Hierarchical Analysis of Visual COVID-19 Features from Chest Radiographs

Shruthi Bannur [1] **Ozan Oktay** [1] **Melanie Bernhardt** [1] **Anton Schwaighofer** [1] **Rajesh Jena** [1] **Besmira Nushi** [2] **Sharan Wadhwani** [3] **Aditya Nori** [1] **Kal Natarajan** [3] **Shazad Ashraf** [3] **Javier Alvarez-Valle** [1] **Daniel C. Castro** [1]

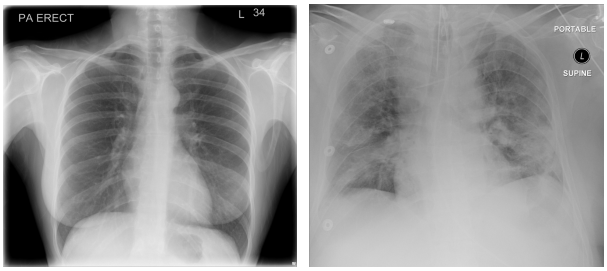
Abstract

Chest radiography has been a recommended procedure for patient triaging and resource management in intensive care units (ICUs) throughout the COVID-19 pandemic. The machine learning efforts to augment this workflow have been long challenged due to deficiencies in reporting, model evaluation, and failure mode analysis. To address some of those shortcomings, we model radiological features with a human-interpretable class hierarchy that aligns with the radiological decision process. Also, we propose the use of a data-driven error analysis methodology to uncover the blind spots of our model, providing further transparency on its clinical utility. For example, our experiments show that model failures highly correlate with ICU imaging conditions and with the inherent difficulty in distinguishing certain types of radiological features. Also, our hierarchical interpretation and analysis facilitates the comparison with respect to radiologists' findings and inter-variability, which in return helps us to better assess the clinical applicability of models.

1. Introduction

Patients affected with COVID-19 frequently experience an upper respiratory tract infection or pneumonia that can rapidly progress to acute respiratory failure, multiple organ failure and death (Zhou et al., 2020). Chest radiography (chest X-ray; CXR) is a front-line tool that is used in screening and triaging varieties of pneumonia due to the diagnostic role of imaging features (Toussie et al., 2020) and its quick turnaround time (Wong et al., 2020), which makes it convenient for patient management in intensive care units. However, in public healthcare systems (e.g. UK NHS) it can take several hours from CXR acquisition until a reporting radiologist is available to provide such a reading or RT-PCR test results become available. Thus, in this clinical context, it is especially relevant to build automated CXR image analysis systems that can benefit front-line patient management processes to decide on the clinical pathway for each patient by providing radiological feedback at point of care. This way, hospitals can better allocate resources and reduce COVID-19 contamination risks before other clinical data (e.g. lab tests) is made available.

There have been several efforts to leverage machine learning models to automate this CXR reading process, reviewed in Wynants et al. (2020). For example, Wehbe et al. (2021) benchmarked deep networks on a multi-site dataset comprised of thousands of CXR scans, where the models were trained to predict RT-PCR test results. Instead, in this study, we aim to predict a radiologist's impression of a chest image



(a) 'NORMAL' (b) 'CLASSIC' COVID-19

(c) 'INDET.' for COVID-19 (d) 'OTHER' abnormality

Figure 1. Example chest X-rays from each category. (a) Normal lungs appear mostly black. (b) Bilateral, peripheral opacities (white 'clouds') in lower lung lobe. (c) Non-typical COVID features, with peripheral coarse white lines. (d) Here, a right-sided basal pleural effusion (fluid accumulation in the left of the image). Images reproduced with permission.

[1]Microsoft Research, Cambridge, UK [2]Microsoft Research, Redmond, USA [3]University Hospitals Birmingham, UK. Correspondence to: Daniel C. Castro <dacoelh@microsoft.com>.

Workshop on Interpretable ML in Healthcare at International Conference on Machine Learning (ICML).

by identifying radiological features associated with COVID-19 pneumonia, rather than lab results or patient outcomes. CXR alone is known not to have enough diagnostic power for predicting RT-PCR results (Cleverley et al., 2020), hence the proposed tool is *not* intended for diagnosis or prognosis on its own. Rather, it provides supporting evidence in addition to other readings and test results.

