Diagnosing COVID-19 Severity from Chest X-Ray Images Using ViT and CNN Architectures

Luis Lara [1 2] **Lucia Eve Berger** [1 2] **Rajesh Raju** [1 2] **Shawn Whitfield** [1 2]

Abstract

The COVID-19 pandemic strained healthcare resources and prompted discussion about how machine learning can alleviate physician burdens and contribute to diagnosis. Chest x-rays (CXRs) are used for diagnosis of COVID-19, but few studies predict the severity of a patient's condition from CXRs. In this study, we produce a large COVID severity dataset by merging three sources and investigate the efficacy of transfer learning using ImageNet and CXR-pretrained models and vision transformers (ViTs) in both severity regression and classification tasks. A pretrained DenseNet161 model performed the best on the three class severity prediction problem, reaching 80% accuracy overall and 77.3%, 83.9%, and 70% on *mild*, *moderate* and *severe* cases, respectively. The ViT had the best regression results[1], with a mean absolute error of 0.5676 compared to radiologist-predicted severity scores. The project's source code is publicly available[2].

1. Introduction

The SARS-CoV2 (COVID-19) pandemic that began in late 2019 had a seismic effect on social patterns, technologies and health-care. COVID-19 often presents as pneumonia (COVID-pneumonia) and can quickly become life-threatening. Medical establishments across the world are still feeling the effects of increased disease burdens and shortages of qualified personnel. For healthcare practitioners needing to triage and decide where best to allocate their resources, cheap, scalable and easy-to-administer tests are essential. Molecular tests indicate whether or not a patient has COVID-19 but do not give an indication of its severity. Chest X-rays (CXRs) are favoured when diagnosing suspected COVID-19 (Rubin et al., 2020) because they are relatively cheap, can be taken bedside and can give an indication of severity.

Machine Learning (ML) technologies have been suggested since the beginning of the pandemic, and applied in emergency rooms (Shamout et al., 2021), as a means of reducing the burden on healthcare workers by increasing accuracy, accelerating diagnosis and suggesting outcomes. At this point, many ML studies have classified COVID-19 vs. non-COVID-19 patients with high accuracy (for example, (Khan et al., 2020; Minaee et al., 2020; Gupta et al., 2021; Jin et al., 2020; Oh et al., 2020; Tabik et al., 2020; Chakraborty et al., 2022; Wang et al., 2020a; Elkorany & Elsharkawy, 2021; Bhattacharyya et al., 2022; Nasir et al., 2023; Wang et al., 2021a)) but far fewer studies stratify patients by disease severity (Cohen et al., 2020a; Le Dinh et al., 2022; Shamout et al., 2021; Signoroni et al., 2021; Wong et al., 2021; Zandehshahvar et al., 2021) despite it being identified early as a critical task (Cohen et al., 2020c). This is partly because obtaining radiologist scores of severity adds an additional cost barrier, so fewer COVID CXR datasets contain this knowledge and are suitable for the task.

The contributions of the current study are:
1) Building a large COVID severity dataset by combining datasets with different severity scores into a single score.
2) Improving on state-of-the-art predictions of COVID severity for the regression task.
3) Comparing the performance of models pretrained on general images (e.g. ImageNet) with models specifically trained on CXRs.

We obtain competitive results in the classification (80% overall accuracy) and regression (0.5676 mean absolute error) tasks using pretrained models fine-tuned on our dataset. Development of ML models that accurately predict severity for COVID-19, while providing graphical explanations of its decisions, is useful for physicians to prioritize patients with the best chance of recovery or to intervene at an earlier time point, saving the health system money, helping patients and unburdening health care professionals.

---

[1]Université de Montréal, Canada [2]Mila Quebec AI Institute, Canada. Correspondence to: Luis Lara <luis.lara@mila.quebec>.

[1]https://huggingface.co/ludolara/vit-COVID-19-severity
[2]https://github.com/stwhitfield/covid-severity

2. Related work

Severity scoring

X-rays of patients with COVID-19 show opacity correlating with the degree of severity of the disease (Wong et al., 2020). Fluid build-up in lungs with COVID-pneumonia reduces the transmittance of the x-ray, resulting in cloudier/more opaque lung images Figure 1. Different scoring systems have been developed to attempt to quantify this severity; the most relevant to the current study are the Brixia score (Borghesi & Maroldi, 2020; Signoroni et al., 2021) and the opacity score (Cohen et al., 2020c;a).
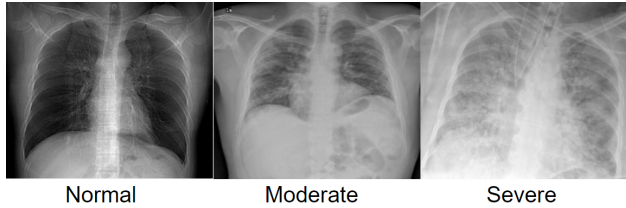

Figure 1. **Lung opacity often increases with COVID severity** Left to right: normal, moderate and severe levels of opacity in lung CXRs of patients with COVID.

Early in the pandemic, medical services in Brescia, Italy adopted the Brixia score (Borghesi & Maroldi, 2020), a semi-quantitative metric where lungs are divided into six regions, with each region assigned a rating from 0 to 3 based on severity as assessed by a radiologist (Borghesi & Maroldi, 2020; Signoroni et al., 2021). A modified version of the Brixia score, with higher spatial resolution, is additionally associated with severity (Jensen et al., 2022).

