8-4-2023

Training certified detectives to track down the intrinsic shortcuts in COVID-19 chest x-ray data sets

Ran Zhang

Dalton Griner

John W. Garrett

Zhihua Qi

Guang-Hong Chen

scientific reports

OPEN

Training certified detectives to track down the intrinsic shortcuts in COVID-19 chest x-ray data sets

Ran Zhang[1,4], Dalton Griner[1,4], John W. Garrett[1,2], Zhihua Qi[3] & Guang-Hong Chen[1,2]✉

Deep learning faces a significant challenge wherein the trained models often underperform when used with external test data sets. This issue has been attributed to spurious correlations between irrelevant features in the input data and corresponding labels. This study uses the classification of COVID-19 from chest x-ray radiographs as an example to demonstrate that the image contrast and sharpness, which are characteristics of a chest radiograph dependent on data acquisition systems and imaging parameters, can be intrinsic shortcuts that impair the model's generalizability. The study proposes training certified shortcut detective models that meet a set of qualification criteria which can then identify these intrinsic shortcuts in a curated data set.

Deep learning has been incredibly successful in object detection, image classification, and natural language processing over the past 10 years due to its ability to learn complex features from data. However, despite its success on benchmark datasets, there are limitations and practical issues when using these models in real-world scenarios. A major challenge is poor generalizability, where performance significantly drops when applied to external datasets[1–3]. This limits the translation and deployment of deep learning models for high-stakes tasks, such as in healthcare applications. This lesson was evident during the COVID-19 pandemic, where many machine learning models were developed, but very few performed well on real-world clinical tests[4–6].

The concept of shortcut learning has recently been explored in deep learning studies[7]. It has been discovered that poor model generalizability can be attributed to shortcut learning when the training dataset has hidden shortcuts, meaning there are spurious correlations between irrelevant image features and the corresponding training labels. This causes models to quickly pick up these spurious correlations instead of the desired image features, establishing incorrect connections between input image data and output labels. For instance, early studies have shown that deep learning models can differentiate chest x-rays from different hospitals and patient groups[1]. This suggests that different data sources and patient characteristics like gender, age, and race could also become shortcuts.

To illustrate the issue of shortcut learning in real-world clinical scenarios, let's take the example of COVID-19 classification using chest x-ray radiographs (CXRs). DeGrave et al.[8] discovered that if COVID-19 positive and negative training data were collected from two different sources, the model would only learn the source label as a shortcut for prediction. As a result, the model would not have the desired prediction power in real-world clinical scenarios. The authors also found that the trained model used extrinsic image features, such as lead markers, for prediction, even though these markers only indicated the orientation of patients in x-ray image acquisitions and did not correspond to any disease features. Some suggestions have been made to remove these extrinsic shortcuts by a segmentation step[9]. However, even with these markers removed from the training dataset, some other shortcuts still existed in the segmented lung tissue-only training dataset[10]. As a result, the nearly perfect performance of the trained model[9] is still not generalizable to real-world clinical datasets. The remaining shortcuts may be attributed to the inherent defining features of CXRs, such as image contrast and sharpness, which can vary from hospital to hospital due to different types of imaging systems, generations of x-ray imaging equipment, hardware components, image post-processing methods used by vendors, and imaging protocols used

[1]Department of Medical Physics, School of Medicine and Public Health, The University of Wisconsin in Madison, Madison, WI 53705, USA. [2]Department of Radiology, School of Medicine and Public Health, The University of Wisconsin in Madison, Madison, WI 53792, USA. [3]Department of Radiology, Henry Ford Health, Detroit, MI 48202, USA. [4]These authors contributed equally: Ran Zhang and Dalton Griner. ✉email: gchen7@wisc.edu

by technologists. All these factors can impact the digital representation of the acquired image data in terms of variations in image contrast and sharpness. Figure 1 demonstrates these variations.

Unlike other shortcuts that have been previously studied, such as age, gender, race, and markers, which are extrinsic and can be removed through careful data collection and cleaning, contrast- and sharpness-related shortcuts are more difficult to detect and mitigate. This is because desired image features are also represented as image contrast and spatial correlations, which are similar to the features of contrast- and sharpness-related shortcuts. This entanglement between desired image features and shortcut features makes studying contrast and sharpness-related shortcuts particularly challenging. Consequently, these shortcuts are referred to as intrinsic shortcuts in this work.

