Improving the generalisation of radiographic AI using automated data curation to mitigate shortcut learning

Ian Selby[1,2], Eduardo González Solares[3], Anna Breger[4,5], Michael Roberts[5,6], Lorena Escudero Sánchez[1], James Rudd[1,6], Nicholas Walton[3], Judith Babar[1,2], Fiona Gilbert[1,2], Carola-Bibiane Schönlieb[5], Evis Sala[7], Jonathan Weir-McCall[1,2]

Department of Radiology, University of Cambridge[1]; Cambridge University Hospitals NHS Trust, Cambridge[2]; Institute of Astronomy, University of Cambridge[3]; Center of Medical Physics and Biomedical Engineering, Medical University of Vienna[4]; Department of Applied Mathematics and Theoretical Physics, University of Cambridge[5]; School of Clinical Medicine, University of Cambridge[6]; Advanced Radiodiagnostics Centre, Fondazione Policlinico Universitario Agostino Gemelli and Università Cattolica del Sacro Cuore[7]

UNIVERSITY OF CAMBRIDGE

NHS Cambridge University Hospitals NHS Foundation Trust

1. Introduction

- Several popular deep learning (DL) models have been shown to rely on spurious associations rather than medical pathology (1-7), a phenomenon known as shortcut learning.
- For example, correlations have been shown to exist between diagnostic COVID-19 chest X-ray (CXR) model predictions and patient positioning, radiographic projection, text annotations, image quality and post-processing, the presence of support apparatus, and more (1-3, 8, 9).
- The result often manifests as a significant drop in performance on datasets 'external' to the training data, i.e. unseen by the DL model during development.
- This study investigated whether automated data pipelines to improve data curation of chest radiographs can improve deep learning model performance on previously unseen datasets.

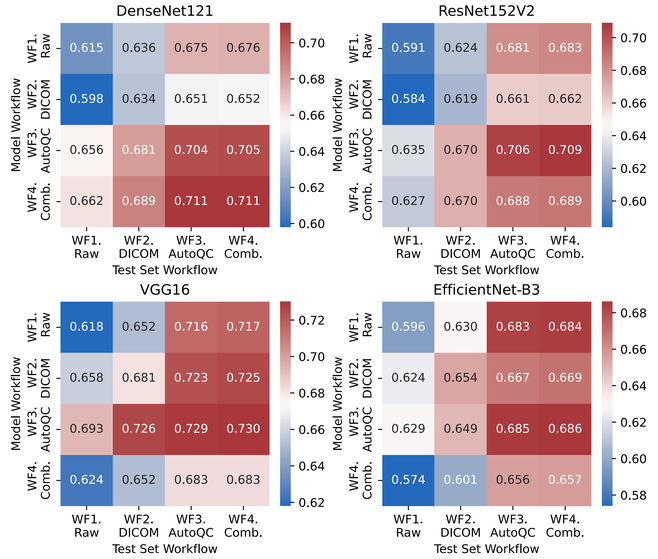


Figure 3: Heatmaps of AUCs for all models tested on external hold-out data curated using all four WFs.

2. Methods

- Two publicly available datasets, MIDRC-1A and MIDRC-R1, were used to develop diagnostic COVID-19 models using four architectures (DenseNet121, ResNet152V2, VGG16, and EfficientNet-B3).
- The ground truth diagnosis was inferred from available COVID-19 laboratory test results.
- Each was trained four times using a different data curation workflow (WF). The four WFs are illustrated in Figure 1 and WFs 3 and 4 utilised an open-source data-cleaning pipeline, *AutoQC* (Available at: https://gitlab.developers.cam.ac.uk/maths/cia/covid-19-projects/autoqc).
- The *AutoQC* pipeline standardises CXRs by detecting and correcting common issues such as grayscale inversion, rotation, and text annotations, while identifying confounders like radiographic projection and the presence of pacemakers.
- The performance of the trained models was assessed using five publicly available datasets (NCCID, BIMCV-COVID-19, BriXia, CUH and RICORD-1C).
- Each model was evaluated on data curated from its own workflow and the other three.
- The difference in the area under the curve (ΔAUC) between the internal hold-out and external test sets was calculated.

