Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 18(6): 9264–9293.
DOI: 10.3934/mbe.2021456
Received: 04 August 2021
Accepted: 12 October 2021
Published: 27 October 2021

Research article

Factors determining generalization in deep learning models for scoring COVID-CT images

Michael James Horry[1], Subrata Chakraborty[1,*], Biswajeet Pradhan[1,2,3,*], Maryam Fallahpoor[1], Hossein Chegeni[4] and Manoranjan Paul[5]

[1] Center for Advanced Modelling and Geospatial Information Systems (CAMGIS), Faculty of Engineering and Information Technology, University of Technology Sydney, Australia
[2] Center of Excellence for Climate Change Research, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[3] Earth Observation Center, Institute of Climate Change, Universiti Kebangsaan Malaysia, Selangor 43600, Malaysia
[4] Fellowship of Interventional Radiology Imaging Center, IranMehr General Hospital, Iran
[5] Machine Vision and Digital Health (MaViDH), School of Computing, Mathematics, and Engineering, Charles Sturt University, Australia

* **Correspondence:** Email: subrata.chakraborty@uts.edu.au, biswajeet.pradhan@uts.edu.au.

Abstract: The COVID-19 pandemic has inspired unprecedented data collection and computer vision modelling efforts worldwide, focused on the diagnosis of COVID-19 from medical images. However, these models have found limited, if any, clinical application due in part to unproven generalization to data sets beyond their source training corpus. This study investigates the generalizability of deep learning models using publicly available COVID-19 Computed Tomography data through cross dataset validation. The predictive ability of these models for COVID-19 severity is assessed using an independent dataset that is stratified for COVID-19 lung involvement. Each inter-dataset study is performed using histogram equalization, and contrast limited adaptive histogram equalization with and without a learning Gabor filter. We show that under certain conditions, deep learning models can generalize well to an external dataset with F1 scores up to 86%. The best performing model shows predictive accuracy of between 75% and 96% for lung involvement scoring against an external expertly stratified dataset. From these results we identify key factors promoting deep learning generalization, being primarily the uniform acquisition of training images, and secondly diversity in CT slice position.

Keywords: COVID-19 scoring; computed tomography; deep learning; external validation; model generalization; image pre-processing

1. Introduction

The COVID-19 pandemic remains a serious worldwide problem with around 2.6 million deaths attributable to this disease as of March 11, 2021 [1]. Whilst vaccination efforts are now underway, the virus is still circulating, and many countries remain locked down with new variants of the disease recently emerging [2]. Pandemics such as COVID-19 are considered by expert panels to be a recurring threat with the expectation that respiratory viruses causing severe pneumonia will continue to mutate and emerge [3].

From the start of this pandemic around December 2019 the research community has worked tirelessly to detect, measure, and treat COVID-19 and its devastating health and economic consequences more effectively. Data scientists and medical AI specialists have recognized the promise of automated computer vision diagnosis and stratification of COVID-19 from medical images [4] including computed tomography (CT) [5–17], Chest X-ray [18–30] and point of care ultrasound [31,32] imaging modes. These computer vision methods almost universally employ a trainable convolutional neural network (CNN) as an image feature extractor/classifier, in a process referred to as "deep learning". The CT imaging mode has received particular focus following results of a large study early in the COVID-19 pandemic showing that chest CT outperformed polymerase chain reaction (PCR) based lab testing in the diagnosis of COVID-19 [33]. CT imaging also has the advantage of showing the extent of a patient's COVID-19 pneumonia lung involvement, thereby providing information about infection progression and severity [34] which can be used to appropriately manage the patient and clinical resources for best outcomes.

Despite this research effort, applications of computer vision assisted diagnosis and stratification of COVID-19 has not yet reached maturity due to concerns such as dataset bias and limited validation against external datasets. These concerns are justified when one considers the many possible sources of bias that could be introduced by systematic differences in medical image acquisition pipelines, apparatus and operating methods, or differences in patient morphology between sample and target populations [35]. Other potential sources of sample bias include differences in the down-sampling method employed to convert high resolution medical images in the form of multiple slices contained in digital imaging and communications in medicine (DICOM) files into smaller numbers (often a single slice) of lower resolution compressed images suitable for training deep learning models [35]. These concerns are compounded by disease positive and disease negative samples frequently being independently sourced in the publicly available COVID-19 CT image corpuses.

A comprehensive meta-analysis of published papers and preprints from January 1, 2020 to October 2020 conducted by [35] found that of the 2212 papers initially identified, only 62 were of sufficient quality for systematic review and of these none were of possible clinical use due to serious problems with methods or underlying bias as assessed using the PROBAST [36] tool. Lack of external testing along with biases introduced by differential contribution sources for disease positive and disease negative image sets were highlighted as particular causes of concern.

This study seeks to provide a clearer understanding of the factors affecting deep learning model generalization in relation to COVID-19 detection from CT images. Identifying and addressing these factors is critical to translating the immense body of research in this field into well engineered clinical tooling. We present the results and interpretation of an inter-dataset study across four publicly available COVID-19 CT image datasets and one high-quality privately curated image dataset which is included as an unbiased data corpus of known provenance. The models described in the dataset source studies are closely replicated using a consistently trained off-the-shelf Densenet-121 convolution neural network (CNN). This creates a comparable base set of trained computer vision COVID-19 diagnostic models.

Systematic differences in brightness and contrast between disease positive and disease negative datasets are potential confounding variables in deep learning training. For this reason, application of histogram equalization is a very commonly used pre-processing technique applied to computer vision algorithms with the objective of removing these differences and promoting generalization [38]. To investigate the effectiveness and limits of these techniques, each COVID-19 diagnostic model is assessed against each other's test data partition with histogram equalization [39] and contrast limited adaptive histogram equalization (CLAHE) [40] image pre-processing techniques applied, these being the most employed image histogram equalization methods [41]. Each test is repeated with the addition of a learning Gabor filter [42] which has been shown by some studies to improve CNN accuracy in thoracic imaging applications by means of feature reinforcement. From these tests we gain insights into the ability of each model to generalize to external datasets, along with an indication of the usefulness of image histogram pre-processing and application of a learning Gabor filter.

The predictive capability of each trained model is then assessed against an independent dataset for which COVID-19 lung disease involvement has been measured and labelled by expert radiologists [43]. This final step reveals which training corpus and pre-processing options result in trained deep learning models that are best able to predict COVID-19 lung involvement scores and therefore be considered successfully generalized. Finally, we combine our two best models and use a min-max algorithm to achieve very high predictive scores for lung involvement against the expertly labelled control data set. Our discussion considers factors promoting generalization which should be considered in the development of generalizable deep learning computer vision models for COVID-19 and other thoracic diseases.

This study represents the first systematic investigation into factors affecting the generalization of models trained on COVID-19 CT datasets. Our investigation into the effect of commonly employed image pre-processing/filtering techniques on model generalization in the context of CT imaging and machine learning is also unique. The outcome of this investigation leads us to promising and, to our knowledge, state-of-the-art results in applying a deep learning model trained on one dataset to successfully inference the lung involvement strata for a completely independent dataset. Our method for combining models using differentiated pre-processing techniques for disease positive and disease negative prediction is novel and foreseeably applicable to a range of computer vision applications.

2. Related work

Although there are many published studies in the use of computer vision in the diagnosis and stratification of COVID-19 very few of these investigate the inter-dataset generalization behavior of the described models. In our review we came across only two studies that externally validated the performance of their model against an independent dataset [13,44]. In [13] a cross dataset generalization study into COVID-19 CT deep learning models by Silva et al. exhibited a drop in accuracy for the presented model from 87.68% to 56.16% concluding that the generalization power of the deep learning models under consideration is "far from acceptable" for a realistic scenario. In contrast Guarrasi et al. [44] achieved good external generalization results for the CXR imaging mode using an optimized combination of deep learning classifier ensembles with fused output layers. The authors of this study attribute their good generalization results to increased diversity in CNN architectures over single network architectures, resulting in better sensitivity to disease positive image features.

The question of deep learning generalization in the context of automated pneumonia detection from the chest X-ray imaging mode was comprehensively investigated by Zech et al. [45]. This study trained and evaluated a CNN-based deep learning model for pneumonia detection on data from three independent corpuses. It was found that internal model performance was superior to external model performance in all three single dataset comparisons. Where datasets were combined into supersets, the superset trained models outperformed the single dataset trained models–however that performance improvement did not generalize to a third dataset that was not included in the training data supersets. Interestingly, the discussion in this study is focused on AUC scores declining in generalization tests by 3–10% on the single dataset tests but does not pay particular attention to the decline in specificity associated with these tests – being in the range of 3–34% in single dataset tests and up to 47% on the superset trained models against independent dataset. Sensitivity results for the experiments remained in the 95–98% range throughout. Since specificity is a measure of the ability of the model to distinguish disease negative samples from the full sample set, it follows that the poor generalization results from this study flow from high numbers of false positive predictions on the external test set. In other words, the CNN models tested were very poor at separating disease negative sample and tended to classify all samples as disease positive on external testing. This study notes that CNNs may learn so called "confounding factors" being non-pathological variables associated to image classes, for example sample images from portable scanners are more likely to be disease positive–since these machines are usually used at the bedside of more seriously ill patients [45]. This study concludes that "Estimates of CNN performance based on test data from hospital systems used for model training may overstate their likely real-world performance".

The effect of confounding factors is of particular concern in deep learning COVID-19 studies where disease positive and disease negative image samples have been separately sourced. In a systematic survey of COVID-19 machine learning detection algorithms Roberts et al. [35] notes that no studies reach the threshold of robustness and reproducibility to support clinical use, due in large to limitations of public datasets used such as traceability/confirmation of COVID-19 diagnosis, differential DICOM extraction and compression image pre-processing resulting in a lack of consistency and comparability. Concerningly, this review found that 16 out of 62 papers had obliviously used a pediatric pneumonia dataset [46] as a control group for separation from COVID-19 image samples, with Maguolo et al. [47] demonstrating separation of the Cohen COVID-19 dataset [48] from this control group achieving AUC up to 0.9997 with the lung field entirely removed from images, thereby representing a perfectly confounded study.

Such concerns in relation to the CXR imaging mode led DeGrave et al. [49] to hypothesize that COVID-19 dataset-specific confounding would lead to poor external generalization results, finding both pre-trained deep learning models, and simpler logistic regression based models, to be equally affected by confounding when using the Cohen COVID-19 dataset [48] against the NIH ChestX-ray14 [50] dataset "No Finding" and "Pneumonia" subsets as a control. This study also found strong evidence of confounding even in higher quality datasets where disease positive and disease negative image samples were sourced from the same region [51] and even from the same clinical institution [52]. The confounding cause in this study was identified using saliency mapping techniques, and included radiological text markup at image edges, as well as the appearance of the diaphragm and cardiac silhouette differing due to differences in radiographic projection. Other studies into the use of the CXR imaging mode for COVID-19 diagnosis, such as [53,54] have reached the same conclusion as [49] showing classification metrics gaps of up to 62% under external testing.