In order to develop an effective strategy for mitigating contrast- and sharpness-related shortcuts, it is necessary to first develop reliable methods for detecting their presence and severity within a carefully curated dataset. While post-hoc model interpretability methods, such as class activation maps[11] and expected gradient[12], have been developed to identify relevant image features used by trained deep learning models for prediction, these methods are unable to detect intrinsic shortcuts within a curated training dataset prior to model training. Furthermore, studies have suggested that these methods may not be effective in diagnosing poor generalization performance of the model[13].

In this paper, we present a novel approach for detecting contrast- and sharpness-related intrinsic shortcuts using certified shortcut detective models. Our approach involves establishing qualification standards for suspected intrinsic shortcuts, designing a training curriculum for training the shortcut detectives to detect these shortcuts, performing certification tests on the trained detectives, and finally deploying them to curated datasets to examine the suspected shortcuts. We applied this approach to the available COVID-19 datasets to assess their quality. Our results demonstrate the effectiveness of this approach in detecting and mitigating intrinsic shortcuts.

Methods

Datasets. Figure 2 provides an overview of the datasets utilized in this study. The MIMIC-CXR dataset served as the training data for the shortcut detectives, while the HF-train dataset, a privately curated COVID-19 chest x-ray dataset, was utilized for certification tests of the trained detectives. The trained shortcut detectives were then applied to a variety of public and private chest x-ray datasets. Specific information about each dataset is provided below.

MIMIC dataset. MIMIC-CXR[14] chest x-ray dataset consists of 377,110 CXRs from 65,379 patients presenting to the Beth Israel Deaconess Medical Center Emergency Department between 2011 and 2016. In this work, 46,894 frontal-view (AP/PA) normal CXRs (cases with "No Finding" labels) were used to train shortcut detectives.

