FINAL REPORT FOR A DEEP LEARNING MODEL TO CLASSIFY COVID-19 CHEST X-RAYS

Justin Semelhago
justin.semelhago@mail.utoronto.ca

Paritosh Jain
paritosh.jain@mail.utoronto.ca

Rahem Ahmed
rahem.ahmed@mail.utoronto.ca

Alvin Shih
alvin.shih@mail.utoronto.ca

ABSTRACT

This report presents the development of an artificial intelligence (AI)-optimized classifier for differentiating chest X-rays (CXRs) containing COVID-19, viral pneumonia, or no disease present, along with its testing and potential applications. Patients, especially those with comorbidities, require accurate and quick tests to discern between these two infections that look very similar on CXRs but require very different triages. Hence due to the large demand for the use of AI in healthcare, two AI models were created to assist radiologists in classifying CXRs. The team obtained training data from publicly available CXRs from Italy and Spain, employed undersampling to balance the dataset, contrast limited adaptive histogram equalization (CLAHE) to amplify unique features in the CXRs, and shuffled and split the data into a train, validation, and testing sets. The baseline model was a random forest classifier which took 31 minutes to train and tended to overfit the data which after performing a grid search produced a train, validation, and test accuracy of 1.00, 0.882, and 0.875, respectively. For the deep learning model, the team started with ResNet-18 and after performing a grid search, the optimal model was found to be a ResNet-50 with 3 fully connected layers and an Adam optimizer requiring 16 minutes to train on an NVIDIA T4 GPU yielding a train, validation, and test accuracy of 0.994, 0.944, and 0.971, respectively. Moreover, 12 hand-curated CXRs which have more diverse demographics outside of the original dataset were also tested to examine the potential of biases due to limited data sources. Overall, the ResNet model yielded a better recall, accuracy, and training time compared to the baseline model. However, more testing on local data may be necessary before the model can be used clinically.

1 INTRODUCTION

1.1 BACKGROUND & MOTIVATION

The effects of COVID-19 on those with comorbidities has proven to be particularly lethal with a morbidity rate of 20% of COVID-19 cases in Canada alone (CIHI). Over-the-counter testing kits have shown to have a high false positive rate which can add safeguards to those with comorbidities but can burden public health systems with people who think they have COVID-19 (Latif et al., 2022). Furthermore, once in a hospital, a patient's chest X-ray (CXR) risks not being classified properly by radiologists thereby affecting their triage as Cozzi et al. (2020) found a classification accuracy of 83.75% when analyzing how successful radiologists were in classifying CXRs. This can be particularly problematic when comparing CXRs with COVID-19 versus ones with viral pneumonia whose morbidity rate to those with comorbidities is 5.8% thus changing the treatment course drastically

(Piazza et al., 2021). Therefore, there exists a need to classify CXRs with COVID-19 and viral pneumonia both accurately and quickly.

Artificial intelligence (AI) can be used to assist radiologists to achieve both of these goals. With AI expected to cut healthcare costs by $150 billion in the United States by 2026, its application across the healthcare industry has become more widespread (Bohr & Memarzadeh, 2020). In particular, with the introduction of widespread graphics process unit (GPU) usage as well as novel and efficient algorithms, deep learning has risen in popularity within the medical imaging space increasing the accuracy of tasks to be done (Lee et al., 2017). Moreover to accomplish the speed goal, using the well-tuned pre-trained model weights developed in this project can result in instantaneous classifications (normally, the time it takes to analyze a CXR can be 14.8 hours across multiple days according to Baltruschat et al. (2020)). Fig. 1 highlights a patient's journey with the introduction of a machine learning model to accomplish the above goals.

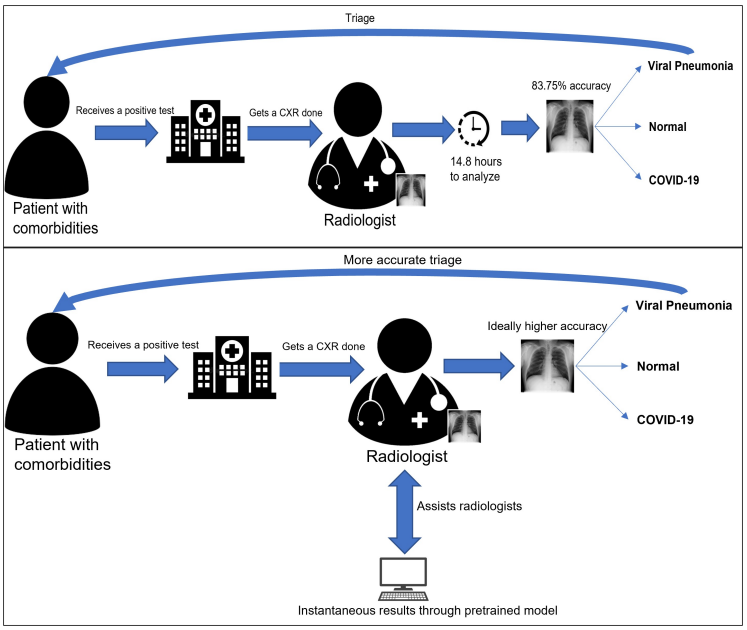

Figure 1: (*top*) Current flow of how a patient would receive a positive test and get a CXR with the associated statistics from Section 1.1. (*bottom*) Updated workflow including the deep learning algorithm providing instant results that would aid the radiologists in classification.

