Clinical Radiology 74 (2019) 338–345

Contents lists available at ScienceDirect

Clinical Radiology

journal homepage: www.clinicalradiologyonline.net

Review

How far have we come? Artificial intelligence for chest radiograph interpretation

K. Kallianos [a], J. Mongan [a], S. Antani [b], T. Henry [a], A. Taylor [a], J. Abuya [c], M. Kohli [a,*]

[a] Department of Radiology and Biomedical Imaging, University of California San Francisco, San Francisco, CA 94143, USA

[b] National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

[c] Department of Radiology, School of Medicine, College of Health Sciences, Moi University, Eldoret, Kenya

Due to recent advances in artificial intelligence, there is renewed interest in automating interpretation of imaging tests. Chest radiographs are particularly interesting due to many factors: relatively inexpensive equipment, importance to public health, commonly performed throughout the world, and deceptively complex taking years to master. This article presents a brief introduction to artificial intelligence, reviews the progress to date in chest radiograph interpretation, and provides a snapshot of the available datasets and algorithms available to chest radiograph researchers. Finally, the limitations of artificial intelligence with respect to interpretation of imaging studies are discussed.

Introduction and fundamental concepts

Over the past 5 years, there has been increasing interest in applying novel artificial intelligence (AI) techniques to medical imaging. Although interest in AI has grown recently, it was originally defined as a field of study in 1956 at Dartmouth.[1] AI is very broad and encompasses many techniques applicable to imaging, ranging from traditional computer vision to newer deep learning techniques (Fig 1).

Most computer programs are a set of explicit instructions, which tell the computer how data should be analysed and handled for a particular task. For example, in order to write a program to recognise images of hotdogs, a programmer would craft mathematical expressions to identify features such as a bun, frankfurter, ketchup, mustard, or relish. If the combination of these expressions exceeds a threshold, the computer would recognise an image as a hotdog. Creation of hand-crafted expressions requires not only highly experienced programmers, but also deep domain expertise for the task at hand, both of which are expensive and difficult to scale to larger applications.

In contrast, machine learning techniques use equations with millions of parameters that can be automatically adjusted through a learning process to recognise features from a large set of data without the need to hand-craft rules. The majority of machine learning in medicine uses large volumes of training data with established truth (supervised learning). Algorithms are trained through exposure to large numbers of example images. Returning to the hotdog example, thousands of example hotdog images are required to send through the algorithm. Based on how confidently the algorithm identifies the image as a hotdog, the

* Guarantor and correspondent: M. Kohli, 530 Parnassus Avenue, RM CL-158, San Francisco, CA 94143, USA. Tel.: +1 415 476 5941.
E-mail address: marc.kohli@ucsf.edu (M. Kohli).

https://doi.org/10.1016/j.crad.2018.12.015
0009-9260/

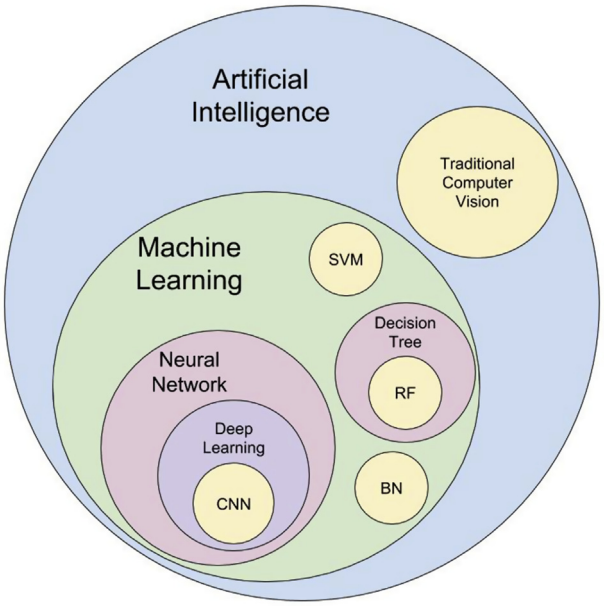

Figure 1 AI terminology map that describes the relationships between terms commonly encountered in medical imaging AI literature, and the lay press. SVM, support vector machine; RF, random forest; BN, Bayesian network; CNN, convolutional neural network.

parameters are adjusted. This process is repeated millions of times, until the algorithm parameters stabilise. New, previously unseen hotdog images are then passed through the algorithm to measure performance.