Cohen et al. compiled an early COVID-19 CXR dataset (Cohen et al., 2020c) that has been hugely influential in the CXR COVID-19 prediction space (Garcia Santa Cruz et al., 2021). A subset of their data (93 images) is accompanied by a score for degree of opacity (DO) (scored 0-6 for both lungs together) reflects the intensity of opacity: 0 = no opacity; 1 = ground glass opacity; 2 = consolidation; 3 = white-out (Cohen et al., 2020a). Since the degree of opacity can be mapped to the Brixia score by linear regression (Signoroni et al., 2021), we chose to use the opacity score as a unifying score for our study.

Prediction tasks

Authors generally take one of two approaches to severity prediction: treat the problem 1) as a classification task, and predict whether patients fall into mild, moderate or severe categories (Le Dinh et al., 2022; Tabik et al., 2020; Zandehshahvar et al., 2021; Gourdeau et al., 2022a), or 2) as a regression task and aim to match as closely as possible

the scores of professional radiologists (Cohen et al., 2020a; Signoroni et al., 2021). Due to the evolving nature of the COVID pandemic, many study results are not directly comparable. Authors will sometimes not release their code or their datasets, and report only one or two favourable metrics. For example, Wong et al. 2021 chose to report only the R^2 values for their regression prediction. Table 1 and Table 2 summarize some notable classification and regression scores.

Table 1. Classification accuracy (%) of previous studies per severity category.

STUDY	MILD	MODERATE	SEVERE
TABIK ET AL.	46	85	97
ZANDEHSHAHVAR ET AL.	80	68	65
LE DINH ET AL.	71	86	-
GOURDEAU ET AL.	49	55	52

Table 2. Regression scores of previous studies. MAE: Mean absolute error; MSE: Mean squared error; R^2: coefficient of determination.

STUDY	MAE	MSE	R^2
COHEN ET AL.	0.78	0.86	0.58
SIGNORONI ET AL.	0.67	0.67	0.53
WONG ET AL.	-	-	0.71
DANILOV ET AL.	0.3	0.66	-

Approaches

One popular approach is segmentation-based cropping and custom-built models. Tabik et al. (2020), Danilov et al. (2022) and Signoroni et al. (2021) all used segmentation-based cropping with a pre-trained segmentation models (often of U-net architecture) to focus on the lungs. They then used the output of these models as input for ResNet50, ResNet18 or MA-Net models, respectively, to predict severity. In contrast, Wong et al. (2021) adapted their previous custom-built COVID-Net architecture (Wang et al., 2020b), consisting of projection-expansion-projection patterns. Their previous model was trained on 14,000 CXRs ranging from healthy patients to those having different forms of pneumonia, making their severity prediction study partly a transfer-learning result.

Transfer learning is another popular approach (Kim et al., 2022), since competitive results can be obtained without the added complexity of custom modules. Le Dinh et al. (2022) tested five different deep learning models (DenseNet121, ResNet50, InceptionNet, Swin Transformer and a hybrid EfficientNet-DOLG) predictions on two different levels of severity, finding that the EfficientNet-DOLG and DenseNet 121 models performed the best. Similarly, Zandehshahvar

et al. (2021) also used a pretrained VGG-16 network to extract representations but then fine-tuned on a radiological pneumonia dataset before transferring these representations to a convolutional neural network that made predictions. Other studies directly pre-train on CXRs: studies by Gourdeau *et al.* (2022b; 2022a) and by Cohen *et al.* (2020a) used DenseNet121 networks pretrained on the CheXpert dataset (Irvin et al., 2019) or on a range of CXR datasets (Cohen et al., 2020b) for feature extraction, before prediction using linear regression models.

3. Methods

Data augmentation and pre-processing

Images were re-sized to 224 x 224 pixels, randomly rotated $\pm$ 30 degrees, randomly horizontally flipped, auto-contrasted, and equalized with probabilities of 0.5 each, and center cropped. Images to be used with torchxrayvision (Cohen et al., 2021b) pretrained models were converted to grayscale and normalized from (0,255) to (-1024,1024) to fit the expected input of these models.

Models

Pretrained models used in this study are given in Table 3. Image processing systems usually use convolutional neural networks (CNNs) which, due to the nature of the convolution operation, permit the model to invariantly learn spatial patterns while keeping the number of parameters relatively low. **AlexNet** (Krizhevsky et al., 2012) was a groundbreaking model in computer vision, showing that convolutions were a powerful tool for image recognition. **SqueezeNet** (Iandola et al., 2016) achieved similar results to AlexNet while having 50x fewer parameters due to architectural choices. **MobileNetv2** (Sandler et al., 2019) is a lightweight CNN with inverted residual structure modules. **VGG-16** (Simonyan & Zisserman, 2015) is a high-performing CNN with a deep architecture, using 3x3 kernel filters but 16 convolutional layers for a relatively simple but expressive network. **DenseNet** (Huang et al., 2018) models have layers that spread their weights over multiple inputs. This means that deeper layers can use features extracted early on, cutting down the total number of parameters and allowing very deep models. **DenseNet-121**, used extensively in this study, contains 120 convolutional layers and a final fully-connected layer.