The structure of model outputs is directly informed by radiologists' decision-making process and was developed in close collaboration with a hospital trust, in contrast to past COVID-19 ML studies performed without clinical involvement (Tizhoosh & Fratesi, 2021). In detail, we propose a post-hoc interpretation of model outputs representing the clinically meaningful hierarchical relationships between target classes, as illustrated in Fig. 2 (cf. Chen et al., 2019).

Our study also focuses on the potential issues and varying practices around model evaluation and biases across different hospital settings (Roberts et al., 2021). Such pitfalls can cast a doubt on the clinical applicability of the models evaluated in previous studies. As a step in this direction, we propose the use of a dedicated error analysis methodology (Nushi et al., 2018) to understand and communicate the dependency between model's failure modes and sample attributes, such as sample difficulty and image acquisition settings. In this way, users can be made aware of reliable operating regimes of such models and ensure clinical safety whilst offering actionable solutions to address such issues.

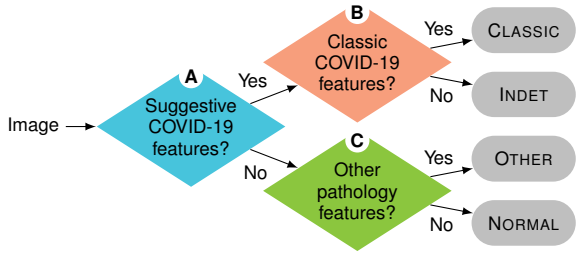
2. Methodology

2.1. Label definitions

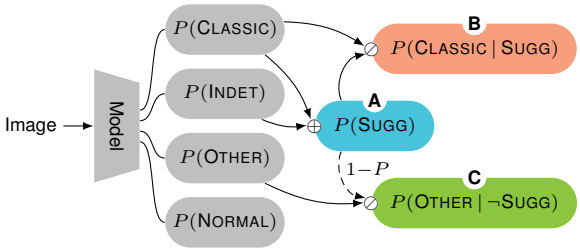
The prediction targets were adapted from the British Society for Thoracic Imaging (BSTI)'s reporting guidelines for COVID-19 findings from chest radiographs (BSTI, 2020):

- NORMAL: No radiological features of COVID-19 nor of other abnormalities.
- CLASSIC: Stereotypical presentation, e.g. lower lobe, peripheral predominant, multifocal, bilateral opacities.
- INDET(ERMINATE): Findings that are compatible with COVID-19 presentation but nonspecific.
- OTHER: Abnormalities that are not suggestive of COVID-19, e.g. pneumothorax, lobar pneumonia, pleural effusion, pulmonary oedema, etc.

Examples are shown in Fig. 1. The labelling process of our partner radiologists is visualised in Fig. 2(a) and can be summarised as follows: if there are any visual features suggestive of COVID-19, the decision is narrowed to CLASSIC vs INDET—even if the image presents other findings like heart failure. In other words, an image is labelled OTHER only if it shows no signs of COVID-19. Note that these categories refer exclusively to the presented radiological features, and labels are assigned with no access to other relevant information such as the patient record or past scans.



(a) Radiologists' decision process



(b) Hierarchical interpretation of multi-class predictions

Figure 2. Clinically informed model interpretation. The model's outputs are mapped to the radiologists' decision branches (A/B/C); branch probabilities are computed according to Eq. (1).

2.2. Hierarchical analysis of COVID-19 features

To enable analysing model outputs at clinically meaningful levels of abstraction, we propose to hierarchically aggregate class probabilities reflecting the radiologists' decision process (Fig. 2). Unlike the related work of Chen et al. (2019), who modelled a taxonomy of (non-COVID) thoracic abnormalities with a hierarchical model architecture, we focus on post-hoc interpretation of a vanilla multi-class predictor.

Specifically, the model is a conventional CNN classifier trained with cross-entropy loss (details in Appendix B). We then use the four-class model outputs to compute normalised binary probabilities for each of the decision branches:

$$\begin{aligned} P(\text{SUGG}) &= P(\text{CLASSIC}) + P(\text{INDET})\,, \\ P(\text{CLASSIC}\,|\,\text{SUGG}) &= P(\text{CLASSIC})/P(\text{SUGG})\,, \\ P(\text{OTHER}\,|\,\neg\text{SUGG}) &= P(\text{OTHER})/(1-P(\text{SUGG}))\,, \end{aligned} \tag{1}$$

where 'SUGG' refers to 'suggestion of COVID-19' (i.e. 'CLASSIC or INDET').

