Emergency Radiology (2022) 29:107–113
https://doi.org/10.1007/s10140-021-01954-x

ORIGINAL ARTICLE

Limited generalizability of deep learning algorithm for pediatric pneumonia classification on external data

Kevin Z. Xin[1] · **David Li**[2,3] · **Paul H. Yi**[3,4]

Received: 25 March 2021 / Accepted: 8 June 2021 / Published online: 14 October 2021

Abstract

Purpose (1) Develop a deep learning system (DLS) to identify pneumonia in pediatric chest radiographs, and (2) evaluate its generalizability by comparing its performance on internal versus external test datasets.

Methods Radiographs of patients between 1 and 5 years old from the Guangzhou Women and Children's Medical Center (Guangzhou dataset) and NIH ChestXray14 dataset were included. We utilized 5232 radiographs from the Guangzhou dataset to train a ResNet-50 deep convolutional neural network (DCNN) to identify pediatric pneumonia. DCNN testing was performed on a holdout set of 624 radiographs from the Guangzhou dataset (internal test set) and 383 radiographs from the NIH ChestXray14 dataset (external test set). Receiver operating characteristic curves were generated, and area under the curve (AUC) was compared via DeLong parametric method. Colored heatmaps were generated using class activation mapping (CAM) to identify important image pixels for DCNN decision-making.

Results The DCNN achieved AUC of 0.95 and 0.54 for identifying pneumonia on internal and external test sets, respectively ($p < 0.0001$). Heatmaps generated by the DCNN showed the algorithm focused on clinically relevant features for images from the internal test set, but not for images from the external test set.

Conclusion Our model had high performance when tested on an internal dataset but significantly lower accuracy when tested on an external dataset. Likewise, marked differences existed in the clinical relevance of features highlighted by heatmaps generated from internal versus external datasets. This study underscores potential limitations in the generalizability of such DLS models.

Keywords Pneumonia · Deep learning · Machine learning · Chest radiograph

Introduction

Deep learning has shown great promise for automated diagnosis of acute conditions on medical imaging approaching or even exceeding the performance of human radiologists for conditions ranging from intracranial hemorrhage to pneumothorax [1–4], with one primary proposed use case of triaging cases for expedited radiologist review and subsequent care. Another proposed use case for deep learning is in the diagnosis of pneumonia in pediatric patients, for which several proof-of-concept studies have demonstrated diagnostic accuracy exceeding 90% [1, 5]. Pneumonia and other respiratory illnesses place a relatively high burden on pediatric emergency departments, with these conditions accounting for 10% of pediatric emergency department visits and 20% of all pediatric hospital admissions [6, 7]. As a result, automated tools for diagnosis of pediatric pneumonia could be particularly useful in the emergency room setting, which has recently been inundated with respiratory disease burden amidst the COVID-19 pandemic [5].

Prior to deploying deep learning systems (DLS) for medical image diagnosis, it is important to evaluate the generalizability of these systems on data that the algorithms have

✉ Paul H. Yi
paul@intelligentimaging.org

1 Transitional Year Program, Mount Carmel Health System, Grove City, OH, USA

2 University of Ottawa Faculty of Medicine, Ottawa, ON, Canada

3 University of Maryland Intelligent Imaging (UMII) Center, Department of Diagnostic Radiology and Nuclear Medicine, University of Maryland School of Medicine, Baltimore, MD, USA

4 Malone Center for Engineering in Healthcare, Johns Hopkins University, Baltimore, MD, USA

never "seen" before, as DLS have been shown to often perform worse on external datasets than on data used to develop them [7–9]. This observed drop in performance of DLS is related to overfitting, which is when an algorithm essentially memorizes training data as opposed to general features of the predicted label/disease of interest (which would lead to poor performance on data from an outside institution); this overfitting is a complex phenomenon, having been observed when there is variation in the prevalence of certain conditions across different datasets or hospital systems [8]. Although prior studies have shown promising results for DLS for pediatric pneumonia diagnosis [1, 10, 11], these studies have largely used a single dataset from China for both development and testing of their algorithms without testing on an external dataset from a different hospital [1]. Evaluating performance of such algorithms on external data is critical to ensuring the safe and responsible use of these potentially useful technologies.