VGG-16, AlexNet, DenseNet, MobileNet_v2 and SqueezeNet models were imported from pytorch with pretrained weights. DenseNet121 variants were imported from the torchxrayvision library (Cohen et al., 2021b) with pretrained weights (all, chex, pc, mimic_nb, mimic_ch, nih, and rsna). Each of these models are pretrained on a different source of CXRs, with datasets of varying sizes

(see Table 3 for sizes). All model weights other than those in classification layers were frozen. The final output layers of ImageNet-pretrained models were adjusted to output 1 value for the regression task, or 3 values for the classification task. For torchxrayvision models we added an extra two linear layers with a ReLU activation function in between to adapt the 18-class output of the model to our regression or classification tasks.

A transformer is a deep learning model that adopts the mechanism of self-attention, differentially weighting the significance of each part of the input data. Transformers found their initial applications in natural language processing (NLP) tasks (Devlin et al., 2019; Brown et al., 2020). The **Vision Transformer (ViT)** (Alexey et al., 2020) computes relationships among pixels in various small sections of the image (e.g., 16x16 pixels), at a drastically reduced cost. Images are presented to the model as a sequence of fixed-size patches (resolution 16x16), which are linearly embedded. The result is fed to the transformer and attention mechanisms applied. We imported a ViT from Hugging Face pretrained on 14 million images and fine-tuned it for 20 epochs on our dataset.

Table 3. Models used in this study. ViT: Vision Transformer. Parameters indicated are trainable parameters. M: million.

MODEL	# PARAMS	PRETRAINED ON
VGG-16	134 M	IMAGENET
VIT	86 M	IMAGENET-21K (14M)
ALEXNET	57 M	IMAGENET
DENSENET	29,8 M	IMAGENET
MOBILENET_V2	2,2 M	IMAGENET
SQUEEZENET	1,7 M	IMAGENET
DENSENET-ALL	7 M	CXRs (231K)
DENSENET-CHEX	7 M	CXRs (64K)
DENSENET-PC	7 M	CXRs (62K)
DENSENET-MIMIC_NB	7 M	CXRs (45K)
DENSENET-MIMIC_CH	7 M	CXRs (45K)
DENSENET-NIH	7 M	CXRs (30K)
DENSENET-RSNA	7 M	CXRs (30K)

Saliency maps

Saliency maps were based on the techniques in (Cohen et al., 2020a). Saliency maps are given by computing the gradient of the output prediction with respect to the input image, and indicate how much changing a pixel will change the prediction. We blurred the saliency map using a 5x5 Gaussian kernel to smooth it. Saliency maps have limitations and only offer a restricted view into why a model made a prediction (Ross et al., 2017; Viviano et al., 2019).

4. Experiments

Datasets

We combined three datasets for which there were severity scores available. The ieee8023 Covid-19 Image Data Collectiont (Cohen et al., 2020c), which we will refer to as the Cohen dataset, currently contains 481 images from covid-positive patients in PA (posterior-anterior) and AP (anterior-posterior) modalities, 93 of which have associated severity scores in the form of opacity scores. This dataset was manually curated and collected from sources including journal publications, and contains samples from across the globe. It has been expressly designed to be suitable for ML tasks and contains not only X-ray images but also physician annotations, collection data and metadata useful for prediction tasks. The Brixia score dataset (Signoroni et al., 2021), contains 4707 images of COVID-positive patients from Northern Italy with Brixia severity scores, in AP and PA projection. The data contain all the variability of a real clinical scenario, since they consist of all CXR images taken in sub-intensive and intensive care units during a month-long period of pandemic peak in the ASST Spedali Civili di Brescia. The RALO Stony Brook Medicine dataset (Cohen et al., 2021a) from the northeastern United States contains 2373 covid-positive images and associated opacity scores. All of these datasets have been scored by expert radiologists for severity assessment. Figure 2 shows the distributions of severity scores for each dataset in stacked histogram form. Although the dataset is imbalanced, it is representative of a real-world scenario.
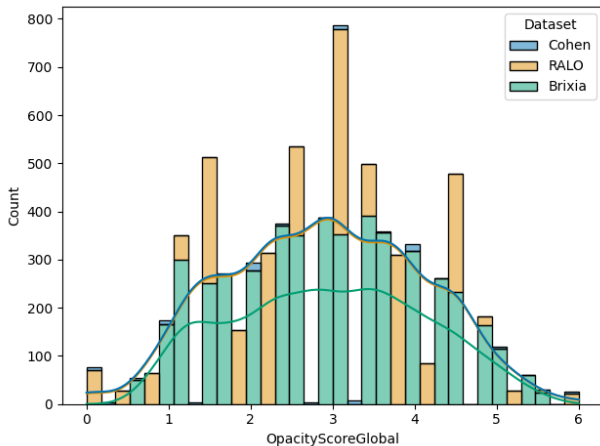

Figure 2. Distribution of opacity scores for each of the three datasets combined in this study, shown in stacked histogram form.

We downloaded each dataset from the link given above, extracting the Brixia images from .dcm format, a common medical imaging format, to .jpg format. These .dcm files have metadata associated with them and we attempted to extract metadata deemed relevant such as AdmittingDiagnosesDescription, PatientBirthDate, PatientSex and PatientAge but most fields were blank. For the Brixia paper, a subset of the Cohen data was scored by two Brixia radiologists to give Brixia scores. There is an overlap of 65 samples between these data and the original Cohen dataset that has severity scores. We used these 65 to train a linear regressor from scikit-learn to convert Brixia scores to opacity scores. The RALO dataset has opacity scores 0-8 while the Cohen dataset has opacity scores 0-6: we rescaled the RALO scores to 0-6. All datasets have multiple radiologist severity scores: we used the mean of these scores for each sample.

