Journal of Digital Imaging
https://doi.org/10.1007/s10278-019-00180-9

Generalizable Inter-Institutional Classification of Abnormal Chest Radiographs Using Efficient Convolutional Neural Networks

Ian Pan[1] · Saurabh Agarwal[2] · Derek Merck[2]

Abstract
Our objective is to evaluate the effectiveness of efficient convolutional neural networks (CNNs) for abnormality detection in chest radiographs and investigate the generalizability of our models on data from independent sources. We used the National Institutes of Health ChestX-ray14 (NIH-CXR) and the Rhode Island Hospital chest radiograph (RIH-CXR) datasets in this study. Both datasets were split into training, validation, and test sets. The DenseNet and MobileNetV2 CNN architectures were used to train models on each dataset to classify chest radiographs into normal or abnormal categories; models trained on NIH-CXR were designed to also predict the presence of 14 different pathological findings. Models were evaluated on both NIH-CXR and RIH-CXR test sets based on the area under the receiver operating characteristic curve (AUROC). DenseNet and MobileNetV2 models achieved AUROCs of 0.900 and 0.893 for normal versus abnormal classification on NIH-CXR and AUROCs of 0.960 and 0.951 on RIH-CXR. For the 14 pathological findings in NIH-CXR, MobileNetV2 achieved an AUROC within 0.03 of DenseNet for each finding, with an average difference of 0.01. When externally validated on independently collected data (e.g., RIH-CXR-trained models on NIH-CXR), model AUROCs decreased by 3.6–5.2% relative to their locally trained counterparts. MobileNetV2 achieved comparable performance to DenseNet in our analysis, demonstrating the efficacy of efficient CNNs for chest radiograph abnormality detection. In addition, models were able to generalize to external data albeit with performance decreases that should be taken into consideration when applying models on data from different institutions.

Keywords Convolutional neural networks · Deep learning · Generalizability · Chest radiographs · Classification

Introduction

Recent advancements in deep learning have enabled the development of image recognition algorithms that rival the accuracy of human interpretation [1, 2]. The vast majority of these algorithms are based on convolutional neural networks (CNNs)—the current state-of-the-art for computer vision—which leverage large amounts of data to solve specific tasks [3]. There has been increasing interest in applying deep learning to healthcare, specifically medical imaging. Researchers have developed algorithms to recognize diabetic retinopathy and other ocular diseases in fundus photographs [4, 5]; detect skin cancer from digital photographs of skin lesions [6]; and analyze pathology slides for lymph node metastases [7].

Radiology in particular is poised to be transformed by deep learning. CNNs have been trained to estimate pediatric bone age from hand radiographs [8–10]; detect critical findings in head CT scans [11]; and detect breast cancer in mammography [12, 13], among other applications [14]. Many deep learning solutions have focused on chest radiography, the most common imaging study performed in the USA and worldwide. Recent successful examples within chest radiography include tuberculosis detection [15], endotracheal tube placement evaluation [16], and classification of various abnormalities [17, 18].

Deep learning applications in medical imaging have been limited due to difficulties with accessing healthcare data within medical institutions. The National Institutes of Health recently released a dataset of 112,120 chest radiographs labeled with 14 different abnormalities to expedite development in this domain [19]. Using this dataset, researchers have trained CNNs to detect the presence of various abnormalities in chest

✉ Ian Pan
ian_pan@brown.edu

[1] Warren Alpert Medical School, Brown University, Box G-9130, Providence, RI 02912, USA

[2] Department of Diagnostic Imaging, Rhode Island Hospital, 593 Eddy St, Main, Floor 3, Providence, RI 02903, USA

Published online: 05 March 2019

radiographs, with one group achieving radiologist-level performance on 11 out of 14 pathologies [20].

High-performing deep learning algorithms have not yet been widely adopted in radiology. Imaging study prioritization would be a particularly useful potential application of deep learning in radiology. In such a workflow, each imaging study would be preliminarily reviewed algorithmically and assigned a score indicating the likelihood of an abnormal finding. Imaging studies with a high probability of an important finding would then be moved to the top of a radiologist's queue. Similarly, CNNs could be used to automatically label imaging studies as normal if they have a sufficiently low abnormality score. In theory, this practice would improve efficiency in the radiology workflow and provide more rapid delivery of imaging findings to the ordering physician.

Deep learning can also provide valuable radiological insight in parts of the developing world where access to trained subspecialist radiologists is limited. Though the diagnostic breadth and depth of these algorithms have yet to approach human capabilities, they can augment the decision-making of healthcare professionals in under-resourced settings by highlighting potential abnormalities in imaging studies. Many of these settings also lack the computational resources and infrastructure to accommodate the use of many deep learning solutions; however, progress has been made in designing computationally efficient CNNs suited to portable, low-power devices (e.g., mobile phones) [21].