The purpose of this study was to (1) develop a DLS for classification of pneumonia in pediatric chest radiographs and (2) compare its performance on internal versus external test datasets. Our hypothesis was that the DLS would perform significantly worse on an external test dataset compared to an internal test dataset.

Materials and methods

This study was a retrospective analysis of de-identified images that are part of the public domain. We aimed to create a DLS model capable of performing binary classification (i.e., pneumonia versus no pneumonia) of pediatric chest radiographs. Subsequently, the DLS model's performance was evaluated on internal versus external test datasets. The study was acknowledged by the Johns Hopkins institutional review board (IRB) as non-human subject research. Accordingly, formal IRB review was not required per our institutional policies.

Dataset curation and re-annotation

Two datasets were used in this study. The first consisted of 5856 frontal pediatric chest radiographs obtained from retrospective cohorts of pediatrics patients between the ages of 1 and 5 years old (specific age distribution not available) from the Guangzhou Women and Children's Medical Center in Guangzhou, China. These images are available on a public online database (Link) [12]. Images in the Guangzhou database were curated and assigned labels previously via the following methods [1]: (1) all chest radiographs were initially screened for quality control by removing any low-quality or unreadable scans, and (2) all images were then graded and assigned ground truth reference standards by two expert physicians (specialty not specified by the dataset curators) before being cleared for use in training AI systems. The Guangzhou database had a designated "training set" containing 5232 chest radiographs, of which 3883 depicted pneumonia (2538 bacterial, 1345 viral) and 1349 depicted normal findings. A holdout test set of 624 images was also provided (hereby referred to as "internal test set"), of which 390 depicted pneumonia (242 bacterial, 148 viral) and 234 depicted normal chest findings.

The second dataset was the ChestXray14 database (Link) [13], which contained a total of 112,120 chest radiographs obtained and curated by the National Institutes of Health (NIH) Health Center. Demographic information (e.g., age, ethnicity) for each image was also present in this dataset. Image labels were assigned by the NIH team via use of natural language processing, with a label accuracy estimated at $\geq$ 90% [2]. We utilized the 383 images in this dataset obtained from children between 1 and 5 years of age (to match the age in the Guangzhou dataset), among which 107 (25%) cases had been labeled as positive for "pneumonia", "infiltration", and "consolidation," whereas 276 (65%) images had no disease labels and depicted normal chest findings. The dataset consisted of 26 radiographs (7%) from 1-year-old children, 70 (18%) from 2-year-old children, 77 (20%) from 3-year-old children, 95 (25%) from 4-year-old children, and 115 (30%) from 5-year-old children. The aforementioned images from the ChestXray14 database served as the "external test set".

DLS development

We developed our DLS using a transfer learning approach with the ResNet-50 deep convolutional neural network (DCNN) pretrained on the ImageNet database. We split the 5232 chest radiographs from the Guangzhou dataset into training and validation sets, with 90% of images assigned to the former and 10% assigned to the latter. Each image was augmented on-the-fly during each training epoch by a random rotation between – 20º and 20º, random cropping, and horizontal flipping. The last linear layer of the DCNN was redefined to yield a binary output of the presence or absence of pneumonia. The solver parameters for our DCNN were 20 epochs and stochastic gradient descent with a learning rate of $5.0*10^{-6}$. At the end of each training epoch, the DCNN was tested on the validation set, and the best-performing DCNN weights were chosen for final testing. The best-performing DCNN configuration was then tested on both the internal test set, which consisted of either the 624 images from the Guangzhou dataset, and the external test set comprised of 383 images from the NIH ChestXray14 dataset.

To identify the features of each image used by the DCNN for its decisions in classifying each radiograph as having pneumonia or not, we produced heatmaps using class

activation mapping (CAM) [14], which provide visual representations of important image pixels for DCNN decision-making by way of a colored heatmap; features with greater importance were assigned more prominent shades of blue according to the visualization color scheme used in our study.

