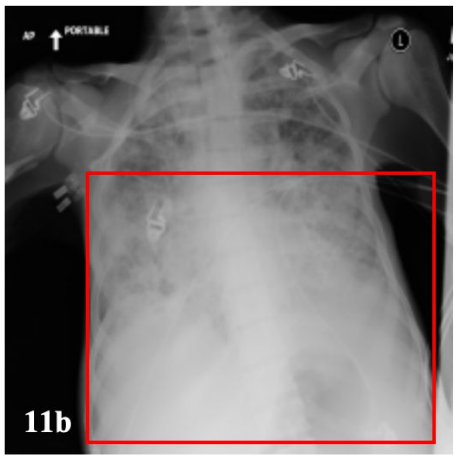


Figure 11. Misclassified images of interest in the test data. **a)** Predicted pneumonia, actually COVID-19. Red arrows indicate potential areas of inflammation. **b)** Predicted COVID-19, actually pneumonia. Red box indicates area of white "haze."

We can also identify areas where the model succeeds by selecting images with high probabilities of correct predictions (Figure 12). We hypothesize that when the X-ray images are clearer and have less "haze," the model can more accurately classify. It is also possible that the model succeeds when images are already well-centered rather than shifted or not filling the entire space.

Figure 12. Correct classifications of normal, pneumonia, and COVID-19 images.

Comparison to Recently Published Paper

Using the same test dataset, we can compare our model to that of Wang et al.'s COVID-Net (L. Wang & Wong, 2020). While COVID-Net is written from scratch, we implement transfer learning to generate our final model. Notably, Wang et al.'s COVID-Net was trained on 13,569 images, while our model is trained on a random subset of 1,260 normal and 1,260 pneumonia from the same larger dataset in addition to the same 105 COVID-19 images. We compare model performance on the same test dataset by analyzing confusion matrices (Figure 13) and class precision and recall (Figure 14). While the overall accuracy is the same at 92.4%, our model performs better in terms of precision and recall for COVID-19. However, overall, both models perform similarly on the test data, with slight differences in precision and recall.

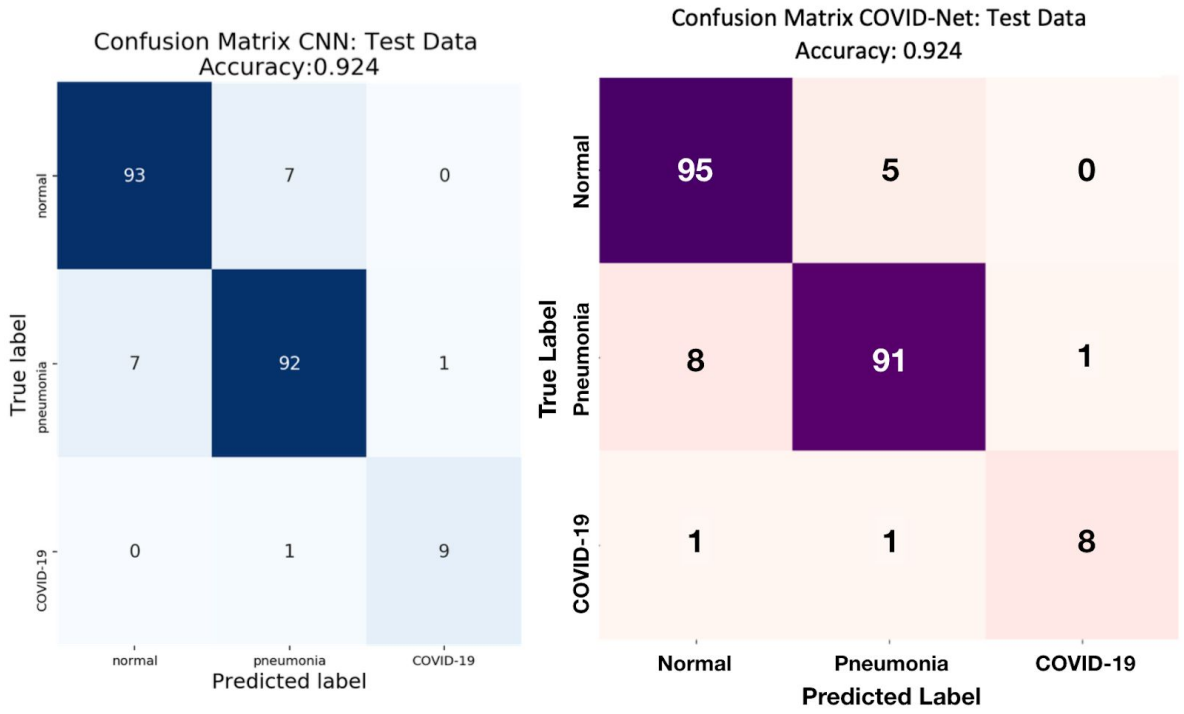


Figure 13. Confusion matrices for our final model (left) and COVID-Net (L. Wang & Wong, 2020) on the same test data. Confusion matrix for COVID-Net is taken from Wang and Wong et al.'s paper.

