Challenges in COVID-19 Chest X-Ray Classification: Problematic Data or Ineffective Approaches?

Muhammad Ridzuan* MUHAMMAD.RIDZUAN@MBZUAI.AC.AE
Ameera Bawazir* AMEERA.BAWAZIR@MBZUAI.AC.AE
Ivo Gollini Navarrete* IVO.NAVARRETE@MBZUAI.AC.AE
Ibrahim Almakky IBRAHIM.ALMAKKY@MBZUAI.AC.AE
Mohammad Yaqub MOHAMMAD.YAQUB@MBZUAI.AC.AE
Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates

Editors: Under Review for MIDL 2022

Abstract

The value of quick, accurate, and confident diagnoses cannot be undermined to mitigate the effects of COVID-19 infection, particularly for severe cases. Enormous effort has been put towards developing deep learning methods to classify and detect COVID-19 infections from chest radiography images. However, recently some questions have been raised surrounding the clinical viability and effectiveness of such methods. In this work, we carry out extensive experiments on a large COVID-19 chest X-ray dataset to investigate the challenges faced with creating reliable solutions from both the data and machine learning perspectives. Accordingly, we offer an in-depth discussion into the challenges faced by some widely-used deep learning architectures associated with chest X-Ray COVID-19 classification. Finally, we include some possible directions and considerations to improve the performance of the models and the data for use in clinical settings.

Keywords: COVID-19, radiology, X-ray, classification.

1. Introduction

On January 30, 2020, the World Health Organization (WHO) declared a global health emergency due to the *coronavirus disease 2019* (COVID-19) outbreak (Burki, 2020). Five times more deadly than the flu, SARS-CoV-2 viral infection's main symptoms are fever, cough, shortness of breath, and loss or change of smell and taste (Struyf et al., 2021). The fast-paced rise in infections and the rate at which it spread around the globe exposed many challenges with diagnosis and treatment. Access to screening strategies and treatment was minimal due to the lack of resources, especially at the start of the crisis (Coccolini et al., 2021).

Polymerase Chain Reaction (PCR) became the closest gold-standard assay for COVID-19 screening. Nevertheless, the limited number of tests and high rate of false negatives i.e. 100% false negative on infection day, which decreases to 38% on day 5 when first symptoms appear, gave radiographers ground to define chest imaging, not as a routine screening standard, but as an integral tool for assessing complications and disease progression (Inui et al., 2021). Chest imaging is especially necessary for symptomatic patients that develop pneumonia, which is characterized by an increase in lung density due to inflammation and fluid in

[*] Contributed equally

the lungs (Cleverley et al., 2020). The Radiological Society of North America (RSNA) developed a standard nomenclature for imaging classification of COVID-19 pneumonia composed by four categories: negative for pneumonia, typical appearance, indeterminate appearance, and atypical appearance of COVID-19 pneumonia (Simpson et al., 2020).

The presence of Ground Glass Opacities (GGOs) and the extent to which they cover lung regions allows radiologists to diagnose COVID-19 pneumonia in chest radiographs. In such manner, the RSNA classifies a case as “typical” if the GGOs are multifocal, round-shaped, present in both lungs, and peripheral with a lower lung-predominant distribution. In an “indeterminate” case, there is an absence of typical findings and the GGOs are unilateral with a predominant distribution in the center or upper sections of the lung. If no GGO is seen and another cause of pneumonia (i.e. pneumothorax, pleural effusion, pulmonary edema, lobar consolidation, solitary lung nodule or mass, diffuse tiny nodules, cavity) is present, the case is categorized as “atypical” (Litmanovich et al., 2020).

The possibility of using Artificial Intelligence (AI) for aiding the fight against COVID-19 motivated researchers to turn to deep learning approaches, especially convolutional neural networks (CNNs), for the detection and classification of COVID-19 infections (Alghamdi et al., 2021). Many studies have reported high performing classification approaches using Chest X-ray Radiographies (CXRs) (Dongsheng et al., 2021) (Pham, 2021), and Computed Tomography (CT) (Barstugan et al., 2020) (Pathak et al., 2020) (Jia et al., 2021). Despite high reported accuracies by these methods (above 96%), questions have been raised regarding their clinical usefulness due to the bias of small datasets, poor integration of multistream data, variability of international sources, difficulty of prognosis, and the lack of collaborative work between clinicians and data analysts (Roberts et al., 2021). Therefore, in this work we delve deep into the obstacles hindering the development of a clinically viable AI based solution for COVID-19 infection classification from chest radiographs.

In this paper, we utilize available large chest X-ray dataset of COVID-19 patients to train deep learning models that have proven effective on computer vision benchmarks, including deep CNNs, Vision Transformers, and most recently ConvMixer. Following from this, the models are evaluated and the results are analysed to identify potential weaknesses in the models or training approaches. The nature of the data and classes are also analysed keeping in mind the clinical needs for the development of such models. Finally, we present an in-depth discussion into the main challenges associated with this task from the architecture, training approach, and data perspectives.

