Is Artificial Intelligence Going to Replace Dermatologists?

Faezeh Talebi-Liasi, MD; Orit Markowitz, MD

PRACTICE **POINTS**

- The use of computer-assisted diagnosis in medicine dates back to the 1960s in radiology.
- New techniques in machine learning, also known as deep learning, were introduced around 2010. Compared to the predecessor forms of computing, these new methods are dynamically changing systems that improve with continuous data exposure and therefore performance is dependent on the quality and generalizability of the training data sets.
- Standardized large data sets and prospective real-life clinical trials are lacking in radiology and subsequently dermatology for diagnosis.
- Artificial intelligence is helpful with triaging and is improving workflow efficiency for radiologists by helping prioritize tasks, which is the current direction for dermatology.

The use of computers or machines in medicine dates back to the 1960s. Deep learning software programming is a subset of artificial intelligence (AI) based on the ability of a machine to learn from data and adaptively change. Deep learning is creating the next industrial revolution across the economy by replacing repetitive low-skilled tasks with learning algorithms. In medicine, image-based fields such as radiology, dermatology, and pathology have seen an increase in the number of studies using deep learning. However, given the current lack of standardized data sets to train these machines, it is difficult to predict if the present results eventually will be translated to real-life clinical settings.

Cutis. 2020;105:28-31.

Artificial intelligence (AI) is a loosely defined term that refers to machines (ie, algorithms) simulating facets of human intelligence. Some examples of AI are seen in natural language-processing algorithms, including autocorrect and search engine autocomplete functions; voice recognition in virtual assistants; autopilot systems in airplanes and self-driving cars; and computer vision in image and object recognition. Since the dawn of the century, various forms of AI have been tested and introduced in health care. However, a gap exists between clinician viewpoints on AI and the engineering world's assumptions of what can be automated in medicine.

In this article, we review the history and evolution of AI in medicine, focusing on radiology and dermatology; current capabilities of AI; challenges to clinical integration; and future directions. Our aim is to provide realistic expectations of current technologies in solving complex problems and to empower dermatologists in planning for a future that likely includes various forms of AI.

Early Stages of AI in Medical Decision-making

Some of the earliest forms of clinical decision-support software in medicine were computer-aided detection and computer-aided diagnosis (CAD) used in screening for breast and lung cancer on mammography and computed tomography.[1-3] Early research on the use of CAD systems in radiology date to the 1960s (Figure), with the first US Food and Drug Administration–approved CAD system in mammography in 1998 and for Centers for Medicare & Medicaid Services reimbursement in 2002.[1,2]

Early CAD systems relied on rule-based classifiers, which use predefined features to classify images into desired categories. For example, to classify an image as a high-risk or benign mass, features such as contour and texture had to be explicitly defined. Although these systems showed on par with, or higher, accuracy vs a radiologist in validation studies, early CAD systems never achieved wide adoption because of an increased rate of false positives as well as added work burden on a radiologist, who had to silence overcalling by the software.[1,2,4,5]

Computer-aided diagnosis–based melanoma diagnosis was introduced in early 2000 in dermatology (Figure) using the same feature-based classifiers. These systems claimed expert-level accuracy in proof-of-concept studies and prospective uncontrolled trials on proprietary devices using these classifiers.[6,7] Similar to radiology, however, real-world adoption did not happen; in fact, the last of these devices was taken off the market in 2017. A recent

From the Department of Dermatology, Icahn School of Medicine at Mount Sinai Medical Center, New York, New York; Department of Dermatology, SUNY Downstate Medical Center, Brooklyn; and Department of Dermatology, New York Harbor Healthcare System, Brooklyn.

The authors report no conflict of interest.

Correspondence: Faezeh Talebi-Liasi, MD (Faezeh.liasi@gmail.com).

meta-analysis of studies using CAD-based melanoma diagnosis point to study bias; data overfitting; and lack of large controlled, prospective trials as possible reasons why results could not be replicated in a clinical setting.[8]

Beyond 2010: Deep Learning

New techniques in machine learning (ML), called deep learning, began to emerge after 2010 (Figure). In deep learning, instead of directing the computer to look for certain discriminative features, the machine learns those features from the large amount of data without being explicitly programed to do so. In other words, compared to predecessor forms of computing, there is less human supervision in the learning process (Table). The concept of ML has existed since the 1980s. The field saw exponential growth in the last decade with the improvement of algorithms; an increase in computing power; and emergence of large training data sets, such as open-source platforms on the Web.[9,10]