In the early days of a novel pandemic such as COVID-19, any large corpus of medical images suitable for the training of deep learning systems will most likely consist of relatively low-quality images contributed from multiple sources and/or scraped from medical journals. Techniques such as lung field segmentation [54] and dataset augmentation [55] have been proposed as potential techniques to improve the signal to noise ratio of these images and thereby improve generalization of deep learning models to unseen data. Using the Cohen COVID-19 dataset, Bassi et al. [54] achieved a mean classification accuracy improvement in external testing of around 4.7% by removing signal outside the lung field. The effect of geometric image augmentation techniques frequently used to reduce overfitting during training by increasing dataset size and variability was tested by Elgendi et al. [55]. Commonly used augmentations include rotation, scaling, reflection, shear, and translation. This study found that augmentation methods were not effective at promoting generalization, and in fact reduced the external generalization capabilities for the tested models. Since CT scans tend to be more uniform than CXR images, the data augmentations in this study are limited to 1-degree of rotation to allow for minor patient position variability, and horizontal flip to augment image volume and better distribute COVID-19 lung involvement signal over both lungs.

3. Materials and methods

3.1. Dataset selection

A comprehensive survey of COVID-19 data sources was produced by Shuja et al. [56] and published in September 2020. This survey was used to identify appropriate data and models for the study. Of these datasets the Cohen dataset [48] and Jun et al. [57] were too small for consideration for this study with only 80 images and 20 CT scans respectively at the time of writing (28 Feb 2021). Wang et al. [14] and Shan et al. [58] with 435 CT images and 549 CT scans respectively were large enough for this study but excluded since the datasets are not publicly available.

The Zhao dataset [60] consisting of 812 CT images that have been scraped from the medRxiv [61] health sciences, and bioRxiv [62] biology preprint archives. This dataset contains a total of 812 CT images, of which 349 images from 216 patients are labelled COVID-19 and 463 images from only 84 patients are labelled as Non-COVID-19. The dataset is accompanied by an associated study that presents a CNN based deep learning model that achieves accuracy of 89% and an F1 measure of 90%. The size of this dataset and an accompanying deep learning model study makes this a good candidate data set for our purposes and is henceforth referred to as the “COVID-CT” dataset in this paper.

Subsequent to the review provided by Shuja et al. [56] two further COVID-19 CT studies have become available, both of which contain sizable datasets with an accompanying CNN based deep learning study focused on COVID-19 diagnosis. Firstly, Soares et al. [63] consisting of 2482 CT scans of which 1252 are COVID-19 positive and 1230 are COVID-19 negative from patients who had presented for other pulmonary diseases. Images were collected from hospitals in Sao Paulo, Brazil. This dataset is accompanied by a study that uses a handcrafted deep learning network achieving an accuracy of 97.38% and an F1 score of 97.31%. The data from this study is publicly available and ideal for our experiments. This dataset is henceforth referred to as the “SARS-COV-2” dataset in this paper. Secondly, Gunraj et al. [7] consisting of 194,922 images from 4,501 patient cases initially derived from CT imaging data collected by the China National Center for Bioinformation (CNCB) [64], but augmented with images from other data sources. The 2A release of this dataset contains 92,268 COVID-19 positive and 30,749 Normal images with the remainder being labelled as non-COVID-19 pneumonia. The dataset is accompanied by a study that utilizes a handcrafted CNN based network utilizing skip-connections. From the published confusion matrix, we can calculate that this network achieved an average accuracy of 99.1% with an F1 score of 99.0% across the three classes. The paper notes in discussion that a generalization study would be needed for this model to be considered for clinical use, as well as a better understanding of how the classifier was able to distinguish between COVID-19 pneumonia cases and other pneumonia cases. The COVID-19 and Normal labelled images are useful for comparison with the other datasets described above and have been included in this study. This subset is henceforth referred to as the “COVIDNet-CT-2A” dataset in this paper.

A large dataset sourced from hospitals in Moscow has recently been made available by Morozov et al. [43]. This dataset contains 1110 CT scans with COVID-19 positive and negative examples. The dataset is unique in that each patient has been assessed by expert radiologists and labelled according to COVID-19 lung involvement stratified at 25% intervals as CT-0 to CT-4 with CT-0 representing 0% lung involvement and CT-4 representing 75–100% lung involvement. This dataset is used as a control in the second experiment of our study to determine the strength of our model’s ability to predict lung involvement in an independent dataset for which involvement stratification labelling is available. Our hypothesis is that a well generalized model should show an increase in disease positive classification score as the level of lung involvement increases, with average predicted scores for the stratified images increasing in accordance with the strata labels. This dataset is henceforth referred to as the “Moscow” dataset in this paper.

Our study introduces one private dataset to provide a very controlled corpus with clear provenance for the purpose of assessing factors affecting generalization performance. This dataset comprises 626 COVID-19 positive images from 626 patients, and 619 COVID-19 negative images from 619 patients. These images were extracted from the Mehr Hospital, Iran PACS before being deidentified, with the middle CT slice extracted, compressed into a 512 × 512 PNG image file, and zoomed in by 40% to remove the circular scan field mask to make the images structurally like the other datasets. Images were originally acquired from a GE Medical Brightspeed 16 detector multislice CT scan machine using a low dose spiral high-resolution CT scan. The source CT scans were independently assessed and matched by two trained and certified radiologists. This dataset is referred to as the “MID-CT” dataset in this paper.

A summary of the qualitative attributes of these selected datasets appears in Supplementary Table A.

An investigation was undertaken to understand the variety of sources underlying each dataset with the key observation being that the datasets do not always have COVID-19 disease positive and disease negative sample images co-sourced. As discussed in the introduction, this would be highly problematic if these independent source provenances were to impart systematic bias upon the dataset. Such a systematic difference would produce artificially high classification metrics for internal testing, but poor results in external testing. Attributes of the datasets that we would expect to be problematic in this manner include:

1) Non-uniform image size, since resizing images to 244 × 244 pixels to suit classifier input requirements will non-uniformly distort the source images due to different source image aspect

ratios. This issue is notable in the COVID-CT and COVIDNetCT-2A datasets.

2) Inconsistent image resolution–for example images that have been extracted from PDF sources tend to be lower resolution than CT slice extracts from DICOM source files. This issue is notable in the COVID-CT and COVIDNetCT-2A datasets.
3) Inconsistent compression artefacts–where disease positive and disease negative samples come from different sources it is important that compression algorithms match to avoid classifiers learning to separate classes based on systematic differences in compression artefacts. This issue is particularly notable on the COVID-CT dataset.
4) Patients' samples with non-COVID-19 pneumonia diagnosis, since the typical radiological findings for COVID-19 pneumonia can be seen in some non-COVID-19 pneumonia [65]. With images downsized to only 244×244 pixels it is questionable that these images still contain enough information for a classifier to separate on the subtle differences between these two types of pneumonia. This is a consideration only for the COVIDNet-CT-2A dataset. For the purposes of our investigation, we have removed the non-COVID-19 pneumonia images from this dataset to support binary classification allowing for comparison with other datasets and models.
5) Inconsistent CT slicing. Some CT slices show very little lung field since due to the image being dominated by the heart as shown in Figure1. These images therefore contain very little, perhaps no COVID-19 disease related pathological information. CT scans represent a moving image locus down the body, and the image properties vary greatly by slice location. A fair training data set would select the same slice, or slices, for each patient for each class to avoid having the classifier learning to separate classes by anatomical features related to slice position rather than pathological finding. In the case of pulmonary disease, the selected slice or slices should maximize the view of lung field and minimize the obfuscation of the image by the diaphragm at lower CT slices, however images such as those shown in Figure 1 (extracted from the COVIDNet-CT-2A dataset) are very common in the publicly available datasets. This is a feature that is present to some extent in each of the datasets except but is particularly notable in the COVIDNetCT-2A dataset, which includes a variable number of CT slices for some patients, many including such obfuscated views.

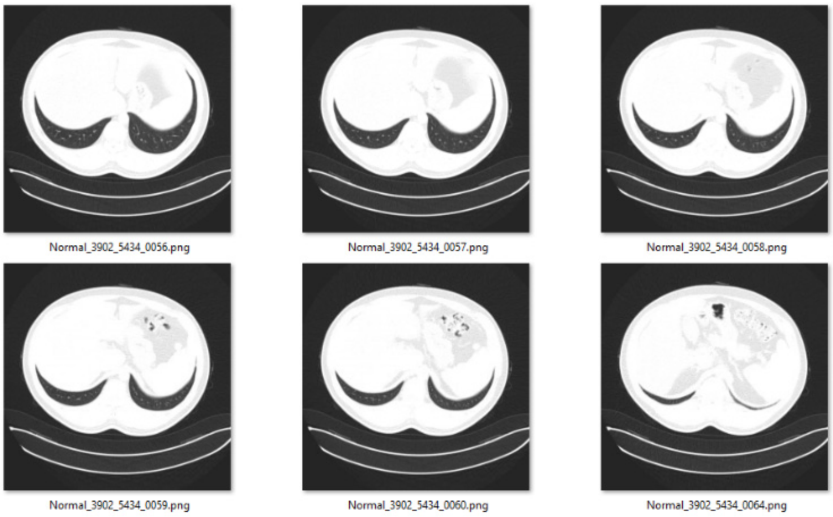
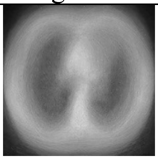
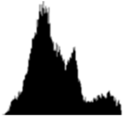
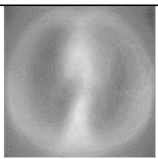
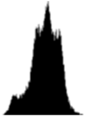
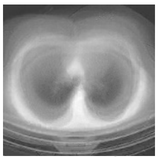
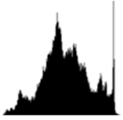
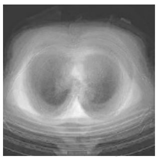
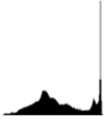
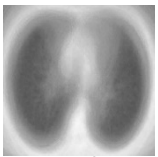
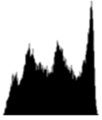
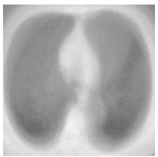
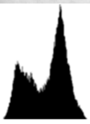
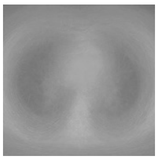
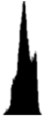
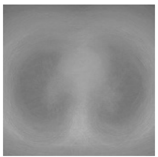
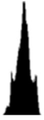
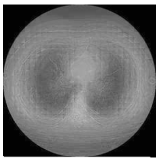
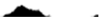
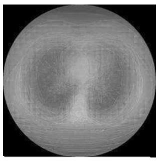
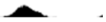

Normal_3902_5434_0056.png
Normal_3902_5434_0057.png
Normal_3902_5434_0058.png
Normal_3902_5434_0059.png
Normal_3902_5434_0060.png
Normal_3902_5434_0064.png

Figure 1. Sample CT images with lung field obfuscated reducing disease positive signal.