1.2 Related Work

There are several examples of both papers and commercial software that exist to classify CXRs. Some examples of frequently cited papers include one by Ismael & Şengür at Firat University in Turkey, who used pre-trained convolutional neural networks (CNNs) and were able to achieve a 91.6% accuracy using support vector machines (SVMs) to subsequently classify the features as having COVID-19 or not (Ismael & Şengür, 2021). Additionally, Yoo et al. (2020) from Hong Kong Sanatorium & Hospital integrated a CNN with three deep learning-based binary decision tree classifiers to classify CXRs depending on their presence of pneumonia, symptoms, and COVID-19 achieving a 95% accuracy. Some examples of commercial software used to classify CXRs include AI-Rad Companion developed by Siemens Healthineers which uses CNNs in its backend and in a study used to detect lung cancer had an accuracy of 90.4% (Chamberlin et al., 2021). Likewise, Fujifilm's CXR-AID has shown to detect abnormal radiological findings at a 97–99% accuracy using a ResNet-34 in its backend (Yoo et al., 2022). Finally, General Electric (GE) Healthcare Critical Care Suite which similarly uses a ResNet-34 can detect pneumothoraces at an accuracy of 96% (GE). From the papers, it is easy to see that this subject is global in nature and widely studied. It has been found that there has been 2,212 papers published in 2020 alone regarding deep learning models to classify CXRs but Roberts et al. (2021) concluded that only 37 papers had: how to reproduce & validate the results, proper code documentation, and followed Checklist for

Artificial Intelligence in Medical Imaging (CLAIM) criteria including explaining biases (Mongan et al., 2020). Moreover, oftentimes the commercial software have difficulties getting approval from the Food and Drug Administration (FDA) due to rigorous regulation of these corporate devices and software (Muehlematter et al., 2021). Thus, even with a large number of alternatives already in this domain, there still is a need to create an ethical deep learning model that is both accurate and fast in its classifications for CXRs.

2 DATA

2.1 SOURCES

There were two main sources of data that were used for the CXRs: the Italian Society of Medical and Interventional Radiology (SIRM) and the Valencia Region Image Bank (BIMCV) (SIRM, 2020)(de la Iglesia Vayá et al., 2021). Note that CXR data could not be collected by the team from local hospitals due to potential Personal Information Protection and Electronic Documents Act (PIPEDA) violations; thus, the SIRM and the BIMCV made these sources publicly available for research purposes (PIPEDA, 2019). One downside to these provided data sources is the fact that no demographic data was provided with it. Therefore, the team collected a hand-curated and labeled dataset of 12 CXRs from across the Internet with CXRs from a diverse population. This includes CXRs from men and women, children to the elderly, post-operative patients with lung transplants and surgical equipment inside them, as well as people from across the world from the United States, Australia, and Norway. The final model was tested on this dataset as deep learning models are capable of learning demographic features and discriminating against them which can contribute as a form of bias and unfairness (Ricci Lara et al., 2022).

2.2 UNDERSAMPLING

From the data sources, there was an imbalanced dataset present (as there was a 10:1 ratio of normal CXRs to viral pneumonia). According to Hasanin et al. (2019), this ratio can be enough to affect machine learning models to simply classify samples as the majority class. One way to combat this is through data augmentation which would involve applying random transformations to the COVID-19 and viral pneumonia CXRs such that they would be approximately equal to the normal class size. However, data augmentation is a domain-specific application and cannot simply be applied without first understanding the repercussions. For example, random flipping and rotating would cause the model to learn features about a CXR that it would not see in practice as MRIs are done in a standard direction and format. Furthermore as will be seen in Section 2.3, it is important that features of the lung are clearly visible for learning thus adding random noise could interfere with valuable information that the CXR holds for COVID-19 and viral pneumonia patients. Thus, undersampling on the normal CXRs was performed using a randomly generated number that would put it in a reasonable range with the rest of the classes producing final CXR counts of COVID: 3,616, viral pneumonia: 1,345, and normal: 3,732.

2.3 CLAHE

From the unprocessed scans, it can be difficult to discern features of the lung from each other, so it is important that a form of tensor normalization is applied such that the features of the lung are amplified for the models to learn from. One technique that can be applied is histogram equalization (HE) which equalizes the pixel intensities and is capable of enhancing low contrast images. A challenge of this method is that the global nature of HE can cause over or under-enhancement and can even just amplify noise (Cheng & Shi, 2004). Within the medical community, contrast limited adaptive HE (CLAHE) has become more popular to circumvent this issue (Pizer et al., 1987). Thus, CLAHE was used in this project to process the CXRs. Fig. 2 shows qualitatively how by applying CLAHE, the issue of over-enhancement and the overamplification of noisy regions from HE is reduced and clearer features of the lungs such as the bronchial tubes become clearer to visualize. This in turn should aid the model in extracting features related to diseases that affect the bronchial tubes including COVID-19 and viral pneumonia (Ye et al., 2020).