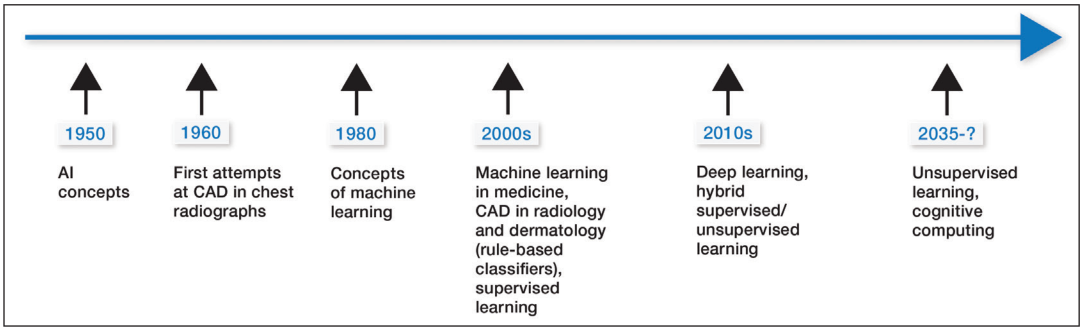
Most ML methods today incorporate artificial neural networks (ANN), computer programs that imitate the architecture of biological neural networks and form dynamically changing systems that improve with continuous data exposure. The performance of an ANN is dependent on the number and architecture of its neural layers and (similar to CAD systems) the size, quality, and generalizability of the training data set.[9-12]

In medicine, images (eg, clinical or dermoscopic images and imaging scans) are the most commonly used form of data for AI development. Convolutional neural networks (CNN), a subtype of ANN, are frequently used for this purpose. These networks use a hierarchical neural network architecture, similar to the visual cortex, that allows for composition of complex features (eg, shapes) from simpler features (eg, image intensities), which leads to more efficient data processing.[10-12]

In recent years, CNNs have been applied in a number of image-based medical fields, including radiology, dermatology, and pathology. Initially, studies were largely led by computer scientists trying to match clinician performance in detection of disease categories. However, there has been a shift toward more physicians getting involved, which has motivated development of large curated (ie, expert-labeled) and standardized clinical data sets in training the CNN. Although training on quality-controlled data is a work in progress across medical disciplines, it has led to improved machine performance.[11,12]

Recent Advances in AI

In recent years, the number of studies covering CNN in diagnosis has increased exponentially in several medical specialties. The goal is to improve software to close the gap between experts and the machine in live clinical


1950
AI concepts
1960
First attempts at CAD in chest radiographs
1980
Concepts of machine learning
2000s
Machine learning in medicine, CAD in radiology and dermatology (rule-based classifiers), supervised learning
2010s
Deep learning, hybrid supervised/unsupervised learning
2035-?
Unsupervised learning, cognitive computing

Timeline of artificial intelligence (AI) in medicine and dermatology. CAD indicates computer-aided diagnosis.

Comparison of Computer Systems

	Computing System			
Attribute	Rule-Based Classifier	Supervised Learning and Hybrid Systems	Unsupervised Learning	Cognitive Computing
Era	Past	Past and present AI in health care	Near-future AI, not currently applied in health care	Distant future
Input	Labeled and structured data	Labeled and structured data	Unlabeled and unstructured data	Unstructured data across multiple sensory (input) domains
Processing classification method	Rule based (predefined features)	Automated	Automated	Automated
Performance	Limited; recognizes only predefined patterns	Dependent on the training data set, overfitting	Unlimited, though hard to assess performance, given lack of a feedback system	Unlimited data processing, data integration, and analytic power

Abbreviation: AI, artificial intelligence.

settings. The current literature focuses on a comparison of experts with the machine in simulated settings; prospective clinical trials are still lagging in the real world.[9,11,13]

We look at radiology to explore recent advances in AI diagnosis for 3 reasons: (1) radiology has the largest repository of digital data (using a picture archiving and communication system) among medical specialties; (2) radiology has well-defined, image-acquisition protocols in its clinical workflow[14]; and (3) gray-scale images are easier to standardize because they are impervious to environmental variables that are difficult to control (eg, recent sun exposure, rosacea flare, lighting, sweating). These are some of the reasons we think radiology is, and will be, ahead in training AI algorithms and integrating them into clinical practice. However, even radiology AI studies have limitations, including a lack of prospective, real-world clinical setting, generalizable studies, and a lack of large standardized available databases for training algorithms.