Baselines

To test the integrity of the data-pipeline, we replicated published baselines. We first focused on the simpler problem of binary classification (Covid or No-Covid) (Minaee et al., 2020). We used a VGG-16 model, loading the weights from ImageNet. We trained for 30 epochs on a subset of our dataset. While papers reported a high 89.2 accuracy, we reached 75% on the validation set.

We also conducted a binary classification experiment with with a CNN without loading pre-trained weights. With this model, we observed an early and evident overfit. The experiment resulted in 70% accuracy. We pursued the fine-tuning route, adapting large pre-trained image models.

Evaluation Methods

For the classification task, we assessed model loss, accuracy, precision and recall. Precision measures the true positives over the actual results (TP/TP+FP). Recall assesses true positives over true positives plus false negatives. We measured sensitivity, the metric that evaluates a model's ability to predict true negatives of each available category. We also included specificity,the metric that evaluates a model's ability to predict true positives of each available category. In comparing models, total parameters and training time were important real life consideration we collected.

For the regression task, we used mean absolute error (MAE), mean squared error (MSE) and the coefficient of determination (R^2) to compare our model predictions with the average score obtained from radiologists ("ground truth"). MAE is useful to understand the error in units of the predicted scores, while MSE is useful for measuring the quality of the estimator and help us rank and compare different models. The coefficient of determination provides a measure of how much variability of outcome is explained by the model.

Experimental Results

Classification experiments were conducted with a number of state-of-the-art pretrained image networks (VGG-16, VGG-19, DenseNet-161). These models were loaded and then the output layer was modified to fit our three and seven class experiments. In training, an Adam optimiser, with a learning rate scheduler was applied. Early stopping was applied to each of the models to prevent overfit. For most models, training stopped around 100 epochs.

For all experiments in the regression we used a PyTorch Lightning training framework. Images and associated metadata were loaded using a custom-made CovidSeverity_Dataset, with channels adjusted according to the input specifications of the model and transformations applied as described above. Data were randomly (random seed 42) split 70:10:20 into train:validation:test sets. As mentioned above, model weights other than those in the classifier layer were frozen, and the classifier layer was modified to output 1 value. Models were fine-tuned on our training set for 400 epochs using an AdamW optimizer with a learning rate of 0.001 and a batch size of 50. Metrics from torchmetrics were logged with Pytorch Lightning data loggers.

For the ViT we fine-tuned the Vision Transformer (ViT) (Google, 2021) from HuggingFace on our training set for 300 epochs using an AdamW optimizer with a learning rate of 0.00001 and a batch size of 10. Early stopping was applied to the model to prevent over-fitting. For ViT, training stopped after 70 epochs. The only difference between the preprocessed data in the ViT and the other models was that it normalized to a tensor image with mean and standard deviation during data augmentation. Best ViT model is available at https://huggingface.co/ludolara/vit-COVID-19-severity.

Results and Analysis

Results for the severity classification task are shown in Table 4. While we began experiments with seven distinct classes (corresponding to the severity score), we reached only 61% classification accuracy with seven classes. We believe that was largely due to the imbalance of the dataset. When we approached binned the severity labels into three categories, *"mild", "moderate"* and *"severe"* , our models performed better (Table 4). Our model performed best on the dominant class (*"moderate"*). Despite the VGG-16 having more trainable parameters, the DenseNet161 had better accuracy, highlighting the idea that architectural decisions are important.

The DenseNet161 model was able to distinguish *mild* cases quite well, but we saw some confusion between *moderate* and *severe* cases (Figure 3). We attribute these mistakes to the imbalanced nature of the dataset: it is possible these true

Model	#Classes	No FT	FT
VGG-16	7	3%	61%
VGG-16	3	0%	78%
DenseNet161	7	0%	**62%**
DenseNet161	3	5%	**80%**
ViT	7	-	50%

Table 4. Classification Accuracy of Different Pretrained CNNs (No FT refers to no fine-tuning, FT refers to fine tuning).

Model	Mild	Moderate	Severe
VGG-16	75.8%	82%	68%
DenseNet161	77.3%	83.9%	70%

Table 5. Classification Accuracy on the Testset Binned in Mild, Moderate and Severe.

values fall on the extremity of the bins. This is a consideration for future work.

Results for the severity regression task are given in Table 6. We obtained slightly better MAE, MSE and R^2 values than Signoroni *et al.* 2021, which we attribute to using a larger dataset. The agreement between our models' predictions and the radiologist predictions, indicated by the R^2 value, only reached around 50%. Since other studies have reported values similar to ours (Cohen et al., 2020a; Signoroni et al., 2021), it seems likely that this is a very hard problem and that there are other factors that a purely image-based prediction model cannot take into account.

Comparing the fine-tuning of models that are pretrained on CXRs and on general images (ImageNet), it is clear that pretraining on CXRs provides a discriminative boost to the model. Perhaps surprisingly, the SqueezeNet model performed as well or better than some CXR-pretrained models. These models were all trained on fewer than 50,000 images, suggesting that there is an approximate minimum of CXRs that are needed for the domain-specific prediction boost we see here. Still, pretraining on CXRs clearly improved results more than having an architecture with a high capacity: VGG-16 and AlexNet models did poorly in the regression task despite having a high number of trainable parameters.