It should be noted that CLAHE requires some hyperparameter tuning. Firstly, it requires a clip limit which represents where to clip the histogram thus limiting the range of values it normalizes across.

It also requires a tile grid size which represents the approximate size of the areas that CLAHE is to be applied on the CXR. Various values for these hyperparameters were tested on until one could both discern the bronchial tubes while also limiting the noise outside the lungs. Hyperparameter values were also tested based on the results of figures such as in Fig. 3 which shows the pixel intensity range before and after HE and CLAHE. Optimal hyperparameter values in this domain would produce a figure similar to the original pixel intensity range while also extending around the peaks which represent the lungs in this case allowing for more lung features to protrude. These values were thus determined to be at a clip limit of three and a tile grid size of four.

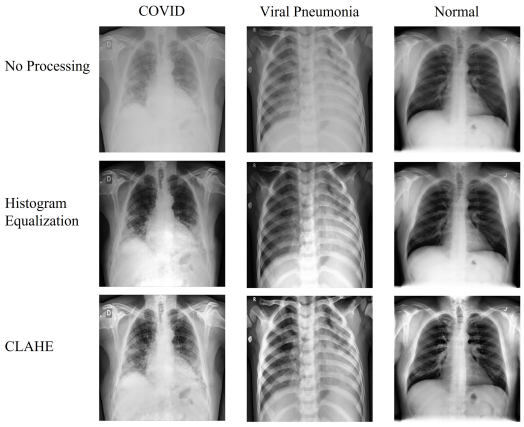

Figure 2: Qualitative comparison of the CXR types with different equalization strategies.

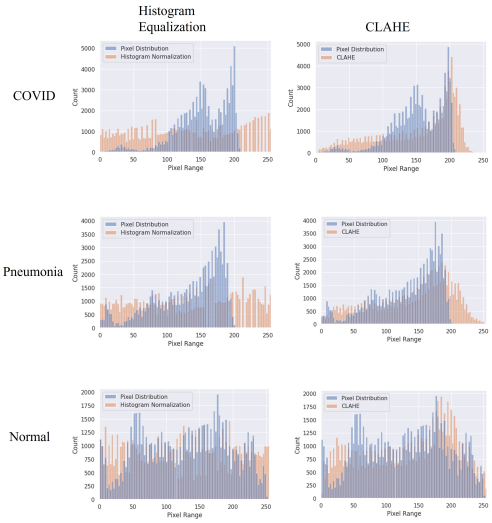

Figure 3: Comparison of the CXR pixel intensity graphs with different equalization strategies.

2.4 Data Splitting

After processing, the data was shuffled and split into training (to train the weights), validation (to tune the hyperparameters), and testing (to test the final model) sets. Initially it was split 80-20 with 80% of the data going into training and validation while 20% of the data goes into testing (not to be touched until the final model is ready). Then the training data was split again into 80-20 splits producing the training and validation sets. This produced final CXR counts of 5,564 for the train set, 1,390 for validation, and 1,739 for testing. These split values were chosen because in deep learning models, these splits tend to not result in overfitting while also achieving a good accuracy (Gholamy et al., 2018).

3 Baseline Model - Random Forest

3.1 Method

To establish a baseline for our deep learning model, the team utilized a random forest classifier from the `scikit-learn` library. A random forest classifier was chosen due to its increased popularity in the medical imaging community as they do not require GPUs unlike the deep learning model presented in this report (Hartmann et al., 2021). A challenge faced was creating the datasets used for the model as they would be different from the data loaders for the deep learning model. Before fitting the model on the data, the images were first flattened creating a number of features equal to one dimension of the CXR with a corresponding number of rows for the other dimension of the CXR. For the model, the image shape (1x299x299) returned a dataframe with 89,401 features in each row.

After flattening the images, an initial random forest classifier was fit with arbitrary parameters (`n_estimators`=10, `criterion`='entropy', `max_features`='sqrt', `max_depth`=None). Then, to optimize the train and validation accuracy of the model, a grid search was conducted to find the best set of hyperparameters. The grid search was conducted on four hyperparameters: `n_estimators`, `criterion`, `max_features`, and `max_depth`.

The `n_estimators` hyperparameter represents the number of decision trees used in the random forest classifier. In the grid search, these values were set to be 10, 40, and 100 as a higher value can result in a better accuracy (compared to the arbitrary parameters set) but a slower performance. The `criterion` hyperparameter specifies the metric used to evaluate the quality of each split. These values were simply set to all of the possible options that `scikit-learn` uses: Gini, entropy, and log-loss. The `max_features` hyperparameter indicates the maximum number of features considered for splitting at each node and similarly to `criterion`, was also set to all the values that `scikit-learn` uses (square root and log base 2). The `max_depth` hyperparameter sets the maximum depth of the decision trees in the forest which was set to either 12 or 50 to prevent the model from training for too long. Note that there are other hyperparameters to tune but these chosen values have shown to have significant effect on the architecture of the random forest (Probst et al., 2019). By varying these hyperparameters and selecting the combination that produces the highest accuracy, an optimal configuration for the random forest classifier was identified.