		Normal	Pneumonia	COVID-19
Precision	Our CNN	0.93	0.92	0.90
	COVID-Net	0.91	0.94	0.89
Recall	Our CNN	0.93	0.92	0.90
	COVID-Net	0.95	0.91	0.80

Figure 14. Precision and recall table for each class and comparing our CNN to Wang et al.'s COVID-Net. These metrics are based on the confusion matrices (Figure 13) on the same test data.

Conclusion

In the current pandemic, the unavailability and long turn-around times for conventional RT-PCR testing has prevented early detection, treatment, and isolation of infected people. Furthermore, CDC guidelines prioritize high risk and symptomatic people for testing, despite knowing that there can be asymptomatic carriers (CDC, 2020). New methods of diagnostic COVID-19 testing could enable more widespread testing, earlier detection and isolation, and potentially slow the spread of COVID-19.

Researchers have explored machine learning algorithms to diagnose COVID-19 based on the idea that COVID-19 acutely affects the respiratory system. Hall et al. and Narin et al. used

deep learning models to identify COVID-19 using X-ray images, but their normal and pneumonia X-rays came from pediatric patients (Hall et al., 2020; Narin et al., 2020). This is not ideal because COVID-19 most often affects the elderly and adults, while young children are less likely to contract the virus (CDC, 2020).

Using adult chest X-rays, we aim to develop a CNN to classify normal, pneumonia and COVID-19 X-rays. We used a pre-trained model, ResNet50, then added layers and tuned hyperparameters to maximize accuracy. The final model has a test accuracy of 92.4% with pneumonia and COVID-19 class recall of 92% and 90%, respectively. This performed much better than the SVM, with an accuracy of 83.8%, and pneumonia and COVID-19 class recall of 88% and 20%, respectively. Comparing the COVID-19 class recalls for the CNN and SVM, we can clearly identify the CNN's strength in identifying COVID-19, which is one of the primary goals of this project.

We present a deep learning model to distinguish amongst normal, pneumonia, and COVID-19 X-rays, with the goal of diagnosing COVID-19 in the current pandemic. Despite high accuracy and class recall for the final CNN, there is still significant room for improvement. Most importantly, more COVID-19 images should be gathered to better balance the dataset. Second, the CNN struggles with images with large amounts of hazy white matter. Future work should seek advice from a radiologist to better understand these X-rays to improve model performance. However, despite these issues, this project shows the potential for using deep learning models to diagnose COVID-19 and pneumonia using X-rays.

Roles

Vanessa Tang: I wrote the initial VGG16 and ResNet50, developed the SVM, and helped write the notebook used to assess model performance. I wrote, animated, narrated, and created the video. I also did extensive research on similar projects to find out that we had initially used pediatric patient images, then found new data sources on adults. Lastly, I wrote the Abstract, Introduction, Background, Data, and Conclusion in addition to part of the Methods and most of the Results sections of the report.

Sang-Jyh Lin: I preprocessed data and shared it with other members. I build ResNet50, DenseNet, and InceptionV3 models and trained on VGG16, ResNet50, DenseNet, and InceptionV3. Through numerous experiments, I developed the best model for our project. I also studied optimization work, including data splitting, hyperparameter selection. For the report, I wrote part of Method and Results sections, and helped modify the final report.

Sangseok Lee: I drew graphs by visualizing the flowchart and confusion matrix and tried various deep learning models including XGboosting. Finally, I Proofread the final report.

Julio Portella: Helped in the edition of the final report. Tried some linear models, including data augmentation and using Wang's dataset.