The ViT had the best predictions in the regression task, with an MAE of 0.5676. We believe that this number is adequate in giving physicians a good indication of the severity of the COVID-pneumonia experienced by their patients. It is unclear to us whether the impressive results of the ViT compared to other models are due to the sheer number of images upon which the ViT was trained (14 million) or whether attention mechanisms provide a particular boost. Using a ViT pretrained from scratch or on a smaller dataset might help to answer this question. We also cannot rule out the possibility that a good subset of the 14 million images in the ViT's dataset are medical images, which would also help

Figure 3. Confusion Matrix of the DenseNet161 Testset (20% of dataset).

the model perform better without this improvement being due to attention-based and architectural factors.

Table 6. Regression results from this study, ranked by MSE. MAE: Mean absolute error; MSE: Mean squared error; R^2: coefficient of determination.

MODEL	MAE ↓	MSE ↓	R^2 ↑
VIT (BASE-SIZED MODEL)	**0.5676**	**0.5135**	**0.5378**
DENSENET-ALL	0.639	0.647	0.528
DENSENET-PC	0.679	0.716	0.489
DENSENET	0.676	0.723	0.478
DENSENET-CHEX	0.715	0.784	0.439
SQUEEZENET	0.725	0.794	0.430
DENSENET-NIH	0.723	0.796	0.430
DENSENET-MIMIC_NB	0.733	0.811	0.422
DENSENET-MIMIC_CH	0.748	0.850	0.389
MOBILENET_V2	0.750	0.868	0.382
DENSENET-RSNA	0.802	0.958	0.315
VGG-16	0.796	0.962	0.308
ALEXNET	0.790	0.963	0.315

Being able to explain an ML model's decisions is a key part of ensuring stakeholder confidence in the model. We generated saliency maps (Figure 4) using our DenseNet-all torchxrayvision pretrained model to help us understand which regions of the image are influencing the model's decisions. Different images had different focal points, suggesting that the model is looking at features of the images and not purely shortcut learning. Without consulting experts (radiologists), however, it is difficult to judge whether the model is focusing on the right regions.

Conclusion

In conclusion, we have leveraged the power of machine learning methods for predicting the severity of the COVID-19 condition of a patient from the chest X-rays (CXRS) by
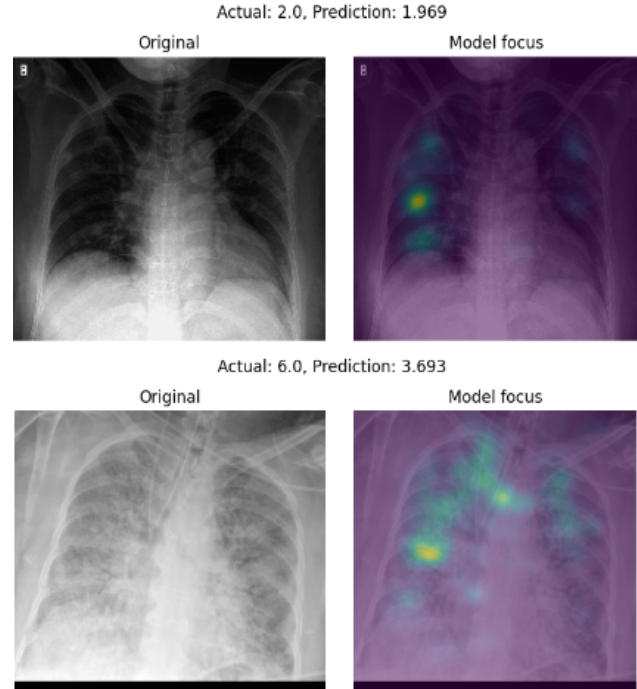

Figure 4. **Examples of saliency maps generated in this study.** Examples of better (top) and worse (bottom) regression predictions show that model gradients are different for different images, suggesting that the model is focusing on different regions. Yellow denotes zones that affect the model's gradients more when the image is run through the model. Actual and predicted opacity scores are given above the image. Images shown here were generated with the DenseNet-all model.

generating a large single dataset from three different data sources. We have approached the problem as regression and classification severity prediction tasks and used different pre-trained ImageNet-models VGG-16, AlexNet, DenseNet, MobileNet_V2, SqueezeNet, vision transformer (ViT) and chest-X ray pre-trained model DenseNet-121.

DenseNet161 and VGN16 pretrained models with fine tuning achieved an accuracy of 80% and 78% with three classes (low, medium, high) and 62% and 61% with seven classes (0-6 scale). We have found that all models exhibit worst performances without fine tuning. Moreover, we did not obtain remarkable results for seven classes for ViT. Further dissection of "edge cases" is necessary to understand how many categories of severity our models should be able to reliably predict. For regression tasks, we found X-ray pre-trained models generally exhibited superior performance over ImageNet pre-trained models.

Our results are competitive with others in the COVID severity prediction space (Zandehshahvar et al., 2021; Tabik et al., 2020; Signoroni et al., 2021; Cohen et al., 2020a) without

training from scratch or using more complicated architectures. Further work could build on our framework and determine exactly how much of an improvement is added by lung segmentation, as has been done in other studies (Danilov et al., 2022; Tabik et al., 2020; Wang et al., 2021b). The ViT has the best result for regression with the lowest MSE value of 0.5135. Further work could determine what properties of the ViT made it more successful than other models for our task. We also developed saliency maps using the DenseNet-all torchxrayvision pre-trained model which help us to identify which regions of the image are influencing the model's decisions, although other approaches such as GRAD-CAM (Selvaraju et al., 2016) could be explored as complementary ways of improving stakeholder confidence in machine learning predictions in radiology.