To have a broad impact in healthcare, deep learning algorithms must generalize to unseen data from different populations (e.g., across different scanners and multiple institutions). A recent study demonstrated variable generalization performance of a deep learning model to detect pneumonia in chest radiographs [22]; however, the majority of recent work reports only internal validation results where performance is evaluated on a holdout test set that originates from the same source as the training and validation data; though these analyses are rigorous, the results can be optimistically biased due to data commonalities that can propagate throughout training, validation, and test sets.

Objective

The data, methods, and results presented here were motivated by three objectives:

1) Evaluate the effectiveness of efficient CNN architectures for abnormality detection in chest radiographs.
2) Demonstrate model generalizability across independently collected datasets.
3) Compare performances of locally trained models (trained on the same data source as the test set) versus externally trained models (train on different data sources from the test set).

Materials and Methods

We used the NIH ChestX-ray14 (NIH-CXR) and Rhode Island Hospital chest radiograph (RIH-CXR) datasets in our experiments. NIH-CXR consists of 112,120 frontal view chest radiographs from 30,805 patients, each of which is labeled with the presence or absence of 14 different findings. We defined normal as the absence of any of the 14 findings, which does not preclude the existence of other findings in these radiographs. We excluded chest radiographs with no findings if at least one of the patient's other radiographs contained a finding to ensure the integrity of our negative labels for a total of 75,555 radiographs. RIH-CXR consists of 17,202 frontal view chest radiographs with a binary class label for normal vs. abnormal collected between September 28, 2017 and March 14, 2018. Labels for this dataset were determined by reviewing the radiology report for an indication of a normal vs. priority result [23]. Normal RIH-CXR radiographs were studies that were read as completely normal without the presence of any pathology (e.g., mild atelectasis, mild scoliosis). We excluded radiographs that contained lines or tubes because chest radiographs performed at RIH for the purpose of verifying line or tube placement would often be associated with a priority result even if the study was normal. This included radiographs with central catheters, drain and chest tubes, endotracheal tubes, and cardiac assist devices. Radiographs with other iatrogenic material, such as coronary stents, sternal wires, or prosthetic valves, were not expressly excluded. Table 1 summarizes basic characteristics of the two datasets.

Image Preprocessing and Network Architectures

We compared the efficient MobileNetV2 CNN architecture [21] (2.3 million parameters) with a 121-layer DenseNet architecture (7 million parameters) [24]. Both architectures were initialized with ImageNet pretrained weights [25]. Images

Table 1 Basic characteristics of the NIH ChestX-ray14 and RIH-CXR datasets

	NIH ChestX-ray14	RIH-CXR
Number of radiographs (patients)	75,555 (30,805)	17,202 (14,471)
Normal	23,798 (31.5%)	9030 (52.5%)
Abnormal	51,759 (68.5%)	8172 (47.5%)
Mean age (SD), years	46.8 (17.0)	49.8 (24.9)
Female	46.0%	52.6%

were padded and resized to 256 × 256 pixels. The single-channel images were then converted to 3-channel images by duplicating the channel 3 times and preprocessed by rescaling the pixel values from [0, 255] to [0, 1] and subtracting the normalized, channel-wise ImageNet means and standard deviations.

Experimental Setup

We conducted two experiments in order to examine model generalizability across datasets (i.e., performance of NIH-CXR-trained model on RIH-CXR and vice versa).

Experiment 1

We divided NIH-CXR into random 70%/10%/20% (training/validation/test) stratified splits with no patient overlap and used the RIH-CXR test set as an external validation dataset.

Our primary evaluation metric was the area under the ROC curve (AUROC). On the NIH-CXR test set, AUROCs were calculated for each of the 14 findings as well as classification of normal vs. abnormal radiographs. On the RIH-CXR test set, AUROC was calculated only for classification of normal vs. abnormal radiographs.

Experiment 2

We divided RIH-CXR into 75%/10%/15% splits by study date (training and validation: September 28, 2017 to February 18, 2018; test: February 18, 2018 to March 14, 2018) and used the NIH-CXR test set as an external validation set. Patients were removed from the test set if they were also present in the training or validation sets.

Our primary evaluation metric was the area under the ROC curve (AUROC). For models trained on RIH-CXR, AUROC was calculated only for classification of normal vs. abnormal radiographs. In RIH-CXR, a small percentage of the studies consisted of multiple frontal view radiographs; in these cases, the predictions were averaged across images.

Model Training

Models were trained using PyTorch 1.0 (https://pytorch.org) [26] in the Python 2.7 programming language (Python Software Foundation, https://www.python.org) using a NVIDIA Titan V 12GB GPU. In experiment 1, models were trained on NIH-CXR to predict 15 different labels, one for each of the 14 findings and an additional label for the presence of any finding (i.e., abnormal vs. normal label). We optimized the weighted sum of class-individual binary cross-entropies, where each class was weighted by the inverse of its frequency in the training set. In experiment 2, models were trained on RIH-CXR to predict only a binary class label for abnormal vs. normal chest radiographs. For these models, we optimized the weighted binary cross-entropy.