Computer hardware and software

All image processing and DCNN development were performed online in Google Colaboratory (Google, Mountain View, CA) using an NVIDIA K80 graphics processing unit (GPU). All coding was performed using the Keras deep learning framework (Version 2.3.0, https://keras.io/).

Statistical analysis

Statistical analyses were performed using Microsoft Excel (Microsoft, Redmond, WA) and GraphPad Prism (GraphPad Prism Inc., San Diego, CA). Receiver operator characteristic (ROC) curves were generated for the model after being tested on both internal and external holdout test sets using GraphPad Prism. Optimal diagnostic thresholds to calculate sensitivity and specificity were determined via Youden's J-statistic. Area under the curve (AUC), 95% confidence interval (CI), and DeLong test for comparison of AUCs (significance defined as $p < 0.05$) were also calculated using GraphPad Prism.

Results

DCNN performance on internal test set versus external test set

On the internal test set, the best performing DCNN for pneumonia detection achieved an AUC of 0.95 (0.94–0.96, 95% CI). At the optimal diagnostic threshold, sensitivity was 87%, and specificity was 90%.

On the external test set, this same DCNN achieved an AUC of 0.54 (0.51–0.57, 95% CI), which was significantly lower than the performance on the internal test set ($p < 0.0001$). At the optimal diagnostic threshold, the sensitivity was 100%, and specificity was 0%.

Comparison of CAM heatmaps

The localization ability of the DCNN on internal test data was demonstrated by CAM heatmap images that focused on clinically relevant features of interest, such as consolidations with (Fig. 1A) and without air bronchograms (Fig. 1B), as well as perihilar interstitial opacities and peribronchial cuffing (Fig. 1C, D).

In contrast, the DCNN CAM heatmaps on the external test set showed poor localization ability without focus on clinically relevant features of interest in the lung parenchyma, but rather to extrapulmonary areas, such as the skull (Fig. 2), upper mediastinum (Fig. 2B), and abdomen (Fig. 2C).

Discussion

Although DLS have shown high diagnostic performance for identification of disease on medical images, caution has been recommended in deployment of such algorithms due to drops in performance on external datasets [7–9]. This study aimed to develop a DLS model for classification of pneumonia in pediatric chest radiographs and subsequently evaluate the model's performance on internal versus external test sets. Similar to prior works, we found that our DLS performed strongly when tested on the internal set, achieving an AUC of 0.95, but had significantly worse performance when tested on an external set, yielding an AUC of only 0.54. Likewise, there was marked difference in the heatmaps generated from the internal versus external sets, with the former showing emphasis of clinically relevant regions of disease and the latter showing emphasis of clinically irrelevant regions [1, 15].

The ability of our DLS model to accurately diagnose pediatric pneumonia on an internal set is consistent with findings from prior studies in both the pediatric and adult populations. In pediatric chest radiographs, prior studies trained DCNNs to identify pneumonia using the same dataset that we used and achieved AUCs ranging from 0.82 to 0.99 [1, 10, 11], which is comparable to our findings. In the adult population, Kermany et al. and Rajpurkar et al. demonstrated the ability of DLS models to diagnose a variety of thoracic diseases on adult chest radiographs, ranging from pneumonia to cardiomegaly, achieving AUCs ≥ 0.85 [8]. Altogether, these findings suggest that deep learning may enable automated diagnosis of thoracic disease at levels approaching that of radiologists.

Despite the promising results of deep learning for detection of pneumonia, skepticism and caution have been suggested towards these algorithms, due to concerns over poor generalizability of these models to external data not used to train them [7, 8, 11]. For example, Zech et al. previously showed that a DCNN trained on over 150,000 chest radiographs achieved an AUC of 0.931 for detection of pneumonia when tested on an internal dataset, but that this declined to 0.815 when tested on an external dataset [16]. Similarly, studies conducted on other imaging modalities (e.g., cardiac magnetic resonance imaging) have shown that while DLS performance is high when the training and testing images come from the same domain (e.g., scanner, site), performance may degrade significantly on images from other