Widespread application of machine learning for photographic image characterisation has resulted from the inexpensive parallel computing in the form of graphical processing units (GPUs) and two scientific advances: convolutional neural networks (a type of deep learning), and the availability of large image datasets with established ground truth.[2,3] With millions of publicly available digital photographs available on the internet, ImageNet researchers used crowdsourcing to label them with everyday objects like dogs, cats, and mountains. These very large datasets, due to their size, provide the ability to train very complex (or deep) neural networks.[4]

Convolutional neural networks (CNN) are a common network architecture that is particularly adept at image recognition tasks. Readers interested in a more detailed explanation of neural networks and architectures are directed to the following reviews.[5,6] When the layers of a CNN are interrogated to identify what causes activation, the layers nearer to the input learn to recognise edges and shadows. In the middle of the network, these shapes are grouped together into more complex patterns. For example, a middle neuron might recognise an eye, or the nose of a cat. The output neurons associate the image features with the final class assignment (e.g., housecat). Today, there are not yet millions of medical images to train deep learning from scratch, typical datasets are in the order of thousands. Fortunately, researchers have identified that algorithms that are initially trained on non-medical images (ImageNet) can be fine-tuned to work with medical images, which decreases the need for very large training datasets. This technique is called transfer learning and has been proven beneficial for several different types of medical images[7–11] including chest radiographs.[12]

Chest radiography is a particularly appealing problem for AI for a number of reasons: it is important to public health, it is one of the most commonly performed imaging procedures throughout the world, it uses relatively inexpensive equipment, and although seemingly simple, it requires expertise to master. In the following paragraphs we will review progress in clinical applications, describe publicly available datasets and algorithms, and discuss limitations of existing techniques.

Clinical applications in chest radiography

Image orientation

Early applications of machine learning for the assessment of chest radiographs were published as early as the 1990s, with neural networks tasked with correctly classifying the orientation of chest radiographs.[13] In this early study, the algorithm was accurate in 88.8% of images, with <1% on images incorrectly aligned along their caudal–cranial axis. From these relatively humble beginnings, machine learning in chest radiograph interpretation has expanded to define the thoracic anatomy with high accuracy. In a more recent companion study, classification of a large series of chest radiographs as either frontal or lateral in orientation was rapid, with 38 images classified per second, with 100% accuracy.[14] Given the dramatic improvements in this field in recent years, machine learning applications have been investigated to aid in the diagnosis for a variety of clinical indications, as described in the sections below.

Segmentation of chest radiographs

Segmentation or automatically separating the lung parenchyma into distinct regions is important in the application of artificial intelligence to the analysis of chest radiographs. There are several challenges in this process, which are relevant to the diagnosis of pulmonary parenchymal disease, specifically related to the edges of the ribs and clavicles, which can obscure evaluation of the underlying lung apices, and small regions, such as the costophrenic angles, where a pleural effusion may be detected. Another challenge is that the abnormalities may be relatively subtle or localised rather than widespread, which may necessitate different approaches for detection.

The United States National Library of Medicine (NLM) recently developed a chest radiograph screening system, and described their process of lung segmentation in an article by Candemir *et al.*, which includes patient-specific adaptive models to detect lung boundaries.[15] This model matches a training image that is most similar to the individual patient radiograph and creates a patient-specific model of the individual lung anatomy, with further refinement of lung boundaries. The reported accuracy (defined as

overlap coefficients, Ω, comparing the ground truth and the segmentation mask) for this method ranged from 91.7% to 95.4% across three chest radiograph databases. Candemir *et al.* achieved results comparable to a human observer (Ω of 94.6%) with this method, improving upon other algorithms reported in the literature. The authors note the importance of evaluating lung segmentation methods on data sets including both normal and abnormal radiographs, as many diseases such as pleural effusions or atelectasis can significantly alter the lung shape.

Localisation of abnormalities

Several algorithms have been designed to not only identify pathology, but also localise abnormalities on chest radiographs. Li *et al.* emphasised that localisation of abnormalities is essential for the clinical applicability of the labels applied to radiographs through deep learning techniques, and aimed to improve localisation of the broad range of chest radiograph abnormalities. After training their algorithm on a subset of images including a high percentage with mark-up localising the relevant finding, the authors saw improved overall detection performance despite a lower total number of training cases. For example, the ability to detect pulmonary oedema was slightly greater (area under the receiver operating characteristic curve [AUC] of 0.86 versus 0.85) using approximately half as many total training images (45,200 versus 88,892 images) simply by the addition of 704 annotated images. When comparing the model-generated localisation of chest radiograph abnormalities to the ground truth, training sets with a higher percentage of annotated images improved localisation ability, while simply increasing the percentage of non-annotated images in the training set did not improve localisation, and even degraded performance for some chest radiograph pathologies. The work of Li *et al.* highlighted that the burden of training deep learning models can be partially alleviated by the incorporation of annotated medical images, and may not necessitate a very large number of images assuming that relevant annotated images are available.