On-the-fly data augmentation (magnification, blurring, contrast adjustment, left-right flips, rotations) with probability, 0.5 was used for regularization. Models were fine-tuned using the Adam optimizer [27] with an initial learning rate of 1e-4, weight decay of 1e-6, and default parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$). No layers were frozen during the fine-tuning process. The learning rate was halved after the validation AUROC plateaued for two epochs; early stopping was initiated after 6 epochs of no improvement in the validation AUROC.

Our final model was a 3-model ensemble of the models with the highest validation AUROC for each of three training sessions for differentiation of abnormal vs. normal chest radiographs.

Statistical Analysis

All statistical analyses were performed using the Python 2.7 programming language. The receiver operating characteristic curve and AUROC were calculated using the scikit-learn package, and the bootstrap method was used to calculate 95% confidence intervals for differences in model performance [28]. We considered any difference where the 95% confidence interval did not contain 0 to be statistically significant (i.e., $p < 0.05$).

Implementation

All code used for model training and evaluation is publicly available on our GitHub repository [https://github.com/i-pan/chestxrays/].

Results

MobileNetV2 models trained on RIH and NIH are referred to as RIH-Mobile and NIH-Mobile, respectively. The same naming convention is used for DenseNet models.

MobileNetV2 vs. DenseNet

MobileNetV2 demonstrated comparable performance to DenseNet on both datasets for the binary classification task. RIH-Mobile and RIH-Dense achieved AUROCs of 0.951 and 0.960, respectively on RIH-CXR, with AUROCs of 0.847 and 0.855 on NIH-CXR. NIH-Mobile and NIH-Dense achieved AUROCs of 0.893 and 0.900 on NIH-CXR, with AUROCs of 0.917 and 0.924 on RIH-CXR. Differences and 95% confidence intervals are reported in Table 2. We also present CheXNet results as reported in Rajpurkar et al. [29] for comparison.

On NIH-CXR, differences between NIH-Mobile and NIH-Dense were not statistically significant ($p > 0.05$) for 6 out of

Table 2 AUROCs for each of the 14 findings in the NIH ChestX-ray14 dataset for MobileNetV2 and DenseNet models with performance differences and bootstrapped 95% confidence intervals. Number of radiographs refers to the total number of radiographs across training, validation, and test sets. CheXNet results from the original paper for each finding are included for comparison. Bolded values indicate the highest AUROC for each finding

Finding	Number of radiographs	CheXNet AUROC	NIH-Dense AUROC	NIH-Mobile AUROC	Mean difference (95% CI)
Atelectasis	11,559	0.809	0.833	0.820	0.013 (0.009, 0.017)*
Cardiomegaly	2776	0.925	0.920	0.913	0.007 (0.001, 0.013)*
Consolidation	4667	0.790	0.804	0.799	0.005 (−0.001, 0.010)
Edema	2303	0.888	0.910	0.915	−0.005 (−0.011, 0.001)
Effusion	13,317	0.864	0.892	0.888	0.004 (0.002, 0.007)*
Emphysema	2516	0.937	0.928	0.919	0.009 (0.002, 0.016)*
Fibrosis	1686	0.805	0.804	0.797	0.006 (−0.007, 0.020)
Hernia	227	0.916	0.939	0.957	−0.019 (−0.047, 0.009)
Infiltration	19,894	0.735	0.722	0.716	0.009 (0.004, 0.013)*
Mass	5782	0.868	0.859	0.833	0.026 (0.018, 0.034)*
Nodule	6331	0.780	0.763	0.733	0.030 (0.020, 0.040)*
Pleural Thickening	3385	0.806	0.813	0.795	0.018 (0.009, 0.026)*
Pneumonia	1431	0.768	0.734	0.740	−0.005 (−0.021, 0.011)
Pneumothorax	5302	0.889	0.898	0.883	0.002 (−0.003, 0.006)

*$p < 0.05$

14 findings. All AUROC differences were ≤ 0.03, and the mean absolute difference across the 14 findings was 0.01. The largest decreases in performance were for mass, nodule, and pleural thickening detection whereas consolidation, edema, fibrosis, hernia, pneumonia, and pneumothorax detection were statistically equivalent.

Locally Trained Models vs. Externally Trained Models

Models trained on NIH-CXR achieved lower AUROCs on the RIH-CXR test set than models trained on RIH-CXR and vice versa. NIH-Dense and NIH-Mobile experienced decreases in performance of 3.7% and 3.6% versus RIH-Dense and RIH-Mobile on RIH-CXR, respectively (AUROCs: 0.924 vs. 0.960; 0.917 vs. 0.951). Similarly, RIH-Dense and RIH-Mobile experienced decreases in performance of 5.0% and 5.2% versus NIH-Dense and NIH-Mobile on NIH-CXR (AUROCs: 0.855 vs. 0.900; 0.847 vs. 0.893). All performance differences were found to be significant at a 0.05 level (Table 3).