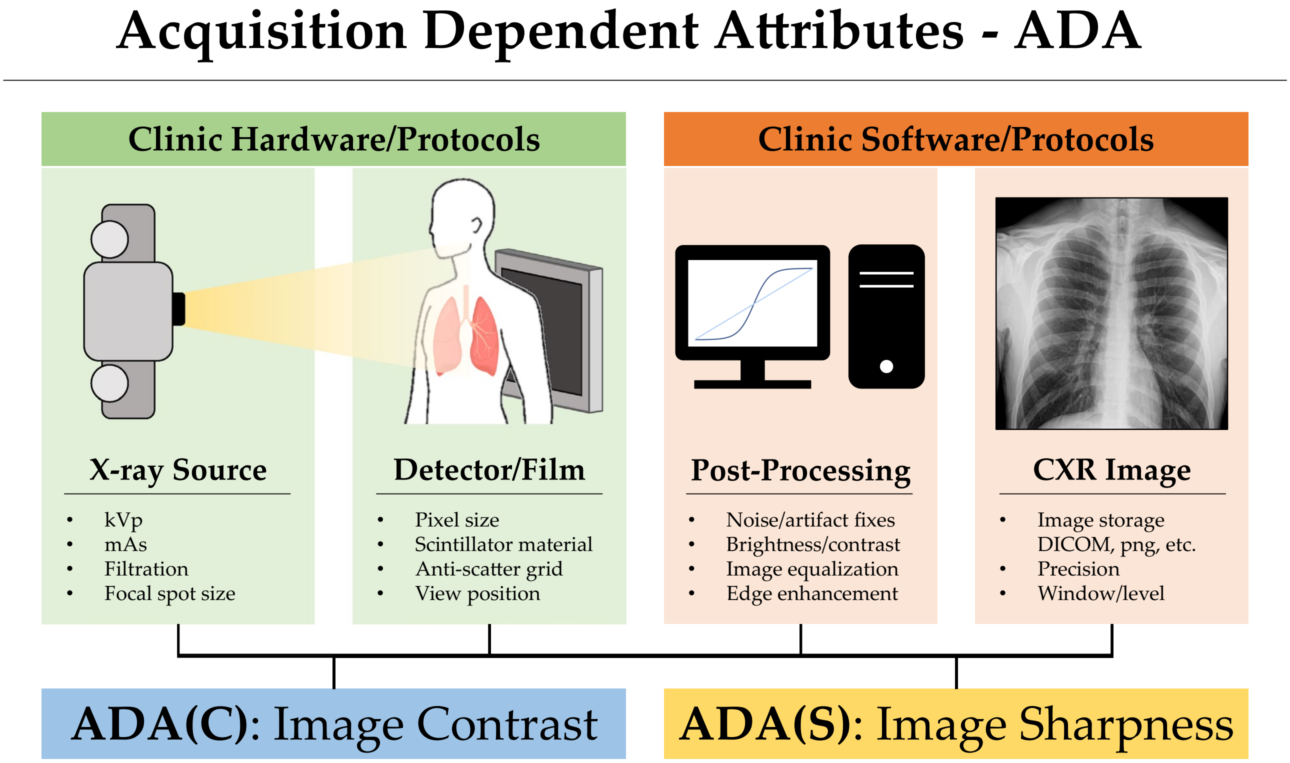

Figure 1. Acquisition dependent attributes (ADA) in chest x-ray images.

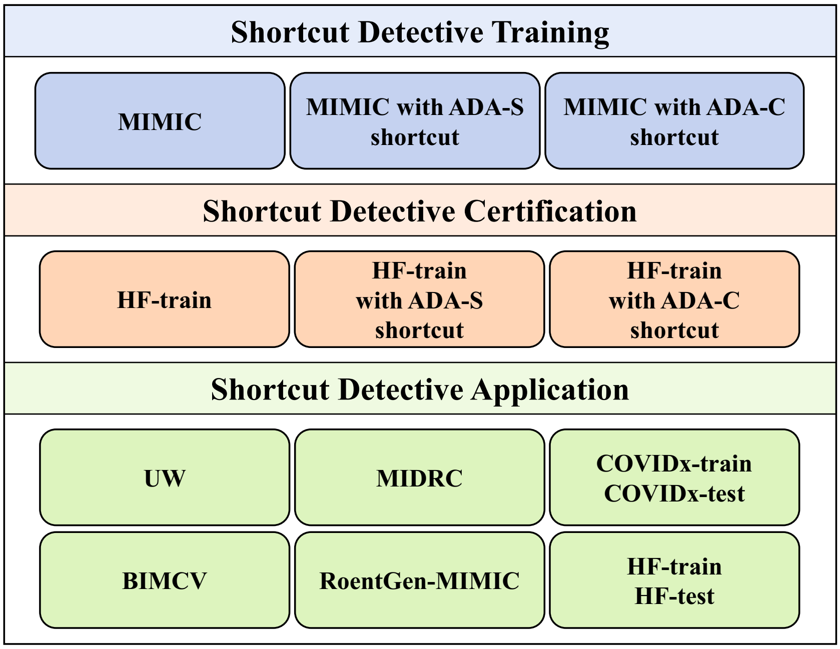
Shortcut Detective Training
MIMIC
MIMIC with ADA-S shortcut
MIMIC with ADA-C shortcut
Shortcut Detective Certification
HF-train
HF-train with ADA-S shortcut
HF-train with ADA-C shortcut
Shortcut Detective Application
UW
MIDRC
COVIDx-train COVIDx-test
BIMCV
RoentGen-MIMIC
HF-train HF-test

Figure 2. An overview of the datasets used in this work.

HF dataset. This is a privately curated COVID-19 CXR dataset from patients presenting to the Henry Ford Health between March 1, 2020, and October 31, 2020. The COVID-19-positive and COVID-19-negative cohorts are collected within the same time range, from the same hospitals, and labeled by their most recent RT-PCR test result seven days before or after the imaging study. For model training and internal testing, two data partitions are generated: HF-train consisted of 8733 COVID-19-positive CXRs from 4383 patients and 16,584 COVID-19-negative CXRs from 8733 patients; HF-test consisted of 695 COVID-19-positive CXRs from 526 patients and 8878 COVID-19-negative CXRs from 6081 patients.

BIMCV dataset. This is a public COVID-19 CXR dataset collected in Spain[15]. This dataset was collected from 11 hospitals in the Valencian Region, Spain, between February and April 2020. After data curation, the dataset consisted of 4169 COVID-19-positive CXRs from 2663 patients and 5050 COVID-19-negative CXRs from 3710 patients.