3.2 Results

The initial random forest model achieved a training accuracy of 0.995 and a validation accuracy of 0.840. After conducting a grid search by fitting on the train data and evaluating on the validation set, it was found that the best hyperparameters for the random forest model were `n_estimators`=100, `criterion`='log_loss', `max_features`='sqrt', and `max_depth`=50. This optimized model achieved a training accuracy of 1.00 and a validation accuracy of 0.882 on the validation set. The final model architecture and grid search process is summarized in Fig. 4.

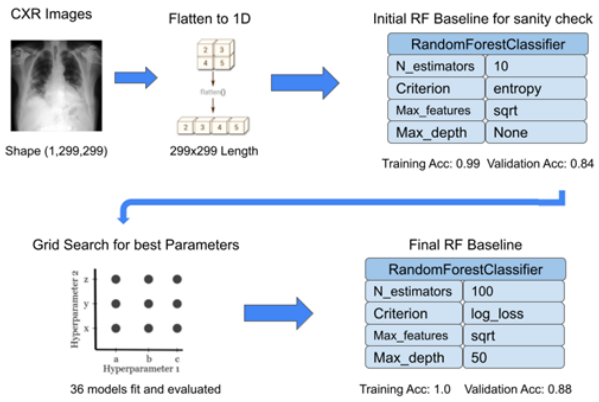

Figure 4: Model architecture diagram for random forest baseline model.

The final baseline model from Fig. 4 yielded a test accuracy of 0.875. The corresponding confusion matrix for the model can be shown in Fig. 5 with the precision and recall for each class found in Table 1. Note that these results will be discussed in more depth in Section 6.

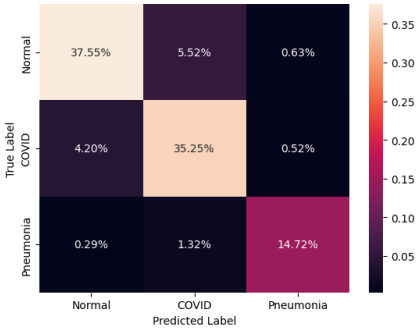

Figure 5: Confusion matrix for the best baseline model on testing set.

	Precision	Recall
Normal	0.859	0.893
Viral Pneumonia	0.882	0.837
COVID	0.901	0.928

Table 1: Precision and recall for the best baseline model.

4 Deep Learning Model - ResNet

4.1 Method

Initially, the team developed a ResNet-18 model (with 18 convolutions) as the primary model. It was chosen because it is the ResNet with the lowest number of convolutions in PyTorch in Python where transfer learning can be used to save the computational resources required to train a large network. Additionally, ResNets employ layer skipping which can address the vanishing gradient problem typically encountered in deep neural networks (He et al., 2016). The model was trained using an Adam optimizer and a cross-entropy loss function, as this is a multiclass classification problem. To prevent retraining of the feature extraction layers, the model weights were frozen to the default weights trained on ImageNet. A fully connected layer was connected to the resulting ResNet, which had three output nodes corresponding to the three classes. The output layer performs the classification, determining whether the image is categorized as a patient with normal, COVID-19, or viral pneumonia features. One challenge encountered was with the input value for the ResNet as the first convolution requires RGB channels while the CXRs only have one channel. This was resolved by overlaying three grayscale images over each other to simulate the same input.

For this model, the team initially trained it with five epochs. To identify a clearer presence of convergence in the model accuracy and loss (as the accuracy was still increasing at a high rate), the number of epochs was increased to 15. Initially, the 15 epoch ResNet-18 model achieved a training accuracy of 0.847 and a validation accuracy of 0.853, shown in Fig. 6. The training and validation loss values were quite low as well (0.003) as seen in Fig. 7. Note that the validation accuracy being strictly greater than the train accuracy was at such a small scale that this can be attributed to random noise. Based on the results of Figs. 6 and 7, the team only performed a grid search on different model architectures rather than hyperparameters. This is because the model did not appear to be over or underfitting and was relatively stable thus not indicating a need to tune the hyperparameters but rather trying to achieve a better accuracy through overarching architecture tuning. The architecture grid search was performed on: the number of convolutions in the ResNet (18, 34, 50, 101, 152), the number of fully connected layers (1, 2, 3), and the optimizer (Adam or Stochastic Gradient Descent (SGD)). Then, once the model with the best validation accuracy was found (with minimal to no overfitting), this selected model was tested on the test and out-of-the-box dataset.

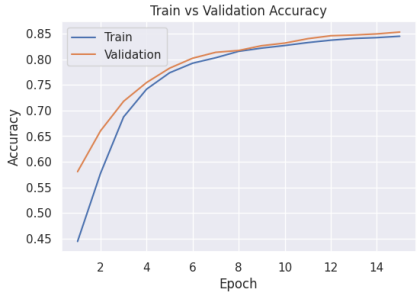

Figure 6: Training and validation accuracy curves of the primary ResNet-18 model (15 epochs).