Correlations between prediction scores from locally trained and externally trained models ranged from 0.84 to 0.86. NIH-trained models consistently predicted higher scores: the median differences across the four comparisons ranged from 0.12 to 0.19 (Table 4). In general, we observed that on NIH-CXR, the largest absolute score differences between NIH-trained and RIH-trained models occurred when RIH models predicted low abnormality scores for positive images. Conversely, on RIH-CXR, the largest score differences occurred when NIH models predicted high abnormality scores for negative images. Figure 1 depicts examples of images with the highest score differences.

We evaluated the F1 score at various thresholds for the four models on the NIH-CXR and RIH-CXR test sets (Fig. 2). Table 5 shows the optimal thresholds and corresponding F1 scores for the eight model-test set pairings. The optimal thresholds for NIH-Dense and NIH-Mobile are increased by 0.20 and 0.25 from NIH-CXR to RIH-CXR. Conversely, the optimal thresholds for RIH-Dense and RIH-Mobile are decreased by 0.30 and 0.25 from

Table 3 AUROCs for the binary classification task of normal vs. abnormal chest radiograph for locally and externally trained models. Mean differences are reported as local model AUROC – external model AUROC

	NIH-Dense AUROC	RIH-Dense AUROC	Mean difference (95% CI)	NIH-Mobile AUROC	RIH-Mobile AUROC	Mean difference (95% CI)
NIH-CXR test	0.900	0.855	0.045 (0.041, 0.049)*	0.893	0.847	0.046 (0.041, 0.050)*
RIH-CXR test	0.924	0.960	0.036 (0.026, 0.046)*	0.917	0.951	0.035 (0.024, 0.044)*

*$p < 0.05$

Table 4 Median prediction scores over all test images of NIH-trained vs. RIH-trained models. IQR interquartile range

	NIH-CXR prediction scores, median (IQR)	RIH-CXR prediction scores, median (IQR)
NIH-Mobile	0.814 (0.453, 0.979)	0.684 (0.467, 0.916)
RIH-Mobile	0.561 (0.251, 0.890)	0.368 (0.142, 0.803)
Difference: NIH – RIH	0.121 (0.035, 0.251)	0.215 (0.041, 0.359)
NIH-Dense	0.852 (0.454, 0.984)	0.612 (0.385, 0.901)
RIH-Dense	0.518 (0.176, 0.924)	0.314 (0.110, 0.839)
Difference: NIH – RIH	0.131 (0.023, 0.306)	0.163 (0.010, 0.319)

RIH-CXR to NIH-CXR. When applying the locally-determined optimal threshold on the external test data, the decrease in F1 scores ranged from 5.3 to 11.8% as compared to the externally-determined optimal threshold.