2. Dataset

The SIIM-FISABIO-RSNA COVID-19 Detection dataset was curated by an international group of 22 radiologists (Lakhani et al., 2021). It includes data from the Valencian Region Medical ImageBank (BIMCV) (de la Iglesia Vayá et al., 2020) and the Medical Imaging Data Resource Center (MIDRC) - RSNA International COVID-19 Open Radiology Database (RICORD) (Tsai et al., 2021). The dataset available for training is composed of 6,054 individual cases (6,334 radiographs), with each case being labelled as negative, typical, indeterminate, or atypical appearance of pneumonia.

Class imbalance is a challenging aspect of this dataset, but it reflects the distribution of cases in reality. The two bigger classes negative and typical account for about 75% of the total number of samples, with 1,676 and 2,855 samples respectively. Indeterminate and atypical samples account for the remaining 25% of samples with 1,049 and 474 samples respectively. In this work, we discuss the impact of this data imbalance on classification performance. The results in this paper are reported on a stratified train-test data split of 80-20% afterwhich the best performing models were tested on a 5-fold stratified cross-validation split. In addition to the SIIM-FISABIO-RSNA dataset, the CheXpert dataset (Irvin et al., 2019) was also used for pre-training. It is composed of 224,316 chest radiographs of 65,240 patients with the presence of 14 chest abnormalities. With such large number of chest radiographs containing various manifestations in the lungs, opicities, pleural effusion, and consolidation, this dataset was chosen for pre-training.

3. Methods

In this work, we trained different deep model architectures to classify each input X-ray image into one of four classes: negative for pneumonia or typical, indeterminate, or atypical for COVID-19. This section details the data preprocessing, augmentations, and various models trained in both supervised and self-supervised approaches.

3.1. Data preprocessing and augmentation

Medical images are inherently different from natural images: radiographs are larger, grayscale, and present similar spatial structures across images. Therefore, not all traditional augmentations are appropriate (Eaton-Rosen et al., 2018). Starting with the most commonly used and clinically-validated data preprocessing and augmentations for chest X-rays, we experimentally determined the best preprocessing and augmentations to be winsorization at 92.5-percentile, horizontal flip, rotation up to ±10 degrees, and scaling up to 20%.

A mirrored lung replacement strategy (generating new images with a mirrored lung that presents GGOs) is proposed (Figure 1a). We also developed a left and right (L/R) lung replacement strategy (replacing the L/R lung with the opposite lung of a different patient within the same class) (Figure 1b). Another approach to tackle the imbalanced categories problem is to perform over and under-sampling of the set. We particularly aim to balance all the classes to the negative class. Hence, we undersample the typical and oversample the atypical and indeterminate class to the size of the negative class. Finally, adding class weight to the loss function allows the model to assign higher weight to the minority classes (Appendix A Table 5).

3.2. Deep convolutional neural network

The baseline was chosen from four CNN architectures to explore the performance of lightweight models like MobileNet (Howard et al., 2017) and EfficientNet (Tan and Le, 2020) against dense models such as ResNet (He et al., 2015) and DenseNet (Huang et al., 2018). DenseNet-121 was selected for comparison and evaluation of the different approaches due to its balance between accuracy and training speed (Appendix A Table 4). The model was trained using the following hyperparameters: image size of 224×224, batch size of 16, cross-entropy loss, ADAM optimizer, and learning rate of 0.001.

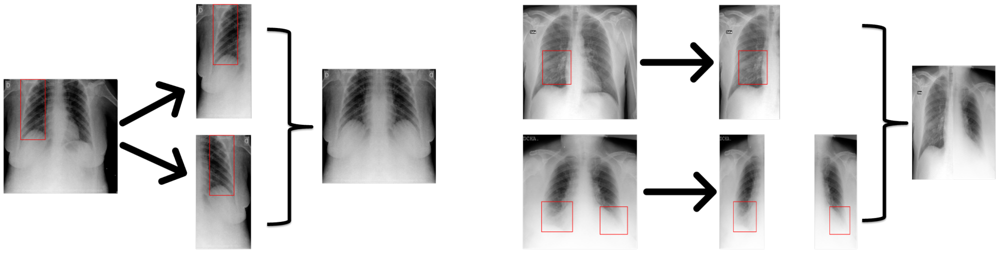


Figure 1: A sample of an X-ray image generated by mirror augmentation strategy in (a) and lung replacement strategy in (b).

3.3. Self-supervised pre-training

Self-supervised pre-training has proven effective in numerous medical tasks with a scarcity of labelled training data. Self-supervised deep CNN models have also been employed to classify COVID-19 cases from chest X-ray images and to deal with the class imbalance problem (Gazda et al., 2021). In this section, we discuss our SSL approaches using deep CNN models on a large unlabelled dataset, CheXpert(Irvin et al., 2019), and then fine-tune the model to classify the above-mentioned four classes.

Adding to the work of (Sowrirajan et al., 2021), we increased the augmentations used in the MoCo architecture by adding horizontal translation, random scaling, and decreasing the color temperature value. Further, we also pre-trained the DenseNet-121 model using the modified MoCo-CXR approach (Sowrirajan et al., 2021), MoCo-V2 (Chen et al., 2020), and MoCo-V2 with balanced data.