Background on tuberculosis screening

Chest radiographs are frequently used to screen for tuberculosis infection, both in asymptomatic patients with positive purified protein derivative (PPD) tests and also in patients with acute or chronic symptoms of respiratory infection. Artificial intelligence may be of particular benefit in the assessment for pulmonary tuberculosis in resource limited settings, where there may be a lack of trained radiologists or healthcare practitioners to interpret the examinations.

The characteristic anatomical location of post-primary tuberculosis, commonly involving the lung apices or superior segments of the lower lobes, is used by radiologists to suggest the presence of sequela of tuberculosis infection. Classic findings include those of airway infection and cavitation in approximately half of patients. Calcification of the lung parenchyma or lymph nodes is suggestive of sequela of healed infection; however, primary tuberculosis, as well as post-primary tuberculosis in immunosuppressed patients, can present with a variety of other manifestations, including lymphadenopathy, consolidation in any portion of the lung, pleural effusions, and as small randomly distributed pulmonary nodules (miliary pattern) in patients with disseminated disease.

Tuberculosis detection

Jaeger *et al.* investigated screening for tuberculosis using an algorithm developed for AMPATH (Academic Model Providing Access to Healthcare), a partnership between institutions in Kenya and the United States and designed for implementation in a mobile vehicle-based unit.[16] Jaeger *et al.* utilised a group of publically available training masks that were most similar to the input radiograph, and defined normal and abnormal features of lung parenchyma through evaluating shape, edge, and texture. In this analysis, the authors demonstrated AUC/accuracy rates of 86.9/78.3% and 88/82.5% in two sample data sets for the detection of tuberculosis. Jaeger *et al.* compared the results of their algorithm with human performance utilising the consensus diagnosis of two practicing radiologists, referenced with ground truth tuberculosis status based on clinical data. The radiologists' consensus had a sensitivity of 100% and specificity of 68%, compared to a sensitivity and specificity of 74.1% and 81.3% respectively for the algorithm output. The authors note that if the algorithm was adjusted to achieve 100% sensitivity, the false-positive rate would rise to double the false-positive rate of the expert consensus; however, the authors emphasise that the combined performance of the experts and the machine learning approach had a substantially lower error rate (4.3%) compared to either the machine only (21.7%) and human consensus (18.1%) approaches, suggesting a complementary role of machine learning with human image interpretation.

Sivaramakrishnan *et al.* assessed the accuracy for tuberculosis detection on frontal chest radiographs using several deep learning models including one customised and five pre-trained CNNs.[17] The authors evaluated radiographs from four datasets including two publically available datasets as well as additional datasets provided from Kenya and India. Accuracy measurements of the customised model ranged from 0.572 to 0.824, whereas accuracy of the pre-trained CNNs ranged from 0.600 to 0.864. The highest performance for all models was seen with the dataset from India, which the authors postulated may be due to the relatively obvious and diffuse manifestations of disease in the dataset, which may have increased the ability of the model to distinguish abnormal from normal cases.

Given the variety of imaging manifestations of tuberculosis, deep learning models aimed at screening for tuberculosis must be able to assess for a variety of abnormalities in different anatomical locations. Xue *et al.* used a dataset of publicly available tuberculosis radiographs that were then annotated by two radiologists with specific annotations depending on the manifestation of tuberculosis in the image.[18] In addition to simply classifying the images as normal

or abnormal, Xue *et al.* focused on localisation of the imaging abnormality through a variety of steps including segmentation of the image area and classification of the extracted areas using a CNN. The authors observed a classification accuracy of 72.8% on their test set, and identified several factors that may have led to misclassification of images, including intrinsic complexity of the images, limited number of images for certain disease manifestations (such as miliary disease), and the coarse mark-up performed by the expert radiologists. They proposed more complex models, a larger data set, and more refined/detailed expert markings as areas where improvements could be made.

In order to detect more localised or subtle manifestations of tuberculosis infection, several authors have investigated combined local–global classification methods. Using this approach, Hogeweg *et al.* combined a texture analysis system evaluating the lung parenchyma at the pixel level with a shape abnormality detection system at the image level in order to improve the diagnostic performance for the diagnosis of tuberculosis.[19] This local–global classifier was subsequently integrated into a commercial product: CAD4TB (Delft Imaging Systems; The Netherlands), which in a high prevalence validation setting in South Africa was found to have an AUC of 0.71 when compared with Xpert MTB/RIF (Cepheid, Sunnyvale, CA, USA; Xpert) sputum testing as the reference standard.[20] It is important to note that the other work published in this field compares with radiologist consensus as the reference standard rather than Xpert sputum testing, which is the current standard of care for pulmonary tuberculosis diagnosis.