Figure 7: Training and validation loss curves of the primary ResNet-18 model (15 epochs).

4.2 Results

When tuning the optimizer, it was found that SGD took a substantial cut in the accuracy (0.507 validation accuracy) while being 5.3% slower in computation time compared to an Adam optimizer (faster computation times are supposed to be a benefit of SGD (Sirignano & Spiliopoulos, 2017)). Thus, tuning using an SGD was abandoned. This resulted in a grid search for 16 different deep learning models. The best model was the ResNet-50 with 3 fully connected layers and an Adam optimizer as shown in Fig. 8. This model achieved a training accuracy of 0.994 and the highest validation accuracy of 0.944, shown in Fig. 9, and contained minimal overfitting. The training and validation loss values were lower than 0.002 as seen in Fig. 10. Note that the blip at Epoch 14 in Figs. 9 and 10 appeared to be random noise as it could not be reproduced when retraining.

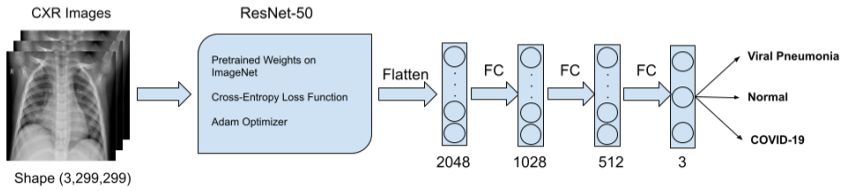

Figure 8: Model architecture diagram for best performing ResNet.

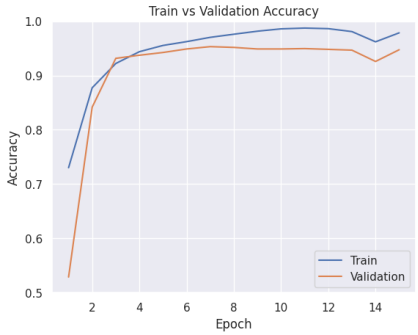

Figure 9: Training and validation accuracy curves of the best ResNet model.

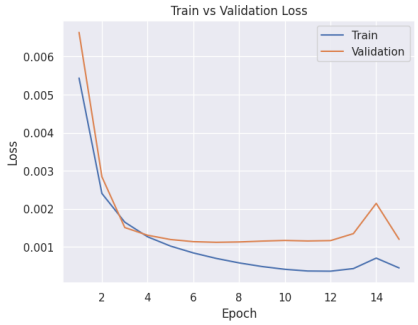

Figure 10: Training and validation loss curves of the best ResNet model.

The best model yielded a final test accuracy of 0.971. The corresponding confusion matrix for the model can be shown in Fig. 11 with the precision and recall for each class found in Table 2. Note that these results will be discussed in more depth in Section 6.

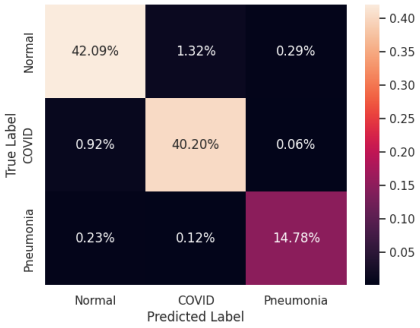

Figure 11: Confusion matrix for the best ResNet model on testing set.

	Precision	Recall
Normal	0.973	0.963
Viral Pneumonia	0.977	0.977
COVID	0.966	0.976

Table 2: Precision and recall for the best ResNet model.

5 Out of the Box Testing

As discussed in Section 2.1, testing was also done on samples outside of the dataset to ensure that the model was not learning demographic features related to the Italian or Spanish data sources. Table 3 shows the CXR dataset that the team hand-curated and labeled along with a set of descriptions highlighting the diversity of the collected CXRs. After loading the pre-trained model weights of the ResNet-50 and applying CLAHE on the CXRs, the model achieved an accuracy of 83.33%. Fig. 12 shows the two misclassifications (the elderly man and the Australian adult) that classified two viral pneumonia CXRs as CXRs containing COVID-19.

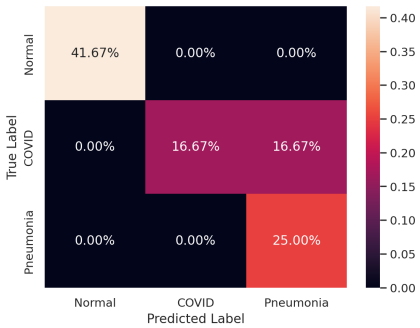

Figure 12: Confusion matrix for the best ResNet model on out of the box samples.