Contributions of Each Team Member

Luis Lara performed dataset preprocessing, creation of testing modules, investigation of efficacy of MLPMixer model and ViT for a seven-class prediction task, produced the results for the ViT model for the regression task, compiled Table 3, contributed the ViT part of the Experimental Results section and contributed ideas to the Results and Analysis regression discussion.

Lucia Eve Berger performed the VGG-16 baseline (binary classification) experiment, severity classification with VGG-16 and densenet CXR models, and contributed the classification part of the following written sections: Baselines, Evaluation Methods, Experimental Results, and Results and Analysis.

Rajesh Raju investigated data preprocessing including image and text augmentation methods, investigated the utility of clinical notes, resized CXR images to 224 x 224, and wrote part of the conclusion.

Shawn Whitfield combined the three CXR datasets, including extracting .dcm files and mapping severity scores into opacity scores, built the PyTorch Lightning framework to facilitate train:validation cycles, performed all regression experiments on pretrained ImageNet and torchxrayvision models (other than the ViT), implemented the saliency mapping, and wrote the following sections of the paper: Abstract; Introduction; Related work; Methods - data augmentation, models, saliency maps; Experiments - datasets, evaluation methods, experimental results (regression), results and analysis (regression, saliency); part of the conclusion.

References

Alexey, D., Lucas, B., Alexander, K., Dirk, W., Xiaohua, Z., Thomas, U., Mostafa, D., Matthias, M., Georg, H., Sylvain, G., Jakob, U., and Neil, H. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

Bhattacharyya, A., Bhaik, D., Kumar, S., Thakur, P., Sharma, R., and Pachori, R. B. A deep learning based approach for automatic detection of COVID-19 cases using chest X-ray images. *Biomed Signal Process Control*, 71: 103182, 2022. doi: 10.1016/j.bspc.2021.103182. URL https://www.ncbi.nlm.nih.gov/pubmed/34580596.

Borghesi, A. and Maroldi, R. COVID-19 outbreak in Italy: experimental chest X-ray scoring system for quantifying and monitoring disease progression. *La radiologia medica*, 125(5):509–513, 2020. doi: 10.1007/s11547-020-01200-3.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. 2020.

Chakraborty, S., Murali, B., and Mitra, A. K. An Efficient Deep Learning Model to Detect COVID-19 Using Chest X-ray Images. *International Journal of Environmental Research and Public Health*, 19(4):2013, 2022. doi: 10.3390/ijerph19042013.

Cohen, J. P., Dao, L., Roth, K., Morrison, P., Bengio, Y., Abbasi, A. F., Shen, B., Mahsa, H. K., Ghassemi, M., Li, H., and Duong, T. Predicting COVID-19 Pneumonia Severity on Chest X-ray With Deep Learning. *Cureus*, 2020a. ISSN 2168-8184. doi: 10.7759/cureus.9448.

Cohen, J. P., Hashir, M., Brooks, R., and Bertrand, H. On the limits of cross-domain generalization in automated x-ray prediction. 2020b. doi: 10.48550/ARXIV.2002.02497. URL https://arxiv.org/abs/2002.02497.

Cohen, J. P., Morrison, P., Dao, L., Roth, K., Duong, T. Q., and Ghassemi, M. Covid-19 image data collection: Prospective predictions are the future. 2020c. doi: 10.48550/ARXIV.2006.11988. URL https://arxiv.org/abs/2006.11988.

Cohen, J. P., Shen, B., Abbasi, A., Hoshmand-Kochi, M., Glass, S., Li, H., Lungren, M. P., Chaudhari, A., and Duong, T. Q. Radiographic assessment of lung opacity score dataset. Mar 2021a. doi: 10.5281/zenodo.4634000.

Cohen, J. P., Viviano, J. D., Bertin, P., Morrison, P., Torabian, P., Guarrera, M., Lungren, M. P., Chaudhari, A., Brooks, R., Hashir, M., and Bertrand, H. Torchxrayvision: A library of chest x-ray datasets and models. 2021b.

Danilov, V. V., Litmanovich, D., Proutski, A., Kirpich, A., Nefaridze, D., Karpovsky, A., and Gankin, Y. Automatic scoring of COVID-19 severity in X-ray imaging based on a novel deep learning workflow. *Scientific Reports*, 12(1), 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-150 13-z.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. 2019.

Elkorany, A. S. and Elsharkawy, Z. F. COVIDetection-Net: A tailored COVID-19 detection from chest radiography images using deep learning. *Optik (Stuttg)*, 231:166405, 2021. doi: 10.1016/j.ijleo.2021.166405. URL https://www.ncbi.nlm.nih.gov/pubmed/3355149 2.

Garcia Santa Cruz, B., Bossa, M. N., Sölter, J., and Husch, A. D. Public Covid-19 X-ray datasets and their impact on model bias – A systematic review of a significant problem. *Medical Image Analysis*, 74:102225, dec 2021. ISSN 1361-8415. doi: 10.1016/J.MEDIA.2021.102225.

Google. google/vit-base-patch16-224-in21k. Hugging Face model hub, 2021. URL https://huggingface.co/google/vit-base-patch16-224-in21k.