Table 1. Average image composites and observations for selected datasets.

Dataset (chi-squared)	COVID Negative Composite Image	COVID Positive Composite Image	Dataset Summary
COVID-CT (7.88)	 	 	Multi-sourced from scraped PDF files. Clear differences in brightness and contrast between disease positive and disease negative classes apparent in average composite images. Histogram comparison shows different frequency distributions between classes with disease negative composite showing a much broader frequency spectrum.
COVIDNet-CT-2A (14.946)	 	 	Multi-sourced from a number of daasets. Some differences in brightness and contrast between disease positive and disease negative classes apparent in average composite images. Histogram comparison different between classes, with the disease negative composite showing higher amplitude across the frequency spectrum resulting in a very high chi-squared value. CT scan artefacts displaying as arcs in lower part of image not present in other datasets.
SARS-COV-2 (7.383)	 	 	Sourced from multiple Brazillian clinics but acquired and processed centrally. Brightness and contrast are a systematic match across classes. Histogram comparison relatively similar in amplitide between classes, with disease negative composite showing a slightly broader spectrum.
MID-CT (1.490)	 	 	Single sourced from Iranian Hospital using consistent CT scanning apparatus and acquisition pipeline. Zoomed from 512×512 to 350×350 and centered to remove circular mask from extracted middle slice. Brightness and contrast systematic match across classes. Histograms almost identical between classes with low chi-squared value.
MOSCOW (0.139)	 	 	Sourced from multiple Moscow based clinics but acquired and processed centrally. Directly extracted from Source DICOM image with Circular image frame shared only with MID-CT. Histogram almost identical between class with very low chi-squared value. Compression artefacts are consistent across classes, but reinforced by averaging and clearly visible on composite image.

It should also be noted that COVIDNet-CT-2A is a superset of images that contains positive samples from the Moscow dataset and negative samples from the COVID-CT dataset, and is therefore initially polluted for the purposes of our generalization study. In assembling the COVIDNet-CT-2A dataset we have removed the COVID-CT and Moscow image samples to remove cross-contamination of the datasets to allow for an unpolluted generalization test. This has the effect that a direct comparison between the internally trained COVIDNet-CT-2A results and ours is invalid although our internal testing results remain very close those of the source study.

Ideally, any systematic variations across datasets of the disease positive and disease negative images would be caused by pathology features only and thereby restricted to brightness differences in the lung field. To test this, we created an average composite image for each dataset and class to provide a visual similarity analysis across the classes. This was achieved by looping through the image files for each class, and recursively joining pairs of images using a weighting function to ensure that every image in the sequence was equally represented in the average composite image. The image frequency histogram for each composite image was calculated and charted in Table 1. Whilst the composite images/histograms for Moscow, MID-CT and SARS-COV-2 datasets are very consistent across classes, both COVID-CT and COVIDNet-CT-2A exhibit significant differences for the different classes. In the case of COVID-CT the corners of the composite image have very different pixel values, with the negative classes being black and the positive classes being a light grey. The image histogram for the COVID-CT disease negative class covers a much broader spectrum than that for the disease positive class. For COVIDNet-CT-2A the average image histograms for the two classes cover a similar frequency spectrum, but the disease positive class amplitude is much lower than that on the disease negative class histogram. From these systematic differences visible between classes in these datasets we hypothesize that deep learning models trained on COVID-CT and COVIDNet-CT-2A are most likely to train on non-pathological features which will limit external generalization.

There are notable significant differences in zoom level and CT image acquisition artefacts between the datasets as evident in Table 1. For example, the MID-CT, COVID-CT, and SARS-COV-2 datasets are both zoomed to maximize the lung field in the frame. The Moscow images appear within a circular mask and are of a different scale to the other datasets. Since the Moscow images are used as a control dataset for lung involvement inferencing, we did not attempt to match the zoom level of this dataset to our other datasets preferring to leave this control dataset completely unprocessed. The COVIDNet-CT-2A average composite images contain a series of horizontal arcs in the lower quadrant of the image that are not present in the other image sets. The extent to which these structural differences between the datasets may limit external generalization is considered in the results and discussion sections of this paper.

3.2. Pre-processing pipelines

The study scope covers both the generalization behavior of trained deep learning models and the effect that commonly used pre-processing techniques have on that generalization behavior. To meet this objective the models have been trained and tested not only against raw data, but also using simple histogram equalization using parameters matching ImageNet [66] and CLAHE as successfully employed in a number of studies into deep learning based lung pathology detection from medical images [67,68]. We were also interested in the effect of handcrafted feature extraction layers, often implemented as wavelet filters such as Gabor filters, which have been shown to improve the accuracy of deep learning classification in thoracic disease imaging applications.

Each combination of image histogram function was trained and tested with and without a learning Gabor filter replacing the first convolution layer in the Densenet-121 network. These experiments reveal which, if any, of these techniques have a positive effect on the generalization behavior of the deep learning model under investigation.

3.3. Model development

The torchvision [69] DenseNet121 [70] network architecture was selected for the study which is consistent with other studies successfully using DenseNet121 as a deep learning classifier for COVID-19 image datasets [11], and thoracic radiology deep learning studies generally [71–74].

The DenseNet neural network architecture consists of a sequence of dense blocks comprising stacked convolution layers separated by transition layers composed of a 1 × 1 convolution down-sampling prior to average pooling as input to the next dense block. Each convolution layer in the dense blocks receives as input the feature maps output of all preceding convolution layers in the dense block. This feature map reuse characteristic means that lower levels in the dense block do not need to relearn features from higher levels in the dense block. This architecture is shown in Figure 2.

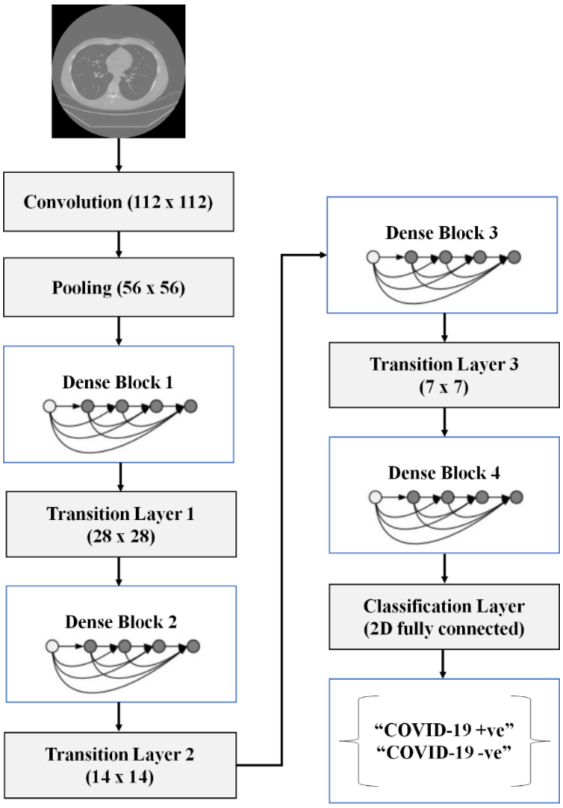

Convolution (112 x 112)
Pooling (56 x 56)
Dense Block 1
Transition Layer 1 (28 x 28)
Dense Block 2
Transition Layer 2 (14 x 14)
Dense Block 3
Transition Layer 3 (7 x 7)
Dense Block 4
Classification Layer (2D fully connected)
"COVID-19 +ve"
"COVID-19 -ve"

Figure 2. Modified DenseNet-121 CNN used for training and inference.

Since our experiments are all binary classifications (COVID-19 disease positive versus COVID-19 disease negative) we replaced the 1000 neuron DenseNet output fully connected layer with 2 neurons as shown in Figure 2 above. Following a number of studies showing the effectiveness of pre-trained network models in thoracic disease classification [75], we employed transfer learning to pretrain the

DenseNet model in recognizing natural shapes from the ImageNet [76] training corpus. The entire network was then fine-tuned against the target datasets at a learning rate of 1e-5 using the Adam optimizer [77] with standard parameters. Augmentations of 1-degree rotation with expansion and horizontal flip were applied. Class imbalance between disease negative and disease positive classes was addressed using a balancing random sampler. For each model a single training run was executed with the trained model captured for the purposes of external testing and involvement scoring.

To ensure consistent training across the datasets and models, the DenseNet network was trained until three consecutive epochs showed no reduction in validation loss, implementing a regularization technique known as early stopping [78]. At this point each model was tested against a holdout test set for each dataset constructed from the recommended test splits for the COVID-CT and COVIDNet-CT-2A datasets and a random selection of 20% of images for the private MID-CT and SARS-COV-2 datasets. This guaranteed that there was no patient overlap for the internal testing of COVID-CT, COVIDNet-CT-2A, and MID-CT (being composed of a single image per patient). The SARS-COV-2 dataset did not provide a recommended test split or patient level metadata, making it impossible to guarantee no patient overlap between training and testing splits for this dataset. This is consistent with the associated study for the SARS-COV-2 dataset which does not mention use an independent cohort of patients for holdout testing [63]. Internal testing results for SARS-COV-2 could therefore be expected to be optimistic, although external testing results for this dataset would not be impacted. Consistently with each source study, classification tasks have been performed at the slice rather than the patient level, since large numbers of patient specific images and patient level metadata are not uniformly provided by the publicly available datasets.

3.4. External testing

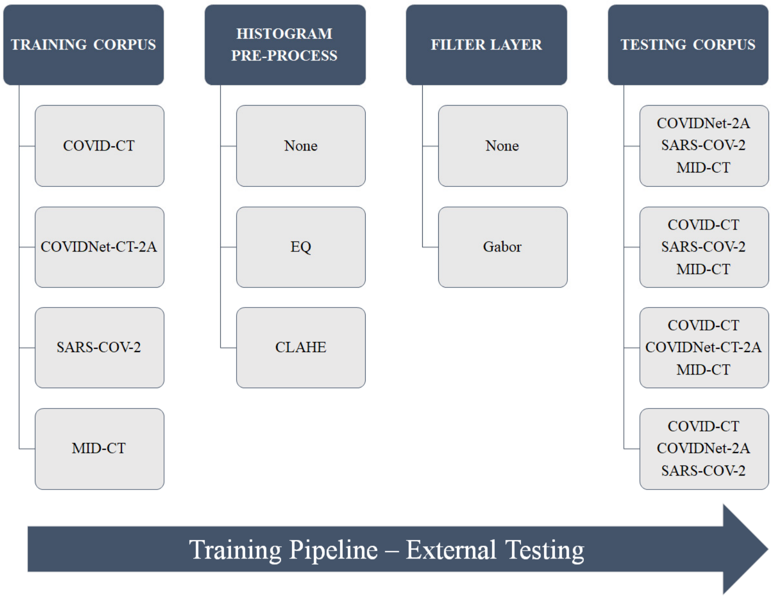

TRAINING CORPUS
COVID-CT
COVIDNet-CT-2A
SARS-COV-2
MID-CT
HISTOGRAM PRE-PROCESS
None
EQ
CLAHE
FILTER LAYER
None
Gabor
TESTING CORPUS
COVIDNet-2A SARS-COV-2 MID-CT
COVID-CT SARS-COV-2 MID-CT
COVID-CT COVIDNet-CT-2A MID-CT
COVID-CT COVIDNet-2A SARS-COV-2
Training Pipeline – External Testing

Figure 3. Experimental flow for model external testing.