Cook *et al.* created a model aimed at detecting subtle or localised abnormalities using six image subregions, similar to those used for facial recognition, combined with a global assessment of the image, followed by fusion of the two techniques with the weight of each component based upon the confidence probability of each method.[21] Using this method, Cook *et al.* demonstrated AUCs of 0.949, 0.982, and 0.76 in all three evaluated datasets, with increased accuracy of the fused image classification in the three datasets compared to the local or global classification alone.

Chest radiograph abnormalities other than tuberculosis

Artificial intelligence has been applied to the interpretation of chest radiographs for multiple other clinical indications. The diagnosis of pneumonia is one area to which machine learning has been applied; a diagnosis that can be challenging as a result of the often non-specific imaging findings that may overlap with a variety of other disease entities such as atelectasis, aspiration, haemorrhage, and acute lung injury. Given the complexity of the image interpretation task, as well as a large public dataset (ChestX-ray8), the Radiologic Society of North America (RSNA) and Google are hosting a Kaggle challenge on pneumonia detection.[22] For the challenge, six RSNA volunteers annotated 30,000 chest radiographs for the presence and location of infiltrates (submitted for publication) to be used as training data. For the test dataset, Society of Thoracic Radiology (STR) volunteers created consensus annotations for 4,000 chest radiographs. At the time of submission of this article, over 900 international teams are participating in the challenge. Detection of pulmonary nodules on radiographs is another area well suited to the application of AI, as radiographs are reportedly the source of 90% of the errors in lung cancer diagnosis.[23] The miss rate for proven lung cancer on chest radiographs in the literature has been reported at nearly 20% in one study with <30% of <1 cm nodules detected.[24] The reasons for this low cancer detection rate are many, and include small lesion size, ill-defined appearance of lung cancers, (specifically pulmonary adenocarcinoma), and overlapping anatomical structures such as hilar vascularity, clavicles, and spine.[25]

Detecting other chest radiograph abnormalities

Rajpurkar *et al.* described the creation of CheXNet, a CNN trained on the ChestX-ray14 dataset, which outputs a probability of pneumonia as well as a heat map localising the most indicative region of the image.[26] When comparing the performance of four individual radiologists and CheXNet algorithm, the investigators demonstrated performance of the CheXNet that was within the range of individual radiologist F1 scores, and higher than the average human F1 score; however, it is worth noting that in this analysis ground truth was defined by the labels placed on the image by the group of radiologists participating in the study, with each individual radiologist and the CheXNet algorithm tested against the other four labels in turn rather than via consensus diagnosis or clinical information. The authors also reported higher AUC for the CheXNet algorithm compared to other algorithms in the diagnosis of 14 common lung pathologies including emphysema, pleural effusions, masses/nodules, and pneumothorax.

Wang *et al.* created the ChestX-ray8 dataset through the use of natural language processing, extracting eight disease labels from the original reports of 108,948 frontal radiographs including atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, and pneumothorax. The authors then demonstrated the ability to detect multiple chest radiograph abnormalities and also generate bounding boxes highlighting the relevant findings using CNNs.[27] In this analysis, Wang *et al.* showed that both pneumothorax and cardiomegaly were well recognised (AUCs of 0.79 and 0.81, respectively), whereas more heterogeneous findings such as masses (AUC of 0.56) and smaller findings such as nodules (AUC of 0.72) were less frequently recognised even with the most successful model. Thus the challenge of detecting lung cancer on chest radiographs currently appears to persist for AI techniques as it does for human observers.

It is worth highlighting again that although the authors above utilised databases of chest radiograph images to train their algorithms, Bar *et al.* demonstrated that training on a non-medical image database (ImageNet, discussed in detail below) could still allow detection of abnormalities on chest radiographs, with a AUC of 0.93 for detection of a right

pleural effusion and 0.89 for detection of cardiomegaly,[12] although the applicability across a broad range of chest radiograph disease entities with such a training approach may be limited.