Source	Label	Description
(Taschetta-Millane, 2021)	COVID	Adult from Michigan
(Lyon, 2020)	COVID	Adult from Northeast US
(Aaløkken et al., 2020)	COVID	Adult from Norway
(Sorrentino, 2022)	Viral Pneumonia	Adult with lung transplant from Australia
(Dominguez, 2022)	Viral Pneumonia	Elderly man from US
(Benmalek et al., 2021)	Viral Pneumonia	Young boy
(Luong, 2021)	Viral Pneumonia	Adult from Australia
(Gaillard, 2022)	Normal	Adult from Australia
(Koksal, 2018)	Normal	Teenage woman from US
(Lloyd-Jones, 2019)	Normal	Woman
(Lloyd-Jones, 2019)	Normal	Teenage girl
(Nabulsi et al., 2021)	Normal	Woman with surgical equipment on her

Table 3: Hand-curated dataset to test the model on CXRs from outside the given data sources.

6 Discussion

While accuracy was presented for both models on the testing set, the most important metric to evaluate the models is recall. This is because recall penalizes the false negative rate which, in a highly sensitive environment such as healthcare, is very important as a missed diagnosis of a patient with comorbidities can result in death. With this in mind, the ResNet model performed much better than the baseline random forest because the recall for each class in Table 2 is strictly better than the recall in Table 1. What is of particular concern is the low recall of viral pneumonia and COVID-19 for the random forest as in this case, part of the calculation means that the model is potentially predicting normal when the CXR actually has a disease. These results are confirmed by the confusion matrices in Figs. 5 and 11 where the random forest dangerously predicted a patient not having any disease when they had COVID-19 4.20% while the ResNet predicted the same at 0.92%. Moreover, while it may be tempting to continue to pursue using a random forest despite the low recall as it does not require a GPU, the train time for the random forest was greater than 30 minutes whereas the best ResNet model weights can simply be loaded and ready to use for a fast inference time (and even if retraining is desired, it takes around 16 minutes on an NVIDIA T4 GPU to do so). Thus, the ResNet model was better than the baseline model.

Fig. 13 highlights some qualitative results of the ResNet model. It shows the classification of COVID, viral pneumonia, and normal CXRs demonstrating the difficulty of being able to classify these CXRs without any additional software. While these samples were taken from the data sources described in Section 2.1 and have visually different demographics from each other, Section 5 was an important step to test the model on data from seemingly random sources with guaranteed known demographics. Given that it only had two misclassifications and that they resulted in no patient being declared as having no disease when they do indeed have a disease, the model was successful on this dataset. The misclassifications of viral pneumonia are also not unique to the ResNet as the lowest recall for the baseline model was also for viral pneumonia. This could be due to the fact that viral pneumonia, when spread in a given community, can spread with SARS-CoV-2 (the strain of coronavirus causing COVID-19) (Pagliano et al., 2021). Thus qualitatively, the misclassifications of viral pneumonia as COVID-19 could be due to both diseases coexisting in the CXR.

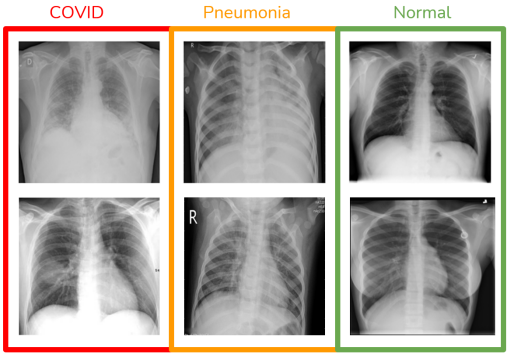

Figure 13: Sample qualitative results of the best ResNet model.

It is important to acknowledge the limitations of this work as it is highly sensitive in nature. The CLAHE hyperparameters were chosen on a qualitative basis that performed well on the given dataset but may note translate generally for all CXRs, so caution must be exercised when applying this model on different CXRs especially if they are zoomed in or out. Furthermore, the team learned how quickly overly complicated models can overfit in both the baseline and deep learning models so if this work is to be used in a clinical practice, it will be necessary to test the model on the local hospital's CXR data before using it in practice. Finally with regards to ethical considerations, the undersampling performed may have increased the variance with the produced model as information could have been lost from the normal CXRs. Also should the end-user wish to retrain the model, collecting data from a diverse source is critical to preventing the model from learning demographic features as mentioned in Section 2.1. This may require hand-tuning weights to shift the decision boundary to account for marginalized groups that the AI model may overlook which has been known to occur in other sensitive environments (Fish et al., 2016).

REFERENCES

Trond Mogens Aaløkken, Anagha P. Parkar, Tom-Vegard Markussen, Haseem Ashraf, Georg Karl Mynarek, Harald Nes, Fredrik Müller, Michael Schubert, Arve Jørgensen, Siri Marie Blomberg, and et al. Bildediagnostikk av pasienter med covid-19. *Tidsskrift for Den norske legeforening*, May 2020. doi: 10.4045/tidsskr.20.0332.

Ivo Baltruschat, Leonhard Steinmeister, Hannes Nickisch, Axel Saalbach, Michael Grass, Gerhard Adam, Tobias Knopp, and Harald Ittrich. Smart chest x-ray worklist prioritization using artificial intelligence: A clinical workflow simulation - european radiology. *SpringerLink*, Nov 2020. URL https://link.springer.com/article/10.1007/s00330-020-07480-7.