2.3. Self-supervised pre-training

The scarcity of large, annotated datasets in chest radiology has been one of the major drivers for leveraging self-supervised model pre-training (Sriram et al., 2021). As self-supervision does not require image labels, it significantly reduces the expert annotation burden and enables local clinical sites to build on pre-trained models using much smaller

private datasets. Further, because the external datasets' labels that may not align with the target task, self-supervised learning partly mitigates widespread issues in COVID-19 studies around inappropriate ground-truth labels and misuse of external datasets (Roberts et al., 2021).

In our study, we employ the BYOL algorithm (Grill et al., 2020) and experiment with two large chest X-ray datasets for model pre-training: NIH-CXR (Wang et al., 2017), with 112,120 frontal-view CXR scans, and CheXpert (Irvin et al., 2019), with 224,316. Both datasets were acquired from large cohorts of subjects diagnosed with chest pathologies including consolidation, pneumonia, lung nodules, etc. Even though these external datasets contain only pre–COVID-19 scans, self-supervision enables the model to learn generic characteristics of lung lobes, opacities, and nodules, which are useful in quantifying COVID-19–related features.

3. Datasets

COVID-19 CXR dataset: The labelled images used for supervised fine-tuning come from a retrospective dataset comprising de-identified chest radiographs, collected in the UK across multiple sites of the University Hospitals Birmingham NHS Foundation Trust during the first COVID-19 wave (1st March to 7th June 2020). The study participants were consecutive patients who had CXR taken for suspected COVID-19 infection, presented in the emergency department, acute medical unit, or inpatient unit. The initial dataset with 6125 images was curated by excluding duplicates and poor-quality scans, resulting in 4,940 usable images (NORMAL: 1154, CLASSIC: 1778, INDET: 1093, OTHER: 915). Of the 3639 unique subjects in the curated dataset, 515 (14.2%) were aged below 40, 2931 (80.5%) were in the 40–89 age group, and the remaining 193 (5.3%) were over 90; 1622 (44.6%) were female patients.

Class labels were extracted from reports by consultant radiologists or specialist radiographers, then blindly reviewed by a radiologist based on the images alone to mitigate biases due to availability of clinical side-information. Cases for which the assigned label disagreed with the original reports were further reviewed by a clinician to reach a consensus. Although not used as a prediction endpoint, RT-PCR test status at imaging time was also recorded for evaluation.

Multi-label test dataset: A subset of this data was held out from training to evaluate the model's predictive performance against a diverse panel of front-line radiology reporters, in an inter-observer variability (IOV) study. It contains 400 images acquired in April 2020 (approx. 100 consecutive patients from each of the four categories), with age and gender distributions similar to the training set. This test dataset was labelled into the four categories separately by three annotators with varying levels of experience in chest radiology: a consultant (*Ann. 1*) and a trainee chest radiologists (*Ann. 2*), and a non-specialist clinician (*Ann. 3*). They had access only to these 400 images, without any other clinical context.

Table 1. Test accuracies of the model and clinicians ($N = 400$; NORMAL: 100, CLASSIC: 101, INDET: 98, OTHER: 101; κ: Fleiss' kappa statistic, indicating inter-annotator agreement)

Classification task	Model	Ann. 1	Ann. 2	Ann. 3	(κ)
SUGG. COVID	**.800**	.775	.730	.700	(.491)
CLASSIC vs INDET	**.724**	.608	.482	.588	(.245)
NORMAL vs OTHER	**.791**	.731	.706	.721	(.488)
Multi-class	**.588**	.563	.463	.488	(.408)

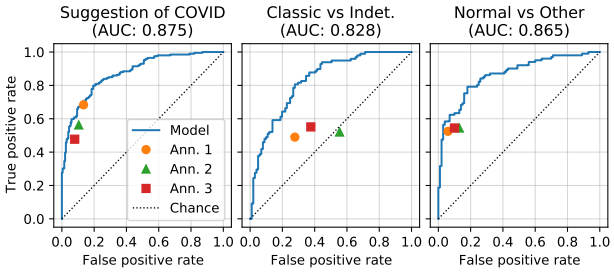


Figure 3. ROC curves for hierarchically aggregated model predictions, compared to clinicians' performance