All combinations of no equalization, simple histogram equalization, application of CLAHE and use of a learning Gabor filter replacing the first convolution layer were scripted, with training completed for all four datasets (reserving the Moscow dataset for the involvement scoring study) resulting in 12 trained models (4 datasets×3 pre-processing options) without Gabor filtering and 12 trained modes with Gabor filtering. Ablation results are provided by the model combination without histogram pre-processing or Gabor filter layer. Each of these trained models was then cross validated against the test data sets corresponding to each other model as an external generalization study. This study was performed for each combination of pre-processing option and with/without Gabor filtering as shown in Figure 3.

3.5. Lung involvement scoring

Each trained model was used to score COVID-19 positive and COVID-19 negative lung involvement against each image in the Moscow expertly stratified dataset. For each stratification in the Moscow dataset an average involvement predictive score for each model was calculated. The correlation between the average predicted score on the y-axis and involvement strata on the x-axis was plotted for each model to test our hypothesis that a generalized COVID-19 diagnostic model would be predictive of lung involvement strata. Once again, this experiment was performed for each combination of pre-processing option and with/without Gabor filtering as shown in Figure 4.

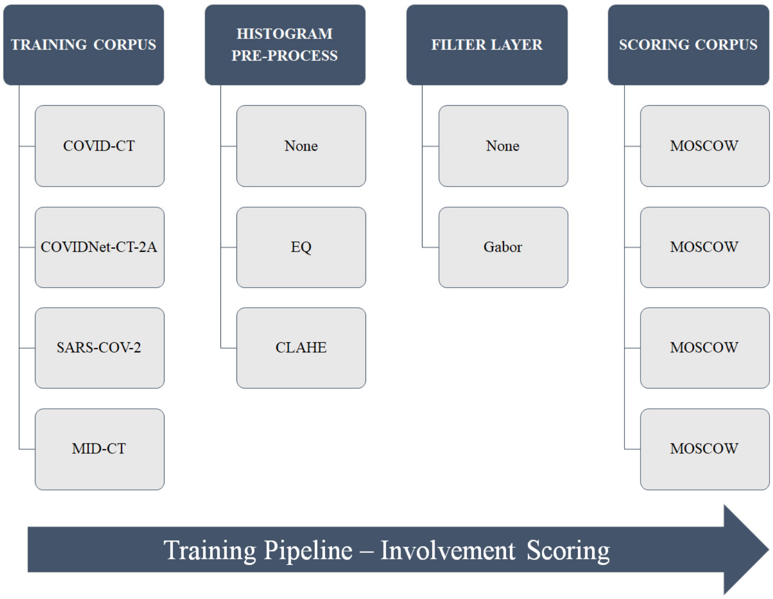


Figure 4. Experimental flow for lung involvement scoring.

4. Results

4.1. Internal testing results

For the purposes of comparison and validation of our DenseNet-121 trained models a summary of results from the source dataset papers is presented in Table 2. Although sensitivity and specificity metrics are usually used in the medical context to assess the effectiveness of predictive models, these were not provided by all the source studies and our summary is thereby limited to Accuracy, Precision, Recall and calculated F1 score. Per class metrics were only documented for the COVID-Net-CT-2A study and we have averaged these metrics to make them comparable with the COVID-CT and

SARS-COV-2 studies.

Table 2. Source dataset results compared.

Study Reference	Source Study Results (%)			
	Accuracy	F1	Precision	Recall
COVID-CT	89.0	90.0	N/A	N/A
COVIDNet-CT-2A	99.1	99.0	98.8	99.2
SARS-COV-2	97.4	97.3	99.1	95.5

Given the very large number of comparison experiment variations we present results in a tabular format over Tables 3−11 summarizing the results for each dataset and trained model. Each cell in the table presents three numbers, from top to bottom these are to be read in order as the results of no histogram pre-processing, standard histogram equalization pre-processing and CLAHE pre-processing. Separate tables have been created showing the results with and without a learning Gabor filter applied to allow for easier comparison.

Table 3 presents internal testing results for all models without the application of a Gabor filter replacing the first convolution layer. The first observation is that there is no clear pattern that image pre-processing such as histogram equalization and CLAHE have any significant effect on internal classification metrics with results being very close regardless of pre-processing technique. The second observation is that results using an off-the-shelf classifier are comparable to those of the original studies, with a small improvement on the SARS-COV-2 study results from our single training run and slightly lower results against the COVIDNet-CT-2A dataset which may be the result of removal of shared Moscow and COVID-CT images from this set. Results for the COVID-CT dataset, using the same pipelines, classifier and hyperparameters as used for other datasets are substantially lower than then the original study, whose results we have been previously unable to replicate using off-the-shelf networks [79].

Table 3. Internal results without Gabor filter layer.

Study Reference		Internal results without learning Gabor filter (%)			
	Variation	Accuracy	F1	Precision (COVID/Normal)	Recall (COVID/Normal)
COVID-CT	No Preprocess	69.0	69.0	63/83	89/50
	Histogram Eq	70.9	70.0	64/85	90/53
	CLAHE	64.0	64.0	60/70	74/54
COVIDNet-CT-2A	No Preprocess	98.9	98.5	97/100	99/99
	Histogram Eq	98.9	98.5	97/100	99/99
	CLAHE	98.8	98.5	97/100	99/98
SARS-COV-2	No Preprocess	97.5	97.5	98/97	97/98
	Histogram Eq	97.9	98.0	98/98	98/97
	CLAHE	98.3	98.0	98/99	99/97
MID-CT	No Preprocess	80.0	80.0	83/78	76/84
	Histogram Eq	79.0	79.0	80/78	78/80
	CLAHE	78.0	78.0	80/76	74/82

Table 4 presents internal testing results with the application of a learning Gabor filter replacing the first convolution layer. This resulted in improved classification metrics for the COVID-CT, COVIDNet-CT-2A and MID-CT datasets. For COVID-CT, the key improvement in classification metrics observed is improved recall for the disease negative class leading to an accuracy gain of 4% when using histogram equalization. This result is still not as good as the original study but matches our previous best result against this dataset using pre-trained networks [80]. For COVIDNet-CT-2A we found that using a combination of Gabor filtering and CLAHE improved our internal accuracy for COVID-Net-CT-2A to 99.1% which is equivalent to the original study, bearing in mind that we have removed the COVID-CT and Moscow data sets from our version of COVIDNet-CT-2A to avoid polluting our external generalization study. For MID-CT we found that the combination of Gabor filtering and CLAHE resulted in an internal accuracy of 86% which is our best internal score for this dataset by a significant margin of 6%.

Table 4. Internal results with Gabor filter layer.

Study Reference		Internal results with learning Gabor filter (%)			
	Variation	Accuracy	F1	Precision (COVID/Normal)	Recall (COVID/Normal)
COVID-CT	No Preprocess	63.5	62.5	59/74	83/46
	Histogram Eq	74.9	75.0	69/83	86/65
	CLAHE	70.0	70.0	65/78	83/58
COVIDNet-CT-2A	No Preprocess	98.6	98.5	97/100	99/98
	Histogram Eq	99.0	98.5	97/100	100/99
	CLAHE	99.1	99.0	97/100	100/99
SARS-COV-2	No Preprocess	97.9	98.0	97/99	99/97
	Histogram Eq	96.3	96.0	95/97	97/95
	CLAHE	95.8	96.0	97/95	95/97
MID-CT	No Preprocess	79.0	79.0	81/77	76/82
	Histogram Eq	79.0	79.0	81/77	76/82
	CLAHE	86.0	86.0	89/83	82/90

4.2. External testing results

Having trained each model for each combination of pre-processing technique with and without a learning Gabor filter replacing the first convolution layer we proceeded to externally test each model against each other training corpus test data partition.

4.2.1. External testing of COVIDNet-CT-2A models

Tables 5 and 6 show the results of external testing of the COVIDNet-CT-2A trained models.

Table 5. COVIDNet-CT-2A external results without Gabor filter layer.

Model Dataset:		Internal results without learning Gabor filter (%)				
COVIDNet-CT-2A		Variation	Accuracy	F1	Precision	Recall
					(COVID/Normal)	(COVID/Normal)
Test Dataset	COVID-CT	No Preprocess	38.7	37.0	33/67	83/17
		Histogram Eq	40.0	38.5	33/69	83/19
		CLAHE	62.3	61.0	45/79	67/60
	MID-CT	No Preprocess	62.8	63.0	47/91	90/49
		Histogram Eq	34.7	22.5	33/92	99/3
		CLAHE	43.1	40.0	36/95	98/16
	SARS-COV-2	No Preprocess	75.7	75.5	58/96	94/67
		Histogram Eq	60.9	60.5	46/98	98/43
		CLAHE	63.0	61.0	47/98	98/46

Table 6. COVIDNet-CT-2A external results with Gabor filter layer.

Model Dataset:		Internal results without learning Gabor filter (%)				
COVIDNet-CT-2A		Variation	Accuracy	F1	Precision	Recall
					(COVID/Normal)	(COVID/Normal)
Test Dataset	COVID-CT	No Preprocess	60.2	60.0	44/82	77/52
		Histogram Eq	54.2	54.0	39/77	73/45
		CLAHE	44.7	44.5	34/71	74/30
	MID-CT	No Preprocess	50.2	50.0	39/82	86/33
		Histogram Eq	60.9	61.0	46/95	95/44
		CLAHE	69.6	69.5	52/94	92/59
	SARS-COV-2	No Preprocess	77.9	70.0	84/77	41/96
		Histogram Eq	84.4	83.0	73/91	83/85
		CLAHE	77.9	73.0	73/79	53/90

The COVIDNet-CT-2A trained models did not generalize well to either the COVID-CT or MID-CT datasets regardless of pre-processing histogram function or presence of a learning Gabor filter. These datasets are significantly different from COVIDNet-CT-2A in terms of imaged quality (COVID-CT is scraped from PDF) and diversity, with MID-CT only containing a single CT slice location and COVIDNet-CT-2A having a variety of CT slice locations.

External testing against the SARS-COV-2 dataset was successful using a combination of learning Gabor filter with histogram equalization resulting in accuracy of 84.4% with reasonably well balanced precision and recall metrics. SARS-COV-2 and COVIDNet-CT-2A have similar image quality and CT slice diversity. Histogram equalization and Gabor filter have improved results due to COVIDNet-CT-2A having been assembled from a wide variety of sources – with these techniques bringing uniformity to the training corpus and promoting generalization when equivalent pre-processing was applied to the SARS-COV-2 test data.