Elmehdi Benmalek, Jamal Elmhamdi, and Abdelilah Jilbab. Comparing ct scan and chest x-ray imaging for covid-19 diagnosis. *Biomedical Engineering Advances*, 1:100003, Jun 2021. doi: 10.1016/j.bea.2021.100003.

Adam Bohr and Kaveh Memarzadeh. The rise of artificial intelligence in healthcare applications. *Artificial Intelligence in Healthcare*, pp. 25–60, Jun 2020. doi: 10.1016/b978-0-12-818438-7.00002-2.

Jordan Chamberlin, Madison R. Kocher, Jeffrey Waltz, Madalyn Snoddy, Natalie F. Stringer, Joseph Stephenson, Pooyan Sahbaee, Puneet Sharma, Saikiran Rapaka, U. Joseph Schoepf, and et al. Automated detection of lung nodules and coronary artery calcium using artificial intelligence on low-dose ct scans for lung cancer screening: Accuracy and prognostic value. *BMC Medicine*, 19 (1), Mar 2021. doi: 10.1186/s12916-021-01928-3.

H.D. Cheng and X.J. Shi. A simple and effective histogram equalization approach to image enhancement. *Digital Signal Processing*, 14(2):158–170, Mar 2004. doi: 10.1016/j.dsp.2003.07.002.

CIHI. Covid-19 hospitalization and emergency department statistics, Feb 2023. URL https://www.cihi.ca/en/covid-19-hospitalization-and-emergency-department-statistics.

Andrea Cozzi, Simone Schiaffino, Francesco Arpaia, Gianmarco Della Pepa, Stefania Tritella, Pietro Bertolotti, Laura Menicagli, Cristian Giuseppe Monaco, Luca Alessandro Carbonaro, Riccardo Spairani, and et al. Chest x-ray in the covid-19 pandemic: Radiologists' real-world reader performance. *European Journal of Radiology*, 132:109272, Nov 2020. doi: 10.1016/j.ejrad.2020.109272.

Maria de la Iglesia Vayá, Jose Manuel Saborit-Torres, Joaquim Angel Montell Serrano, Elena Oliver-Garcia, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, and et al. Bimcv-covid19. Apr 2021. URL https://bimcv.cipf.es/bimcv-projects/bimcv-covid19/. Accessed: 2023-02-17.

Moises Dominguez. Community-acquired pneumonia, Jul 2022. URL https://step2.medbullets.com/infectious-dis/120676/community-acquired-pneumonia. Accessed: 2023-03-24.

Benjamin Fish, Jeremy Kun, and Ádám D. Lelkes. A confidence-based approach for balancing fairness and accuracy. *Proceedings of the 2016 SIAM International Conference on Data Mining*, Jan 2016. doi: 10.1137/1.9781611974348.17.

Frank Gaillard. Normal chest x-ray, Nov 2022. URL https://radiopaedia.org/cases/normal-chest-x-ray. Accessed: 2023-03-24.

GE. Critical care suite 2.0, Mar 2020. URL https://www.gehealthcare.com/products/radiography/critical-care-suite. Accessed: 2023-04-04.

Afshin Gholamy, Vladik Kreinovich, and Olga Kosheleva. Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation. *Semantic Scholar*, 2018.

Dennis Hartmann, Dominik Muller, Iñaki Soto-Rey, and Frank Kramer. Assessing the role of random forests in medical image segmentation. *Image and Video Processing Arxiv*, Mar 2021. doi: 10.48550/arXiv.2103.16492.

Tawfiq Hasanin, Taghi M. Khoshgoftaar, Joffrey L. Leevy, and Richard A. Bauder. Severely imbalanced big data challenges: Investigating data sampling approaches. *Journal of Big Data*, 6(1), Nov 2019. doi: 10.1186/s40537-019-0274-4.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016. doi: 10.1109/cvpr.2016.90.

Aras M. Ismael and Abdulkadir Şengür. Deep learning approaches for covid-19 detection based on chest x-ray images. *Expert Systems with Applications*, 164:114054, Feb 2021. doi: 10.1016/j.eswa.2020.114054.

Ozlem Koksal. How to read chest x-rays, Aug 2018. URL https://iem-student.org/how-to-read-chest-x-rays/. Accessed: 2023-03-24.

Ghazanfar Latif, Hamdy Morsy, Asmaa Hassan, and Jaafar Alghazo. Novel coronavirus and common pneumonia detection from ct scans using deep learning-based extracted features. *Viruses*, 14 (8):1667, Jul 2022. doi: 10.3390/v14081667.

June-Goo Lee, Sanghoon Jun, Young-Won Cho, Hyunna Lee, Guk Bae Kim, Joon Beom Seo, and Namkug Kim. Deep learning in medical imaging: General overview. *Korean Journal of Radiology*, 18(4):570, 2017. doi: 10.3348/kjr.2017.18.4.570.

Graham Lloyd-Jones. Chest x-ray - quality - normal chest x-ray, Oct 2019. URL https://www.radiologymasterclass.co.uk/gallery/chest/quality/chest-x-ray-normal-female. Accessed: 2023-03-24.