UW dataset. This is a privately curated COVID-19 CXR dataset. It includes consecutive patient cases from the University of Wisconsin Hospitals and Clinics (UW Health) from March 2020 to September 2021. The dataset comprised 1025 COVID-19-positive CXRs from 658 patients and 8774 COVID-19-negative CXRs from 5953 patients.

MIDRC dataset. This is a large, multi-institution public COVID-19 CXR dataset curated and released by the Medical Imaging & Data Resource Center (MIDRC). A total of 6453 COVID-19-positive CXRs from 5199 patients and 20,072 COVID-19-negative CXRs from 9947 patients were pulled from the MIDRC Data Commons (https://data.midrc.org/, date accessed: December 7th, 2022).

COVIDx dataset. This is a public COVID-19 CXR dataset released by the COVID-Net Open Source Initiative[16]. A total of 29,986 CXRs from 16,648 patients are included in the training dataset (COVIDx-train), and 400 CXRs are included in the test dataset (COVIDx-test). (Data downloaded from https://www.kaggle.com/datasets/andyczhao/covidx-cxr2?select=competition_test, date accessed: December 7th, 2022).

RoentGen-MIMIC dataset. This dataset contains 943 synthetic CXRs generated by the RoentGen model[17] and 1,000 real "No Finding" CXRs from the MIMIC dataset. The RoentGen model, which is trained using the MIMIC dataset, is able to generate visually convincing synthetic CXRs with different pathologies. To generate the synthetic CXRs used in this work, a text prompt of "No finding" was used as the input, only frontal view CXRs are included.

Training and certification of shortcut detectives

Overview of the proposed framework. An overview of the shortcut detective training and certification process is shown in Fig. 3.