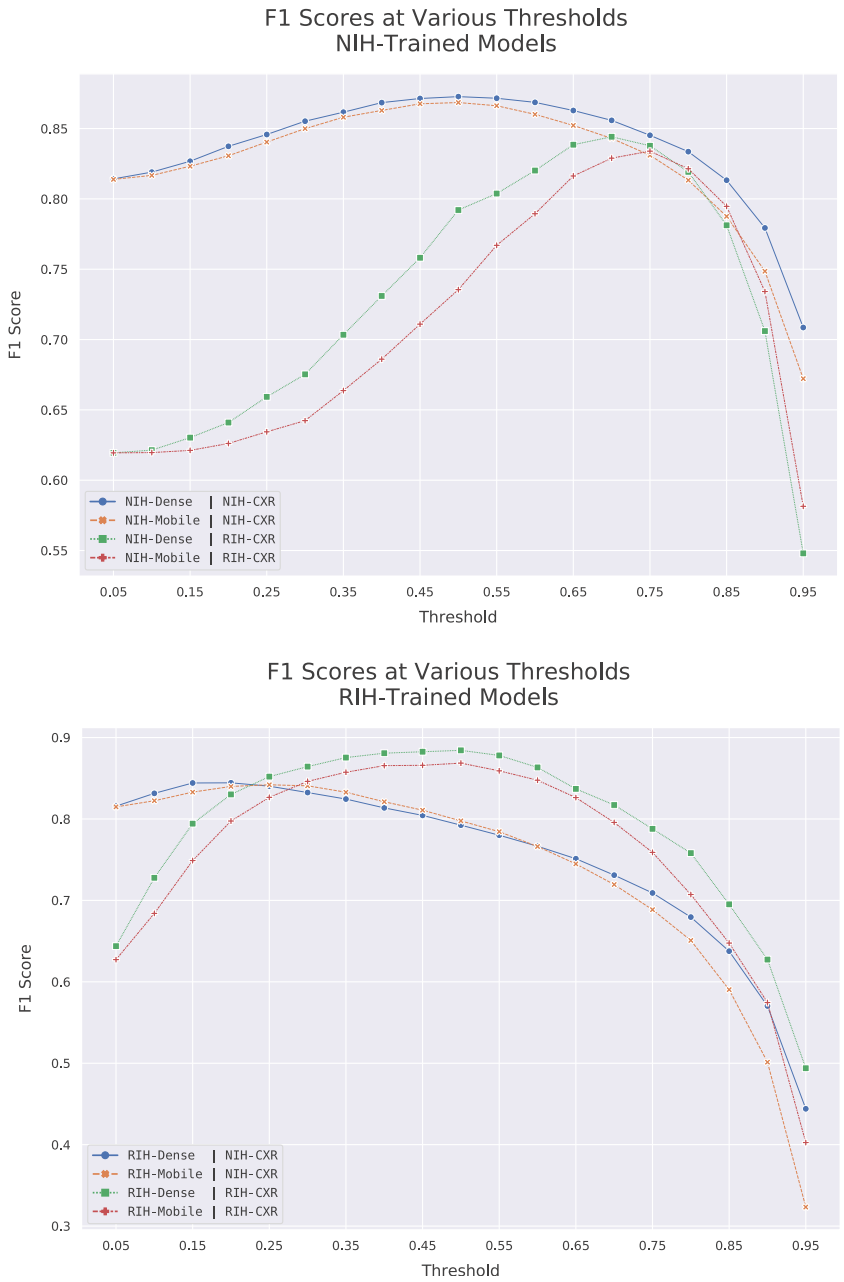

Fig. 1 F1 scores at various thresholds for NIH-Dense/NIH-Mobile (top) and RIH-Dense/RIH-Mobile (bottom) on NIH-CXR and RIH-CXR test sets