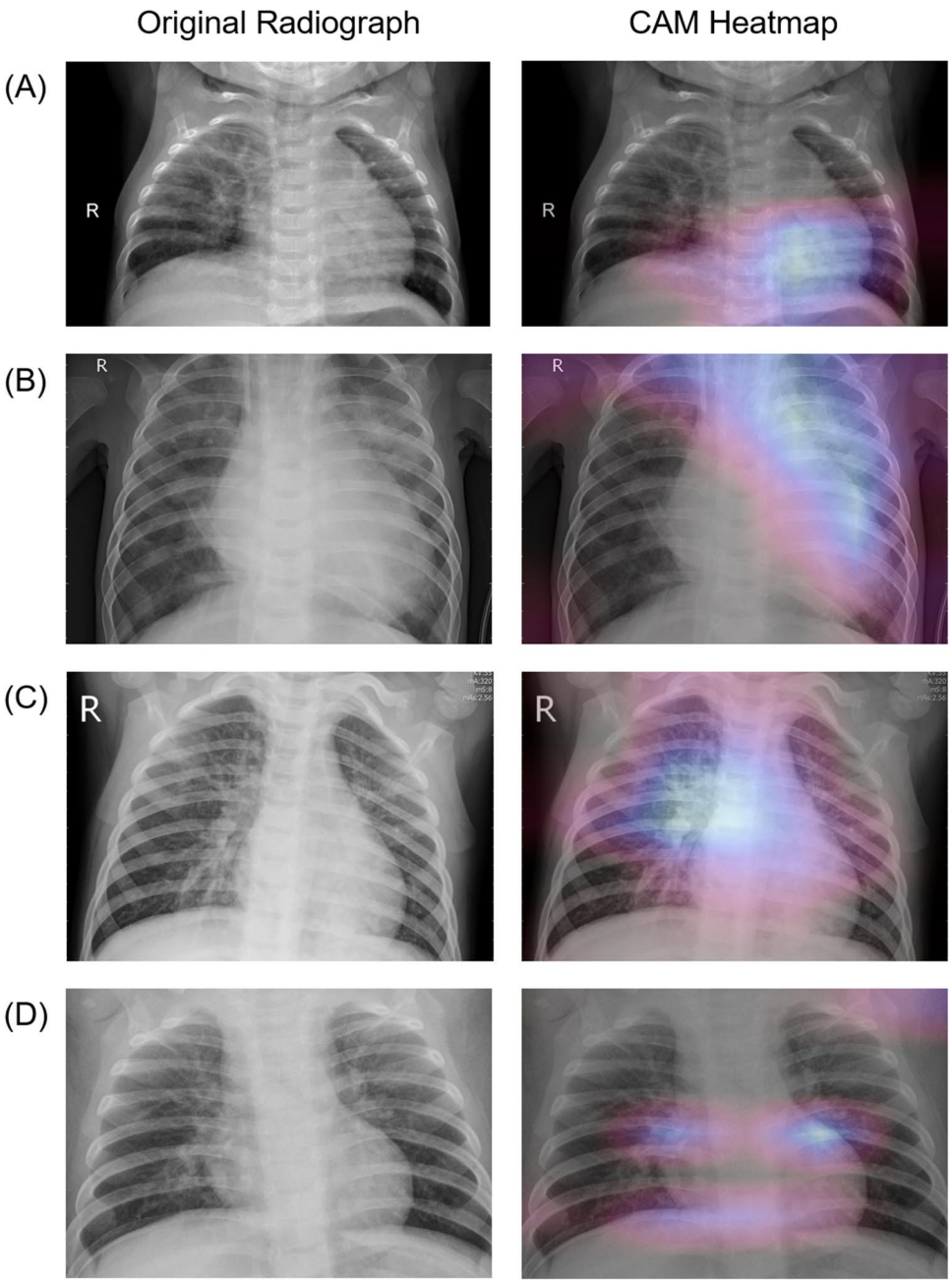

Fig. 1 Class activation mapping (CAM) heatmaps of radiographs from internal test set. **A**, **B** Radiographs demonstrate bacterial pneumonia, and CAM heatmaps showed appropriate emphasis on consolidation within the lung fields. **C**, **D** Radiographs demonstrate viral pneumonia, and CAM heatmaps showed appropriate emphasis on regions of perihilar thickening and peribronchial cuffing

scanners or clinical sites [16]. Our findings are consistent with these prior cautionary results, as we found that our DLS significantly declined in performance when tested on external data, achieving an AUC of 0.54. In other words, the DLS performed only slightly greater than chance.