Inspired by the work of (Pathak et al., 2016), we explored the impact of focused lung masking inpainting on the model's ability to learn effective representations to identify chest abnormalities. We applied targeted lung masking by approximating its location for both lungs with varying sizes up to 32 × 32. Using this as a pretext task, we substituted the original AlexNet encoder with DenseNet-121, used the Mean Squared Error (MSE) loss, and omitted the adversarial loss to focus on transferability rather than fine reconstruction. Additionally, center inpainting was also explored, where a center 100 × 100 mask is created on the X-ray images, and the model is tasked with reconstructing the original masked region. Figure 5 in Appendix A shows the center mask and left and right targeted lung mask with the reconstructed images.

3.4. Vision Transformer

Vision Transformer (ViT) models (Dosovitskiy et al., 2021) emerged recently to outperform CNN models on many vision tasks including medical ones in some settings (Matsoukas et al., 2021). In this work, we explored the performance of ViT models, where we fed 16 × 16 patches of the chest X-ray images to both pre-trained and fine-tuned ViT models. For pre-training, a wide varieties of X-ray dataset were used including CheXpert (Irvin et al., 2019), tuberculosis radiography data (Rahman et al., 2020), and NIH data (Wang et al., 2017), while SIIM-FISABIO-RSNA dataset is used for fine-tuning. For pre-training, we created a mask over the input image and tasked the ViT with reconstructing the masked region patch.

Table 1: Experimental results for 5-fold cross-validation.

Experiment	F_1 **Score**	**Acc. (%)**
Baseline: DenseNet121	0.4205 ± 0.0149	57.34 ± 2.79
MoCo-V2	0.4583 ± 0.0168	58.60 ± 1.69
L/R targeted inpainting	**0.4794 ± 0.0327**	**61.77 ± 2.28**

3.5. ConvMixer

Exploration of the recent ConvMixer model (Trockman and Kolter, 2022) was also done, where standard convolutions operate directly on the patches as input to achieve mixing between spatial and channel dimensions. We explore various architectures of ConvMixer, ConvMixer-1024/20, ConvMixer-1536/20, and ConvMixer-768/32, with different patches and kernel sizes while starting from ImageNet pre-trained weights and from random initialization.

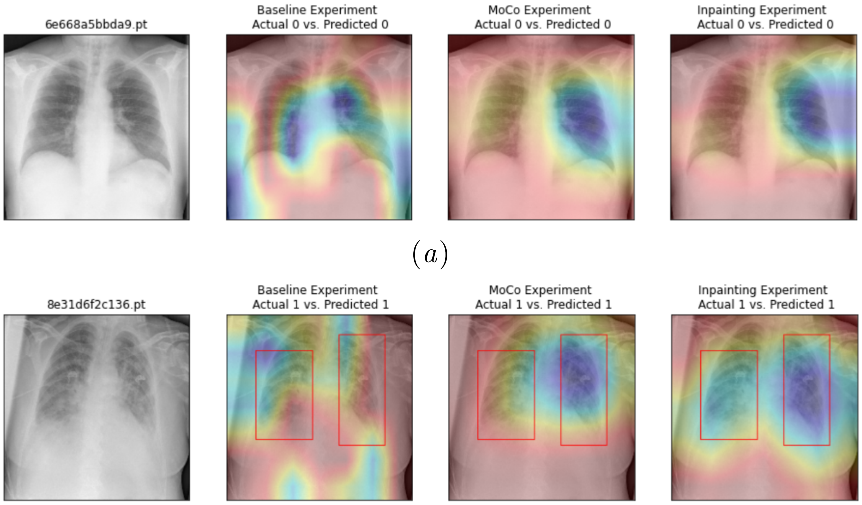
4. Results and Discussion

Tables 1 and 2 summarize our experimental results on the SIIM-FISABIO-RSNA dataset. Assessing the best performing models from the 80-20% split (Table 2) on the 5-fold cross-validation set (Table 1) allows for better judgment surrounding the generalization ability of the model. We also focus on the F_1-macro score to ensure fair comparison between the models considering the unbalanced nature of the classes. Our best performing baseline architecture was DenseNet-121, consistent with its reported success with CXRs in literature (Rajpurkar et al., 2017). Interestingly, the lung augmentation strategies to address class imbalance lead to a deterioration of the performance of the baseline model (Appendix A Table 6). This can be explained with unnatural-looking X-rays that were introduced to the dataset (Appendix A Figure 6).

The best performing model uses the MoCo and inpainting weights pre-trained on CheXpert with an average F1-score of 0.4583 and 0.4794, respectively. ViT performs the worst, likely due to the dissimilarity between ImageNet and CXRs and the lack of a large enough X-ray dataset used for pre-training. For a qualitative comparison of our results, we present the GradCAM (Selvaraju et al., 2016) heatmap outputs of the baseline against the SSL models (Figure 2). The effectiveness of self-supervised learning is evident in that the model is better able to focus on the lung regions when using the MoCo and inpainting pre-trained weights, while the decision-making appears to be more sporadic on the baseline model. This suggests that using a SSL model pre-trained on a larger, related dataset may result in better predictions and fewer false positives than self-supervised. Comparing MoCo and inpainting SSL methods, we have not observed a consistent trend distinguishing the qualitative outputs of the two. Nevertheless, the gains from these methods are still inadequate for practical clinical use in the classification of COVID-19 appearances, with the maximum F_1-macro score not exceeding 0.5. We outline below some challenges with the image, class, and labels of this dataset for consideration.

Table 2: Summary experimental results for 80-20% train-test split.