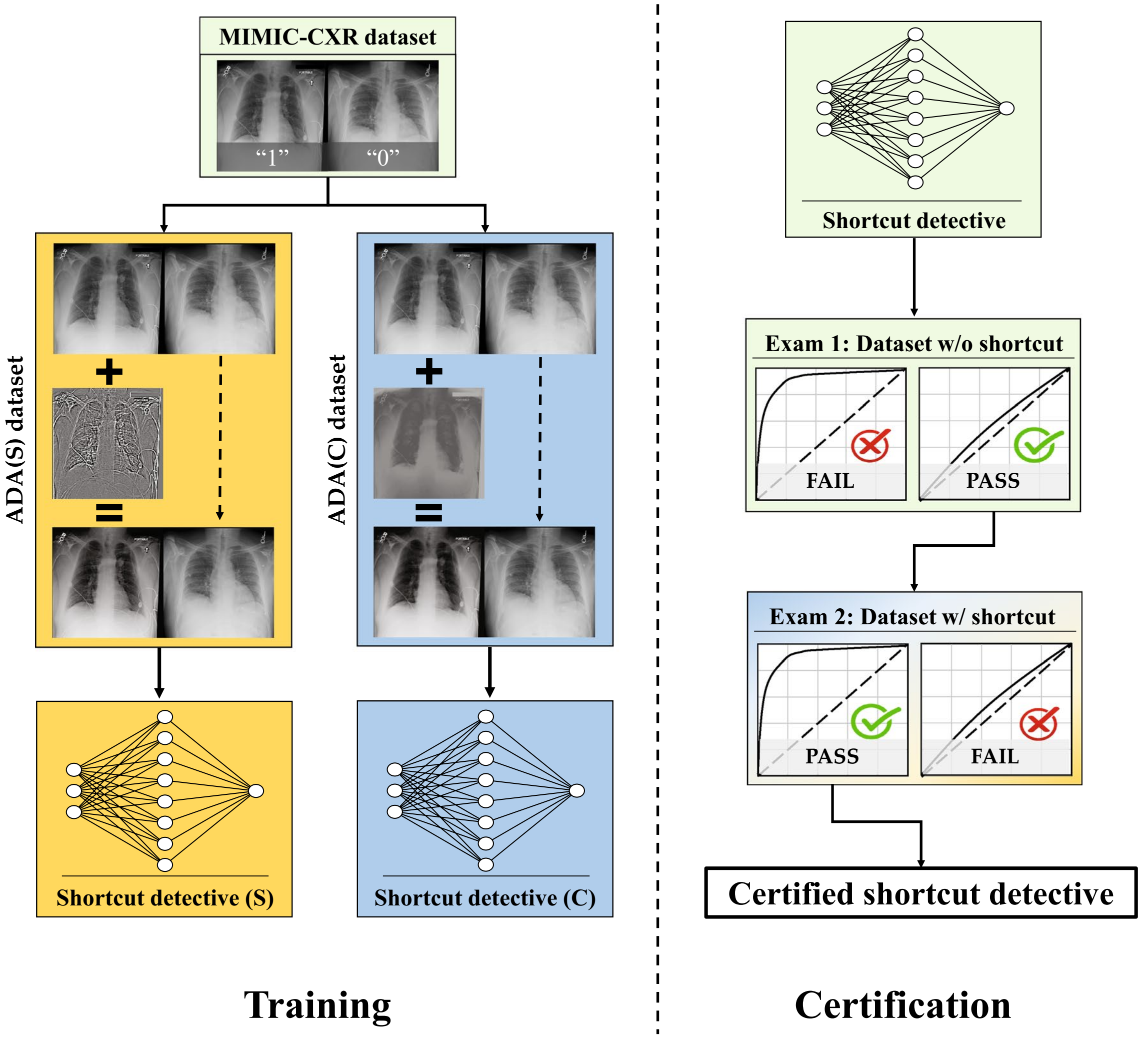

Figure 3. A framework to train and certify shortcut detectives.

To begin, 46,894 normal CXRs from the MIMIC dataset were selected and randomly divided into two equal groups: 23,447 of the CXRs were assigned as the positive class (“1”) while the rest were assigned as the negative class (“0”).

Then, to construct the training dataset for the shortcut detective, the image contrast or the image sharpness of the positive class were adjusted using the approach shown in Appendix A1. Since only normal CXRs were included, there were no disease-specific features present that could be used to distinguish between the two classes. Therefore, if a model was able to differentiate between the two classes, it would be because the model had learned the corresponding global image contrast or sharpness characteristics, rather than disease features.

Subsequently, shortcut detectives (neural network models for binary classification) were trained using the constructed training datasets with contrast or sharpness shortcuts. Details of the training are discussed in the following section and in Appendix A2.

To assess the efficacy of the shortcut detectives on COVID-19 CXR dataset, two types of examinations are necessary. Firstly, when the shortcut detective is deployed on a COVID-CXR dataset without the corresponding shortcut, it should not be able to differentiate between the COVID-positive and COVID-negative classes. Hence, an Area Under the Receiver Operating Characteristics curve (AUC) close to 0.5 is expected. This examination is crucial to ensure that the image features utilized by the shortcut detectives are not interwoven with the original imaging task. Secondly, when the shortcut detective is applied to a COVID-CXR dataset with a known shortcut, it should demonstrate superior classification performance, ideally with an AUC close to 1. For the first examination, the HF-train dataset is utilized. As the positive and negative cohorts are gathered within the same timeframe and from the same hospitals, it is expected that no contrast and sharpness shortcut exists. This has also been corroborated recently[18], where a model trained on this dataset demonstrated consistent test performance on various external COVID-19 clinical test datasets. For the second examination, known shortcuts are integrated into the COVID-positive class or COVID-negative class of the HF-train dataset using the same procedures outlined in Appendix A1.

Finally, if the trained shortcut detectives pass the two exams, i.e. AUC of close to 0.5 on the shortcut-free dataset and AUC of close to 1.0 on the shortcut-present dataset, they are referred to as certified shortcut detectives.

Image preprocessing and model architecture. For the MIMIC, HF, BIMCV, UW Health and MIDRC datasets, the original DICOM images are converted to 8-bit png format with a size of 224-by-224 using the default window level and window width. For the COVIDx dataset, images are directly resized to 224-by-224.

To train the shortcut detectives, five different model architectures (Table 1) that are broadly used for image classification with state-of-the-art performance on the ImageNet[19] classification tasks are investigated in this work. Although we cannot exhaust all the possible model architectures for this purpose, the models we investigated include classic and modern Convolutional Neural Networks (CNN) and the recently introduced Swin Transformer. These models vary in architectural design and complexity (number of model parameters and floating-point operations, FLOPs). For each model architecture, an ensemble of five individually trained models with different training-validation splits are used. More technical details on the model training are shown in Appendix A2.