In an effort to better understand potential reasons for the difference in performance of our DLS on internal versus external test sets, we generated CAM heatmaps to visualize portions of the images emphasized by the DCNNs for decision-making. We found that the DCNN model focused on different portions of a chest radiograph when evaluating images from the internal compared to the external datasets. Specifically, internal test data heatmaps showed that the DLS focused on clinically relevant areas of pneumonia, such as consolidations and perihilar opacities, while the external test data heatmaps showed that the DLS focused on clinically irrelevant areas, such as the skull and abdomen, which were infrequently included in the field-of-view in the training data. We note that there appeared to be systematic differences in the field-of-view of the CXRs between the two datasets, apparently related to institutional differences in radiograph acquisition protocols. Generally, the internal CXRs generally have a more focused field-of-view including the thorax and upper abdomen and the external CXRs

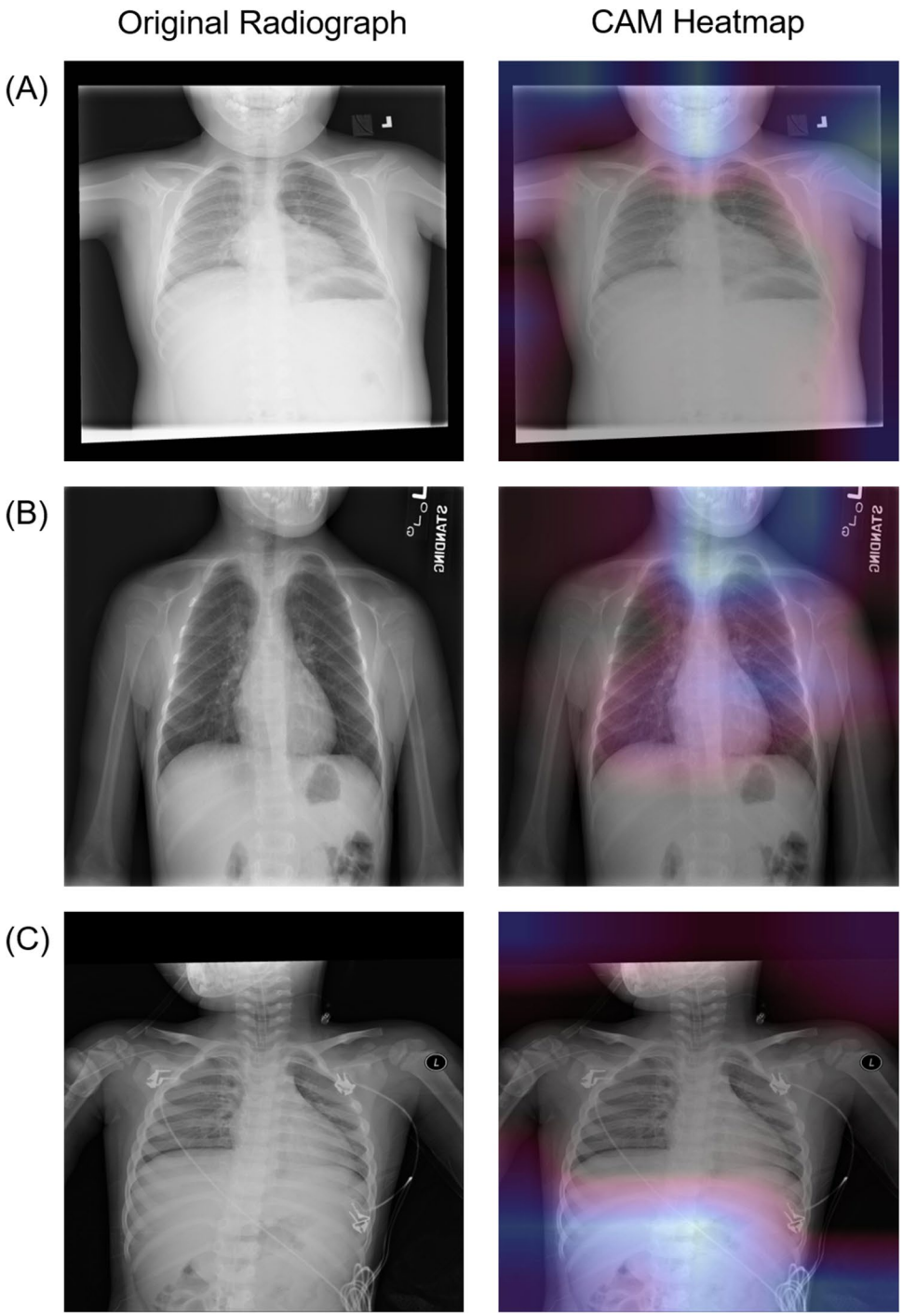

Fig. 2 Class activation mapping (CAM) heatmaps of radiographs from external test set. There is a lack of focus on clinically relevant features of interest within the lung parenchyma. Instead, extrapulmonary regions such as the skull (**A**), mediastinum (**B**), and abdomen (**C**) are given emphasis

being more variable in coverage, often including portions of the neck and skull, as well as larger portions of the abdomen. It is thus possible that because extra-thoracic regions of anatomy were generally not included in the training data, the DLS could have become "confused" when seeing them in the external test set. For instance, the DLS may have interpreted the "areas of opacity" that represent the skull and intraabdominal organs as signs of pneumonia. We do caution, however, that this is a speculation, as CAM heatmaps show parts of the image that are emphasized in decision-making, but they do not explain precisely what it is about those areas that make them important. Whatever the reason, it is clear that the DLS in our study interpreted the internal and external test images differently, as evidenced by the drop in performance.

There are several limitations to our study. First, the datasets used to train our DLS were small, with 5232 images, compared to much larger adult chest radiograph datasets