David Luong. Pneumonia: Radiology reference article, Jun 2021. URL https://radiopaedia.org/articles/pneumonia. Accessed: 2023-03-24.

Scott Lyon. Ai tool gives doctors a new look at the lungs in treating covid-19, May 2020. URL https://www.princeton.edu/news/2020/05/21/ai-tool-gives-doctors-new-look-lungs-treating-covid-19. Accessed: 2023-03-24.

John Mongan, Linda Moy, and Charles E. Kahn. Checklist for artificial intelligence in medical imaging (claim): A guide for authors and reviewers. *Radiology: Artificial Intelligence*, 2(2), Mar 2020. doi: 10.1148/ryai.2020200029.

Urs J Muehlematter, Paola Daniore, and Kerstin N Vokinger. Approval of artificial intelligence and machine learning-based medical devices in the usa and europe (2015–20): A comparative analysis. *The Lancet Digital Health*, 3(3), Mar 2021. doi: 10.1016/s2589-7500(20)30292-2.

Zaid Nabulsi, Andrew Sellergren, Shahar Jamshy, Charles Lau, Edward Santos, Atilla P. Kiraly, Wenxing Ye, Jie Yang, Rory Pilgrim, Sahar Kazemzadeh, and et al. Deep learning for distinguishing normal versus abnormal chest radiographs and generalization to two unseen diseases tuberculosis and covid-19. *Scientific Reports*, 11(1), Nov 2021. doi: 10.1038/s41598-021-93967-2.

P. Pagliano, C. Sellitto, V. Conti, T. Ascione, and Silvano Esposito. Characteristics of viral pneumonia in the covid-19 era: An update. *Infection*, 49(4):607–616, Mar 2021. doi: 10.1007/s15010-021-01603-y.

Thais Piazza, Daniela Pena Moreira, Hugo André Rocha, Agner Pereira Lana, Ilka Afonso Reis, Marcos Antônio Santos, Augusto Afonso Guerra-Júnior, and Mariangela Leal Cherchiglia. Comorbidities and in-hospital death of viral pneumonia adults admitted to sus (2002–2015). *Revista de Saúde Pública*, 55:43, 2021. doi: 10.11606/s1518-8787.2021055003109.

PIPEDA. Pipeda in brief, May 2019. URL https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/pipeda_brief/. Accessed: 2023-03-14.

Stephen M. Pizer, E. Philip Amburn, John D. Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B. Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, 39(3):355–368, Sep 1987. doi: 10.1016/s0734-189x(87)80186-x.

Philipp Probst, Marvin N. Wright, and Anne-Laure Boulesteix. Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery*, 9(3), Feb 2019. doi: 10.1002/widm.1301.

María Agustina Ricci Lara, Rodrigo Echeveste, and Enzo Ferrante. Addressing fairness in artificial intelligence for medical imaging. *Nature Communications*, 13(1), Aug 2022. doi: 10.1038/s41467-022-32186-3.

Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, and et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. *Nature Machine Intelligence*, 3(3):199–217, Mar 2021. doi: 10.1038/s42256-021-00307-0.

Justin Sirignano and Konstantinos Spiliopoulos. Stochastic gradient descent in continuous time. *SSRN Electronic Journal*, Apr 2017. doi: 10.2139/ssrn.2954149.

SIRM. Covid-19 database. May 2020. URL `https://sirm.org/category/covid-19/`. Accessed: 2023-02-17.

Sajoscha A. Sorrentino. Pneumonia - right middle lobe: Radiology case, Nov 2022. URL `https://radiopaedia.org/cases/pneumonia-right-middle-lobe-1`. Accessed: 2023-03-24.

Melinda Taschetta-Millane. How does covid-19 appear in the lungs?, Oct 2021. URL `https://www.itnonline.com/content/how-does-covid-19-appear-lungs`. Accessed: 2023-03-24.

Zheng Ye, Yun Zhang, Yi Wang, Zixiang Huang, and Bin Song. Chest ct manifestations of new coronavirus disease 2019 (covid-19): A pictorial review. *European Radiology*, 30(8):4381–4389, Mar 2020. doi: 10.1007/s00330-020-06801-0.

Hyunsuk Yoo, Eun Young Kim, Hyungjin Kim, Ye Ra Choi, Moon Young Kim, Sung Ho Hwang, Young Joong Kim, Young Jun Cho, and Kwang Nam Jin. Artificial intelligence-based identification of normal chest radiographs: A simulation study in a multicenter health screening cohort. *Korean Journal of Radiology*, 23(10):1009, Sep 2022. doi: 10.3348/kjr.2022.0189.

Seung Hoon Yoo, Hui Geng, Tin Lok Chiu, Siu Ki Yu, Dae Chul Cho, Jin Heo, Min Sung Choi, Il Hyun Choi, Cong Cung Van, Nguen Viet Nhung, and et al. Deep learning-based decision-tree classifier for covid-19 diagnosis from chest x-ray imaging. *Frontiers in Medicine*, 7, Jul 2020. doi: 10.3389/fmed.2020.00427.