Experiment	F_1 **Score**	**Acc. (%)**
Baseline: DenseNet121	0.4345	58.19
MoCo-CXR	0.4315	59.54
Modified MoCo-CXR	0.4279	60.41
MoCo-V2	0.4534	58.19
Center inpainting	0.4472	59.62
L/R targeted inpainting	**0.4993**	**62.48**
ConvMixer-1024/20; ks9, p14; pre-trained	0.4478	58.04
ViT base pool cls	0.3152	46.16



(a)

(b)

Figure 2: GradCAM heatmaps for "Negative" (Class 0) and "Typical" (Class 1) appearance of pneumonia (importance increases from red to blue). Bounding boxes show the ground truth radiologists' annotations. More examples on Appendix A Figure 9

.

4.1. Lack of visual cues

Given the complexity of chest X-rays where 3D anatomical features are superimposed in 2D format, and the abstract appearance of diseases, it is difficult even for the experienced radiologists to precisely distinguish different pathological patterns on CXR, particularly due to the presence of ground-glass opacities (Hansell et al., 2008; Bai et al., 2020; Cozzi et al., 2021). The pathologies usually do not have well-defined shapes, sizes, or edges, but rather are characterized by intensity variations and locations relative to other organs.

4.2. Label inconsistencies

An important consideration that has been made in the curation of this dataset is to differentiate between the detection of visual symptoms versus the inference of a disease. For

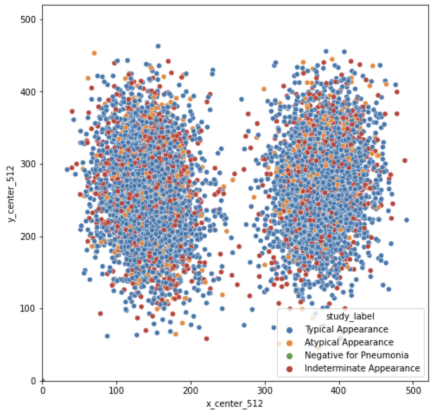


Figure 3: Distribution of bounding box center points after rescaling to 512 x 512. While an assumption is made on the positions of the lungs, the pattern shows a wide distribution in the positions of the bounding boxes.

typical and indeterminate, the descriptions of their appearance suggest that location is a determining factor to distinguish between the classes, where typical is bilateral and primarily found in the lower lung, and indeterminate is unilateral but primarily found in the central or upper lung (Litmanovich et al., 2020). However, some inconsistencies are demonstrated in the labeling of the dataset (Figure 3; Appendix A Figure 7), resulting in uncertainty of the ground truth labels that may hinder the learning of the model and contribute to poorer performance.

4.3. Loaded class

As for the atypical class, the challenge comes from the loaded terminology. "Atypical" is an umbrella term that consists of an array of abnormalities uncommonly reported for COVID-19 pneumonia, including "pneumothorax or pleural effusion, pulmonary edema, lobar consolidation, solitary lung nodule or mass, diffuse tiny nodules, [and] cavity" (Litmanovich et al., 2020). Given the multiple possible appearances of atypical, coupled with the lack of available data belonging to this class (8%), the model may not have adequate samples of each abnormality to learn meaningful representations from. This may be the reason the self-supervised models pre-trained on CheXpert are also able to perform better than fully supervised, especially since CheXpert contains images with some of these abnormalities. For future consideration, it is suggested that such loaded class label is avoided or untangled to preserve the integrity of the ground truths.

4.4. Indeterminate class

It is common in clinical practice to designate "indeterminate", "suspected", or the likes to uncertain cases. According to (Simpson et al., 2020), "indeterminate features are those that have been reported in COVID-19 pneumonia but are not specific enough to arrive at a relatively confident radiologic diagnosis." Typically, such cases warrant further tests or

Table 3: Comparison of binary classifications for COVID-19 pneumonia appearances.

Binary classes	F_1 **Score**	**Acc. (%)**
Typical-Indeterminate	0.5679	73.27
Atypical-Indeterminate	0.5891	67.97
Typical-Atypical	0.5749	82.09

investigation to verify the status and condition of the patient due to insufficient visual cues to confidently reach the true diagnosis.

From a machine learning perspective, the incorporation of such classes is counteracting its need to have highly accurate ground truths for effective loss calculation and performance evaluation. We hypothesized the removal of such class would improve the predictive power of the model. However, Table 3 shows similar performance in the binary classifications of all pairs of positive classes, including typical-atypical; while the accuracy is higher, this is due to class imbalance and the F_1-score remains similar. A recent study suggests that over 25% of COVID-19 patients exhibit co-occurring symptoms, thus blurring the distinction between classes, adding variability and uncertainty to the ground truth labels (Kim et al., 2020).

5. Conclusions and Recommendations

Most SOTA methods have been developed for use with natural images. Our work has shown that the current state of data curation and SOTA machine learning architectures are still insufficient for the accurate classification of COVID-19 pneumonia from CXRs. We have outlined several challenges in creating a viable AI solution for COVID-19 classification. While it is easy to suggest better curation of datasets, such effort is time-consuming and challenging especially with the lack of available resources (Allyn, 2020). As long as manual annotators are involved, there will always be room for errors and subjectivity. It is thus important for the community, both clinicians and machine learning researchers, to acknowledge that unless a golden standard is used to annotate the labels, a 100% accuracy is unlikely and in fact undesired. In the case of COVID-19, a golden standard is yet to exist (Hernández-Huerta et al., 2021).