4.2.2 External testing of COVID-CT models

Tables 7 and 8 show the results of external testing of the COVID-CT trained models. The challenging nature of this dataset has been previously described and our expectation that these models would not generalize is supported by the external testing results with classification metrics that show no significant class separation regardless or pre-processing option or presence of a learning Gabor filter. Results are provided as evidence that the quality of the training image corpus is critical to model generalization capability. In this case the COVID-CT images have been scraped from pdf sources with variable resolution, image size and quality. Other image sets are composed of high-quality conversions from DICOM source. This experiment provides evidence that models must be trained on high-quality source images to promote generalization.

Table 7. COVID-CT external results without Gabor filter layer.

Model Dataset: COVID-CT		Internal results without learning Gabor filter (%)				
		Variation	Accuracy	F1	Precision (COVID/Normal)	Recall (COVID/Normal)
Test Dataset	COVIDNet-CT-2A	No Preprocess	51.2	40.5	50/75	97/9
		Histogram Eq	55.7	49.5	52/78	94/20
		CLAHE	53.2	45.0	51/75	95/14
	MID-CT	No Preprocess	50.2	41.5	49/60	92/11
		Histogram Eq	47.8	34.0	48/40	97/2
		CLAHE	48.3	32.5	48/0	100/0
	SARS-COV-2	No Preprocess	52.7	52.5	51/55	56/50
		Histogram Eq	43.3	41.0	44/41	66/22
		CLAHE	47.8	47.5	47/49	58/38

Table 8. COVID-CT external results with Gabor filter layer.

Model Dataset: COVID-CT		Internal results without learning Gabor filter (%)				
		Variation	Accuracy	F1	Precision (COVID/Normal)	Recall (COVID/Normal)
Test Dataset	COVIDNet-CT-2A	No Preprocess	51.2	43.5	50/62	91/14
		Histogram Eq	51.2	42.5	50/67	94/11
		CLAHE	51.7	48.5	50/58	81/25
	MID-CT	No Preprocess	53.7	43.5	51/100	100/10
		Histogram Eq	52.7	42.0	51/91	99/10
		CLAHE	47.8	32.5	48/0	99/0
	SARS-COV-2	No Preprocess	48.3	47.5	46/50	37/59
		Histogram Eq	45.8	45.0	46/46	62/30
		CLAHE	49.8	49.5	48/52	54/46

4.2.3 External testing of MID-CT models

Tables 9 and 10 show the results of external testing of the privately acquired MID-CT trained models. Despite the uniformity and high-quality of this dataset, the models based on this dataset have failed to generalize to any other dataset with classification results close to chance. The unique property of the MID-CT dataset is that it is composed of CT slices exclusively from the midpoint of the CT scan. Other datasets have CT slices taken from diverse CT scan positions. Therefore, the MID-CT dataset does not present an anatomical match for the other datasets leading to extremely poor generalization.

Table 9. MID-CT external results without Gabor filter layer.

Model Dataset: MID-CT		Internal results without learning Gabor filter (%)				
		Variation	Accuracy	F1	Precision (COVID/Normal)	Recall (COVID/Normal)
Test Dataset	COVID-CT	No Preprocess	50.0	33.5	0/50	0/100
		Histogram Eq	50.0	33.5	0/50	0/100
		CLAHE	50.0	33.5	0/50	0/100
	COVIDNet-CT-2A	No Preprocess	50.0	35.0	50/50	98/2
		Histogram Eq	53.0	42.5	52/71	96/10
		CLAHE	59.0	57.0	56/67	82/36
	SARS-COV-2	No Preprocess	53.0	45.5	52/62	90/16
		Histogram Eq	50.0	33.5	50/0	100/0
		CLAHE	53.0	51.0	55/52	32/74

Table 10. MID-CT external results with Gabor filter layer.

Model Dataset: MID-CT		Internal results without learning Gabor filter (%)				
		Variation	Accuracy	F1	Precision (COVID/Normal)	Recall (COVID/Normal)
Test Dataset	COVID-CT	No Preprocess	50.0	33.5	0/50	0/100
		Histogram Eq	50.0	33.5	0/50	0/100
		CLAHE	50.0	33.5	0/50	0/100
	COVIDNet-CT-2A	No Preprocess	51.0	35.5	51/100	100/2
		Histogram Eq	50.0	33.5	50/0	100/0
		CLAHE	50.0	38.0	50/50	94/6
	SARS-COV-2	No Preprocess	50.0	33.5	0/50	0/100
		Histogram Eq	68.0	66.5	82/62	46/90
		CLAHE	67.0	63.5	95/60	36/98

4.2.4 External testing of SARS-COV-2 models

Tables 11 and 12 show the results of external testing of the SARS-COV-2 trained models. Although performance of these models against the COVID-CT and MID-CT datasets was poor for reasons already discussed, generalization of these models against the COVIDNet-CT-2A was very good with best performance from the histogram equalized model achieving an F1 score of 86% along with 85% and 86.7% from sensitivity and specificity respectively. The improves upon the good performance of the COVIDNet-CT-2A model against the SARS-COV-2 data previously described, and in contrast to those results a Gabor filter was not needed due to the high consistency of the SARS-COV-2 dataset. We interpret these findings as evidence that a smaller (SARS-COV-2 n = 812), high-quality dataset has produced a more generalizable model than a larger diversified dataset (COVIDNet-CT-2A n = 163,630 images) that has been built from diverse sources.

Table 11. SARS-COV-2 external results without Gabor filter layer.

Model Dataset:		Internal results without learning Gabor filter (%)				
SARS-COV-2		Variation	Accuracy	F1	Precision (COVID/Normal)	Recall (COVID/Normal)
Test Dataset	COVID-CT	No Preprocess	50.0	34.5	50/50	98/2
		Histogram Eq	51.3	37.5	51/71	98/4
		CLAHE	50.4	37.0	50/56	97/4
	COVIDNet-CT-2A	No Preprocess	82.9	83.0	83/83	83/82
		Histogram Eq	85.8	86.0	86/85	85/87
		CLAHE	83.8	84.0	80/88	89/78
	MID-CT	No Preprocess	61.3	60.5	59/66	76/47
		Histogram Eq	48.8	38.5	49/43	90/7
		CLAHE	51.7	39.0	51/70	97/6

Table 12. SARS-COV-2 external results with Gabor filter layer.

Model Dataset:		Internal results without learning Gabor filter (%)				
SARS-COV-2		Variation	Accuracy	F1	Precision (COVID/Normal)	Recall (COVID/Normal)
Test Dataset	COVID-CT	No Preprocess	50.4	35.0	50/67	99/2
		Histogram Eq	51.3	37.5	51/71	98/4
		CLAHE	50.0	36.0	50/44	96/2
	COVIDNet-CT-2A	No Preprocess	81.3	81.0	81/82	82/81
		Histogram Eq	85.0	85.0	86/84	83/87
		CLAHE	84.6	84.5	82/87	88/81
	MID-CT	No Preprocess	64.2	63.5	61/68	76/53
		Histogram Eq	50.0	41.0	50/50	89/11
		CLAHE	50.8	37.0	50/62	97/4

4.3 Lung involvement scoring results

Having shown some success in external generalization testing we proceeded to test each trained model as an inferencing classifier against the Moscow data set which had been held out as a control. The Moscow dataset is unique in the publicly available COVID-19 CT datasets in that it is a high-quality dataset that has been labelled by expert radiologists with a lung involvement score from 0 to 100% as defined in Table 13.

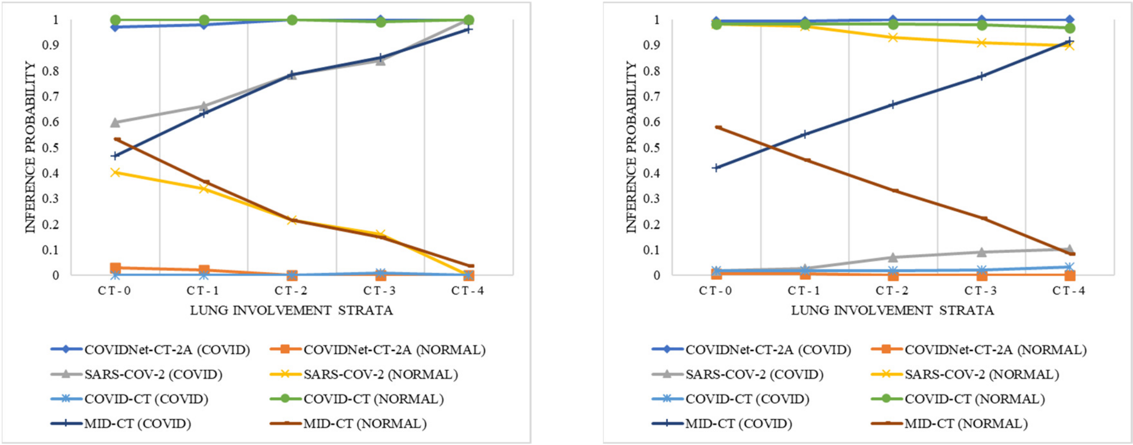
Table 13. Lung involvement strata definitions.

Stratification	Lung involvement (%)
CT-0	0
CT-1	0–25
CT-2	2 –50
CT-3	50–75
CT-4	75–100

Since COVID-19 pneumonia is indicated on CT imagery as lung involvement [80] it follows that generalization of the trained models against the Moscow data set would be evidenced by an increasing inference COVID-19 disease positive inference probability corresponding to an increasing lung involvement stratum. To test this hypothesis, we used each trained model, with each pre-processing option and with/without learning Gabor filters, to score each of the Moscow dataset images for COVID-19 probability. Average predicted score for each stratification was plotted on the y-axis against ground truth lung involvement stratifications on the x-axis.

The results of these experiments are presented in Figures 5–10. We expect a line representing COVID-19 disease positive inference probability to have a positive gradient as involvement score increases from CT-0 to CT-4, and conversely the line representing COVID-19 disease negative inference probability should have a negative gradient. These gradients should be more pronounced the better a model has performed against the Moscow data set.