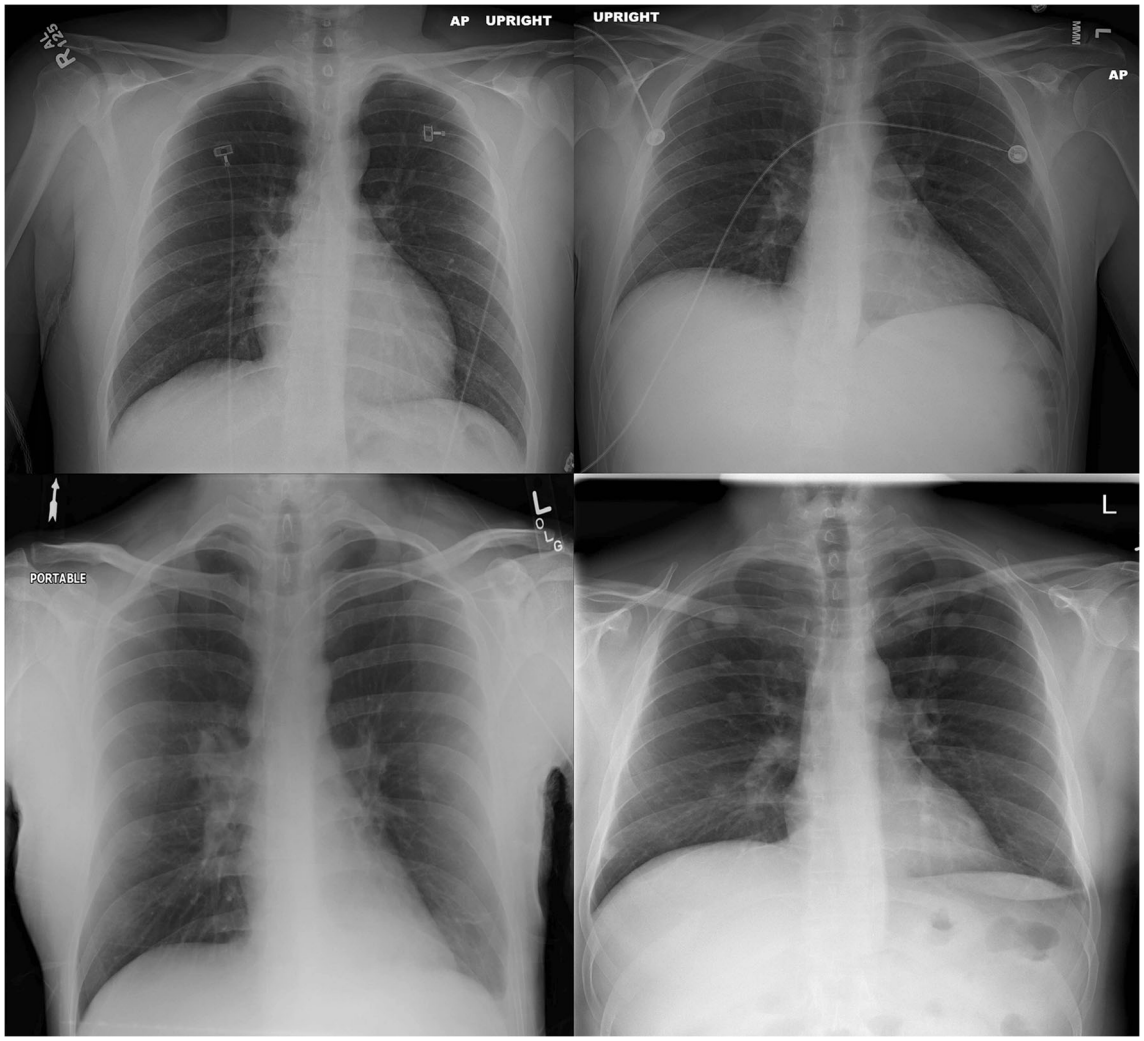

Fig. 2 Images with the largest score difference between locally-trained and externally-trained models. Top left: normal radiograph from RIH-CXR (RIH-Mobile, 0.06; NIH-Mobile, 0.86). Top right: normal radiograph from RIH-CXR (RIH-Dense, 0.10; NIH-Dense, 0.89). Bottom left: abnormal radiograph with findings of atelectasis, effusion, pleural thickening from NIH-CXR (NIH-Mobile, 0.92; RIH-Mobile, 0.09). Bottom right: abnormal radiograph with findings of nodule from NIH-CXR (NIH-Dense, 0.97; RIH-Dense, 0.13)

Discussion

Our analysis demonstrates the effectiveness of mobile CNN architectures for detecting abnormalities in chest radiographs. This opens the door to low-power CNN applications that can be directly embedded into the imaging apparatus itself to streamline the radiology AI workflow. These results also suggest that the computational requirements of larger, deeper networks may not justify the performance gain for certain medical imaging classification tasks, particularly in settings with limited resources.

In our experiments, NIH-trained models achieved higher AUROCs on the RIH-CXR test set versus the NIH-CXR test

Table 5 Optimal thresholds and corresponding F1 scores for the four models on the two test sets. Local-to-external threshold change is the change in the optimal threshold when predicting on local vs. external data. Local-to-external performance change is the percent change in model performance when using the locally determined optimal threshold vs. the externally determined optimal threshold on the external test data

	NIH-CXR test threshold (F1 score)	RIH-CXR test threshold (F1 score)	Local-to-external threshold change	Local-to-external performance change (%)
NIH-Mobile	0.50 (0.869)	0.75 (0.834)	+ 0.25	− 11.8
NIH-Dense	0.50 (0.873)	0.70 (0.844)	+ 0.20	− 6.2
RIH-Mobile	0.25 (0.842)	0.50 (0.869)	− 0.25	− 5.3
RIH-Dense	0.20 (0.844)	0.50 (0.884)	− 0.30	− 6.2

set, which may seem counterintuitive; however, we emphasize that we are comparing performance differences in locally trained versus externally trained models, as opposed to performance differences of the same model on different datasets. We postulate that this result is due to the relative difficulty level and noise of the two datasets. The NIH data labels were mined using natural language processing on radiology reports and can be considered weaker labels, whereas the RIH data were explicitly labeled as normal or priority. As described, normal NIH radiographs may have contained findings that were not included in the 14 provided labels, whereas normal RIH radiographs were entirely normal. Thus, it is likely more difficult for the model to distinguish between normal and abnormal radiographs in the NIH data than in the RIH data. Despite these differences in the data annotation process, externally trained models were able to perform within 5.2% of locally trained models.

Model bias was clearly demonstrated in our experiments. Due to differences in data distribution, the NIH-trained models were more likely to predict higher scores for both positive and negative images than RIH-trained models. When introducing binary decision rules, thresholds determined on local data resulted in further decreased performance on external data. This is noteworthy because different institutions likely have different prior distributions of abnormal images. Thus, score thresholds which were determined on one institution's data will likely not transfer another institution's data. Dataset size also influences generalizability; NIH-trained models showed smaller generalization gaps than RIH-trained models, despite having weaker labels, which can be partly explained by the NIH dataset's larger sample size.

In previous work, Lakhani and Sundaram achieved near-perfect performance for tuberculosis detection (AUROC = 0.99) using two CNN architectures, AlexNet and GoogLeNet. Cicero et al. demonstrated excellent performance for abnormality detection in chest radiography, which included abnormal vs. normal (AUROC = 0.964) and five specific findings (pleural effusion, pulmonary edema, cardiomegaly, pneumothorax, consolidation) using GoogLeNet.

Rajpurkar et al. developed CheXNeXt using NIH ChestX-ray14, which detected the presence of 14 findings in chest radiographs, based on the same 121-layer DenseNet CNN architecture as our NIH-Dense and RIH-Dense models. They conducted a rigorous study comparing the algorithm's performance with that of nine radiologists demonstrating equivalent performance for 11 out of 14 findings.