Narrowing our discussion to studies of mammography—given the repetitive nature and binary output of this modality, which has made it one of the first targets of automation in diagnostic imaging[1,2,5,13]—AI-based CAD in mammography, much like its predecessor feature-based CAD, has shown promising results in artificial settings. Five key mammography CNN studies have reported a wide range of diagnostic accuracy (area under the curve, 69.2 to 97.8 [mean, 88.2]) compared to radiologists.[15-19]

In the most recent study (2019), Rodriguez-Ruiz et al[15] compared machines and a cohort of 101 radiologists, in which AI showed performance comparability. However, results in this artificial setting were not followed up with prospective analysis of the technology in a clinical setting. First-generation, feature-based CADs in mammography also showed expert-level performance in artificial settings, but the technology became extinct because these results were not generalizable to real-world in prospective trials. To our knowledge, a limitation of radiology AI is that all current CNNs have not yet been tested in a live clinical setting.[13-19]

The second limitation of radiology AI is lack of standardization, which also applies to mammography, despite this subset having the largest and oldest publicly available data set. In a recent review of 23 studies on AI-based algorithms in mammography (2010-2019), clinicians point to one of the biggest flaws: the use of small, nonstandardized, and skewed public databases (often enriched for malignancy) as training algorithms.[13]

Standardization refers to quality-control measures in acquisition, processing, and image labeling that need to be met for images to be included in the training data set. At present, large stores of radiologic data that are standardized within each institution are not publicly accessible through a unified reference platform. Lack of large standardized training data sets leads to selection bias and increases the risk for overfitting, which occurs when algorithm models incorporate background noise in the data into its prediction scheme. Overfitting has been noted in several AI-based studies in mammography,[13] which limits the generalizability of algorithm performance in the real-world setting.

To overcome this limitation, the American College of Radiology Data Science Institute recently took the lead on creating a reference platform for quality control and standardized data generation for AI integration in radiology. The goal of the institute is for radiologists to work collaboratively with industry to ensure that algorithms are trained on quality data that produces clinically useable output for the clinician and patient.[11,20]

Similar to initial radiology studies utilizing AI mainly as a screening tool, AI-driven studies in dermatology are focused on classification of melanocytic lesions; the goal is to aid in melanoma screening. Two of the most-recent, most-cited articles on this topic are by Esteva et al[21] and Tschandl et al.[22] Esteva et al[21] matched the performance of 21 dermatologists in binary classification (malignant or nonmalignant) of clinical and dermoscopic images in pigmented and nonpigmented categories. A CNN developed by Google was trained on 130,000 clinical images encompassing more than 2000 dermatologist-labeled diagnoses from 18 sites. Despite promising results, the question remains whether these findings are transferrable to the clinical setting. In addition to the limitation on generalizability, the authors do not elaborate on standardization of training image data sets. For example, it is unclear what percentage of the training data set's image labels were based on biopsy results vs clinical diagnosis.[21]

The second study was the largest Web-based study to compare the performance of more than 500 dermatologists worldwide.[22] The top 3–performing algorithms (among a pool of 139) were at least as good as the performance of 27 expert dermatologists (defined as having more than 10 years' experience) in the classification of pigmented lesions into 7 predefined categories.[22] However, images came from nonstandardized sources gathered from a 20-year period at one European academic center and a private practice in Australia. Tschandl et al[22] looked at external validation with an independent data set, outside the training data set. Although not generalizable to a real-world setting, looking at external data sets helps correct for overfitting and is a good first step in understanding transferability of results. However, the external data set was chosen by the authors and therefore might be tainted by selection bias. Although only a 10% drop in algorithmic accuracy was noted using the external data set chosen by the authors, this drop does not apply to other data sets or more importantly to a real-world setting.[22]

Current limitations and future goals of radiology also will most likely apply to dermatology AI research. In medicine and radiology, the goal of AI is to first help users by prioritizing what they should focus on. The concept of comparing AI to a radiologist or dermatologist is potentially shortsighted. Shortcomings of the current supervised or semisupervised algorithms used in medicine underscore the points that, first, to make their outputs clinically usable, it should be clinicians who procure and standardize training data sets and, second, it appears logical that the performance of these category of algorithms requires constant monitoring for bias. Therefore, these algorithms cannot operate as stand-alone diagnostic machines but as an aid to the clinician—if the performance of the algorithms is proved in large trials.

Near-Future Directions and Projections

Almost all recent state-of-the-art AI systems tested in medical disciplines fall under the engineering terminology of narrow or weak AI, meaning any given algorithm is trained to do only one specific task.[9] An example of a task is classification of images into multiple categories (ie, benign or malignant). However, task classification only works with preselected images that will need substantial improvements in standardization.

Although it has been demonstrated that AI systems can excel at one task at a time, such as classification, better than a human cohort in simulated settings, these literal machines lack the ability to incorporate context; integrate various forms of sensory input such as visual, voice, or text; or make associations the way humans do.[9] Multiple tasks and clinical context integration are required for predictive diagnosis or clinical decision-making, even in a simulated environment. In this sense, CNN is still similar to its antiquated linear CAD predecessor: It cannot make a diagnosis or a clinical decision but might be appropriate for triaging cases that are referred for evaluation by a dermatologist.