Deployment of the shortcut detectives. Certified shortcut detectives are deployed to detect shortcuts in real-world datasets, including BIMCV, UW, MIDRC, COVIDx, and RoentGen-MIMIC. We also trained two COVID-19 classification models using HF-train dataset and COVIDx-train dataset and compared their generalizability using internal and external tests.

Statistics. The 95% confidence intervals (CI) for the AUC were calculated using the statistical software R (version 4.0.0) with the pROC package. CIs were calculated using the bootstrap method with 2000 bootstrap replicates.

Results

Certification of the shortcut detectives. As shown in Table 2, the trained detectives successfully passed the two exams. Note that an AUC of 0.0 is simply due to the assignment of class labels, which is equivalent to an AUC of 1.0. Both indicate perfect classification performance. It is also shown in Table 2 that all five neural network architectures achieve a similar performance level. This result aligns well with our understanding that contrast and sharpness shortcuts are intrinsic to the dataset and the task.

Shortcut detection in real-world datasets. Using the certified shortcut detectives, we investigated several curated COVID-CXR datasets for possible shortcuts. UW Health is a single-site, privately curated dataset; BIMCV and MIDRC are both multi-institutional public datasets; COVIDx is the first open access COVID-CXR dataset made available by the experts in the computer science community.

Model	Number of parameters	FLOPs	ImageNet accuracy (%)	Year developed
VGG-16[20]	138 M	31 B	73.4	2014
DenseNet-121[21]	8 M	5.7 B	74.4	2017
EfficientNet[22]	54 M	24 B	85.1	2020
Swin Transformer[23]	88 M	15 B	83.6	2021
ConvNeXt[24]	89 M	15 B	84.1	2022

Table 1. List of different model architectures studied in this work. Note: ImageNet accuracy data obtained from: https://pytorch.org/vision/stable/models.html#classification.

	VGG	DENSE	Eff	Swin	Conv
ADA(S) shortcut detective					
Exam 1: HF-train	**0.49** [0.48, 0.50]	**0.56** [0.56, 0.57]	**0.56** [0.55, 0.57]	**0.53** [0.52, 0.53]	**0.54** [0.53, 0.54]
Exam 2a: With ADA-S(+)	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]
Exam 2b: With ADA-S(−)	**0.00** [0.00, 0.00]	**0.00** [0.00, 0.00]	**0.00** [0.00, 0.00]	**0.00** [0.00, 0.00]	**0.00** [0.00, 0.00]
ADA(C) shortcut detective					
Exam 1: HF-train	**0.47** [0.46, 0.48]	**0.48** [0.47, 0.49]	**0.47** [0.46, 0.48]	**0.49** [0.48, 0.49]	**0.48** [0.47, 0.48]
Exam 2a: With ADA-C(+)	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]
Exam 2b: With ADA-C(−)	**0.00** [0.00, 0.00]	**0.00** [0.00, 0.00]	**0.00** [0.00, 0.00]	**0.00** [0.00, 0.00]	**0.00** [0.00, 0.00]

Table 2. Certification of the shortcut detectives. Significant values are in [bold]. Note: (+) means the shortcut is added to the COVID-positive CXRs; (−) means the shortcut is added to the COVID-negative CXRs.

Similar to the detective certification process, for a COVID-CXR dataset where COVID-19-positive cases are assigned a label “1” and negative cases are given a label “0”, if the shortcut detective can differentiate the two classes (AUC significantly deviates from 0.50), it indicates the existence of a corresponding shortcut. For the RoentGen-MIMIC dataset, real CXRs are labelled as “1” and synthetic CXRs are labelled as “0”.

The results presented in Table 3 were obtained using shortcut detectives based on the DenseNet model. Results for other model architectures are presented in Appendix A3. The results confirm the presence of image sharpness and contrast shortcuts in the COVIDx dataset, which can be exploited by models trained on such data and compromise their generalizability in real clinical settings. Conversely, the other three datasets that were curated by medical professionals exhibit no such shortcuts. We conducted a performance comparison between two models trained on COVIDx and Henry Ford datasets, respectively, which are of similar size, but the former has both sharpness and contrast shortcuts as identified by the shortcut detectives. The results displayed in Table 4 indicate that the COVIDx model exhibits poor generalization performance, as evidenced by the large AUC gap between internal and external tests. In contrast, the HF model exhibits consistent performance on both internal and external tests. Additionally, as shown in Table 3, some contrast and sharpness differences were also detected in the RoentGen-MIMIC dataset. While the generated synthetic CXRs appear visually realistic, caution must be exercised when using them for AI model development due to the potential learning of shortcuts caused by the inherent contrast and sharpness differences between real and synthetic data.