Existing datasets and algorithms

CNN deep learning algorithms are heavily dependent on large training data sets to compute weights within the network. With insufficient data, the models rapidly over-fit to pixel patterns in the training collection adversely impacting model use on other images with the similar pathology, more significantly affecting its generalisability. With the advent of AI-based services offered by commercial vendors, such as IBM Watson, there has been a significant push for organisations to develop a data policy and appoint a data steward for sharing image data sets.[28] This peer-to-peer sharing and monetisation of data has led to data being dubbed as *the new oil*, a currency whose value is ever increasing with the expanse of AI-based tools. Peer-to-peer agreements limiting wide distribution of data adversely affect the opportunity for scientists at large to contribute toward improvements in AI-based techniques for healthcare.

As mentioned above, transfer learning is useful in cases where a small number of medical images are available. Fortunately, medical imaging researchers can benefit from either training algorithms with general purpose images from ImageNet[2] or starting with an off-the-shelf pre-trained network. ImageNet has over 14 million images in 22,000 categories, a subset of these (1.2 million images in 1,000 categories) have been used to train networks top-performers in the ILSVRC (ImageNet Large Scale Visual Recognition Challenge).[29] Due to the open-source culture of computer science research, the algorithms used to win ILSVRC are commonly made available for further modification and training: AlexNet,[3] VGG16,[30] Inception,[31] ResNet,[32] and its natural extension DenseNet.[33] These networks differ in their architecture as well as number of layers, and are being employed currently in imaging research.

A general medical image-related model can also be developed using NLM's PubMedCentral (PMC) Open Access data.[34] It offers images and related text from the freely available open access biomedical literature. These images and other metadata can also be directly accessed through NLM's Open-i(R) website using the API.[35] Open-i includes both radiological and other biomedical imaging data sets. An ongoing field of research is using these collections to train networks for transfer learning. Although not specifically related to chest radiography, another important archive of image data for research is The Cancer Imaging Archive (TCIA).[36] TCIA includes thousands of images across numerous imaging methods and malignancy types.

Other data sets that are specific to cardiopulmonary diseases and chest radiology are the two image collections from NLM,[37] a large chest X-ray image dataset from the Clinical Center at the National Institutes of Health (NIH) commonly referred to as ChestX-ray8,[27] and another hospital acquired dataset of chest X-ray images from Indiana (USA) and made available via NLM's Open-i website.[38] A collection of paediatric chest X-rays with examples of bacterial and viral pneumonia are also available.[39]

Limitations

There are several barriers to bringing the advances in AI described in the works reviewed here to fruition in a clinical setting, including limits on generalisation, available training data, lack of explainablity, and a lack of publication standards for reproducibility.

As the "knowledge" in a machine learning-based algorithm is derived from a training set, these algorithms are inherently limited by the size, quality, and type of training data. Deep learning algorithms are notable for their ability to generalise or apply their knowledge to new, previously unseen images. Although this generalisation is often substantially greater than other machine learning techniques, it is far less than human capability. Due to this limited ability to generalise, algorithms typically perform best on data that are very similar to training data. It is common for investigators to report performance against a test set that was obtained from the same sources and using the same methods as the training set. As the test set is much more similar to the training set than data encountered in clinical application would be, the performance reported on such test sets substantially overestimate real-world performance. Methodological decisions that lead to greater similarity between training and test data will tend to amplify the divergence between test-set reported and real-world performance. For instance, even though there may be no image data shared between the training and test sets, if both sets include images from the same patient, or worse from the same study, the performance on the test set is likely to be substantially artificially inflated. As it can be difficult to anticipate all the sources of similarity that may impact the performance of the algorithm, external validation of an algorithm against a test set of data obtained from a different source (or ideally multiple different sources) should in general be considered more reliable than internal validation against a test set similar to the training set. Although external validation provides a more reliable assessment of model performance, improving model performance on the range of inputs likely to be encountered across different clinical practices requires increasing the diversity of sources of the training set. There are no direct technical challenges to either using diverse training data or performing external validation, but the administrative and financial barriers to assembling large multi-institutional data sets can be formidable. It is anticipated that as organisations and professional societies come together to generate large public or licensable data sets, these barriers will diminish.

Another substantial limitation based on training data relates to the type of training data available. Radiologists typically interpret studies in the clinical context, making

use of data beyond what is contained in the images. It is possible, if challenging, to incorporate such data into deep learning models, but this capability can only be realised if the training data set contains relevant clinical information. At present, the cost and challenges of collecting non-image-based clinical data associated with each image in a structured fashion at the scale needed for deep learning is sufficiently difficult and resource intensive that most algorithms use little or no input data other than images and learn from annotations created by radiologists in the absence of clinical context. It is expected that as the field matures and tools and techniques for extracting clinical data are further refined, incorporation of clinical context data into training data sets will become more common.

