Deep Learning Lends a Hand to Pediatric Radiology[1]

Ronald M. Summers, MD, PhD

Machine learning in radiology is a hot topic. A part of computer science, machine learning is a field in which systems can be designed and trained to learn concepts from data to make predictions. Machine learning, and in particular a subtype called deep learning, has shown high accuracy in performing difficult tasks, such as object recognition in images and speech recognition, and is now of great interest for medical image analysis (1).

Machine learning can be used for a number of applications in radiology, including automated detection of disease, segmentation of lesions, and quantitation. Radiologists are anxious to learn whether and how machine learning will affect their practices. In diverse fields of medical image analysis, including nonradiologic tasks such as diagnosis of skin lesion and retinal photographs (2,3), evidence indicates that machine learning can diagnose disease on images at a level comparable to that of skilled physicians. There are few diagnostic applications in which machine learning performs comparably to board-certified radiologists. In this issue of *Radiology*, Larson et al (4) present one such example.

Larson et al (4) developed and validated a machine learning system for the assessment of skeletal maturity (ie, bone age) on pediatric hand radiographs. By using a type of machine learning called a deep-learning neural network, they trained their computer model on 12 611 images and validated their model on 1425 images. Next, they tested their computer model on two different data sets. The first test set consisted of 200 hand radiographs from the authors' institution. The second test set consisted of 913 images from the publicly available Digital Hand Atlas. The images in the first test set were evaluated independently by four radiologists, one of whom wrote the original clinical report. The authors determined the differences in the bone ages calculated by the computer model and those of the human observers by using a pairwise analysis. They calculated these differences two ways. The first, called the root mean square error, was the square root of the sum of the squares of the paired differences. The second, the mean absolute difference, was calculated as the mean of the absolute values of the difference between the estimates provided by the reviewer and model and those of the reference standard bone age. On the test set that consisted of 200 hand radiographs, the root mean square of the paired interobserver difference ranged from 0.93 to 1.17 years. When applied to the second test data set, the computer model had a root mean square error of 0.73 years. Among the 200 test-case examinations, a fraction would be reclassified to a different diagnosis (advanced, normal, or delayed bone age), ranging from 15.5% for the computer model and 14.0%–18.5% for the four human observers. The authors concluded that their deep-learning convolutional neural network model could estimate skeletal maturity with accuracy similar to that of an expert radiologist and also similar to that of existing automated methods.

Saliency maps, which demonstrate parts of the radiograph for which the model output was most sensitive, highlighted the proximal interphalangeal joints, the metacarpal-phalangeal joints, and the carpal bones, anatomic areas that correspond with the maturity indicators used in the standards of Gruelich and Pyle. This is interesting because the computer model learned the importance of the joints on its own by using only the training data.

The authors assessed the importance of training sample size on the performance of the computer model

https://doi.org/10.1148/radiol.2018172898

Content code: PD

Radiology 2018; 287:323–325

[1] From the Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Radiology and Imaging Sciences Department, Clinical Center, National Institutes of Health, Building 10, Room 1C224D, MSC 1182, Bethesda, MD 20892-1182. Received December 11, 2017; accepted December 20. The intramural research programs of the National Institutes of Health, Clinical Center, supported this work. **Address correspondence to** the author (e-mail: *rms@nih.gov*).

R.M.S. supported by Intramural Research Program of the National Institutes of Health, Clinical Center (1Z01 CL040004).

Conflicts of interest are listed at the end of this article.

See also the article by Larson et al in this issue.

on the second test set. By using about an eighth of the data, the root mean square error of the interobserver difference was 1.08 years, decreasing to 0.73 years when the entire training sample set was used. These results provide some insight into the tradeoff between training sample size and model accuracy. One possible conclusion from these results is that further gains in model accuracy from increasing training sample size will be limited for automating this particular clinical task.