4. Results

A DenseNet-121 backbone (Huang et al., 2017) was first pre-trained with self-supervision on NIH-CXR (cf. comparison with CheXpert in Appendix A), then fine-tuned on the private COVID-19 training dataset with 5-fold cross-validation. It was ensured that all images from the same patient were either in the training or in the validation set (see Appendix B). The results discussed below are based on an ensemble of the 5 trained models for the best hyperparameter configuration. Our open-source implementation is available at `https://aka.ms/innereyeoss`.

4.1. Classification performance

Table 1 presents the test accuracies for multi-class and for each binary classification branch illustrated in Fig. 2 (A, B, and C). The corresponding ROC analysis of these binary sub-tasks is shown in Fig. 3. Reported results for 'CLASSIC vs INDET' and 'NORMAL vs OTHER' include only images with the relevant labels.

We see that the model outperforms the clinicians across all hierarchical and multi-class tasks, with respect to the reference labels defined as in the training dataset. Our model achieves lower scores on 'CLASSIC vs INDET', seemingly the hardest sub-task of the three, and the clinicians perform closer to chance level. However, this is an inherently

ambiguous problem—discussions with the annotators (including the original labellers) revealed they tended to use different thresholds for distinguishing these two classes, which is also reflected in the lowest κ value (Table 1).

4.2. Model failure analysis

To better understand the model's failure patterns, we employed a semi-automatic error analysis tool (Nushi et al., 2018) that trains a decision tree to identify partitions of data on which the model underperforms, according to attributes that are most predictive of mistakes. Such a principled analysis of prediction errors can significantly benefit predictive models in healthcare settings by identifying potential biases and opportunities for model improvement.

The attributes considered in our analysis (see Fig. 4) included RT-PCR test status for SARS-CoV-2 (positive, negative, or unknown) and X-ray acquisition direction (posteroanterior or anteroposterior view; PA/AP). For this analysis, we focused on the top-level binary classification task ('SUGG. COVID'), which is the most relevant for patient management in a hospital. The analysis identified that the acquisition direction had the strongest association with model errors; in particular, AP images showed a higher error rate than PA. This is consistent with the clinical context, as PA imaging is used as standard-of-care with higher diagnostic quality, whereas AP is reserved for cases when the patient is too ill to stand upright—correlating strongly with being in ICU and presenting much higher variations in image appearance and layout of the patient anatomy. In addition, AP scans with negative or unknown RT-PCR status display elevated error rates, conceivably due to the higher prevalence of INDET and OTHER in this group.

To further investigate this effect, we analysed error patterns across classes and views (Table 2). As expected, mistakes are notably more frequent in the ambiguous INDET class and extremely diverse OTHER class. On the other hand, NORMAL lungs in standard PA view were the most accurately predicted by our model as well as by the annotators. Lastly, we note that INDET–PA appears extremely challenging not only for the model but also for the annotators, who attained error rates of 52%, 81%, and 95% for these images. We observe similar error patterns for *Ann. 1* in Table C.2, and a more detailed analysis of the clinicians' performance and disagreements is presented in Appendix C.

5. Discussion

In this study, we developed and evaluated a clinically informed hierarchical interpretation of a ML model for detecting signs of COVID-19 in chest X-rays. By aligning with the experts' decision-making process, this formulation led to more transparent engagement with the clinical partners and helped created trust in the ML model. Furthermore, this enabled systematic evaluation on clinically relevant predictive sub-tasks, which suggested that the model performs at least as accurately as clinicians on these challenging problems. While not attempted here, we also envision that the operating points for each sub-task can be chosen independently, and appropriate confidence thresholds could be set for deferring decisions to human experts.

Moreover, we conducted a detailed data-driven analysis of model failures to understand in which circumstances the model's predictions may be less reliable. Although this kind of error analysis is not often found in healthcare-related ML studies, we believe it is crucial for providing transparency and actionable insights about a model's behaviour. For example, we may consider additional inputs to the model (here, AP/PA view) and/or complementing the training set with more data from underperforming strata. The analysis may also be useful after deployment if presented as reliability information alongside the model's predictions.