Discussion

Shortcut learning has been a topic of interest in the machine learning community, particularly in computer vision (CV) and natural language processing (NLP). Researchers have explored shortcut learning behavior from different perspectives, such as underspecification[3], shortcut learning in various NLP tasks[25], and mitigation strategies for domain-knowledge agnostic models[26]. Notably, it has been observed that convolutional neural networks tend to rely on content with high spatial frequency or strong local correlations to establish connections between input and labels in CV and NLP[27,28]. However, it remains unclear whether these observations can be extended to medical diagnosis tasks, where clinical datasets have distinct characteristics from those in ImageNet. Therefore, studying shortcut learning for well-defined, clinically relevant tasks using real-world clinical datasets is crucial for medical AI applications.

In this study, we demonstrated that acquisition-dependent attributes (ADAs), such as image contrast and sharpness differences arising from the entire image generation pipeline, can serve as intrinsic shortcuts during clinical diagnosis learning tasks. Inadequate quality control procedures during data collection can allow these shortcuts to inadvertently infiltrate the curated dataset. If shortcuts contaminate the dataset, neural networks can easily exploit them during training, thereby impairing their ability to generalize to other real-world datasets.

Thus, it is imperative to identify possible shortcuts in the training dataset prior to model development. In this study, we present a methodical framework for training and validating shortcut detectives for chest X-ray classification, with emphasis on image contrast and sharpness—two essential intrinsic characteristics of chest X-ray images. However, it should be noted that these are not the only possible shortcuts that may exist in chest X-ray datasets. If other intrinsic shortcuts are suspected in a collected dataset, the general framework presented in this work can be utilized to construct similar shortcut detectives and identify alleged intrinsic shortcuts.

Once the intrinsic shortcuts are identified, it is essential to develop strategies to mitigate their impact on the learned models. One possible approach is to develop standardization and normalization techniques for image contrast and sharpness to adjust these attributes without affecting the disease features. Alternatively, examining proven intrinsic shortcut-free datasets, such as the baseline dataset (Henry Ford Health) and the three additional datasets (UW Health, BIMCV, and MIDRC) shown to be free of intrinsic shortcuts in this study, can provide further insight on how to avoid these shortcuts in the data curation process. However, it is worth noting that this is a limitation of the present work, and future research should investigate the development of mitigation strategies for the identified intrinsic shortcuts.

	COVIDx	**RoentGen-MIMIC**	**UW**	**BIMCV**	**MIDRC**
ADA(S) shortcut	**0.84 [0.83, 0.84]**	**0.37 [0.34, 0.39]**	0.50 [0.48, 0.52]	0.47 [0.45, 0.48]	0.45 [0.45, 0.46]
ADA(C) shortcut	**0.81 [0.80, 0.81]**	**0.05 [0.04, 0.06]**	0.53 [0.51, 0.55]	0.56 [0.55, 0.57]	0.53 [0.52, 0.54]

Table 3. Shortcut detection on real-world COVID-CXR datasets (DenseNet model result). Significant values are in [bold].

	Internal	**UW**	**BIMCV**	**MIDRC**
COVIDx	1.00 [1.00, 1.00]	0.60 [0.58, 0.62]	0.61 [0.60, 0.62]	0.57 [0.56, 0.58]
HF	0.78 [0.76, 0.80]	0.77 [0.76, 0.79]	0.79 [0.78, 0.80]	0.75 [0.74, 0.75]

Table 4. Performance of two trained models.

Data availability

BIMCV (https://bimcv.cipf.es/bimcv-projects/bimcv-covid19/), MIDRC (https://data.midrc.org/) and MIMIC-CXR (https://physionet.org/content/mimic-cxr/2.0.0/) datasets are publicly available. Henry Ford Dataset and UW Health Dataset are private datasets with protected patient information. The python script used for model training and evaluation; weights of all trained models can be found on the GitHub repository: https://github.com/uw-ctgroup/shortcut.