In terms of machine learning, the challenge surrounding visual cues calls for a shift in approach where effort has to be put in developing methods that focus on intensity variations rather than edge detections. The superiority of DenseNet-121 over other CNNs suggests that a method that exploits and propagates earlier low-level features to later layers in a CNN may be of particular benefit for the development of a successful medical machine learning model.

The presence of co-infections calls for a re-evaluation of the labeling strategy and perhaps the use of multilabel classification for COVID-19 appearances. Additionally, to pave a path to certainty for the indeterminate class, we suggest emulating the clinicians' workflow and integrating the clinical context of the patients along with other relevant information through multimodal learning. Such solution will diffuse the uncertainty in the labelling and have the added benefit of being able to increase the model's explainability and the potential end-user clinicians' confidence.

References

Hanan S. Alghamdi, Ghada Amoudi, Salma Elhag, Kawther Saeedi, and Jomanah Nasser. Deep learning approaches for detecting covid-19 from chest x-ray images: A survey. *IEEE Access*, 9:20235–20254, 2021. doi: 10.1109/ACCESS.2021.3054484.

Jennifer Allyn. International radiology societies tackle radiologist shortage. 2020.

Harrison Bai, Ben Hsieh, Zeng Xiong, Kasey Halsey, Ji Whae Choi, Linh Tran, Ian Pan, Lin-Bo Shi, Dong-Cui Wang, Ji Mei, Xiao-Long Jiang, Qiu-Hua Zeng, Thomas Egglin, Ping-Feng Hu, Saurabh Agarwal, Fangfang Xie, Sha Li, Terrance Healey, Michael Atalay, and Wei-Hua Liao. Performance of radiologists in differentiating covid-19 from viral pneumonia on chest ct. *Radiology*, 296:200823, 03 2020. doi: 10.1148/radiol.2020200823.

Mucahid Barstugan, Umut Ozkaya, and Saban Ozturk. Coronavirus (covid-19) classification using ct images by machine learning methods, 2020.

Talha Burki. Outbreak of coronavirus disease 2019. *The Lancet Infectious Diseases*, 20(3): 292–293, 2020. ISSN 1473-3099. doi: https://doi.org/10.1016/S1473-3099(20)30076-1. URL https://www.sciencedirect.com/science/article/pii/S1473309920300761.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020.

Joanne Cleverley, James Piper, and Melvyn M. Jones. The role of chest radiography in confirming covid-19 pneumonia. *BMJ*, 370, 7 2020. ISSN 1756-1833. doi: 10.1136/BMJ.M2426.

Federico Coccolini, Enrico Cicuttin, Camilla Cremonini, Dario Tartaglia, Bruno Viaggi, Akira Kuriyama, Edoardo Picetti, Chad Ball, Fikri Abu-Zidan, Marco Ceresoli, et al. A pandemic recap: lessons we have learned. *World journal of emergency surgery*, 16(1):1–8, 2021.

Diletta Cozzi, Edoardo Cavigli, Chiara Moroni, Olga Smorchkova, Giulia Zantonelli, Silvia Pradella, and Vittorio Miele. Ground-glass opacity (ggo): A review of the differential diagnosis in the era of covid-19. *Japanese journal of radiology*, 39(8):721–732, 2021.

Maria de la Iglesia Vayá, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, Francisco García-García, Marisa Caparrós, Germán González, and Jose María Salinas. Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients, 2020.

Dongsheng, Zhujun Zhang, Yanzhong Zhao, and Qianchuan Zhao. Research on classification of covid-19 chest x-ray image modal feature fusion based on deep learning. *Journal of Healthcare Engineering*, 2021, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain

Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Zach Eaton-Rosen, Felix J. S. Bragman, Sébastien Ourselin, and M. Jorge Cardoso. Improving data augmentation for medical image segmentation. 2018.

Matej Gazda, Ján Plavka, Jakub Gazda, and Peter Drotar. Self-supervised deep convolutional neural network for chest x-ray classification. *IEEE Access*, 2021.

David M. Hansell, Alexander A. Bankier, Heber Macmahon, Theresa C McLoud, Nestor Luiz Müller, and Jacques Remy. Fleischner society: glossary of terms for thoracic imaging. *Radiology*, 246 3:697–722, 2008.

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

M. T. Hernández-Huerta, L. Pérez-Campos Mayoral, L. M. Sánchez Navarro, G. Mayoral-Andrade, E. Pérez-Campos Mayoral, E. Zenteno, and E. Pérez-Campos. Should rt-pcr be considered a gold standard in the diagnosis of covid-19? *Journal of Medical Virology*, 93(1):137–138, 2021. doi: 10.1002/jmv.26228.

Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017.

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

Shohei Inui, Wataru Gonoi, Ryo Kurokawa, Yudai Nakai, Yusuke Watanabe, Keita Sakurai, Masanori Ishida, Akira Fujikawa, and Osamu Abe. The role of chest imaging in the diagnosis, management, and monitoring of coronavirus disease 2019 (covid-19). *Insights into imaging*, 12(1):1–14, 2021.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, 2019.