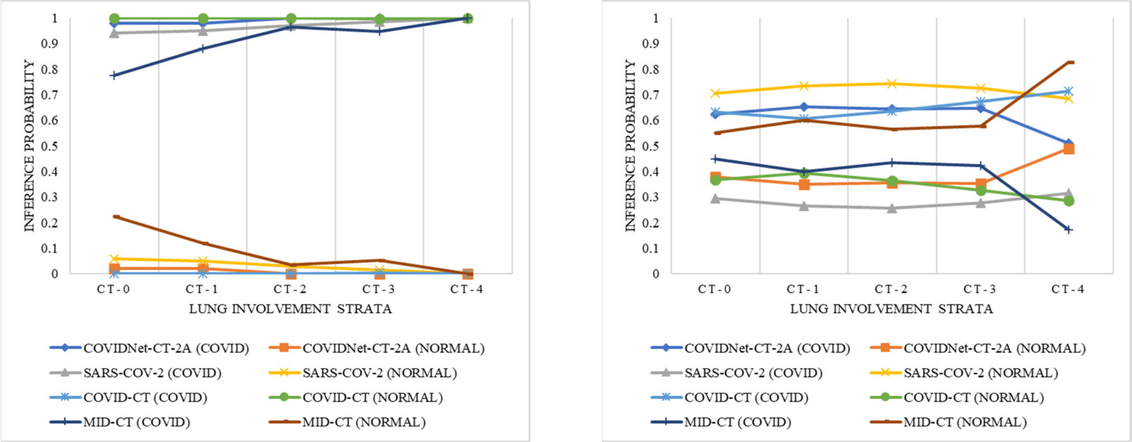
Figures 5(a),(b) represent results from trained models without histogram equalization or CLAHE pre-processing, with Figure 5(b) showing results of adding a learning Gabor filter in place of the first convolution layer of the network. The expected pattern for lung involvement prediction is evident for both the SARS-COV-2 and MID-CT models, although the performance of the MID-CT based model is considered better as it is more effective than the SARS-COV-2 based model in predicting the normal condition represented by data points at CT-0. Figure 5(b) shows that addition of a Gabor filter had the effect of reducing the performance of both MID-CT and SARS-COV-2 models.



(a) **(b)**

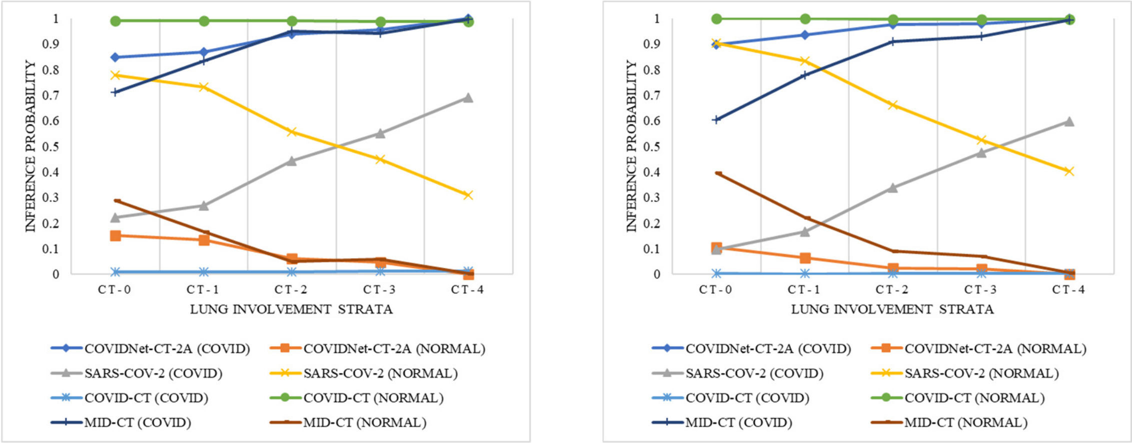
Figure 5. Lung involvement scoring with no histogram pre-processing: (a) Without a learning Gabor filter layer; (b) with a learning Gabor filter layer.

Figures 6(a),(b) represent results from trained models with standard histogram equalization. Figure 6(b) shows the effect of a learning Gabor filter in the position of the first convolution layer of the network. Histogram equalization has eroded the good generalization performance of MID-CT and SARS-COV-2 based models without histogram equalization shown at Figure 5(a). Application of a Gabor filter to the models further eroded these results as shown in Figure 6(b). From this, one could conclude that standard histogram equalization alone, and the combination of standard histogram equalization and learning Gabor filter is not an effective combination of techniques to promote generalization.



(a) **(b)**

Figure 6. Lung involvement scoring with histogram pre-processing: (a) Without a learning Gabor filter layer; (b) with a learning Gabor filter layer.



a) **(b)**

Figure 7. Lung involvement scoring with CLAHE pre-processing: (a) Without a learning Gabor filter layer; (b) with a learning Gabor filter layer.

Figures 7(a),(b) represent results from trained models with CLAHE image processing. Figure 7(b) shows the effect of CLAHE image processing with a learning Gabor filter in the position of the first convolution layer of the network. CLAHE has enhanced the good generalization performance of SARS-COV-2 based model over no pre-processing, and allowed the COVIDNet-CT-2A to show minimal generalization where none was previously evident in Figure 6(a), but eroded the generalization performance of MID-CT. The SARS-COV-2 based model is now able to correctly predict a low probability for disease positive at CT-0 and a high probability for disease positive at CT-4 with a linear gradient. With the Gabor filter included the models, we achieved better results only for SARS-COV-2 based model at the lower stratifications of lung involvement (CT-0 and CT-1) as shown by Figure 7(b).

Confirming the results of the first experiment, models based on SARS-COV-2 exhibited superior generalization performance despite this being a much smaller dataset than COVIDNet-CT-2A. The high-quality, image consistency and slice location diversity present in SARS-COV-2 have promoted good generalization properties in the model trained on this dataset.

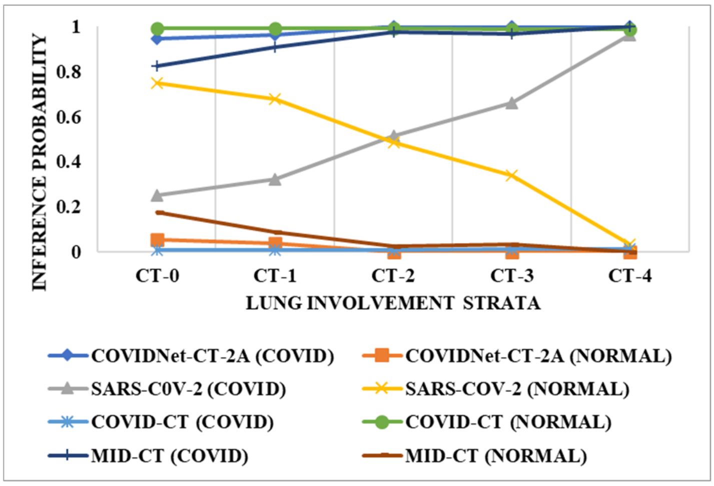


Figure 8. Lung involvement scoring with CLAHE and Min-Max Gabor ensemble.

Table 14. SARS-COV-2 CLAHE model with Min-Max Gabor ensemble average prediction per strata.

Involvement Strata	Ground Truth Involvement %	Average Prediction SARS-C0V-2 (COVID) %	Average Prediction SARS-COV-2 (NORMAL) %
CT-0	0	25.184	74.816
CT-1	0–25	32.240	67.760
CT-2	25–50	51.353	48.647
CT-3	50–75	65.917	34.083
CT-4	75–100	96.380	3.620

This promising result led us to focus on which combination of techniques would maximize the predictive value of these models against the Moscow dataset. Inspection of the results graphs for the SARS-COV-2 models showed that application of CLAHE provided the best results, with a Gabor filter providing highest accuracy at CT-0 (minimum lung involvement) representing an improvement in sensitivity and absence of the Gabor filter providing highest accuracy at CT-4 (maximum lung involvement) representing an improvement in specificity. We combined these SARS-COV-2 based models to use CLAHE as a pre-process and the minimum predicted probability from the Gabor/non-Gabor models as the COVID-19 disease negative score and the maximum probability from the Gabor/non-Gabor models as the COVID-19 disease positive inference probability. Figure 8 shows the results of this min-max Gabor ensemble. The ensemble results from this model are excellent, achieving an average score of 75% for the Moscow images without lung involvement and an average score of 96% for the Moscow images with 75–100% lung involvement, with an approximately continuous gradient for the stratifications between these extremes. The SARS-COV-2 based models have generalized well to the Moscow dataset when using this ensembling technique providing very accurate predictive scoring of lung involvement. Table 14 provides the average predictions by involvement stratification for this ensemble model.

5 Discussion

5.1 Effect of differential image class provenance and patient image diversity

The experiments showed that attributes of the training datasets had a significant impact on the predictive capability of generated models against external datasets. Our poorest performing models were models based on the COVID-CT dataset. These models did not generalize in either external testing or involvement scoring. Internal testing showed that both histogram equalization and Gabor filter had a positive effect on metrics for this dataset, with these techniques combined achieving a best F1 score of 75%. We interpret the positive effect of histogram equalization on these results to a reduction in the variability of brightness and contrast amongst source images that was very evident in this dataset. Applying a Gabor filter would have the effect of reducing dimensionality manifesting in these experiments as an improvement in specificity since the Gabor filter reinforces pneumonia related features that would be absent from disease negative images. It was previously noted that this dataset has different sources for disease positive and disease negative images, and we interpret the observed poor external generalization of models trained on COVID-CT to the classifier learning confounding differences between these classes instead of pathological features. Additionally, we noted previously that the disease positive images for this dataset are sourced from 216 patients, whilst the disease negative images are sourced from only 84 patients. This very low number of disease negative patients would be insufficient for a machine learning classifier to learn generalized disease features. Given these considerations it is unsurprising that models based on this dataset were not predictive of lung involvement against the Moscow stratified dataset.

5.2 Anatomical matching and effect of single vs multiple CT slice locations

The performance of the MID-CT dataset-based models in external testing was very poor, with the only tests improving upon blind chance doing so at the expense of sensitivity or specificity, i.e., a tendency to classify the majority of samples as either disease positive or disease negative regardless of ground truth. The MID-CT dataset did have disease positive and disease negative classes co-sourced and subjected to a consistent acquisition pipeline, and the number of patients in disease positive and disease negative classes was relatively balanced. However, the MID-CT dataset only has a single CT image slice per patient taken from the median index. This is different to the other datasets used in external testing which all have image slices taken from multiple CT scan locations. We attribute the poor generalization performance of MID-CT based models to this anatomical mismatch and lack of CT scan location diversity.

5.3 Large diverse multi-sourced image corpus vs smaller uniformly acquired image corpus

The COVIDNet-CT-2A based models showed good generalization in external testing against the SARS-COV-2 dataset. In contrast, under involvement testing against the Moscow dataset, the COVIDNet-CT-2A trained models had little to no predictive value. This contrasts with the SARS-

COV-2 dataset-based models which all showed good generalization against the Moscow dataset, with particularly good results when CLAHE pre-processing was used. These results establish that a smaller, higher quality dataset with disease positive and negative samples co-sourced and consistently acquired (SARS-COV-2) is more effective in generalization to external datasets than a very large, more diverse, dataset assembled from multiple sources with a variety of acquisition pipelines (COVIDNet-CT-2A).

5.4 Effect of image pre-processing and learning Gabor filter

The study found that image histogram equalization and CLAHE and Gabor filtering had a negative effect on MID-CT based model external classification results despite the combination of CLAHE and Gabor filtering providing significantly better results in internal testing for models based on this dataset. We suspect that image pre-processing and Gabor filtering have most likely reinforced signal noise in the manner noted by [81], rather than reinforcing pathological signal only. The MID-CT dataset was single sourced from with fixed CT apparatus operating parameters and therefore very consistent in terms of brightness and contrast. In our study, image histogram pre-processing and wavelet filtering techniques such as Gabor filtering improved classification metrics only where differences in image brightness and contrast have been imparted on the data corpus by different CT apparatus and operational parameters. This is the case with the SARS-COV-2 dataset which has been sourced from multiple Brazil located clinics and shows moderate variability in terms of image size, brightness, and contrast.