References

Ai, T., Yang, Z., Hou, H., Zhan, C., Chen, C., Lv, W., Tao, Q., Sun, Z., & Xia, L. (2020). Correlation of Chest CT and RT-PCR Testing in Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases. *Radiology*, 200642. https://doi.org/10.1148/radiol.2020200642

CDC. (2020, February 11). *Coronavirus Disease 2019 (COVID-19)*. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-criteria.html

Cohen, J. P., Morrison, P., & Dao, L. (2020). *COVID-19 Image Data Collection*. https://arxiv.org/abs/2003.11597v1

COVID-19 Map—Johns Hopkins Coronavirus Resource Center. (n.d.). Retrieved April 9, 2020, from https://coronavirus.jhu.edu/map.html

Fang, Y., Zhang, H., Xie, J., Lin, M., Ying, L., Pang, P., & Ji, W. (2020). Sensitivity of Chest CT for COVID-19: Comparison to RT-PCR. *Radiology*, 200432. https://doi.org/10.1148/radiol.2020200432

Gautret, P., Lagier, J.-C., Parola, P., Hoang, V. T., Meddeb, L., Mailhe, M., Doudier, B., Courjon, J., Giordanengo, V., Vieira, V. E., Dupont, H. T., Honoré, S., Colson, P., Chabrière, E., La Scola, B., Rolain, J.-M., Brouqui, P., & Raoult, D. (2020). Hydroxychloroquine and azithromycin as a treatment of COVID-19: Results of an open-label non-randomized clinical trial. *International Journal of Antimicrobial Agents*, 105949. https://doi.org/10.1016/j.ijantimicag.2020.105949

Hall, L., Goldgof, D., Paul, R., & Goldgof, G. M. (2020). *Finding COVID-19 from Chest X-rays using Deep Learning on a Small Dataset*. https://doi.org/10.36227/techrxiv.12083964.v2

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *ArXiv:1512.03385 [Cs]*. http://arxiv.org/abs/1512.03385

Hermes, L. (1999, January). *Support vector machines for land usage classification in Landsat TM imagery*. ResearchGate. https://www.researchgate.net/publication/3803218_Support_vector_machines_for_land_usage_classification_in_Landsat_TM_imagery

Liu, Y., Gayle, A. A., Wilder-Smith, A., & Rocklöv, J. (2020). The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of Travel Medicine*, *27*(2). https://doi.org/10.1093/jtm/taaa021

Narin, A., Kaya, C., & Pamuk, Z. (2020). *Automatic Detection of Coronavirus Disease (COVID-19) Using X-ray Images and Deep Convolutional Neural Networks*. https://arxiv.org/abs/2003.10849v1

Pan, F., Ye, T., Sun, P., Gui, S., Liang, B., Li, L., Zheng, D., Wang, J., Hesketh, R. L., Yang, L., & Zheng, C. (2020). Time Course of Lung Changes On Chest CT During Recovery From 2019 Novel Coronavirus (COVID-19) Pneumonia. *Radiology*, 200370. https://doi.org/10.1148/radiol.2020200370

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *ArXiv:1711.05225 [Cs, Stat]*. http://arxiv.org/abs/1711.05225

Simpson, S., Kay, F. U., Abbara, S., Bhalla, S., Chung, J. H., Chung, M., Henry, T. S., Kanne, J. P., Kligerman, S., Ko, J. P., & Litt, H. (2020). Radiological Society of North America

Expert Consensus Statement on Reporting Chest CT Findings Related to COVID-19. Endorsed by the Society of Thoracic Radiology, the American College of Radiology, and RSNA. *Radiology: Cardiothoracic Imaging*, *2*(2), e200152. https://doi.org/10.1148/ryct.2020200152

Subbaraman, N. (2020). Coronavirus tests: Researchers chase new diagnostics to fight the pandemic. *Nature*. https://doi.org/10.1038/d41586-020-00827-6

Wang, L., & Wong, A. (2020). *COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest Radiography Images*. https://arxiv.org/abs/2003.09871v2

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). *ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases*. 10.

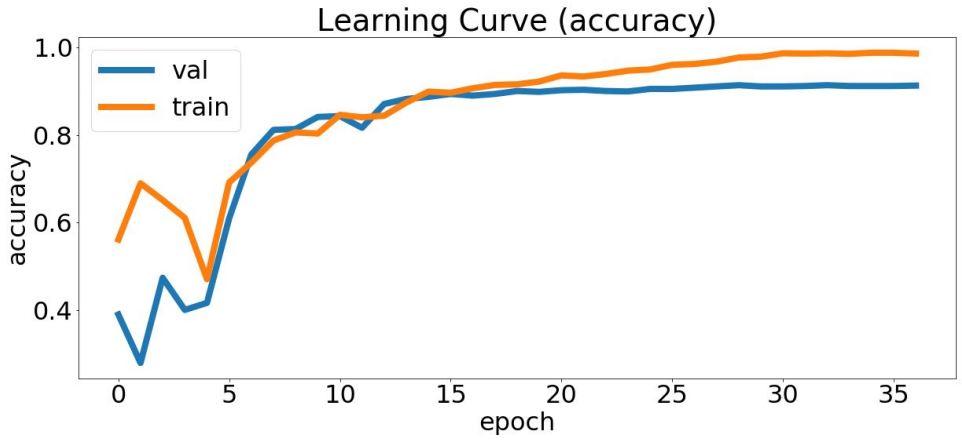
WHO. (2020). *Q&A on coronaviruses (COVID-19)*. https://www.who.int/news-room/q-a-detail/q-a-coronaviruses