which have over 100,000 images, and which might be expected to limit generalizability of our algorithm. However, to our knowledge, this is the largest publicly available pediatric chest radiograph dataset, and one that has been used in several prior studies developing DLS with high-performance levels exceeding 90% accuracy. Furthermore, our intention was, in fact, to show that highly performing DCNNs should be interpreted with caution when there has not been evaluation on external test data; we thus recommend caution when training models on small datasets from a single site. Second, we evaluated only binary classification of chest radiographs into the presence or absence of pneumonia, without further evaluation of type of pneumonia (e.g., bacterial or viral). We elected not to divide the classification into bacterial versus viral pneumonia due to the overlapping radiographic appearance of these etiologies. Additionally, since classifying pneumonia into smaller sub-categories is a more difficult task, we would expect that the differences between internal and test data would be accentuated. Third, without more detailed information about the differences in patient population and imaging protocols of the internal and external datasets than that provided by the dataset curators, we are unable to fully assess specific demographic factors that may be contributing to the discrepancies and biases of our model. Fourth, we evaluated only a single DCNN architecture for our study; it is possible that other DCNN architectures may generalize better, although prior work has shown comparable performances for detection of abnormalities on chest radiographs between different DCNN architectures [17].

Conclusion

In conclusion, our DLS model for identifying pneumonia on pediatric chest radiographs performed exceptionally well when tested on an internal dataset but had a far lower accuracy when tested on an external dataset. Likewise, there were marked differences in the clinical relevance of features highlighted by heatmaps generated from internal versus external datasets. Future recommended areas of research and work include the curation of larger and more diversified pediatric CXR datasets curated from multiple clinical sites (as there are few of these compared to datasets for adult CXRs) and evaluating the impact of a preprocessing pipeline to exclude extra-thoracic regions of CXRs (such as the head/neck) on generalizability of CNNs for pneumonia detection, as these extra-thoracic regions appeared to be confounders for our CNN. Given the difference in performance between internal and external data, we recommend caution when evaluating DCNNs for medical image diagnosis that have been evaluated only on internal test data, and we propose that such algorithms should be evaluated on external test data prior to clinical deployment.