A common critique of machine learning techniques in general and deep learning in particular is that the algorithms they produce are not explainable: a classification algorithm outputs a class assignment for the input image (e.g. “normal” or “pneumothorax”) and perhaps a level of confidence in the assignment, but there is currently no feasible approach for evaluating the interaction of thousands or millions of input pixels with millions or tens of millions of weights to determine why the algorithm yielded the output that it did. This problem is less obvious for segmentation or detection algorithms, because the output highlights portions of the image, but the issue remains that it is unclear how the algorithm arrived at the output. This is problematic in a medical setting as the absence of an explanation may undermine clinician confidence in correct outputs and limit the ability to improve performance by addressing incorrect outputs. These concerns are underscored by examples of adversarial images, in which a correctly classified input image may be altered in ways that are imperceptible to the human eye, but cause the altered image to be misclassified by the algorithm. Although it is unlikely that adversarial images would be encountered in a clinical setting, their existence emphasises that the failure modes of AI-based algorithms may be very different from the human visual system and difficult for humans to understand.

Concerns about explainability are not unique to medical applications of AI; this is an active area of research in which several important advances have been made. An important class of explainability techniques highlights the portions of the input image that have greatest impact on the output. Two examples of these are saliency maps and class activation maps (Fig 2).[40,41] In an anthropomorphised sense, these techniques provide explainability at the level of what the algorithm is looking at in the input image. Although this falls short of truly explaining why the output was produced, knowing what inputs are important may be enough to allow reasonable inferences to be made about why they are important. For instance, Zech *et al.* used class activation maps to try to understand a classifier they created to determine the institution at which a chest radiograph was obtained.[42] They found that the maps consistently highlighted the area of the laterality marker (“what”). Inspecting these laterality markers, they saw that the institutions in their data set used differently shaped markers, so the institution could generally be determined from the appearance of the marker (“why”). We expect that further advances in explainability will be rapidly incorporated into radiological applications, and that these will help to address the limited explainability of current techniques.

The majority of the papers reviewed for this article do not describe their computational methods in sufficient detail to enable the work to be reasonably reproduced. This incompleteness hinders advances in the field in at least two ways: readers are limited in their ability to fully evaluate the quality of the work and other investigators are limited in their ability to build on the work described. There are several factors that contribute to this deficiency. Foremost is likely inexperience with these new techniques: both authors and reviewers are often unaccustomed to all of the

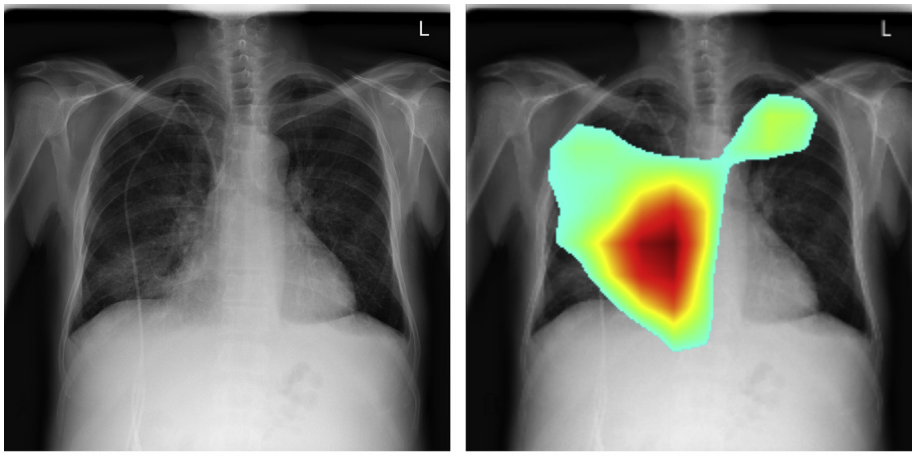

Figure 2 (a) Image from the ChestX-ray8 dataset in a patient with a right lower lobe infiltrate. This image was classified as abnormal by a CNN trained to detect infiltrates on chest radiographs. (b) Class activation map (Grad-CAM) that indicates which input pixels were important in classifying the radiograph, correctly identifies the infiltrate in the right lower lobe.

choices that must be made in applying AI techniques, so authors neglect to completely describe them and reviewers fail to recognise their omission. Manuscript length limits may also be a factor in important details being omitted, as complete description of a computational method is often substantially longer than a typical clinical research methods section. This situation will improve as authors and reviewers become more familiar with the details of these techniques and journals promote guidelines describing minimum standards for publication of AI research as they have with other types of research. We further expect use of supplemental methods sections that fully describe the details of the computational approach will become more common.