Automated bone age assessment is not new. The authors identified more than 15 documented attempts to perform automated bone age assessment. However, the article by Larson et al (4) brings new insights. First, it shows that a machine learning computer model for bone age assessment can be developed without the need for computing features crafted by experts. Whereas earlier research required the computer scientist to write software to locate the joints in the hand on the radiograph, as mentioned earlier the current system learned the importance of joints from the data itself. This means that development of machine learning software for this and other similar radiology applications can be more time efficient, less costly, and require less domain-specific expertise. From a broader perspective, the ability to learn important features without human direction is transformative and an essential reason for the success of the latest machine learning models. Second, the authors have shown that their machine learning computer model can perform comparably to trained pediatric radiologists. Whereas similar performance without machine learning was previously reported for bone age assessment (5), the work in this issue is one of the earliest to show human expert level performance with machine learning on any radiology interpretive task. Other recent examples of human expert level performance or high sensitivity and specificity include identification of hemorrhage, mass effect, and hydrocephalus on head computed tomographic (CT) images; colitis on abdominopelvic CT images; and tuberculosis on chest radiographs (6–8). Third, the authors outlined an approach that with limited modification could be adapted to develop computer models that learn how to perform other quantitative radiology analyses given sufficient training data.

It is notable that the main technical components of the model (deep residual network architecture, TensorFlow machine learning library [version 0.9.0; Google, Mountain View, Calif] and ImageNet pretrained weights) are freely and publicly available. Thus, the precise technical details of the computer model are mostly widely available, thereby accelerating technical development by anyone interested in similar automated clinical tasks. The limited availability of labeled data will therefore be one of the main obstacles for future research and development.

The method did have some important limitations. Because bone age is determined subjectively, the reference standard was subject to the variability of the readings from the human observers who participated in the study. Second, the computer model could not detect hand disorders such as dysplasias, rickets, and congenital syndromes (ie, the model could only determine the bone age). Third, the model was not effective in predicting bone age for children under 2 years of age, possibly relating to the lack of sufficient training examples.

What are the other implications of this and similar studies? First, we are likely to see many more examples of the application of machine learning to radiology because the number of potential opportunities for these technologies is vast (9). Such examples may be highly specific, for example bone age assessment as in this study. Or they could be more comprehensive, attempting to diagnose two or more diseases in the same image or three-dimensional volume (6,10). Many data sets of radiology images and accompanying reference standards have been collected over the years and new ones are appearing (1,11). These data sets will likely be used to develop more advanced, more accurate computer models with the use of machine learning. Whereas regulatory issues have been a concern of machine learning–based medical device developers, deep learning–based medical devices have already received clearance from the U.S. Food and Drug Administration (12).

Whether the clinical practice patterns and economics of current radiology practice will support and encourage the dissemination of computer software such as the one developed by Larson et al (4) remains to be determined. Ideally, such software would be integrated into radiology picture archiving and communication systems to provide seamless access to both the images and the results, and insertion of the results into the radiology report. The benefits of these machine learning systems for their intended use in the clinic will also need to be assessed with appropriate observer trials.

Disclosures of Conflicts of Interest: R.M.S. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed pending and/or awarded patents for automated image analyses; royalty income from iCad, Zebra Medical, Imbio, ScanMed, and Koninklijke Philips; and his laboratory received research support from Ping An Technology Company and Nvidia. Other relationships: disclosed no relevant relationships.

References

1. Greenspan H, van Ginneken B, Summers RM. Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique. IEEE Trans Med Imaging 2016;35(5):1153–1159.

2. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542(7639):115–118.

3. Gulshan V, Peng L, Coram M, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA 2016;316(22):2402–2410.

4. 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 2018;287(1)313–322.

5. van Rijn RR, Thodberg HH. Bone age assessment: automated techniques coming of age? Acta Radiol 2013;54(9):1024–1029.

6. Prevedello LM, Erdal BS, Ryu JL, et al. Automated Critical Test Findings Identification and Online Notification System Using Artificial Intelligence in Imaging. Radiology 2017;285(3):923–931.

7. Liu J, Wang D, Lu L, et al. Detection and diagnosis of colitis on computed tomography using deep convolutional neural networks. Med Phys 2017;44(9):4630–4642.

8. Lakhani P, Sundaram B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. Radiology 2017;284(2):574–582.

9. Summers RM. Progress in Fully Automated Abdominal CT Interpretation. AJR Am J Roentgenol 2016;207(1):67–79.

10. 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. CVPR, 2017; 3462–3471.

11. NIH Clinical Center provides one of the largest publicly available chest x-ray datasets to scientific community. https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community. Published September 27, 2017. Accessed November 17, 2017.

12. FDA letter to Arterys, Inc. https://www.accessdata.fda.gov/cdrh_docs/pdf16/K163253.pdf. Published January 5, 2017. Accessed November 17, 2017.