Data availability Images from the Guangzhou Women and Children's Medical Center (Link) and NIH ChestXray14 (Link) datasets are available for download from public databases.

Code availability The training and testing code used in this study are available upon request.

Declarations

Ethics approval The study was acknowledged by the Johns Hopkins IRB institutional review board (IRB) as non-human subject research. Accordingly, formal IRB review was not required per our institutional policies.

Consent to participate Not applicable.

Consent for publication Not applicable.

Conflict of interest The authors declare no competing interests.

References

1. Kermany DS, Goldbaum M, Cai W et al (2018) Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 172:1122-1131.e9. https://doi.org/10.1016/j.cell.2018.02.010
2. Wang X, Peng Y, Lu L et al (2017) ChestX-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. ArXiv. https://doi.org/10.1109/CVPR.2017.369
3. Yi PH, Kim TK, Yu AC et al (2020) Can AI outperform a junior resident? Comparison of deep neural network to first-year radiology residents for identification of pneumothorax. Emerg Radiol 27:367–375. https://doi.org/10.1007/s10140-020-01767-4
4. Hosny A, Parmar C, Quackenbush J et al (2018) Artificial intelligence in radiology. Nat Rev Cancer 18:500–510. https://doi.org/10.1038/s41568-018-0016-5
5. Borkowski A, Viswanadham N, Thomas LB et al (2020) Using artificial intelligence for COVID-19 chest x-ray diagnosis. Federal Practitioner 37:398–404. https://doi.org/10.12788/fp.0045
6. Katz SE, Williams DJ (2018) Pediatric community-acquired pneumonia in the United States: changing epidemiology, diagnostic and therapeutic challenges, and areas for future research. Infect Dis Clin N Am 32:47–63. https://doi.org/10.1016/j.idc.2017.11.002
7. Sathitratanacheewin S, Sunanta P, Pongpirul K (2020) Deep learning for automated classification of tuberculosis-related chest X-ray: dataset distribution shift limits diagnostic performance generalizability. Heliyon 6:e04614. https://doi.org/10.1016/j.heliyon.2020.e04614
8. Zech JR, Badgeley MA, Liu M et al (2018) Confounding variables can degrade generalization performance of radiological deep learning models. ArXiv. https://doi.org/10.1371/journal.pmed.1002683
9. Yasaka K, Abe O (2018) Deep learning and artificial intelligence in radiology: current applications and future directions. PLoS Med 15:e1002707. https://doi.org/10.1371/journal.pmed.1002707
10. Hashmi MF, Katiyar S, Keskar AG et al (2020) Efficient pneumonia detection in chest x-ray images using deep transfer learning. Diagnostics 10:417. https://doi.org/10.3390/diagnostics10060417

11. Longjiang E, Zhao B, Guo Y et al (2019) Using deep-learning techniques for pulmonary-thoracic segmentations and improvement of pneumonia diagnosis in pediatric chest radiographs. Pediatr Pulmonol 54:1617–1626. https://doi.org/10.1002/ppul.24431
12. Kermany D, Zhang K, Goldbaum M (2018) Large dataset of labeled optical coherence tomography (OCT) and chest x-ray images. https://data.mendeley.com/datasets/rscbjbr9sj/3. Accessed 15 Dec 2020
13. National Institutes of Health NIH chest x-rays: Over 112,000 chest x-ray images from more than 30,000 unique patients. https://www.kaggle.com/nih-chest-xrays/data. Accessed 15 Dec 2020
14. Zhou B, Khosla A, Lapedriza A, et al (2015) Learning deep features for discriminative localization. arXiv:1512.04150
15. Rajpurkar P, Irvin J, Ball RL et al (2018) Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 15:e1002686. https://doi.org/10.1371/journal.pmed.1002686
16. Chen C, Bai W, Davies RH et al (2019) Improving the generalizability of convolutional neural network-based segmentation on CMR images. Front Cardiovasc Med 7:105. https://doi.org/10.3389/fcvm.2020.00105
17. Dunnmon JA, Yi D, Langlotz CP et al (2019) Assessment of convolutional neural networks for automated classification of chest radiographs. Radiology 290:537–544. https://doi.org/10.1148/radiol.2018181422