Guangyu Jia, Hak-Keung Lam, and Yujia Xu. Classification of covid-19 chest x-ray and ct images using a type of dynamic cnn modification method. *Computers in biology and medicine*, 134:104425, 2021.

D. Kim, J. Quinn, B. Pinsky, N. H. Shah, and I. Brown. Rates of co-infection between sars-cov-2 and other respiratory pathogens. *JAMA*, 323(20):2085–2086, 2020. doi: 10.1001/jama.2020.6266.

Paras Lakhani, John Mongan, Chinmay Singhal, Quan Zhou, Katherine P Andriole, William F Auffermann, Prasanth Prasanna, Tessie Pham, Michael Peterson, Peter J Bergquist, and et al. The 2021 siim-fisabio-rsna machine learning covid-19 challenge: Annotation and standard exam classification of covid-19 chest radiographs., Oct 2021. URL osf.io/532ek.

Diana E Litmanovich, Michael Chung, Rachael R Kirkbride, Gregory Kicska, and Jeffrey P Kanne. Review of chest radiograph findings of covid-19 pneumonia and suggested reporting language. *Journal of thoracic imaging*, 35(6):354–360, 2020.

Christos Matsoukas, Johan Fredin Haslum, Magnus Söderberg, and Kevin Smith. Is it time to replace cnns with transformers for medical images?, 2021.

Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. *CoRR*, abs/1604.07379, 2016. URL http://arxiv.org/abs/1604.07379.

Yadunath Pathak, Prashant Kumar Shukla, Akhilesh Tiwari, Shalini Stalin, and Saurabh Singh. Deep transfer learning based classification model for covid-19 disease. *Irbm*, 2020.

Tuan D Pham. Classification of covid-19 chest x-rays with deep learning: new models or fine tuning? *Health Information Science and Systems*, 9(1):1–11, 2021.

Tawsifur Rahman, Amith Khandakar, Muhammad Abdul Kadir, Khandaker Rejaul Islam, Khandakar F. Islam, Rashid Mazhar, Tahir Hamid, Mohammad Tariqul Islam, Saad Kashem, Zaid Bin Mahbub, Mohamed Arselene Ayari, and Muhammad E. H. Chowdhury. Reliable tuberculosis detection using chest x-ray with deep learning, segmentation and visualization. *IEEE Access*, 8:191586–191601, 2020. doi: 10.1109/ACCESS.2020.3031384.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Yi Ding, Aarti Bagul, Curtis P. Langlotz, Katie S. Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *CoRR*, abs/1711.05225, 2017. URL http://arxiv.org/abs/1711.05225.

Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, 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, 2021.

Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016. URL http://arxiv.org/abs/1610.02391.

Scott Simpson, Fernando U Kay, Suhny Abbara, Sanjeev Bhalla, Jonathan H Chung, Michael Chung, Travis S Henry, Jeffrey P Kanne, Seth Kligerman, Jane P Ko, et al. Radiological society of north america expert consensus document 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, 2020.

Hari Sowrirajan, Jingbo Yang, Andrew Y. Ng, and Pranav Rajpurkar. Moco-cxr: Moco pretraining improves representation and transferability of chest x-ray models, 2021.

T Struyf, JJ Deeks, J Dinnes, Y Takwoingi, C Davenport, MMG Leeflang, R Spijker, L Hooft, D Emperador, J Domen, SR A Horn, and A Van den Bruel. Signs and symptoms to determine if a patient presenting in primary care or hospital outpatient settings has covid-19. *Cochrane Database of Systematic Reviews*, (2), 2021. ISSN 1465-1858. doi: 10.1002/14651858.CD013665.pub2. URL https://doi.org//10.1002/14651858.CD013665.pub2.

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.

Asher Trockman and J. Zico Kolter. Patches are all you need?, 2022.

Emily B Tsai, Scott Simpson, Matthew P Lungren, Michelle Hershman, Leonid Roshkovan, Errol Colak, Bradley J Erickson, George Shih, Anouk Stein, Jayashree Kalpathy-Cramer, et al. The rsna international covid-19 open radiology database (ricord). *Radiology*, 299 (1):E204–E213, 2021.

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. doi: 10.1109/cvpr.2017.369. URL http://dx.doi.org/10.1109/CVPR.2017.369.

Appendix A.

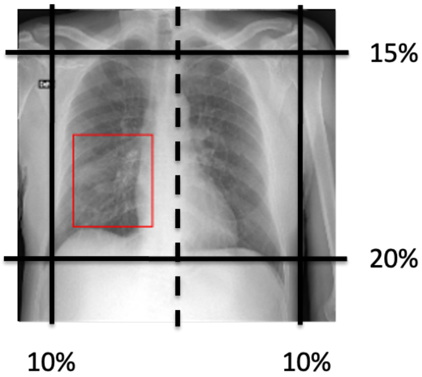


Figure 4: Mask region constraint for targeted L/R inpainting. This is performed by posing the following constraints: 10% from the left and right, 15% from the top, and 20% from the bottom of the chest X-rays.

Table 4: Comparison of CNN architectures to define baseline.