Received: 14 April 2023; Accepted: 1 August 2023
Published online: 04 August 2023

References

1. Zech, J. R. *et al.* Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med.* **15**, e1002683 (2018).
2. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J. & Song, D. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 15262–15271 (2021).
3. D'Amour, A. *et al.* Underspecification presents challenges for credibility in modern machine learning. arXiv:2011.03395 (2020).
4. Wynants, L. *et al.* Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. *BMJ* **369**, 523 (2020).
5. Roberts, M. *et al.* Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Mach. Intell.* **3**, 199–217 (2021).
6. Born, J. *et al.* On the role of artificial intelligence in medical imaging of COVID-19. *Patterns* **2**, 100269 (2021).
7. Geirhos, R. *et al.* Shortcut learning in deep neural networks. *Nature Mach. Intell.* **2**, 665–673 (2020).
8. DeGrave, A. J., Janizek, J. D. & Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Mach. Intell.* **3**, 610–619 (2021).
9. Oh, Y., Park, S. & Ye, J. C. Deep learning COVID-19 features on CXR using limited training data sets. *IEEE Trans. Med. Imaging* **39**, 2688–2700 (2020).
10. Teixeira, L. O. *et al.* Impact of lung segmentation on the diagnosis and explanation of COVID-19 in chest X-ray images. *Sensors* **21**, 7116 (2021).
11. Selvaraju, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* 618–626 (2017).
12. Erion, G., Janizek, J. D., Sturmfels, P., Lundberg, S. M. & Lee, S.-I. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nat. Mach. Intell.* **3**, 620–631 (2021).
13. Viviano, J. D., Simpson, B., Dutil, F., Bengio, Y. & Cohen, J. P. Saliency is a possible red herring when diagnosing poor generalization. arXiv:1910.00199 (2019).
14. Johnson, A. E. *et al.* MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data* **6**, 1–8 (2019).
15. Vayá, M. D. L. I. *et al.* Bimcv covid-19+: A large annotated dataset of rx and ct images from covid-19 patients. arXiv:2006.01174 (2020).
16. Wang, L., Lin, Z. Q. & Wong, A. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Sci. Rep.* **10**, 1–12 (2020).
17. Chambon, P. *et al.* RoentGen: Vision-language foundation model for chest x-ray generation. https://doi.org/10.48550/arXiv.2211.12737 (2022).
18. Zhang, R. *et al.* A generalizable artificial intelligence model for COVID-19 classification task using chest x-ray radiographs: Evaluated over four clinical datasets with 15,097 patients. arXiv:2210.02189 (2022).
19. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848.
20. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. http://arxiv.org/abs/1409.1556 (2014).
21. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 4700–4708 (2017).
22. Tan, M. & Le, Q. V. EfficientNetV2: Smaller models and faster training. https://doi.org/10.48550/arXiv.2104.00298 (2021).
23. Liu, Z. *et al.* Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 10012–10022 (2021).
24. Liu, Z. *et al.* A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 11976–11986 (2022).
25. Niven, T. & Kao, H.-Y. Probing neural network comprehension of natural language arguments. http://arxiv.org/abs/1907.07355 (2019).
26. Du, M. *et al.* Towards interpreting and mitigating shortcut learning behavior of NLU models. http://arxiv.org/abs/2103.06922 (2021).
27. Wang, H., Wu, X., Huang, Z. & Xing, E. P. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 8684–8694 (2020).
28. Jo, J. & Bengio, Y. Measuring the tendency of cnns to learn surface statistical regularities. http://arxiv.org/abs/1711.11561 (2017).

Acknowledgements

This study was partially supported by NIH Grants 3U01EB021183-W4, R01HL153594, R01EB032474, and a grant from the Wisconsin Partnership Program. The imaging and associated clinical data downloaded from MIDRC (The Medical Imaging Data Resource Center) and used for research in this [publication/press release] were made possible by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institutes of Health under contracts 75N92020C00008 and 75N92020C00021. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

R.Z. and G.H.C. conceived the study and designed the experiments. J.W.G. and Z.Q. curated two private clinical datasets for tests. R.Z. prepared the datasets for model training and testing. D.G. and R.Z. wrote the software code. R.Z. and G.H.C. performed the statistical analysis. G.H.C. supervised the study. R.Z. and G.H.C. wrote the manuscript. All authors read and edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-39855-3.

Correspondence and requests for materials should be addressed to G.-H.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.