Conclusion and future directions

AI has a rich history of application to chest radiographs; however, we have not yet made the transition from bench to bedside. Although AI and deep learning in particular, have overcome many of the limitations of hand-crafted computer vision applications, we are still several steps away from having widely generalised AI for chest radiograph interpretation; however, as interpretation is only a small part of the imaging life cycle, there are several other areas where acceptable performance is likely to be more easily attained. For example, worklist prioritisation, automated quality assessment at the modality, and computation of new quantitative image features that are not currently recognised to name a few.

Conflict of Interest

The authors declare no conflict of interest.

References

1. *Dartmouth AI archives*. 2018. Available at: http://raysolomonoff.com/dartmouth/. [Accessed 4 April 2018].
2. Deng J, Dong W, Socher R, *et al*. ImageNet: a large-scale hierarchical image database. In: *CVPR09*; 2009. Miami FL - http://tab.computer.org/pamitc/archive/cvpr2009/index.html.
3. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. *Adv Neural Inf Process Syst* 2012;**25**:1097–105.
4. Chartrand G, Cheng PM, Vorontsov E, *et al*. Deep learning: a primer for radiologists. *RadioGraphics* 2017;**37**:2113–31. https://doi.org/10.1148/rg.2017170077.
5. Kohli M, Prevedello LM, Filice RW, *et al*. Implementing machine learning in radiology practice and research. *AJR Am J Roentgenol* 2017;**208**(4):754–60.
6. Erickson BJ, Korfiatis P, Akkus Z, *et al*. Machine learning for medical imaging. *RadioGraphics* 2017;**37**:505–15. https://doi.org/10.1148/rg.2017160130.
7. Shie C-K, Chuang C-H, Chou C-N, *et al*. Transfer representation learning for medical image analysis. In: *2015 37th annual International conference of the IEEE engineering in medicine and biology society (EMBC), Milan*. Piscataway, NJ: IEEE; 2015. p. 711–4.
8. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clin Radiol* 2018;**73**:439–45. https://doi.org/10.1016/j.crad.2017.11.015.
9. Kim HG, Choi Y, Ro YM. *Modality-bridge transfer learning for medical image classification*. 2017. https://doi.org/arXiv:1708.03111v1.
10. Ravishankar H, Sudhakar P, Venkataramani R, *et al*. *Understanding the mechanisms of deep transfer learning for medical images*. 2017. ArXiv 170406040 Cs.
11. Chang J, Yu J, Han T, *et al*. A method for classifying medical images using transfer learning: a pilot study on histopathology of breast cancer. In: *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)*. Piscataway, NJ: IEEE; 2017. p. 1–4.
12. Bar Y, Diamant I, Wolf L, *et al*. Chest pathology detection using deep learning with non-medical training. In: *2015 IEEE 12th International Symposium on biomedical imaging (ISBI)*. Piscataway, NJ: IEEE; 2015. p. 294–7.
13. Boone JM, Seshagiri S, Steiner RM. Recognition of chest radiograph orientation for picture archiving and communications systems display using neural networks. *J Digit Imag* 1992;**5**:190–3.
14. Rajkomar A, Lingam S, Taylor AG, *et al*. High-throughput classification of radiographs using deep convolutional neural networks. *J Digit Imag* 2017;**30**:95–101. https://doi.org/10.1007/s10278-016-9914-9.
15. Candemir S, Jaeger S, Palaniappan K, *et al*. Lung segmentation in chest radiographs using anatomical atlases with non-rigid registration. *IEEE Trans Med Imag* 2014;**33**:577–90. https://doi.org/10.1109/TMI.2013.2290491.
16. Jaeger S, Karargyris A, Candemir S, *et al*. Automatic tuberculosis screening using chest radiographs. *IEEE Trans Med Imag* 2014;**33**:233–45. https://doi.org/10.1109/TMI.2013.2284099.
17. Sivaramakrishnan R, Antani S, Xue Z, *et al*. Visualizing abnormalities in chest radiographs through salient network activations in deep learning. In: *2017 IEEE life sciences conference (LSC)*. Piscataway, NJ: IEEE; 2017. p. 71–4.
18. Xue Z, Jaeger S, Antani S, *et al*. Localizing tuberculosis in chest radiographs with deep learning. In: *Medical imaging 2018: imaging informatics for healthcare, research, and applications*. International Society for Optics and Photonics; 2018. 105790U.
19. Hogeweg L, Mol C, de Jong PA, *et al*. Fusion of local and global detection systems to detect tuberculosis in chest radiographs. *Med Image Comput Comput Assist Interv* 2010;**13**(Pt 3):650–7.
20. Muyoyeta M, Maduskar P, Moyo M, *et al*. The sensitivity and specificity of using a computer aided diagnosis program for automatically scoring chest X-rays of presumptive TB patients compared with Xpert MTB/RIF in Lusaka Zambia. *PLoS One* 2014;**9**:e93757, https://doi.org/10.1371/journal.pone.0093757.
21. Ding M, Antani S, Jaeger S, *et al*. Local-global classifier fusion for screening chest radiographs. In: Cook TS, Zhang J, editors. *Medical imaging 2017: imaging informatics for healthcare, research, and applications*, vol. 10138. Bellingham, WA: SPIE; 2017. 101380A. Proc. of SPIE.
22. *RSNA pneumonia detection challenge*. 2018. Available at: https://www.kaggle.com/c/rsna-pneumonia-detection-challenge. [Accessed 26 September 2018].
23. del Ciello A, Franchi P, Contegiacomo A, *et al*. Missed lung cancer: when, where, and why? *Diagn Interv Radiol* 2017;**23**:118–26. https://doi.org/10.5152/dir.2016.16187.
24. Quekel LGBA, Kessels AGH, Goei R, *et al*. Miss rate of lung cancer on the chest radiograph in clinical practice. *Chest* 1999;**115**:720–4. https://doi.org/10.1378/chest.115.3.720.
25. Wu M-H, Gotway MB, Lee TJ, *et al*. Features of non-small cell lung carcinomas overlooked at digital chest radiography. *Clin Radiol* 2008;**63**:518–28. https://doi.org/10.1016/j.crad.2007.09.011.
26. Rajpurkar P, Irvin J, Zhu K, *et al*. *CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning*. 2017. ArXiv 171105225 Cs Stat.
27. Wang X, Peng Y, Lu L, *et al*. 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)*. Piscataway, NJ: IEEE; 2017. p. 3462–71.
28. *Battle JC How to share big data for artificial intelligence*. 2018. Available at: https://acrdsi.org/Blog/How-to-Share-Big-Data-for-Artificial-Intelligence. [Accessed 24 September 2018].
29. Russakovsky O, Deng J, Su H, *et al*. *ImageNet large scale visual recognition challenge*. 2014. ArXiv 14090575 Cs.