The study by Zech et al. most closely resembles our own. The authors observed variable generalization performance of a deep learning model to detect pneumonia in chest radiographs. Our findings complement their study by also demonstrating differences in generalization performance when applying a deep learning model trained on one institution's data to another institution. In addition, we show that these differences are similar between two CNN architectures.

Our work differs from the aforementioned papers in that, in addition to evaluating the effectiveness of our models for abnormality detection in chest radiographs, we systematically compared the performance of two CNN architectures (MobileNetV2 and DenseNet). Furthermore, we show how CNNs trained on one dataset of chest radiographs are still effective on independently collected chest radiographs, albeit with a decrease in performance. This gives us confidence that CNNs are learning appropriate, generalizable features during the training process but whose predictions are biased to the training data. Institutions should take care to validate commercial algorithms or algorithms trained on other data sources on internal datasets to accurately assess performance within their own ecosystems. We strongly recommend that algorithms be subject to an initial trial period where performance is first evaluated behind the scenes on local data before clinical deployment. A thorough understanding of differences between the training data and the implementation site's data is essential for proper clinical implementation of deep learning models.

Limitations

Labels for the RIH-CXR dataset were generated via review of the radiology reports, which are each produced by a single radiologist. Similarly, the NIH ChestX-ray14 data were weakly labeled via natural language processing. Given the variation that exists in reads among different radiologists, our labels are not as robust as labels generated via consensus of multiple radiologists or by a dedicated fellowship-trained thoracic radiologist. Furthermore, we define normal radiographs in ChestX-ray14 to be radiographs without any of the 14 findings. However, it is possible that these radiographs may not be entirely normal and contain other findings. One limitation during inference is that only the frontal view is used to determine the prediction; up to 15% of the lung is not visualized on the frontal view and the lateral view is often necessary for accurate interpretation [30]. In addition, our analysis only examined generalizability between two institutions; a more rigorous analysis would compare across several institutions. Finally, we reiterate that our goal is not to replace the diagnostic capabilities of a radiologist but to augment decision-making and expedite the imaging workflow by quickly identifying chest radiographs with potentially urgent findings and suggesting what these findings may be.

Conclusion

In this study, we examined the effectiveness of mobile CNN architectures for abnormality detection in chest radiographs and the generalizability of models trained on different datasets. Our results indicate that mobile architectures achieve performance comparable to larger architectures for abnormal

vs. normal classification of chest radiographs as well as detection of specific abnormalities. We also demonstrate that models can generalize across datasets from different institutions with a 3.6–5.2% performance decrease. Additional performance decreases of 5.3–11.8% can be seen when determining optimal thresholds for decision rules. Further work should focus on investigating clinical implementation of these models for imaging study triage and prioritization with an evaluation of their clinical utility. In addition, studies should explore how much external data is necessary for a deep learning model to adapt to a new institution's data distribution.