Finally, we noted that the use of a Gabor filter following CLAHE pre-processing improved the specificity of the SARS-COV-2 based models, whilst the model with CLAHE but without the Gabor filter showed superior sensitivity. By combining these models with a Min-Max algorithm we produced an ensemble model that accurately predicted lung involvement ground truth against the Moscow dataset being a completely independent dataset. The accuracy of this combined model is remarkable when one considers that the SARS-COV-2 and Moscow datasets have been sourced from the different geographical regions, and that these datasets show significant structural differences on their average image composite.

5.5 Key factors promoting model generalization

These experiments have identified two key factors promoting CT image based deep learning model generalization:

1) Uniform image sample acquisition. Disease positive and disease negative classes should be consistently (though not necessarily identically) sourced and subjected to consistent image acquisition/compression pipelines. This will minimize systematic structural differences between the classes and eliminate bias resulting in model training on non-pathological "confounding factors". Our SARS-COV-2 and MID-CT models met these criteria and generalized to the Moscow corpus even though these datasets and the Moscow images were structurally quite different, suggesting that consistency in the model training data is more important that structural consistency across training and external inferencing datasets. The use of histogram equalization techniques was shown to improve generalization results where there is moderate variability in image corpus brightness and contrast (SARS-COV-2). However, these techniques proved ineffective where the image set is already very consistent (MID-CT) or where there is major variation in image attributes

across the datasets (COVID-CT and COVIDNet-CT-2A).

2) Diversity in CT slice position. The model trained on SARS-COV-2 displayed superior model generalization against the Moscow dataset in comparison the MID-CT trained model. The key difference between these datasets is that SARS-COV-2 has been assembled from a variety of CT slice positions whereas the MID-CT dataset uses only the middle CT slice. From this we can conclude that if image samples are uniformly acquired for each class then diversity of CT slice position in the dataset improves model generalization.

6 Conclusions

Computer vision based medical diagnostic systems cannot be considered clinically mature until they are also shown to be generalizable across populations. However, most of the research to date has focused on achieving higher accuracy internal classification within datasets rather than considering how these models might generalize to external datasets, and thereby to real-world situations.

In these experiments, only models trained on the SARS-COV-2 dataset exhibited successful generalization for both external testing and lung involvement scoring. External testing against the COVIDNet-CT-2A dataset yielded highest accuracy of 85.8% and 85.0% for tests using histogram equalization with/without employment of a learning Gabor filter, respectively. The SARS-COV-2 based models also generalized well to the Moscow dataset, particularly when CLAHE processing was used. An ensemble of SARS-COV-2 models using the Gabor filter only for low lung involvement scores achieved a predictive accuracy of 75% for zero lung involvement and 96% for 75–100% lung involvement with an almost linear relationship between these stratifications.

The SARS-COV-2 image corpus has been consistently sourced with disease positive and negative samples co-sourced and qualitatively very similar in appearance. This is in contrast to the COVID-CT and COVIDNet-CT-2A datasets, which both contain image samples from a wide range of sources. This result may be interpreted as greater dataset consistency having minimized confounding bias, thereby promoting generalization of models trained on SARS-COV-2. Further, the key difference between the successfully generalized SARS-COV-2 models, and the poorly generalized MID-CT based models is that images comprising SARS-COV-2 have been selected from a variety of slice positions whereas MID-CT has images only from the middle CT slice. From this we conclude that CT slice diversity in the training corpus also promotes model generalization.

This study has some limitations. It would be preferable to have segmented the lung field from the CT image to improve the signal-to-noise-ratio. However, initial experiments showed segmentation results using a U-Net [82] to be highly sensitive to the quality of the source images. We were conscious that systematically different segmentation artefacts for different image corpuses would impose bias the datasets and therefore chose not to segment. In the future, additional slices for the MID-CT dataset will be procured to try to improve the generalization results for this dataset.

Ethics approval of research

The study was conducted under University of Technology Sydney approved ethics no: ETH21-6193.

Informed Consent Statement: Patient consent was waived due to all subject images having been stripped of all personally identifying information prior to use in this study.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found at the cited locations. MID-CT data is privately maintained and will be made available on request from the corresponding author.

Acknowledgments

This work was supported by the Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), Faculty of Engineering and IT, University of Technology Sydney (UTS), Sydney, NSW 2007, Australia.

Conflict of interest

The authors no known conflict of interest exist in relation to this work.

References

1. STATISTA, *Coronavirus Deaths Worldwide by Country*, 2021. Available from: https://www.statista.com/statistics/1093256/novel-coronavirus-2019ncov-deaths-worldwide-by-country/.
2. U. S. CDC., *About Variants of the Virus that Causes COVID-19*, 2021. Available from: https://www.cdc.gov/coronavirus/2019-ncov/transmission/variant.html.
3. Global Preparedness Monitoring Board, *A World in Disorder*, 2021. Available from https://www.gpmb.org/annual-reports/overview/item/2020-a-world-in-disorder.
4. A. Ulhaq, J. Born, A. Khan, D. P. S. Gomes, S. Chakraborty, M. Paul, COVID-19 control by computer vision approaches: A survey, *IEEE Access,* **8** (2020), 179437–179456.
5. C. Butt, J. Gill, D. Chun, B. A. Babu, Deep learning system to screen coronavirus disease 2019 pneumonia, *Appl. Intell.*,**1** (2020), 1–7.
6. J. Chen, L. Wu, J. Zhang, L. Zhang, D. Gong, Y. Zhao, et al., Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography, *Sci. Rep.,* **10** (2020), 19196.
7. H. Gunraj, L. Wang, A. Wong, COVIDNet-CT: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest CT images, *Front. Med.,* **7** (2020), 1025.
8. R. Kumar, S. Zhang, W. Wang, W. Amin, J. Kumar, Blockchain-federated-learning and deep learning models for COVID-19 detection using CT imaging, preprint, arXiv:2007.06537.
9. Z. Li, Z. Zhong, Y. Li, T. Zhang, L. Gao, D. Jin, et al., From community-acquired pneumonia to COVID-19: A deep learning-based method for quantitative analysis of COVID-19 on thick-section CT scans, *Eur. Radiol.,* **30** (2020), 6828–6837.
10. Q. Ni, Z. Y. Sun, L. Qi, W. Chen, Y. Yang, L. Wang, et al., A deep learning approach to characterize 2019 coronavirus disease (COVID-19) pneumonia in chest CT images, *Eur. Radiol.,* **30** (2020), 6517–6527.
11. T. D. Pham, A comprehensive study on classification of COVID-19 on computed tomography with pretrained convolutional neural networks, *Sci. Rep.,* **10** (2020), 16942–16942.
12. M. Polsinelli, L. Cinque, G. Placidi, A light CNN for detecting COVID-19 from CT scans of the chest, *Pattern Recognit. Lett.,* **140** (2020), 95–100.

13. P. Silva, E. Luz, G. Silva, G. Moreira, R. Silva, D. Lucio, et al., COVID-19 detection in CT images with deep learning: A voting-based scheme and cross-datasets analysis, *Inform. Med. Unlocked.*, **20** (2020), 100427–100427.
14. S. Wang, B. Kang, J. Ma, X. Zeng, M. Xiao, J. Guo, et al., A deep learning algorithm using CT images to screen for Corona virus disease (COVID-19), *Eur. Radiol.*, **31** (2021), 6096–6104.
15. X. Wang, X. Deng, Q. Fu, Q. Zhou, J. Feng, H. Ma, et al., A Weakly-Supervised Framework for COVID-19 Classification and Lesion Localization From Chest CT, *IEEE Trans. Med. Imaging*, **39** (2020), 2615–2625.
16. T. Akram, M. Attique, S. Gul, A. Shahzad, M. Altaf, S. S. R. Naqvi, et al., A novel framework for rapid diagnosis of COVID-19 on computed tomography scans, *Pattern Anal. Appl.,* **24** (2021), 951–964.
17. M. Mohammadpoor, M. S. Karizaki, M. S. Karizaki, A deep learning algorithm to detect coronavirus (COVID-19) disease using CT images, *PeerJ. Comp. Sci.,* **7** (2021), e345.
18. J. Zhang, Y. Xie, Y. Li, C. Shen, Y. Xia, COVID-19 screening on chest X-ray images using deep learning based anomaly detection, preprint, arXiv:2003.12338.
19. F. Ucar, D. Korkmaz, COVIDiagnosis-Net: Deep Bayes-SqueezeNet based diagnosis of the coronavirus disease 2019 (COVID-19) from X-ray images, *Med. Hypotheses,* **140** (2020), 109761.
20. Y. Oh, S. Park, J. C. Ye, Deep learning COVID-19 features on CXR using limited training data sets, *IEEE Trans. Med. Imaging,* **39** (2020), 2688–2700.
21. S. H. Yoo, H. Geng, T. L. Chiu, S. K. Yu, D. C. Cho, J. Heo, et al., Deep learning-based decision-tree classifier for COVID-19 diagnosis from chest X-ray imaging, *Front. Med.,* **7** (2020), 427.
22. J. Civit-Masot, F. Luna-Perejón, A. Civit, Deep learning system for COVID-19 diagnosis aid using X-ray pulmonary images, *Appl. Sci.,* **10** (2020), 4640.
23. M. Blain, M. T Kassin, N. Varble, X. Wang, Z. Xu, D. Xu, et al., Determination of disease severity in COVID-19 patients using deep learning in chest X-ray images, *Diagn. Interv. Radiol.,* **27** (2020), 20–27.
24. J. P. Cohen, L. Dao, K. Roth, P. Morrison, Y. Bengio, A. F. Abbasi, et al., Predicting COVID-19 pneumonia severity on chest X-ray with deep learning, *Cureus,* **12** (2020), e9448.
25. B. Liu, Y. Zhou, Y. Yang, Y. Zhang, Experiments of federated learning for COVID-19 chest X-ray images, preprint, arXiv:2007.0559.
26. M. E. Karar, E. E. D. Hemdan, M. A. Shouman, Cascaded deep learning classifiers for computer-aided diagnosis of COVID-19 and pneumonia diseases in X-ray scans, *Complex Intell. Syst.,* **7** (2021), 235–247.
27. H. Amin, A. Darwish, A. E. Hassanien, Classification of COVID19 X-ray images based on transfer learning InceptionV3 deep learning model, in *Digital Transformation and Emerging Technologies for Fighting COVID-19 Pandemic: Innovative Approaches*, Springer International Publishing, (2021), 111–119.
28. K. Shankar, E. Perumal, A novel hand-crafted with deep learning features based fusion model for COVID-19 diagnosis and classification using chest X-ray images, *Complex Intell. Syst.,* **7** (2020), 1277–1293.
29. O. M. Elzeki, M. Shams, S. Sarhan, M. Abd Elfattah, A. E. Hassanien, COVID-19: A new deep learning computer-aided model for classification, *PeerJ. Comp. Sci.,* **7** (2021), e358.
30. H. S. Alghamdi, G. Amoudi, S. Elhag, K. Saeedi, J. Nasser, Deep learning approaches for detecting COVID-19 from chest X-ray images: A survey, *IEEE Access,* **9** (2021), 20235–20254.