Medical AI also may use electronic health records or patient-gathered data (eg, apps). However, clinical images are more structured and less noisy and are more easily incorporated in AI training. Therefore, as we are already witnessing, earlier validation and adoption of AI will occur in image-based disciplines, beginning with radiology; then pathology; and eventually dermatology, which will be the most challenging of the 3 medical specialties to standardize.

Final Thoughts

Artificial intelligence in health care is in its infancy; specific task-driven algorithms are only beginning to be introduced. We project that in the next 5 to 10 years, clinicians will become increasingly involved in training and testing large-scale validation as well as monitoring narrow AI in clinical trials. Radiology has served as the pioneering area in medicine and is just beginning to utilize narrow AI to help specialists with very specific tasks. For example, a task would be to triage which scans to look at first for a radiologist or which pigmented lesion might need prompt evaluation by a dermatologist. Artificial intelligence in medicine is not replacing specialists or placing decision-making in the hands of a nonexpert. At this point, CNNs have not proven that they make us better at diagnosing because real-world clinical data are lacking, which may change in the future with large standardized training data sets and validation with prospective clinical trials. The near future for dermatology and pathology will follow what is already happening in radiology, with AI substantially increasing workflow efficiency by prioritizing tasks.

REFERENCES

1. Kohli A, Jha S. Why CAD failed in mammography. *J Am Coll Radiol*. 2018;15:535-537.
2. Gao Y, Geras KJ, Lewin AA, Moy L. New frontiers: an update on computer-aided diagnosis for breast imaging in the age of artificial intelligence. *Am J Roentgenol*. 2019;212:300-307.
3. Ardila D, Kiraly AP, Bharadwaj S, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med*. 2019;25:954-961.
4. Le EPV, Wang Y, Huang Y, et al. Artificial intelligence in breast imaging. *Clin Radiol*. 2019;74:357-366.
5. Houssami N, Lee CI, Buist DSM, et al. Artificial intelligence for breast cancer screening: opportunity or hype? *Breast*. 2017;36:31-33.
6. Cukras AR. On the comparison of diagnosis and management of melanoma between dermatologists and MelaFind. *JAMA Dermatol*. 2013;149:622-623.
7. Gutkowicz-Krusin D, Elbaum M, Jacobs A, et al. Precision of automatic measurements of pigmented skin lesion parameters with a MelaFind™ multispectral digital dermoscope. *Melanoma Res*. 2000;10:563-570.
8. Dick V, Sinz C, Mittlböck M, et al. Accuracy of computer-aided diagnosis of melanoma: a meta-analysis [published online June 19, 2019]. *JAMA Dermatol*. doi:10.1001/jamadermatol.2019.1375.
9. Hosny A, Parmar C, Quackenbush J, et al. Artificial intelligence in radiology. *Nat Rev Cancer*. 2018;18:500-510.
10. Gyftopoulos S, Lin D, Knoll F, et al. Artificial intelligence in musculoskeletal imaging: current status and future directions. *Am J Roentgenol*. 2019;213:506-513.
11. Chan S, Siegel EL. Will machine learning end the viability of radiology as a thriving medical specialty? *Br J Radiol*. 2019;92:20180416.
12. Erickson BJ, Korfiatis P, Kline TL, et al. Deep learning in radiology: does one size fit all? *J Am Coll Radiol*. 2018;15:521-526.
13. Houssami N, Kirkpatrick-Jones G, Noguchi N, et al. Artificial Intelligence (AI) for the early detection of breast cancer: a scoping review to assess AI's potential in breast screening practice. *Expert Rev Med Devices*. 2019;16:351-362.
14. Pesapane F, Codari M, Sardanelli F. Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. *Eur Radiol Exp*. 2018;2:35.
15. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *J Natl Cancer Inst*. 2019;111:916-922.
16. Becker AS, Mueller M, Stoffel E, et al. Classification of breast cancer in ultrasound imaging using a generic deep learning analysis software: a pilot study. *Br J Radiol*. 2018;91:20170576.
17. Becker AS, Marcon M, Ghafoor S, et al. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. *Invest Radiol*. 2017;52:434-440.
18. Kooi T, Litjens G, van Ginneken B, et al. Large scale deep learning for computer aided detection of mammographic lesions. *Med Image Anal*. 2017;35:303-312.
19. Ayer T, Alagoz O, Chhatwal J, et al. Breast cancer risk estimation with artificial neural networks revisited: discrimination and calibration. *Cancer*. 2010;116:3310-3321.
20. American College of Radiology Data Science Institute. Dataset directory. https://www.acrdsi.org/DSI-Services/Dataset-Directory. Accessed December 17, 2019.
21. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115-118.
22. Tschandl P, Codella N, Akay BN, et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *Lancet Oncol*. 2019;20:938-947.