Gourdeau, D., Potvin, O., Archambault, P., Chartrand-Lefebvre, C., Dieumegarde, L., Forghani, R., Gagné, C., Hains, A., Hornstein, D., Le, H., Lemieux, S., Lévesque, M.-H., Martin, D., Rosenbloom, L., Tang, A., Vecchio, F., Yang, I., Duchesne, N., and Duchesne, S. Tracking and predicting COVID-19 radiological trajectory on chest X-rays using deep learning. *Scientific Reports*, 12(1), 2022a. ISSN 2045-2322. doi: 10.1038/s41598-022-09356-w.

Gourdeau, D., Potvin, O., Biem, J. H., Cloutier, F., Abrougui, L., Archambault, P., Chartrand-Lefebvre, C., Dieumegarde, L., Gagné, C., Gagnon, L., Giguère, R., Hains, A., Le, H., Lemieux, S., Lévesque, M.-H., Nepveu, S., Rosenbloom, L., Tang, A., Yang, I., Duchesne, N., and Duchesne, S. Deep learning of chest X-rays can predict mechanical ventilation outcome in ICU-admitted COVID-19 patients. *Scientific Reports*, 12(1), 2022b. doi: 10.1038/s41598-022-10136-9.

Gupta, A., Anjum, Gupta, S., and Katarya, R. InstaCovNet-19: A deep learning classification model for the detection of COVID-19 patients using Chest X-ray. *Appl Soft Comput*, 99:106859, 2021. doi: 10.1016/j.asoc.2020.106859. URL https://www.ncbi.nlm.nih.gov/pubmed/33162872.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks, 2018.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., and Keutzer, K. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and ¡0.5mb model size, 2016.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R. L., Shpanskaya, K. S., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., and Ng, A. Y. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *CoRR*, abs/1901.07031, 2019. URL http://arxiv.org/abs/1901.07031.

Jensen, C. M., Costa, J. C., Nørgaard, J. C., Zucco, A. G., Neesgaard, B., Niemann, C. U., Ostrowski, S. R., Reekie, J., Holten, B., Kalhauge, A., Matthay, M. A., Lundgren, J. D., Helleberg, M., and Moestrup, K. S. Chest x-ray imaging score is associated with severity of COVID-19 pneumonia: the MBrixia score. *Scientific Reports*, 12(1), 2022. doi: 10.1038/s41598-022-25397-7.

Jin, C., Chen, W., Cao, Y., Xu, Z., Tan, Z., Zhang, X., Deng, L., Zheng, C., Zhou, J., Shi, H., and Feng, J. Development and evaluation of an artificial intelligence system for COVID-19 diagnosis. *Nature Communications*, 11(1), 2020. doi: 10.1038/s41467-020-18685-1.

Khan, A. I., Shah, J. L., and Bhat, M. M. CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images. *Comput Methods Programs Biomed*, 196:105581, 2020. doi: 10.1016/j.cmpb.2020.105581. URL https://www.ncbi.nlm.nih.gov/pubmed/32534344.

Kim, H. E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M. E., and Ganslandt, T. Transfer learning for medical image classification: A literature review. *BMC Medical Imaging*, 22(1), 2022. doi: 10.1186/s12880-022-00793-7.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Le Dinh, T., Lee, S.-H., Kwon, S.-G., and Kwon, K.-R. COVID-19 Chest X-ray Classification and Severity Assessment Using Convolutional and Transformer Neural Networks. *Applied Sciences*, 12(10):4861, 2022. doi: 10.3390/app12104861.

Minaee, S., Kafieh, R., Sonka, M., Yazdani, S., and Jamalipour Soufi, G. Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning. *Med Image Anal*, 65:101794, 2020. doi: 10.1016/j.media.2020.101794. URL https://www.ncbi.nlm.nih.gov/pubmed/32781377.

Nasir, N., Kansal, A., Barneih, F., Al-Shaltone, O., Bonny, T., Al-Shabi, M., and Al Shammaa, A. Multi-modal image classification of COVID-19 cases using computed tomography and X-rays scans. *Intelligent Systems with Applications*, 17, 2023. doi: 10.1016/j.iswa.2022.200160.

Oh, Y., Park, S., and Ye, J. C. Deep Learning COVID-19 Features on CXR Using Limited Training Data Sets. *IEEE Transactions on Medical Imaging*, 39(8):2688–2700, 2020. doi: 10.1109/tmi.2020.2993291.

Ross, A. S., Hughes, M. C., and Doshi-Velez, F. Right for the right reasons: Training differentiable models by constraining their explanations. *CoRR*, abs/1703.03717, 2017. URL http://arxiv.org/abs/1703.03717.

Rubin, G. D., Ryerson, C. J., Haramati, L. B., Sverzellati, N., Kanne, J. P., Raoof, S., Schluger, N. W., Volpi, A., Yim, J.-J., Martin, I. B. K., Anderson, D. J., Kong, C., Altes, T., Bush, A., Desai, S. R., Goldin, O., Goo, J. M., Humbert, M., Inoue, Y., Kauczor, H.-U., Luo, F., Mazzone, P. J., Prokop, M., Remy-Jardin, M., Richeldi, L., Schaefer-Prokop, C. M., Tomiyama, N., Wells, A. U., and Leung, A. N. The Role of Chest Imaging in Patient Management during the COVID-19 Pandemic: A Multinational Consensus Statement from the Fleischner Society. *Radiology*, 296(1):172–180, 2020. doi: 10.1148/radiol.2020201365.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks, 2019.

Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., and Batra, D. 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.

Shamout, F. E., Shen, Y., Wu, N., Kaku, A., Park, J., Makino, T., Jastrzębski, S., Witowski, J., Wang, D., Zhang, B., Dogra, S., Cao, M., Razavian, N., Kudlowitz, D., Azour, L., Moore, W., Lui, Y. W., Aphinyanaphongs, Y., Fernandez-Granda, C., and Geras, K. J. An artificial intelligence system for predicting the deterioration of COVID-19 patients in the emergency department. *npj Digital Medicine*, 4(1), 2021. doi: 10.1038/s41746-021-00453-0.

Signoroni, A., Savardi, M., Benini, S., Adami, N., Leonardi, R., Gibellini, P., Vaccher, F., Ravanelli, M., Borghesi, A., Maroldi, R., and Farina, D. BS-Net: Learning COVID-19 pneumonia severity on a large chest X-ray dataset. *Med Image Anal*, 71:102046, 2021. doi: 10.1016/j.media.2021.102046. URL https://www.ncbi.nlm.nih.gov/pubmed/33862337.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. 2015.

Tabik, S., Gomez-Rios, A., Martin-Rodriguez, J. L., Sevillano-Garcia, I., Rey-Area, M., Charte, D., Guirado, E., Suarez, J. L., Luengo, J., Valero-Gonzalez, M. A., Garcia-Villanova, P., Olmedo-Sanchez, E., and Herrera, F. COVIDGR Dataset and COVID-SDNet Methodology for Predicting COVID-19 Based on Chest X-Ray Images. *IEEE Journal of Biomedical and Health Informatics*, 24(12):3595–3605, 2020. ISSN 2168-2194. doi: 10.1109/jbhi.2020.3037127.

Viviano, J. D., Simpson, B., Dutil, F., Bengio, Y., and Cohen, J. P. Underwhelming generalization improvements from controlling feature attribution. *CoRR*, abs/1910.00199, 2019. URL http://arxiv.org/abs/1910.00199.

Wang, D., Mo, J., Zhou, G., Xu, L., and Liu, Y. An efficient mixture of deep and machine learning models for COVID-19 diagnosis in chest X-ray images. *PLOS ONE*, 15(11): e0242535, 2020a. doi: 10.1371/journal.pone.0242535. URL https://www.ncbi.nlm.nih.gov/pubmed/33201919.

Wang, G., Liu, X., Shen, J., Wang, C., Li, Z., Ye, L., Wu, X., Chen, T., Wang, K., Zhang, X., Zhou, Z., Yang, J., Sang, Y., Deng, R., Liang, W., Yu, T., Gao, M., Wang, J., Yang, Z., Cai, H., Lu, G., Zhang, L., Yang, L., Xu, W., Wang, W., Olvera, A., Ziyar, I., Zhang, C., Li, O., Liao, W., Liu, J., Chen, W., Chen, W., Shi, J., Zheng, L., Zhang, L., Yan, Z., Zou, X., Lin, G., Cao, G., Lau, L. L., Mo, L., Liang, Y., Roberts, M., Sala, E., Schönlieb, C.-B., Fok, M., Lau, J. Y.-N., Xu, T., He, J., Zhang, K., Li, W., and Lin, T. A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral and COVID-19 pneumonia from chest X-ray images. *Nature Biomedical Engineering*, 5(6):509–521, 2021a. ISSN 2157-846X. doi: 10.1038/s41551-021-00704-1.

Wang, L., Lin, Z. Q., and Wong, A. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Scientific Reports*, 10(1), 2020b. ISSN 2045-2322. doi: 10.1038/s41598-020-76550-z.

Wang, Z., Xiao, Y., Li, Y., Zhang, J., Lu, F., Hou, M., and Liu, X. Automatically discriminating and localiz-

ing COVID-19 from community-acquired pneumonia on chest X-rays. *Pattern Recognit*, 110:107613, 2021b. doi: 10.1016/j.patcog.2020.107613. URL https://www. ncbi.nlm.nih.gov/pubmed/32868956.

Wong, A., Lin, Z. Q., Wang, L., Chung, A. G., Shen, B., and Abbasi, A. Towards computer - aided severity assessment via deep neural networks for geographic and opacity extent scoring of SARS - CoV - 2 chest X - rays. *Scientific Reports*, pp. 1–8, 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-88538-4. URL https://doi.org/10.1038/s41598-021-88538-4.

Wong, H. Y. F., Lam, H. Y. S., Fong, A. H.-T., Leung, S. T., Chin, T. W.-Y., Lo, C. S. Y., Lui, M. M.-S., Lee, J. C. Y., Chiu, K. W.-H., Chung, T. W.-H., Lee, E. Y. P., Wan, E. Y. F., Hung, I. F. N., Lam, T. P. W., Kuo, M. D., and Ng, M.-Y. Frequency and Distribution of Chest Radiographic Findings in Patients Positive for COVID-19. *Radiology*, 296(2):E72–E78, 2020. doi: 10.1148/radiol.2020201160.

Zandehshahvar, M., Van Assen, M., Maleki, H., Kiarashi, Y., De Cecco, C. N., and Adibi, A. Toward understanding COVID-19 pneumonia: a deep-learning-based approach for severity analysis and monitoring the disease. *Scientific Reports*, 11(1), 2021. doi: 10.1038/s41598-021-90411-3.