References

1. Yu Q, Yang Y, Liu F, Song Y-Z, Xiang T, Hospedales TM: Sketch-A-Net: a deep neural network that beats humans. Int J Comput Vis. 122(3):411–425, 2017. https://doi.org/10.1007/s11263-016-0932-3
2. Dodge S, Karam L. A Study and Comparison of Human and Deep Learning Recognition Performance Under Visual Distortions. arXiv:170502498 [cs]. May 2017. http://arxiv.org/abs/1705.02498
3. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, eds. Advances in Neural Information Processing Systems 25. 2012:1097–1105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf
4. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL, Webster DR: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 316(22):2402–2410, 2016. https://doi.org/10.1001/jama.2016.17216
5. Ting DSW, Cheung CY-L, Lim G et al.: Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA. 318(22):2211–2223, 2017. https://doi.org/10.1001/jama.2017.18152
6. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S: Dermatologist-level classification of skin cancer with deep neural networks. Nature. 542(7639):115–118, 2017. https://doi.org/10.1038/nature21056
7. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, van der Laak JAWM, and the CAMELYON16 Consortium, Hermsen M, Manson QF, Balkenhol M, Geessink O, Stathonikos N, van Dijk MCRF, Bult P, Beca F, Beck AH, Wang D, Khosla A, Gargeya R, Irshad H, Zhong A, Dou Q, Li Q, Chen H, Lin HJ, Heng PA, Haß C, Bruni E, Wong Q, Halici U, Öner MÜ, Cetin-Atalay R, Berseth M, Khvatkov V, Vylegzhanin A, Kraus O, Shaban M, Rajpoot N, Awan R, Sirinukunwattana K, Qaiser T, Tsang YW, Tellez D, Annuscheit J, Hufnagl P, Valkonen M, Kartasalo K, Latonen L, Ruusuvuori P, Liimatainen K, Albarqouni S, Mungal B, George A, Demirci S, Navab N, Watanabe S, Seno S, Takenaka Y, Matsuda H, Ahmady Phoulady H, Kovalev V, Kalinovsky A, Liauchuk V, Bueno G, Fernandez-Carrobles MM, Serrano I, Deniz O, Racoceanu D, Venâncio R: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA. 318(22):2199–2210, 2017. https://doi.org/10.1001/jama.2017.14585
8. Lee H, Tajmir S, Lee J, Zissen M, Yeshiwas BA, Alkasab TK, Choy G, Do S: Fully automated deep learning system for bone age assessment. J Digit Imaging. 30(4):427–441, 2017. https://doi.org/10.1007/s10278-017-9955-8
9. Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP: Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. Radiology. 287(1):313–322, 2017. https://doi.org/10.1148/radiol.2017170236
10. Halabi SS, Prevedello LM, Kalpathy-Cramer J, Mamonov AB, Bilbily A, Cicero M, Pan I, Pereira LA, Sousa RT, Abdala N, Kitamura FC, Thodberg HH, Chen L, Shih G, Andriole K, Kohli MD, Erickson BJ, Flanders AE: The RSNA pediatric bone age machine learning challenge. Radiology.:180736, 2018. https://doi.org/10.1148/radiol.2018180736
11. Chilamkurthy S, Ghosh R, Tanamala S, Biviji M, Campeau NG, Venugopal VK, Mahajan V, Rao P, Warier P: Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet. 392(10162):2388–2396, 2018. https://doi.org/10.1016/S0140-6736(18)31645-3
12. Ribli D, Horváth A, Unger Z, Pollner P, Csabai I: Detecting and classifying lesions in mammograms with Deep Learning. Sci Rep. 8:4165, 2018. https://doi.org/10.1038/s41598-018-22437-z
13. Becker AS, Marcon M, Ghafoor S, Wurnig MC, Frauenfelder T, Boss A: Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. Invest Radiol. 52(7):434–440, 2017. https://doi.org/10.1097/RLI.0000000000000358
14. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI: A survey on deep learning in medical image analysis. Medical Image Analysis. 42:60–88, 2017. https://doi.org/10.1016/j.media.2017.07.005
15. Lakhani P, Sundaram B: Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology. 284(2):574–582, 2017. https://doi.org/10.1148/radiol.2017162326
16. Lakhani P: Deep convolutional neural networks for endotracheal tube position and X-ray image classification: challenges and opportunities. J Digit Imaging. 30(4):460–468, 2017. https://doi.org/10.1007/s10278-017-9980-7
17. Cicero M, Bilbily A, Colak E, Dowdell T, Gray B, Perampaladas K, Barfett J: Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs. Invest Radiol. 52(5):281–287, 2017. https://doi.org/10.1097/RLI.0000000000000341
18. Putha P, Tadepalli M, Reddy B, et al. Can Artificial Intelligence Reliably Report Chest X-Rays?: Radiologist Validation of an Algorithm trained on 1.2 Million X-Rays. arXiv:180707455 [cs]. July 2018. http://arxiv.org/abs/1807.07455
19. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. 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). 2017:3462–3471. https://doi.org/10.1109/CVPR.2017.369
20. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz CP, Patel BN, Yeom KW, Shpanskaya K, Blankenberg FG, Seekins J, Amrhein TJ, Mong DA, Halabi SS, Zucker EJ, Ng AY, Lungren MP: Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLOS Medicine. 15(11): e1002686, 2018. https://doi.org/10.1371/journal.pmed.1002686

21. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. arXiv: 180104381 [cs]. 2018. http://arxiv.org/abs/1801.04381
22. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK: Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med. 15(11):e1002683, 2018. https://doi.org/10.1371/journal.pmed.1002683
23. Swenson DW, Baird GL, Portelli DC, Mainiero MB, Movson JS: Pilot study of a new comprehensive radiology report categorization (RADCAT) system in the emergency department. Emerg Radiol. 25(2):139–145, 2018. https://doi.org/10.1007/s10140-017-1565-8
24. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. arXiv:160806993 [cs]. 2016. http://arxiv.org/abs/1608.06993
25. Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge. arXiv:14090575 [cs]. 2014. http://arxiv.org/abs/1409.0575
26. Paszke A, Gross S, Chintala S, et al. Automatic differentiation in PyTorch. 2017. https://openreview.net/forum?id=BJJsrmfCZ.
27. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv:14126980 [cs]. 2014. http://arxiv.org/abs/1412.6980
28. Efron B, Tibshirani R: Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statist Sci. 1(1):54–75, 1986. https://doi.org/10.1214/ss/1177013815
29. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv:171105225 [cs,* stat*]*. 2017. http://arxiv.org/abs/1711.05225
30. Raoof S, Feigin D, Sung A, Raoof S, Irugulpati L, Rosenow EC: Interpretation of plain chest roentgenogram. Chest. 141(2):545–558, 2012. https://doi.org/10.1378/chest.10-1302