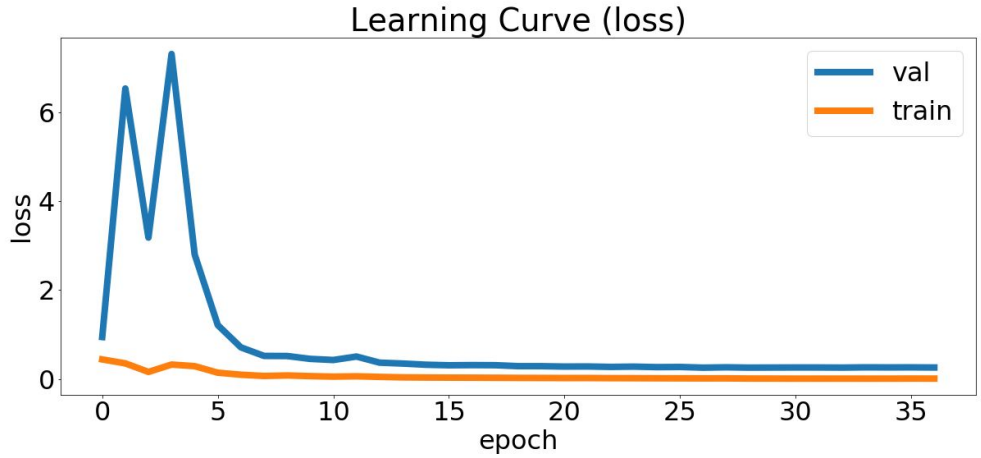
Wu, Z., & McGoogan, J. M. (2020). Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. *JAMA*, *323*(13), 1239–1242. https://doi.org/10.1001/jama.2020.2648

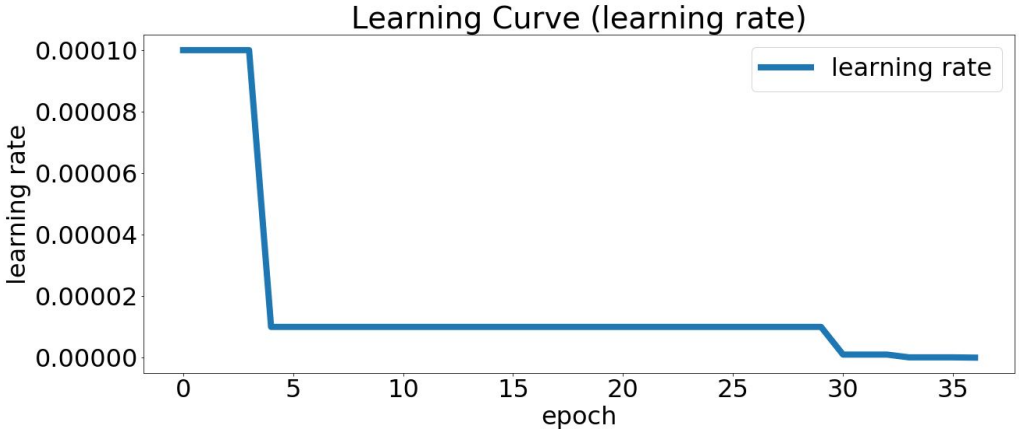
Xu, Z., Shi, L., Wang, Y., Zhang, J., Huang, L., Zhang, C., Liu, S., Zhao, P., Liu, H., Zhu, L., Tai, Y., Bai, C., Gao, T., Song, J., Xia, P., Dong, J., Zhao, J., & Wang, F.-S. (2020). Pathological findings of COVID-19 associated with acute respiratory distress syndrome. *The Lancet Respiratory Medicine*, *8*(4), 420–422. https://doi.org/10.1016/S2213-2600(20)30076-X

Appendix

Word count without figure captions: 2449







At the beginning of training, a higher learning rate can help model find different possibilities and avoid being trapped into local minimums too early. After we searched for several epochs, we saved the best result from these 4 epchs and continue training on that result with lower learning rates and can help the model converge smoothly.