30. Simonyan K, Zisserman A. *Very deep convolutional networks for large-scale image recognition*. 2014. ArXiv 14091556 Cs.
31. Szegedy C, Liu W, Jia Y, *et al*. *Going deeper with convolutions*. 2014. https://doi.org/arXiv:1409.4842v1.
32. He K, Zhang X, Ren S, *et al*. *Deep residual learning for image recognition*. 2015. ArXiv 151203385 Cs.
33. Huang G, Liu Z, van der Maaten L, *et al*. *Densely connected convolutional networks*. 2016. ArXiv 160806993 Cs.
34. *Open access subset*. 2018. Available at: https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/. [Accessed 24 September 2018].
35. *Open-i biomedical image search engine- open-i*. 2018. Available at: https://openi.nlm.nih.gov/. [Accessed 24 September 2018].
36. Clark K, Vendt B, Smith K, *et al*. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imag* 2013;**26**:1045–57. https://doi.org/10.1007/s10278-013-9622-7.
37. Jaeger S, Candemir S, Antani S, *et al*. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quant Imag Med Surg* 2014;**4**:475–7. https://doi.org/10.3978/j.issn.2223-4292.2014.11.20.
38. Demner-Fushman D, Kohli, Rosenman MB, *et al*. Preparing a collection of radiology examinations for distribution and retrieval. *J Am Med Inform Assoc* 2015, https://doi.org/10.1093/jamia/ocv080. ocv080.
39. Rajaraman S, Candemir S, Kim I, *et al*. Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs. *Appl Sci* 2018;**8**:1715. https://doi.org/10.3390/app8101715.
40. Simonyan K, Vedaldi A, Zisserman A. *Deep inside convolutional networks: visualising image classification models and saliency maps*. 2013. https://doi.org/arXiv:1312.6034v2.
41. Rs R, Das A, Vedantam R, *et al*. *Grad-CAM: why did you say that? Visual explanations from deep networks via gradient-based localization*. 2016. https://www.researchgate.net/publication/308964930_Grad-CAM_Why_did_you_say_that_Visual_Explanations_from_Deep_Networks_via_Gradient-based_Localization.
42. Zech JR, Badgeley MA, Liu M, *et al*. *Confounding variables can degrade generalization performance of radiological deep learning models*. 2018. ArXiv 180700431 Cs Stat.