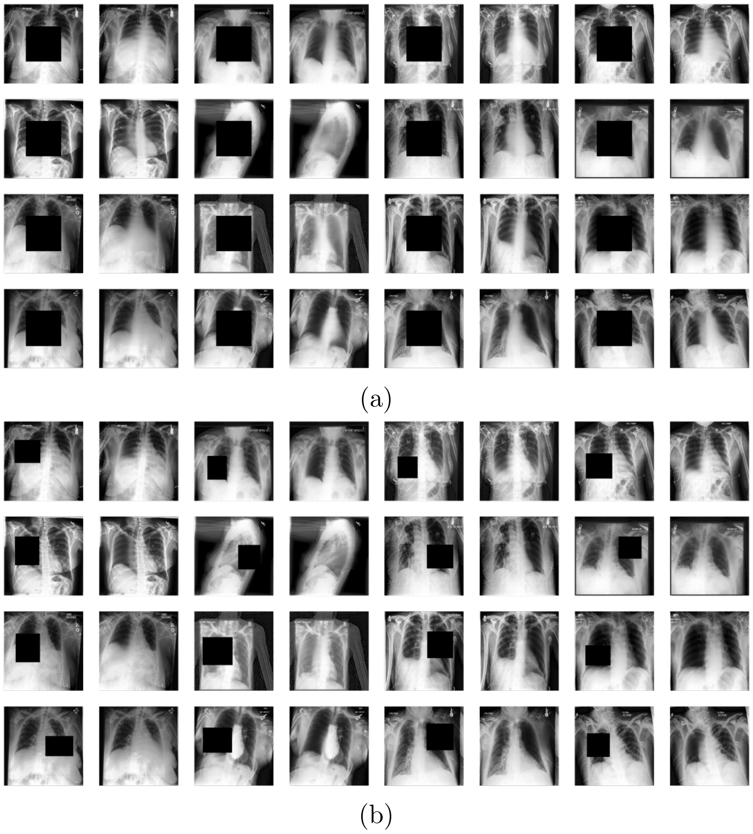
Experiments	F_1 **Score**	**Acc. (%)**
MobileNet	0.3356	57.18
EfficientNet	0.3434	**61.47**
ResNet50	0.1617	47.8
DenseNet121	**0.4345**	58.19

Table 5: Comparison of baseline and augmentation experiments to address data imbalance. The baseline and class weight experiments were performed on the original imbalanced dataset, while the lung and mirror replacement experiments were performed on the augmented balanced dataset.

Experiments	F_1 **Score**	**Acc. (%)**
Baseline: DenseNet121	**0.4345**	**58.19**
Class weights	0.4243	55.42
Lung replacement	0.4213	56.53
Mirror replacement	0.4243	55.42

Table 6: Comparison of mirror and lung replacement strategies for the baseline, MoCo-V2 and L/R targeted inpainting experiments.

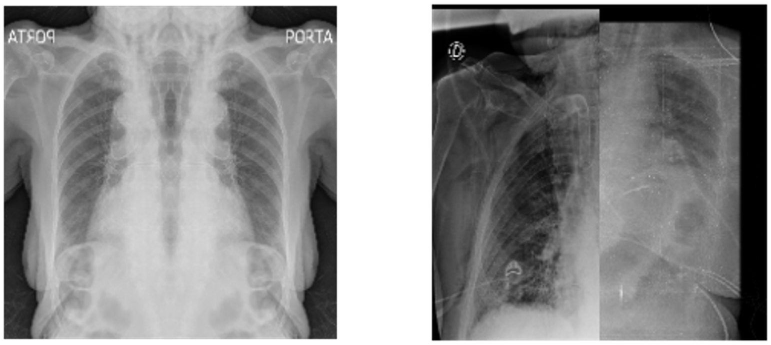
Experiments	F_1 **Score**
Baseline: DenseNet121	**0.4345**
Baseline + mirror aug	0.4243
Baseline + lung replacement	0.4213
MoCo-V2	**0.4534**
MoCo-V2 + mirror aug	0.4236
MoCo-V2 + L/R aug	0.3908
L/R targeted inpainting	**0.4740**
L/R targeted inpainting + mirror aug	0.4441
L/R targeted inpainting + L/R aug	0.4593



(a)

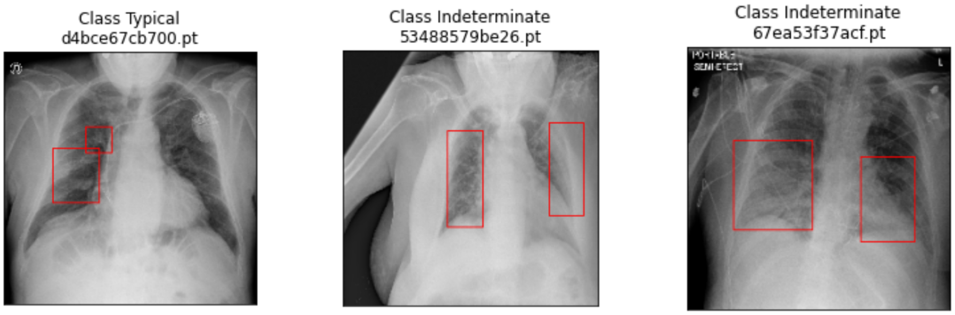
(b)

Figure 5: Visualization of inpainting self supervised pre-training model output, showing image reconstruction from center mask (a) and targeted left and right mask (b).