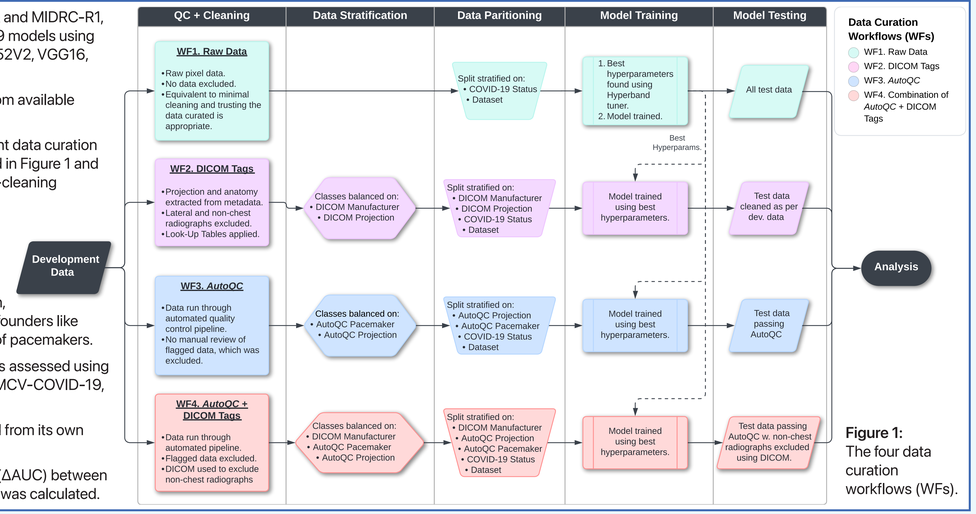


Figure 1: The four data curation workflows (WFs).

3. Results and Discussion

- 43,176 CXRs were included in WF1, with 33.2% (14,328) classified as COVID-positive. The development sets of the other workflows were up to 60% smaller after data curation.
- Models trained using WF1 data experienced the largest fall in generalisation (mean ΔAUC = -0.15 [95%CI:-0.17, -0.14]).
- Models trained using data cleaned using only DICOM metadata (WF2) had a mean ΔAUC = -0.07 [95%CI: -0.09, -0.05].
- Models trained utilising *AutoQC* demonstrated the most consistent performance with mean ΔAUC = -0.04 [95%CI: -0.06, -0.02] and -0.02 [95%CI: -0.04,0.00] respectively for WFs 3 and 4 ($p < 0.05$).
- *AutoQC* reduced the potential for spurious correlations, as demonstrated by smaller ΔAUC between the internal and external test sets, which serve as a proxy for the model's 'generalisability gap'.
- Not only were the ΔAUCs reduced, the AutoQC models exhibited the highest AUC on the external data.

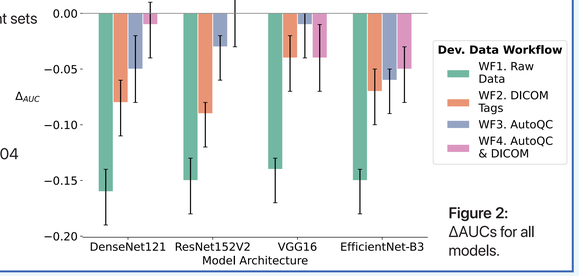


Figure 2: ΔAUCs for all models.

AutoQC

LinkedIn

Thanks to:

NCITA National Cancer Imaging Translational Accelerator

CANCER RESEARCH UK

cmih

intel

DRAGON

NIHR | Cambridge Biomedical Research Centre

References

1. López-Cabrera JD, Orozco-Morales R, Portal-Díaz JA, et al. Current limitations to identify covid-19 using artificial intelligence with chest x-ray imaging (part ii). The shortcut learning problem. Health Technol. 2021 Nov 1;11(6):1331–45. https://pubmed.ncbi.nlm.nih.gov/34660166/
2. DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal. Nat Mach Intell. 2021 May 31;3(7):610–9. https://www.nature.com/articles/s42256-021-00338-7
3. Jabbour S, Fouhey D, Kazerooni E, et al. Deep Learning Applied to Chest X-Rays: Exploiting and Preventing Shortcuts. Journal of Machine Learning Research. 2020 Sep 21;126:1–32. https://arxiv.org/abs/2009.10132v1
4. Duffy G, Clarke SL, Christensen M, et al. Confounders mediate AI prediction of demographics in medical imaging. NPJ Digit Med. 2022 Dec 22;5(1):1–6. https://www.nature.com/articles/s41746-022-00720-8
5. Mukherjee P, Shen TC, Liu J, et al. Confounding factors need to be accounted for in assessing bias by machine learning algorithms. Nat Med. 2022 Jun 16;28(6):1159–60. https://www.nature.com/articles/s41591-022-01847-7
6. Zech JR, Badgeley MA, Liu M, et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs. PLoS Med. 2018 Nov;15(11):e1002683. https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002683
7. Geirhos R, Jacobsen JH, Michaelis C, et al. Shortcut Learning in Deep Neural Networks. Nat Mach Intell. 2020;2:665–73. https://doi.org/10.1038/s42256-020-00257-z
8. Garcia Santa Cruz B, Bossa MN, Sölter J, Husch AD. Public Covid-19 X-ray datasets and their impact on model bias – A systematic review of a significant problem. Med Image Anal. 2021 Dec 1;74(2020):102225. https://www.sciencedirect.com/science/article/pii/S1361841521002070X
9. Driggs D, Selby I, Roberts et al. Machine Learning for COVID-19 Diagnosis and Prognostication: Lessons for Amplifying the Signal While Reducing the Noise. Radiol Artif Intell. 2021 Mar 24;e210011. https://doi.org/10.1148/ryai.2021210011

4. Conclusion

Improved data curation using automated tools can assist developers and enhance the generalisation of deep learning models for chest radiographs while facilitating more consistent performance on data from new locations and equipment.