We envisage the deployment of this model in a front-line hospital setting to automatically identify features of COVID-19 in chest X-rays. This would require validation against radiologists in a prospective multi-site study. A successful model may potentially ease pressures on already stretched radiology services and aid less experienced clinical staff in decision-making. This can be used in conjunction with clinical status and RT-PCR in efficiently distributing patients from the front-line areas to other hospital zones thus avoiding in-hospital bottlenecks.

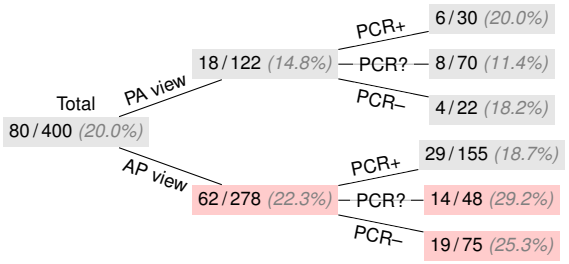


Figure 4. Error analysis tree for the proposed model. Nodes indicate 'errors / instances (ratio)', and ones highlighted in light red have error rates higher than the overall 20.0%. 'PCR+/–/?' correspond to RT-PCR positive / negative / unknown.

Table 2. Model error distribution stratified by original class and PA/AP view ('errors / instances (ratio)')

Class	PA view	AP view	Total
NORMAL	1 / 61 (1.6%)	7 / 39 (17.9%)	8 / 100 (8.0%)
CLASSIC	1 / 12 (8.3%)	6 / 89 (6.71%)	7 / 101 (6.9%)
INDET	13 / 21 (61.9%)	20 / 77 (26.0%)	33 / 98 (33.7%)
OTHER	3 / 28 (10.7%)	29 / 73 (39.7%)	32 / 101 (31.7%)
Total	18 / 122 (14.8%)	62 / 278 (22.3%)	80 / 400 (20.0%)

Acknowledgements

This project was supported by the Microsoft Studies in Pandemic Preparedness program, with Azure credits provided by Microsoft AI for Health. The NIH-CXR and CheXpert datasets were published by the NIH Clinical Center and Stanford University School of Medicine, respectively.

We gratefully acknowledge UHB Radiology Informatics (Carol Payne), UHB Informatics (Suzy Gallier), and PIONEER for the provision of the private UHB dataset. We would also like to thank Matthew Lungren, Kenji Takeda, and Usman Munir for the feedback and support; Omkar More, Shu Peng, and Vijay Kannan for their efforts in implementing DICOM chest X-ray support in the Azure ML labelling tool and facilitating the IOV study; as well as the study participants for their contribution.

References

BSTI. COVID-19 BSTI reporting templates and codes, 2020. URL https://www.bsti.org.uk/covid-19-resources/covid-19-bsti-reporting-templates/.

Chen, H., Miao, S., Xu, D., Hager, G. D., and Harrison, A. P. Deep hierarchical multi-label classification of chest X-ray images. In *Proceedings of the 2nd International Conference on Medical Imaging with Deep Learning (MIDL 2019)*, volume 102 of *PMLR*, pp. 109–120. PMLR, 2019.

Cleverley, J., Piper, J., and Jones, M. M. The role of chest radiography in confirming covid-19 pneumonia. *BMJ*, 370(m2426), 2020. doi: 10.1136/bmj.m2426.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pp. 21271–21284, 2020.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4700–4708, 2017.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):590–597, 2019. doi: 10.1609/aaai.v33i01.3301590.

Kingma, D. P. and Ba, J. L. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR 2015)*, 2015.

Nushi, B., Kamar, E., and Horvitz, E. Towards accountable AI: Hybrid human-machine analyses for characterizing system failure. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 6(1), 2018.

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., Weir-McCall, J. R., Teng, Z., Gkrania-Klotsas, E., AIX-COVNET, Rudd, J. H. F., Sala, E., and Schönlieb, C.-B. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, 3:199–217, 2021. doi: 10.1038/s42256-021-00307-0.

Sriram, A., Muckley, M., Sinha, K., Shamout, F., Pineau, J., Geras, K. J., Azour, L., Aphinyanaphongs, Y., Yakubova, N., and Moore, W. COVID-19 prognosis via self-supervised representation learning and multi-image prediction, 2021. arXiv:2101.04909.