31. J. Born, N. Wiedemann, M. Cossio, C. Buhre, G. Brändle, K. Leidermann, et al., Accelerating detection of lung pathologies with explainable ultrasound image analysis, *Appl. Sci.,* **11** (2021), 672.
32. S. Roy, W. Menapace, S. Oei, B. Luijten, E. Fini, C. Saltori, et al., Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound, *IEEE Trans. Med. Imaging,* **39** (2020), 2676–2687.
33. T. Ai, Z. Yang, H. Hou, C. Zhan, C. Chen, W. Lv, et al., Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in China: A report of 1014 cases, *Radiology,* **296** (2020), E32–E40.
34. A. Bernheim, X. Mei, M. Huang, Y. Yang, Z. A. Fayad, N. Zhang, et al., Chest CT findings in coronavirus disease-19 (COVID-19): Relationship to duration of infection, *Radiology,* **295** (2020), 200463.
35. M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, et al., Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans, *Nat. Mach. Intell.,* **3** (2021), 199–217.
36. R. F. Wolff, K. G. M. Moons, R. D. Riley, P. F. Whiting, M. Westwood, G. S. Collins, et al., Probast: A tool to assess the risk of bias and applicability of prediction model studies, *Ann. Intern. Med.,* **170** (2019), 51–58.
37. Y. Ji, Z. Ma, M. P. Peppelenbosch, Q. Pan, Potential association between COVID-19 mortality and health-care resource availability, *Lancet Glob. Health,* **8** (2020), e480.
38. E. Tartaglione, C. A. Barbano, C. Berzovini, M. Calandri, M. Grangetto, Unveiling COVID-19 from chest X-ray with deep learning: A hurdles race with small data, *Int. J. Environ. Res. Public Health,* **17** (2020), 1–17.
39. OpenCV, *OpenCV: Histograms–2: Histogram equalization*, 2021. Available from: https://docs.opencv.org/master/d5/daf/tutorial_py_histogram_equalization.html.
40. K. Zuiderveld, Contrast limited adaptive histogram equalization, in *Graphics gems IV: Academic Press Professional*, Academic Press, (1994), 474–485.
41. Z. Al-Ameen, G. Sulong, A. Rehman, A. Al-Dhelaan, T. Saba, M. Al-Rodhaan, An innovative technique for contrast enhancement of computed tomography images using normalized gamma-corrected contrast-limited adaptive histogram equalization, *Eurasip J. Adv. Sig. Pr.,* **2015** (2015), 32.
42. A. Alekseev, A. Bobe, Gabornet: Gabor filters with learnable parameters in deep convolutional neural network, preprint, arXiv:1904.13204.
43. S. P. Morozov, A. E. Andreychenko, I. A. Blokhin, P. B. Gelezhe, A. P. Gonchar, A. E. Nikolaev, et al., MosMedData: data set of 1110 chest CT scans performed during the COVID-19 epidemic, *Dig. Diagnostics,* **1** (2020), 49–59.
44. V. Guarrasi, N. C. D'Amico, R. Sicilia, E. Cordelli, P. Soda, Pareto optimization of deep networks for COVID-19 diagnosis from chest X-rays, *Pattern Recognit.,* **121** (2022), 108242.
45. J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, E. K. Oermann, Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study, *PLoS Med.,* **15** (2018), e1002683.
46. P. Mooney, *Chest X-ray images (pneumonia)*. Available from: https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia.

47. G. Maguolo, L. Nanni, A critic evaluation of methods for COVID-19 automatic detection from X-ray images, *Inf. Fusion,* **76** (2021), 1–7.
48. J. Cohen, P. Morrison, L. Dao, COVID-19 image data collection, preprint, arXiv:2003.11597.
49. A. J. DeGrave, J. D. Janizek, S. I. Lee, AI for radiographic COVID-19 detection selects shortcuts over signal, *Nat. Mach. Intell.,* **3** (2021), 610–619.
50. X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R. M. Summers, ChestX-Ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 3462–3471.
51. J. Saborit, J. Montell, A. Pertusa, A. Bustos, M. Cazorla, J. Galant, et al., BIMCV COVID-19+: A large annotated dataset of RX and CT images from COVID-19 patients, preprint, arXiv:2006:01174.
52. A. Bustos, A. Pertusa, J. M. Salinas, M. de la Iglesia-Vayá, Padchest: A large chest X-ray image dataset with multi-label annotated reports, *Med. Imag. Anal.,* **66** (2020), 101797.
53. K. B. Ahmed, G. M. Goldgof, R. Paul, D. B. Goldgof, L. O. Hall, Discovery of a generalization gap of convolutional neural networks on COVID-19 X-rays classification, *IEEE Access,* **9** (2021), 72970–72979.
54. P. R. Bassi, R. Attux, COVID-19 detection using chest X-rays: Is lung segmentation important for generalization?, preprint, arXiv:2104.06176.
55. M. Elgendi, M. U. Nasir, Q. Tang, D. Smith, J.-P. Grenier, C. Batte, et al., The effectiveness of image augmentation in deep learning networks for detecting COVID-19: A geometric transformation perspective, *Frontiers Med.,* **8** (2021).
56. J. Shuja, E. Alanazi, W. Alasmary, A. Alashaikh, COVID-19 open source data sets: A comprehensive survey, *Appl. Intell.,* **51** (2020), 1296–1325.
57. M. Jun, G. Cheng, W. Yixin, A. Xingle, G. Jiantao, Y. Ziqi, et al., *COVID-19 CT lung and infection segmentation dataset (verson 1.0)*, Zenodo, 2020. Available from https://doi.org/10.5281/zenodo.3757476.
58. F. Shan, Y. Gao, J. Wang, W. Shi, N. Shi, M. Han, et al., Lung infection quantification of COVID-19 in CT images with deep learning, preprint, arXiv:2003.04655.
59. J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, M. Ghassemi, COVID-19 image data collection: Prospective predictions are the future, preprint, arXiv:2006.11988.
60. J. Zhao, Y. Zhang, X. He, P. Xie, COVID-CT-dataset: A CT scan dataset about COVID-19, preprint, arXiv:2003.13865.
61. MedRxiv, *the Preprint Server for Health Sciences*, Available from https://www.medrxiv.org.
62. BioRxiv, *the Preprint Server for Biology*, Available from https://www.biorxiv.org.
63. E. Soares, P. Angelov, S. Biaso, M. H. Froes, D. K. Abe, SARS-COV-2 CT-scan dataset: A large dataset of real patients CT scans for SARS-COV-2 identification, preprint, medRxiv:2020.04.24.20078584.
64. K. Zhang, X. Liu, J. Shen, Z. Li, Y. Sang, X. Wu, et al., Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography, *Cell*, **181** (2020), 1423–1433.
65. S. A. Duzgun, G. Durhan, F. B. Demirkazik, M. G. Akpinar, O. M. Ariyurek, COVID-19 pneumonia: The great radiological mimicker, *Insights Imaging,* **11** (2020), 118–118.

66. A. Krizhevsky, I. Sutskever, G. Hinton, Imagenet classification with deep convolutional neural networks, *Commun. ACM,* **60** (2017), 84–90.
67. S. K. Wajid, A. Hussain, K. Huang, W. Boulila, Lung cancer detection using local energy-based shape histogram (LESH) feature extraction and cognitive machine learning techniques, in *2016 IEEE 15th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC)*, (2016), 359–366.
68. R. Sarkar, A. Hazra, K. Sadhu, P. Ghosh, A novel method for pneumonia diagnosis from chest X-ray images using deep residual learning with separable convolutional networks, in *Computer Vision and Machine Intelligence in Medical Image Analysis*, Springer, (2019), 1−12.
69. S. Marcel, Y. Rodriguez, Torchvision the machine-vision package of Torch, in *Proceedings of the 18th ACM International Conference on Multimedia*, (2010), 1485–1488.
70. G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 2261−2269.
71. J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, et al., CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison, preprint, arXiv:1901.07031.
72. H. Pham, T. Le, D. Ngo, D. Tran, H. Nguyen, Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels, *Neurocomputing*, 437 (2021), 186−194.
73. I. Allaouzi, M. Ben Ahmed, A novel approach for multi-label chest X-ray classification of common thorax diseases, *IEEE Access,* **7** (2019), 64279−64288.
74. H. Wang, S. Wang, Z. Qin, Y. Zhang, R. Li, Y. Xia, Triple attention learning for classification of 14 thoracic diseases using chest radiography, *Med. Image Anal.,* **67** (2021), 64279−64288.
75. M. A. Morid, A. Borjali, G. Del Fiol, A scoping review of transfer learning research on medical image analysis using ImageNet, *Comput. Biol. Med.,* **128** (2021).
76. Pytorch.org, *Transfer Learning for Computer Vision Tutorial*. Available from https://pytorch.org/tutorials/beginner/transfer_learning_tutorial.html.
77. D. Kingma, J. Ba, ADAM: A method for stochastic optimization, preprint. arXiv:1412.6980.
78. L. Prechelt, Early stopping-but when?, in *Lecture Notes in Computer Science*, Springer Berlin, (2012), 53−67.
79. M. Horry, S. Chakraborty, M. Paul, A. Ulhaq, B. Pradhan, M. Saha, et al., COVID-19 detection through transfer learning using multimodal imaging data, *IEEE Access,* **8** (2020), 149808−149824.
80. T. C. Kwee, R. M. Kwee, Chest CT in COVID-19: What the radiologist needs to know, *Radiographics,* **40** (2020), 1848−1865.
81. J. L. Lehr, P. Capek, Histogram equalization of CT images, *Radiology,* **154** (1985), 163−169.
82. O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*, Springer Verlag, (2015), 234−241.

Appendix

Table A. Dataset quantitative properties.

Dataset	Image Total COVID Negative	Image Total COVID Positive	Patient Total COVID Negative	Patient Total COVID Positive	Image Format/Extraction	Dimensions (pixels)
COVID-CT	463	349	55	216	Compressed (PNG)	Variable
COVIDNet-CT-2A*	60083	94548	573	3055	Compressed (PNG)	Variable
SARS-COV-2	1230	1252	60	60	Compressed (PNG)	Variable
MID-CT	619	626	619	626	DICOM extracted middle slice to compressed (PNG)	512×512 (uniform)
MOSCOW	254	856	254	856	NifTI extracted slice 20 to Compressed (JPG)	512×512 (uniform)

* Pneumonia images excluded from analysis. MOSCOW and COVID-CT images removed from analysis to avoid inter-dataset pollution.