(a) (b)

Figure 6: Failures in (a) mirrored lung augmentation and (b) lung replacement augmentation.



(a) (b) (c)

Figure 7: Counter-description examples of CXR COVID-19 findings: (a) unilateral opacity classified as typical; (b) bilateral opacity classified as indeterminate; (c) bilateral, lower-to-central lung opacity classified as indeterminate.

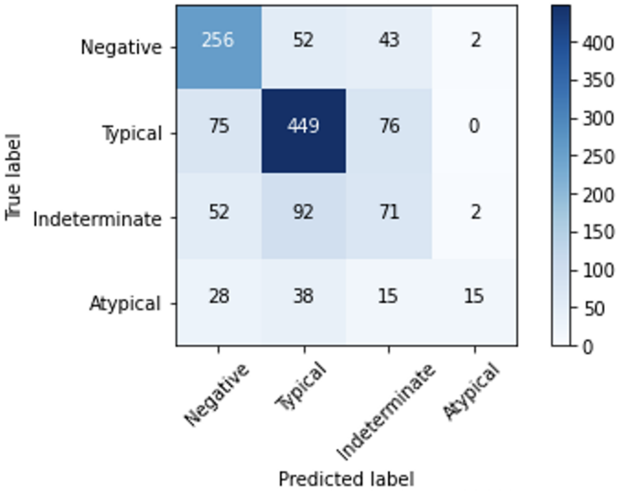


Figure 8: Confusion matrix for the best-performing inpainting model.

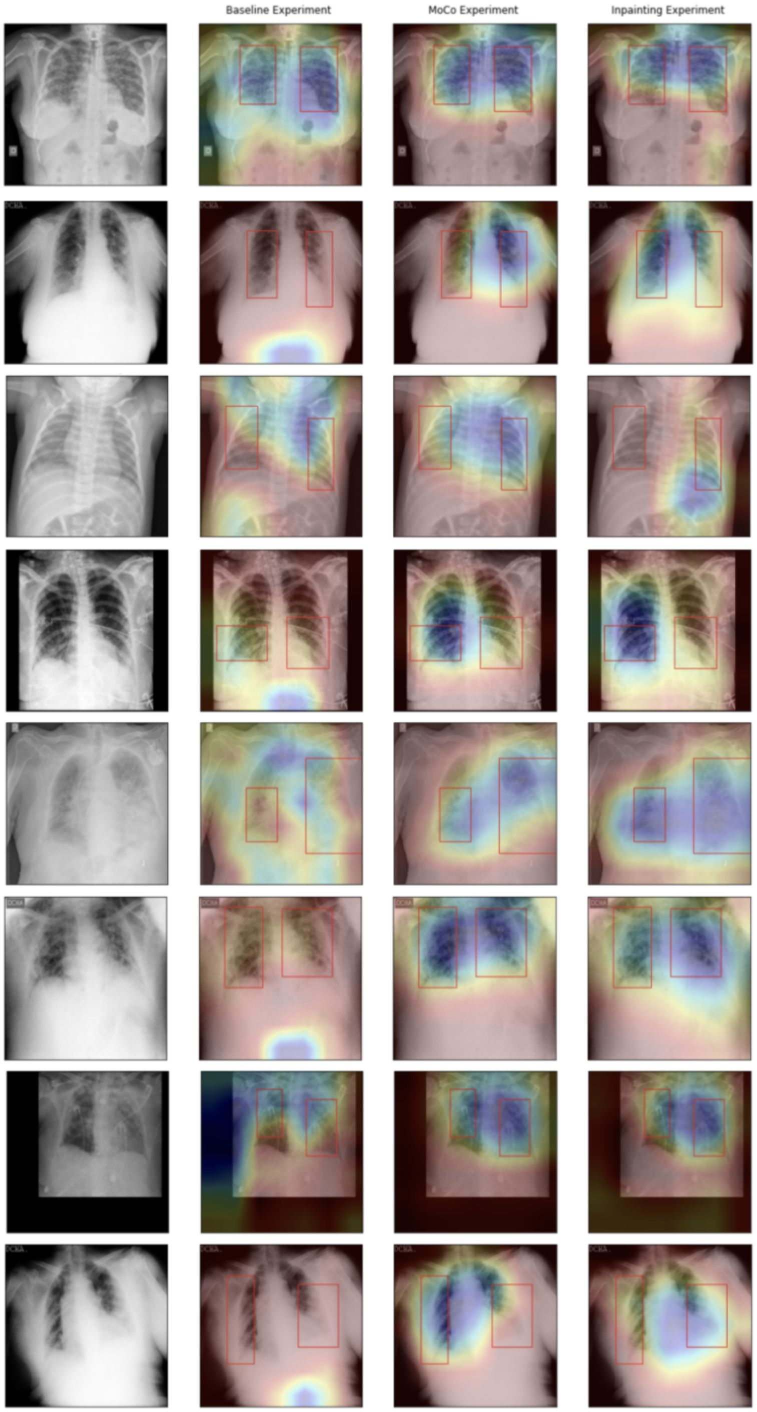


Figure 9: Further GradCAM heatmap outputs for baseline, MoCo and inpaiting experiments. Importance increases from red to blue. Bounding boxes show the groundtruth radiologists' annotations.