Tizhoosh, H. R. and Fratesi, J. COVID-19, AI enthusiasts, and toy datasets: Radiology without radiologists. *European Radiology*, 33:3553–3554, 2021. doi: 10.1007/s00330-020-07453-w.

Toussie, D., Voutsinas, N., Finkelstein, M., Cedillo, M. A., Manna, S., Maron, S. Z., Jacobi, A., Chung, M., Bernheim, A., Eber, C., Concepcion, J., Fayad, Z. A., and Gupta, Y. S. Clinical and chest radiography features determine patient outcomes in young and middle-aged adults with COVID-19. *Radiology*, 297(1):E197–E206, 2020. doi: 10.1148/radiol.2020201754.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. 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)*, pp. 3462–3471, 2017.

Wehbe, R. M., Sheng, J., Dutta, S., Chai, S., Dravid, A., Barutcu, S., Wu, Y., Cantrell, D. R., Xiao, N., Allen, B. D., MacNealy, G. A., Savas, H., Agrawal, R., Parekh, N., and Katsaggelos, A. K. DeepCOVID-XR: An artificial intelligence algorithm to detect COVID-19 on chest radiographs trained and tested on a large US clinical data set. *Radiology*, 299(1):E167–E176, 2021. doi: 10.1148/radiol.2020203511.

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.

Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., Bonten, M. M. J., Damen, J. A. A., Debray, T. P. A., De Vos, M., Dhiman, P., Haller, M. C., Harhay, M. O., Henckaerts, L., Kreuzberger, N., Lohmann, A., Luijken, K., Ma, J., Andaur Navarro, C. L., Reitsma, J. B., Sergeant, J. C., Shi, C., Skoetz, N., Smits, L. J. M., Snell, K. I. E., Sperrin, M., Spijker, R., Steyerberg, E. W., Takada, T., Van Kuijk, S. M. J., Van Royen, F. S., Wallisch, C., Hooft, L., Moons, K. G. M., and Van Smeden, M. Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. *BMJ*, 369(m1328), 2020. doi: 10.1136/bmj.m1328.

Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y., Song, B., Gu, X., Guan, L., Wei, Y., Li, H., Wu, X., Xu, J., Tu, S., Zhang, Y., Chen, H., and Cao, B. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet*, 395(10229):1054–1062, 2020. doi: 10.1016/S0140-6736(20)30566-3.

A. Pre-training comparison

Table A.1 presents the model performance obtained in the classification tasks illustrated in Fig. 2(b) in terms of precision-recall area-under-the-curve (PR-AUC), ROC-AUC, and accuracy metrics, along with multi-class (one-of-four) accuracy, in 5-fold cross-validation runs.

The experimental results demonstrate that model pre-training significantly improves the classification accuracy for every task in the hierarchical decision making process. The same behaviour is observed when we pre-trained the models on two different external datasets: NIH-CXR (Wang et al., 2017) and CheXpert (Irvin et al., 2019). However, there was no significant difference between the compared pre-trained models whilst they consistently outperformed models trained with random weight initialisation. Due to comparable performance of the pre-trained models, we used NIH pre-training for the remaining experiments.

B. Implementation details

Our model uses a DenseNet-121 (Huang et al., 2017) backbone for image feature extraction. The model is trained using cross entropy loss over the 4 classes for 50 epochs with a batch size of 64. We use the Adam optimiser (Kingma & Ba, 2015) and a learning rate of 10^{-5}. The pixel values of each image are linearly normalised between 0 and 255. During training and inference, each image is resized to size 256×256, and then a center crop taken to get an image of size 224×224. When training, we perform data augmentation using random horizontal flips, affine transforms, random crops, and brightness, contrast, saturation and gamma transforms.

Table A.1. Cross-validation classification results aggregated over $K = 5$ folds ($N = 4940$; NORMAL: 1154, CLASSIC: 1778, INDET: 1093, OTHER: 915). Mean ± (std.).

Pre-training	Classification task	PR-AUC	ROC-AUC	Accuracy
None	SUGG. COVID	.922 ± .005	.893 ± .006	.815 ± .012
	CLASSIC vs INDET	.722 ± .023	.835 ± .014	.758 ± .013
	NORMAL vs OTHER	.793 ± .043	.843 ± .029	.772 ± .033
	Multi-class	—	—	.625 ± .018
NIH-CXR	SUGG. COVID	.938 ± .007	.915 ± .009	.837 ± .009
	CLASSIC vs INDET	.746 ± .023	.857 ± .005	.777 ± .012
	NORMAL vs OTHER	.813 ± .031	.855 ± .017	.784 ± .016
	Multi-class	—	—	.655 ± .016
CheXpert	Multi-class	—	—	.653 ± .015
Both	Multi-class	—	—	.658 ± .015

C. Multi-expert labelling process

In this study, we formed an isolated test set with multiple expert labels to assess the clinical applicability of the learnt models. For this purpose, a panel of three clinicians manually labelled a subset of the in-house dataset ($N = 400$) described in Section 3. The labelling process was carried out using a cloud-based annotation tool on an image-by-image basis, where each expert was able to adjust the image intensity window/level and analyse CXR scans in high-resolution. At the same time, the annotation time of each expert was monitored throughout this exercise to create a proxy measure quantifying the difficulty of each labelling task and also the experience level of each annotator.

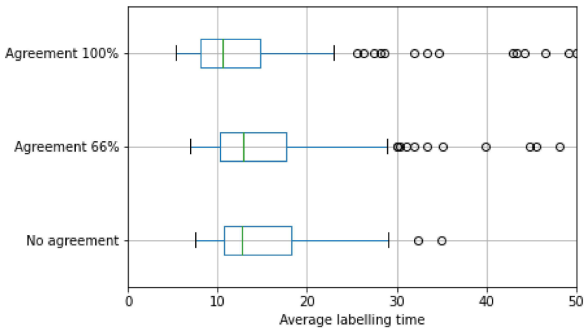


Figure C.1. Distribution of average labelling time (in seconds) on 380 CXR images, broken down by how often the three clinicians agreed on the class assignment. A small subset of of the 400 images ($N = 11$) whose average labelling time exceeded 50 seconds were omitted from the graph to reduce the noise level in the analysis, and a further 9 were skipped by at least one annotator.

Figure C.1 shows a breakdown of average labelling time by agreement of the 3 annotators. We observe that, for the samples where the experts arrive at the same conclusion, the annotation time is consistently lower compared to the (presumably more difficult) samples where they do not agree. Additionally, we observed a systematic correlation between the annotators' disagreement and the model's errors (see Table C.1), suggesting that the model tended to have more difficulty on ambiguous cases. Lastly, Table C.2 reports the breakdown of errors of the most experienced annotator (*Ann. 1*) by class label and CXR view. Comparing to Table 2, it suggests that the model and *Ann. 1* have similar error patterns for this task.

Table C.1. Model error rates versus expert disagreement for each predictive task. We indicate agreement patterns as '3' for full agreement, '2:1' for partial agreement, and '1:1:1' for full disagreement between the three annotators. The model's overall error rates are also included for reference.

	Annotator agreement			
Classification task	3	2:1	1:1:1	Overall
SUGG. COVID	18.6%	22.6%	—	20.0%
CLASSIC vs INDET	19.5%	33.9%	—	27.6%
NORMAL vs OTHER	16.4%	29.9%	—	20.9%
Multi-class	30.3%	47.9%	56.1%	40.8%

Table C.2. Distribution of disagreements between the labels provided by a consultant chest radiologist (*Ann. 1*) and the reference labels collected on the test set. The results are stratified by target class and PA/AP view ('errors / instances (ratio)').

Class	PA view	AP view	Total
NORMAL	1 / 61 *(1.6%)*	6 / 39 *(15.4%)*	7 / 100 *(7.0%)*
CLASSIC	1 / 12 *(8.3%)*	17 / 89 *(19.1%)*	18 / 101 *(17.8%)*
INDET	11 / 21 *(52.4%)*	34 / 77 *(44.2%)*	45 / 98 *(45.9%)*
OTHER	3 / 28 *(10.7%)*	17 / 73 *(23.3%)*	20 / 101 *(19.8%)*
Total	16 / 122 *(13.1%)*	74 / 278 *(26.6%)*	90 / 400 *(22.5%)*